AL-FARABI KAZAKH NATIONAL UNIVERSITY

Tukeyev U.A.

DATA MINING AND ANALYSIS

Educational manual

Almaty

« Qazaq University »

2018

Recommended for publication

Editorial-Publishing Council of Al-Farabi KazNU (Protocol No. 1 of 11.10.2018)

Reviewers:

professor, Doctor of Technical Science Akhmetov B.S.

professor, Doctor of Technical Science Shukayev D.N.

Candidate of Physical and Mathematical Sciences, Associate Professor Makashev E.P.

Tukeyev U.

T 90 Data mining and analysis: educational manual. - Second ed., revised and enlarged / Tukeyev U.A. - Almaty: Qazaq University, 2018. - 105 p.

ISBN 978-601-04-3701-2

The technology of data mining and analysis is used practically in all spheres of human activity, where retrospective data are accumulated. The methods of data mining and analysis were most widely spread in the following sectors: retail trade; the banking sector; insurance; telecommunications; industrial production; stock and currency markets. In this textbook, the tasks of prediction, classifying, clustering and association rules on specific examples are discussed in more detail.

The textbook is intended for undergraduate and graduate students of natural and technical science specialties.

© Tukeyev U.A., 2018

ISBN 978-601-04-3701-2 [©] Al-Farabi KazNU, 2018

CONTENT

Introduction	4
Topic 1. Intelligent data analysis	6
Topic 2. Model of prediction task	13
Topic 3. Predicting using standard functions	17
Topic 4. Calculation of the determination coefficient for a prediction	
function	23
Topic 5. Calculation of confidence intervals for predicted values	28
Topic 6. Recovering missing data for the prediction task	31
Topic 7. Analysis of data releases for the prediction task	37
Topic 8. Selection of factors for the prediction task	43
Topic 9. Estimation of the quality of the data model for the prediction task.	52
Topic 10. Data smoothing for the prediction task	61
Topic 11. Trend functions in prediction tasks	70
Topic 12. Seasonal component in prediction functions	79
Topic 13. Decision trees	93
Topic 14. Clustering	105
Topic 15. Association rules	116

INTRODUCTION

Data mining and analysis is a set of methods for detecting from large quantities of data (big data) previously unknown, practically useful knowledge for decision-making in various spheres of human activity. The peculiarity of the methods of mining and analysis of data, distinguishing them from traditional statistical methods are:

- the discovery of non-obvious regularities: these patterns do not identified by standard methods of information processing or expert by way of;

- the detection of objective laws: the knowledge obtained will be fully correspond to reality;

- finding practically useful regularities: the found knowledge can be found in concrete application in practice;

- reveal regularities without rigid restrictions to initial data and their distribution.

The most common tasks of data development and analysis are:

- classification,
- predicting,
- clustering,
- association.

The basis of methods of data development and analysis is various methods of classification, predicting, clustering, modeling, based on the application of: decision trees; artificial neural networks; genetic algorithms; associative memory; fuzzy logic, etc. Methods of mining and analysis of data also include multidimensional statistical methods: correlation and regression analysis; factor and component analysis; variance analysis; time series analysis and etc.

Data mining and analysis is a multi-stage and very labor-intensive process, which can be divided into three main stages:

- initial research;
- building a model;
- implementation of the model.

The most time-consuming stage of initial data mining involves:

- data cleaning: removal of duplicate observations from the sample, mistakenly

entered data with obvious errors, extreme values (outliers), checking logical rules and conditions;

- analysis and restoration, if necessary, of missing values in data;

- data conversion;
- setting up data properties;
- conducting an exploratory analysis of data using graphic and statistical methods;
- selection of the necessary data for building the model.

At the stage of the model construction, various models of data mining and analysis are considered, and the best of them is selected.

The implementation of the selected model implies its application to new data in order to obtain predictions or estimates of expected results, as well as subsequent monitoring of the quality of the model.

The technology of data mining and analysis is used practically in all spheres of human activity, where retrospective data are accumulated. The methods of development and analysis of data were most widely spread in the following sectors: retail trade; the banking sector; insurance; telecommunications; industrial production; stock and currency markets.

In this educational manual, the tasks of predicting, classifying, clustering and associative rules on specific examples are discussed in more detail. When studying the above problems, Excel is used for predicting tasks and the Rapidminer program for classification, clustering and associative rules tasks. Both programs are easily accessible: Excel is present in almost all local computers, and the Rapidminer program for educational purposes is freely available.

TOPIC 1

INTELLIGENT DATA ANALYSIS

1.1. Methodology of data mining

Data mining (data mining (DM) - Intelligent data analysis) is a process of discovering meaningful new relationships and dependencies by processing large amounts of data stored in repositories, using pattern recognition technology, statistical and mathematical methods.

There are various methods of data mining, among which is considered the de facto standard methodology CRISP-DM (CRoss-Industry Standard Process for DM) [1]. According to a survey in 2014 methodology CRISP-DM holds a leading position.

	logy are you using for your analytics,a science projects ? [200 votes total]2007 poll
CRISP-DM (86)	43% 42%
My own (55)	27.5%
SEMMA (17)	8.5% 13%
Other, not domain- specific (16)	8% 4%
KDD Process (15)	7.5% 7.3%
My organizations' (7)	3.5% 5.3%
A domain-specific methodology (4)	2 % 4 .7%
None (0)	0% 4.7%

Fig. 1.1. The results of the survey on the use of the methodology of data mining (http://www.kdnuggets.com/)

The methodology CRISP-DM has a life cycle that includes six phases (Figure 1.2.) [2]:

Phase 1. "Business-understanding." At this phase in CRISP-DM standard defines the following tasks:

- The objectives and requirements of the project in terms of business or research units.

- Translating the objectives and constraints in the definition of the problem of data mining.

- Development of the preliminary strategy (plan) to achieve their goals.

Phase 2. "Understanding the data." The second phase begins with data collection, the description thereof. Identify problems with data quality, such as, for errors or omissions. A search of interesting data sets that may contain hidden patterns. Thus, this phase comprising the following tasks:

- The collection of baseline data;

Phase 3. "Preparing data" includes the following tasks:

- Starting from the raw data is preparing the final data set that must be used for all subsequent phases. This phase is very time-consuming;

- Selected cases and variables to be analyzed, and are suitable for the analysis;

- Conversion are performed on certain variables, if necessary;

- Is cleaned original data, so that they were ready for modeling tools.

Phase 4. "Modelling" includes the following tasks:

- Selection of appropriate modeling techniques;

- Calibration of the model parameters to optimize results;

- Can be used several modeling techniques for the same data mining.

- If necessary, loop back to the data preparation phase, to bring the shape data in accordance with the specific requirements of a particular means of data mining.

Phase 5. "assessment" includes the following tasks:

- Evaluation of one or more models are available as modeling phase and the effectiveness of their pre-deployment for use in the field;

- Determine whether the model is actually achieves the objectives set for it in the first phase;

- To establish that some important aspect of the business and research tasks are not taken into account adequately;

- A conclusion regarding the use of the results of data mining.

6. The phase "deployment (implementation)" includes the following tasks:

- Creation of a model does not mean the end of the project;

- An example of a simple deployment: report creation;

- An example of a more complex deployment: the implementation of parallel data mining in another department.

- For enterprises, the customer deploys often based on the model chosen.



Fig. 1.2. The phases of the life cycle methodology CRISP-DM [1]

1.2 Tasks of Data Mining

The main tasks of the most common tasks of data mining include:

- classification
- prediction
- clustering
- Association.

Classification task.

Under the classification of objects commonly understood classification (observation events) to one of the previously known classes. For the classification must be present features characterizing the class to which belongs to this or that event or object. The classification relates to the learning strategies with the teacher (supervised learning). The problem of classification is commonly used categorical prediction of the dependent variable (ie, the dependent variable is the category) based on a sample of continuous and / or categorical variables.

Various methods are used for classification. The main ones are:

- Classification using decision trees;
- Bayesian (naive) classification;
- classification using artificial neural networks;
- classification by support vector machines;
- statistical methods such as linear regression;
- classification by using the nearest neighbor method.

Prediction task.

Prediction is classification similar, except that for the prediction, the results are in the future. Examples in the field of business studies and predicting tasks include:

- Predict the stock price for three months in the future;

- Prediction of the percentage increase in the incidence of accidents in traffic next year, if the speed limit is increased.

For prediction may be used as a traditional statistical evaluation methods and evaluation points, the confidence interval simple linear regression and correlation,

and multiple regression techniques and Data Mining and Knowledge Discovery, such as neural networks, k-nearest method for neighbor.

Clustering task

Clustering refers to the grouping of records, observations or instances into classes of similar objects. A cluster is a set of records that are similar to each other and different from the entries in other clusters. Clustering differs from classification in that there is no target variable for clustering. The clustering problem is not trying to classify, evaluate or predict the value of the target variable. Instead, clustering algorithms searching the initial set of data segmentation in a relatively homogeneous subgroups or clusters, where the cluster records maximized similarity and the similarity of the records of the cluster is minimized.

Clustering is often performed as a preliminary step in the process of data mining, and the resulting clusters are used as additional materials in a variety of technical methods in the following stages, such as neural networks.

Association

The task of association for the mining operation is located, which attributes "go hand in hand." Association Tasks are most common in the business world, where it is known as market basket analysis [2]. Association rules have the form of production rules: "if« A », and then« B »" with a measure of support and confidence of the rule. Support call the number or percentage of records containing a particular set of data. The reliability of the rules shows what is the likelihood of that event A implies event B. For example, a particular supermarket can find that out of 1,000 customers in stores Thursday, 200 customers have bought diapers, and of those 200 customers who bought diapers, 50 customers bought beer . Thus, the right of association will be "If you buy diapers, and then buy a beer" with the support of 200/1000 = 20%and reliability rules 50/200 = 25%.

Examples of problems in business and research associations include:

- Study the proportion of cell phone subscribers in the company's plan to respond positively to the proposal to upgrade services;

- The study of the percentage of children whose parents read to them are themselves good readers;

- To find out what items are purchased at the supermarket together and that the details have never purchased together

- Determination of the proportion of cases in which a new drug will exhibit dangerous side effects.

1.3 The phases of "Business Understanding" and " Data Understanding"

As it was said above, the following tasks are defined in the "Businessunderstanding" phase in the CRISP-DM standard:

- presentation of the objectives and requirements of the project in terms of business or research unit.

- translation of goals and limitations into the definition of the problem of data mining.

- development of a preliminary strategy (plan) to achieve the set goals.

In the "Data Understanding "phase, data are collected, described, analyzed and evaluated. This phase includes the following tasks:

- collection and description of source data;

-preliminary, semantic analysis of data on hidden patterns;

- evaluation of data quality.

Let's consider more in detail the tasks of these phases of data mining.

Let's consider on an example of maintenance of quality of cars for manufacturers of cars. Quality assurance is in a top priority for car manufacturers.

The goal of any car company is to reduce the costs associated with warranty claims and improve customer satisfaction. Through interviews with employees of the automotive company, who are technical experts in the production of vehicles, researchers can formulate specific business problems such as:

- Are there interdependencies between warranty claims?

- Are there any previous warranty claims related to similar requirements in the future?

- Is there a link between a certain type of claim and a specific garage?

The data mining plan consists in applying the appropriate data mining techniques to try and identify these and other possible associations.

In the "Understanding Data" phase, the collection and description of the initial data of a particular domain are performed. For the example in question, this is the database of the car manufacturer. How this database is used by consumers, how much this database and its elements are understood by end users. Is it possible to find answers to the questions identified in the previous stage of business understanding on this basis? Preliminary assessment of the quality of the data: the completeness of the data, if there are missing data, if there are data bursts, than what is explained.

In the following sections, the remaining phases of data mining are discussed in detail.

1.2 PRACTICAL LESSON 1 AND IWS (INDIVIDUAL WORK OF STUDENT) 1

Subject. Methodology and objectives of data mining.

Lesson plan.

1) Study the lecture material and methodology problems of data mining, above.

2) Setting the IWS. Make a detailed description of one of the objectives fo intelligent data analysis with examples.

3) A comparative analysis of this problem.

Literature.

1. Marbán Ó., Mariscal G. and Segovia J. A Data Mining & Knowledge Discovery Process Model. Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca, ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

2. Larose D.T. Discovering knowledge in data - an introduction to data mining. Wiley-Interscience, Hoboken, New Jersey, 2005

TOPIC 2

MODEL OF PREDICTION TASK

2.1. The lecture material

The prediction task is solved in two stages:

1) from the observable data $Y, t, X_1, X_2, ..., X_m$ determine the form of the function F;

2) knowing the form of the prediction function to make a prediction, that is, to find the values of the variable Y for new values of independent variables and the time factor.

Solving the prediction task means making a prediction, which consists in finding the value of the variable Y, for which there are no values in the original data set.

The task of a prediction can be presented in the following form [1]:

$Y = F(t, X_1, X_2, \dots X_m; \varepsilon),$

where the variable Y is a predictable variable, the variables $X_1, X_2, ..., X_m$ are independent variables, t is the time factor, and ε is a random variable indicating that Y is also a random variable. A random variable determines the inaccuracy in the measurement of the values of the variable Y, as well as the incompleteness of knowledge about the effects of the time factor and factors $X_1, X_2, ..., X_m$ on the variable Y. If the original data is presented in the form of a table (function graph) then y_i are the values of the function F for the specific values of its arguments $t_i, x_{1i}, x_{2i}, ..., x_{mi}$, where i = 1, 2, ..., n.

The determination of the form of the function F, which is unknown to the prediction problem, is performed on the basis of the available initial data, as well as some a priori considerations on the possible form of the function. The function is selected: $Y = F(b_1, b_2, ..., b_k; t; X_1, X_2, ... X_m)$, where the parameters $b_1, b_2, ..., b_k$ are chosen so that the values of the function F for given The values of the arguments $t_i, x_{1i}, x_{2i}, ... x_{mi}$ correspond to the values y_i as best as possible. Thus, the selected function F approximates the observed data, that is, the function F can be called an approximating function, since it approximates the observed data.

The constructed function F can be used for the second stage, the step of predicting a data of the dependent predicted variable for future periods of time for which there

are not yet observable data values. In this aspect, the approximating function F is called the prediction function.

The influence of the random variable ε is usually determined by adding a random effect to the value of the function *F*, i.e. The scheme of the random action *F* + ε , or the product of the random action and the value of the function *F*, is assumed. A scheme for the random action *F* * ε is adopted.

The predicted value of the predicted variable is calculated as the value of the prediction function for the values of the arguments $t_{n+1}, x_{1n+1}, x_{2n+1}, \dots, x_{mn+1}$. If the prediction function does not explicitly contain the time argument t, then such data models are called cause-and-effect or casual. If the prediction function does not contain factor arguments X_1, X_2, \dots, X_m and depends only on time, then this model is called a time series model.

Models of time series are characterized by a number of concepts, such as a step or projection period, time horizon, the trend in seasonal changes. The prediction period - is the time step, which presents the data in the original data set table $t, X_1, X_2, ..., X_m, Y$. Prediction horizon - is the amount of the prediction period, which will make the prediction. Prediction horizon can be short-term (several periods), medium (about ten periods) or long term (more than ten periods). General trend is a trend in the data table changes the original data set according to the time. Seasonal changes - it is repetitive (periodic) change the values of the factors that influence the behavior of the entire system. Trend and seasonal changes are the components of the prediction function. There are additive model prediction function, when seasonal variations are added to the trend, i.e. f(t) = T(t) + S(t), where T and S represent, respectively, the trend and seasonal components of the function f, and the multiplicative model, where the trend and seasonal variations are multiplied, i.e. f(t) = T(t) * S(t).

2.2 PRACTICAL CLASSES 2 AND IWS 2

Subject. Building a data model and a trend line for the prediction task.

Lesson plan.

1) Formulate the problem of prediction and, by analogy with the table below, make a table of data for its prediction task.

The period	Time t	Estimated	variable	Y
_	(Months)	(sales)		

1	January	89390
2	February	95970
3	March	97850
4	April	80940
5	May	114760
6	June	89190
7	July	105235
8	August	108920

2) Using Excel tools, build a trend line for this prediction task. It is shown below how to do it. Analyze the results.

To build in Excel charts or diagrams for a given set of input data must be:

- Select the required set of data;
- Using the tab "Paste" option in the "Chart" to find the "Graphics" section;
- Click on the desired type of graphics.

It should be noted that the trend line cannot be added in bulk, petal, circular, annular and stacked charts. For the original data table above chart will look like:



To construct a trend line for this set of data, follow these steps in Excel:

1) Click on the chart to select it;

2) Right-click on the chart, click on the pop-up menu option "format trend line" or "Add trend line";

3) in the Properties window of the trend line, specify parameters such as: the type of trend line forecast (forward or backward) to the desired number of periods, display the equation on the chart, showing squared (coefficient of determination).

The figure below shows a diagram of the forecast values of the predicted variable trend line.



However, the predicted values on the chart gives us a general idea, but not the exact values, i.e., schedule can determine the approximate values of the predicted values predicted variable.

The following topic will be considered standard features predicting and how to obtain specific values of the predicted variable.

Literature.

1. Minko A.A. Prediction in business using Excel. M .: Eksmo, 2007, -208 p.

TOPIC 3

PREDICTING USING STANDARD FUNCTIONS

3.1. Lecture material

In Excel, there are a number of standard functions that allow to solve the prediction problem using known mathematical functions, such as: linear function, logarithmic function, polynomial function, power function, exponential and other functions. Excel features allow you to select any function from the standard functions available in Excel to approximate the original data set of a specific task. Further, based on the selected function, it is possible predict the value of the dependent variable for a certain number of periods in the future. The use of standard Excel functions for solving the prediction problem is discussed below with a specific example.

The initial data set with the predicted "sales" variable is set:

Period(month)	Sales
1	89390
2	85970
3	97850
4	80940
5	114760
6	89190
7	135235
8	158920

When selecting a trend line in Excel, you can specify the following types:

- Linear. The linear trend line is described by the equation Y = mX + b, where Xis the variable-factor, m and b are the computable parameters of the trend line, where m - determines the slope of the line, b - defines the point of intersection of the line with the vertical coordinate axis (ordinate).

The diagram of the linear trend line is shown below in the figure.



Fig. 3.1. Linear trend line.

- Logarithmic. The logarithmic trend line is described by the equation $Y = c \ln(X) + b$, where *c* and *b* are the computable parameters of the logarithmic trend line.





Fig. 3.2. Logarithmic trend line.

- Polynomial. The equation of the polynomial trend line is $Y = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_2 x^2 + c_1 x^1 + b$, where $c_n, c_{n-1}, \dots, c_2, c_1$, b are the computable parameters of the polynomial trend line.

The diagram of the polynomial trend line is shown below in the figure.



Fig. 3.3. Polynomial trend line.

- Power. The equation of the power trend line has the form $Y = cX^b$ where *c*, b are computable parameters of the power trend line.



The diagram of the power line of the trend is shown below in the figure.

Fig. 3.4. Power line trend.

- The exponential. The equation of the exponential trend line has the form $Y = ce^{bX}$ where *c*, b are computable parameters of the exponential trend line.

The diagram of the exponential trend line is shown below in the figure.



Fig. 3.5. Exponential trend line.

Let's make a prediction for the exponential trend line. To do this, right-click on the trend line you need to open the "trend line format" and specify how many periods to make a prediction. In this example, the prediction is for 4 periods ahead.



Fig. 3.6. Forecast by exponential trend line.

However, the chart does not allow you to see the exact values of the predicted values of the predicted variable. You can see only approximate values of the forecast. To obtain accurate prediction values, it is necessary to use the obtained trend equation in the diagram. For example, this will be the exponential equation $y = 73753e^{0,0758x}$. Substituting the values of the forecast periods for x, we will obtain the numerical values of the predicted variable for the given four prediction periods.

To do this, you must create columns in other free columns in the Excel workbook for the predicted periods and sales. In the first element of the "sale" column, you must type the trend equation, then copy it and write (extend) it to the other elements of this column. As a result, the following predicted values for the exponential trend line will be obtained.

Period(month)	Sales
9	145900
10	157389,2
11	169783,1
12	183153

Below are screenshots of obtaining numerical values of the predicted variable for exponential and power trend lines.

_		C [≫] ÷ ME INSI	ERT PAGE L	AYOUT	FORMULAS	DATA	REVIEW	VIEW				lab work1 -	Excel										— ♂ × Sign ir
	👗 Cut 🖺 Copy		Calibri	- 11	- A A	= = =	87 -	Wrap Tex	rt	General	÷			Normal	Bad		Good		₩.	🖹 🏹	∑ AutoSum	` <u>A</u> Ţ	A
Paste	Copy	at Painter	B / U -	189-13	ð - A -		(: +: E	Merge &	Center -	₽ -%,	€000	Conditional				ulation	Check Ce		Insert D	elete Format	↓ Fill *	Sort & F	ind &
				Cont.	_		Alianna			Numbe	1 100 410	Formatting *	Table *		Styles				*	* *	✓ Clear *	Filter * S	elect *
							Alignmer	1.	13	Numbe					styles					Cells		Ealang	
61			× √ f.																				~
_	A		C						I	J	K	L	M	N	0	Р	Q	R	S	T	U	V	W 🔺
1		1	89390 85970		125524,6		145900	•	-														
2		2			128454,6		157389,2						Sales		u = 7275	2 n0.0758x							
3		3	97850 80940		131164		169783,1						Juico		y = 7575.	56-,							
4		+ 5	80940	12	133687,4	12	183153			200000				,									
5		5	89190							160000		_	/										
7	,		135235							140000													
8	1		158920							120000		X											
9		5	100920							100000		\wedge				es							
10										80000		-			Exp	on. (Sales)							
11										60000													
12										40000													
13										20000													
14										0		5	10	15									
15 16										-		-											
17																							
18																							
19																							
20																							
21																							
22																							
23																							
24 25																							
25																							
27																							
28																							
29																							
30		0		-																			
_			Лист2 (÷		_	_					_	_	_	1								•
READY	CALCULA	TE										_									B 🗉		+ 100%
			0		w	X				A real of											• - P* (🕴 att 🌒	

Fig. 3.7. Screenshot of obtaining numerical values of the predicted variable for the exponential trend line

	Come o	Calibri		· A A ·				center -	General	• • • • •	Conditional Formatting *	Format as	Normal Neutral	Bad	ulation	Good Check Ce	•	Insert De	k i	∑ AutoSum ↓ Fill * Clear *	Sort & F	Find &
	Format Painter			·			- merge or	-	<u>.</u> , ,0	.00 4.0	Formatting *	Table *		61 Jun			•		• •		Filter * S	elect *
						Alignmer	nt	1 a	Numo	ver is				Styles				0	ens		Editing	
	- E >	< v .	5r =775	580*D1^0,2	19																	
	A B	С		E		G		I	J	K	L	M	N	0	Р	Q	R	S	Т	U	V	W
	1	89390		125524,6	9	145900																
	2	85970		128454,6		157389,2						Salar										
	3	97850		131164		169783,1						Sales		y = 73753	3e ^{0,0758x}							
	4	80940	12	133687,4	12	183153			200000													
	5	114760							180000 -			/	/									
	6	89190							160000		_	/										
	7	135235							140000													
	8	158920							120000	_	M											
									80000		/ ¥											
				C - 1					60000					Exp	on. (Sales)							
				Sale	s			-	40000													
	180000						y = 7758	Ox ^{0,219}	20000													
	160000								0 -													
	140000			_	\sim				0		5	10	15									
	120000							-														
	100000		-	\sim		_	-Sales															
	80000		Y				— Power (()														
	60000						— Power (sales)														
	40000							-														
	20000							-														
	0							-														
		0 2	4	6	8	10		-														
								_														
6	Лист1 Ј	ист2	(+)											1 4								

Fig. 3.8. Screenshot of obtaining the numerical values of the predicted variable for the power line trend

3.2 PRACTICAL LESSON 3 and IWS 3

Subject. Construction of a table of predictive values for the prediction task

Plan of the lesson.

1) Construct various trend lines for the prediction task formulated in the previous lesson.

2) The task of the IWS. Construct predictive values for different predicted variables. Analyze the results.

Literature.

1. Minko A.A. Prediction in business using Excel. M.: Eksmo, 2007, -208 p.

TOPIC 4

CALCULATION OF THE DETERMINATION COEFFICIENT FOR PREDICTION FUNCTION

4.1. Lecture material

In Section 2, we showed a scheme for solving the prediction task, consisting of two stages: the determination of the approximation function and the determination of the values of the predicted variable from the approximation function by the future values of the independent variables. Naturally, the quality of determining the values of the predicted variable by the future values of independent variables essentially depends on the degree of accuracy of the approximation of the input data by the prediction function. In this section, one of the tools for determining the accuracy of the approximation of the input data by the prediction function is considered, namely, the determination coefficient / 1,2 /. The coefficient of determination shows the degree of accuracy of the approximation of the initial data by the prediction function. The coefficient of determination is denoted as R^2 and is calculated by the formula:

$$R^{2} = 1 - \frac{\sum_{i=1}^{n} e_{i}^{2}}{\sum_{i=1}^{n} (y_{i} - \bar{y})^{2}}$$

Here y_i - *the observed values* of the predicted variable *Y*;

 \overline{y} - *the average value* of y_i : $\overline{y} = \frac{1}{n} \sum_{i=1}^n y_i$. This is the average value of the predicted variable for the initial n points of the variable;

 e_i - residuals or prediction errors defined as $e_i = y_i - f(X_i)$ and $f(X_i)$ are the values of the prediction function at the i-th point of the original data, i.e. residuals are the difference between the observed values and the values of the prediction function, or in other words are errors of approximation of the observed values by the prediction function;

The sum $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(X_i))^2$ is the sum of the squares of the residuals, or the sum of the squares of the prediction (approximation) errors at each point of the original data. Denoted as SSE - the sum of squares of the residuals (sum of squares errors);

The sum in the denominator $\sum_{i=1}^{n} (y_i - \bar{y})^2$, is called *the total sum of squares of deviations* from the mean of the predicted variable (sum of squares total-SST).

It should be noted here that there is another component of the measure of the accuracy of approximation of the observed values by the prediction function. This is the sum of the squares of the regression (sum of squares regression - SSR), this is the sum of the squares of the deviations of the prediction function values from the average observed values of the predicted variable Y.

The sum of squares of the regression SSR is determined by the formula $\sum_{i=1}^{n} (f(X_i) - \bar{y})^2$.

And the sum of SSR and SSE is SST:

SST = SSR + SSE.

Then the coefficient of determination can be represented as:

$$R^2 = 1 - \frac{SSE}{SST}$$

However, the coefficient of determination can also be represented as:

$$R^2 = \frac{SSR}{SST}$$

Then the formula for the determination coefficient will be as follows:

$$R^{2} = \frac{\sum_{i=1}^{n} (f(X_{i}) - \bar{y})^{2}}{\sum_{i=1}^{n} (y_{i} - \bar{y})^{2}}$$

From the formula for determining the coefficient of determination, it is clear that the values of $R^2 \leq 1$. The value of the determination coefficient is $R^2 = 1$ if the sum of the squares of the residues is zero, i.e. when all residuals are zero or when the original data is accurately described by the data model.

The coefficient of determination in Excel can be automatically calculated using the standard functions LINEST and LOGEST /1, 3/. The LINEST function is used to calculate the coefficients of multiple linear or polynomial regressions, and the LOGEST function is used to calculate the coefficients of exponential regression. The LINEST function calculates the coefficients $b_{0,}m_1,...,m_k$ in the equation $Y = b_0 + m_1X_1 + \cdots + m_kX_k$ of the linear multiple regression, or the same coefficients in the

equation $Y = b_0 + m_1 X + \dots + m_k X^k$ of the polynomial regression (from one factor). The LOGEST function calculates the coefficients $b_{0,}, m_1, \dots, m_k$

In the equation $Y = b_0 * m_1^{X_1} * ... * m_k^{X_k}$ of exponential regression.

The syntax for calling these functions is as follows:

= LINEST (Values _Y; {Values _X}; Constant; Statistics)

= LOGEST(Values _Y; {Values _X}; Constant; Statistics)

The argument "Values _Y" is a one-dimensional array of values for the variable Y (or a reference to a range of cells containing this array). The optional argument {Values _X} is an array of X factor values (or a reference to a range of cells containing this array). If this argument is omitted, then it is assumed that this is an array of natural numbers {1,2,3, ...} of the same size as the array "Values _Y". The "Constant" argument indicates whether the coefficient b_0 of linear, polynomial or exponential regression should be equal to 0. If this argument is TRUE, 1 or omitted, the coefficient b_0 is calculated as usual. If the argument is FALSE or 0, then b_0 is set to 0, and the values of the coefficients m_i are selected with this in mind.

The "Statistics" argument indicates whether additional statistical characteristics of the regression are to be calculated. If this argument is TRUE or 1, then the function calculates and displays additional characteristics. If the Statistics argument is FALSE, O or omitted, then the function returns only the values of the coefficients mi and b0. Calculated additional statistical characteristics:

- s_1, s_2, \dots, s_k - mean-square deviations for the coefficients m_1, m_2, \dots, m_k ;

- s_b mean-square deviations for the coefficient b_0 ;
- R^2 coefficient of determination;
- s_{ε} residual standard deviation (standard regression error);
- F criterial statistics for checking the significance of the regression equation;
- df is the degree of freedom;
- SSR the sum of squares of regression (sum of squares regression);
- SSE is the sum of the squares of the residuals (sum of squares error).

Additional statistical characteristics of regression are located in the output array of standard functions LINEST and LGRF as follows:

m_k	m_{k-1}	•••	m_2	m_1	Ьо
S_k	S_{k-1}	•••	S_2	S_1	S_b
R^2	Se				
F	df				
SSR	SSE				

The remaining cells of the output array of additional statistical characteristics of the regression are filled with the values # H / D.

An example of the calculation of the LINEST function of additional statistical characteristics of the regression is shown in Fig.4.1.

To perform the calculations by the LINEST function it is need:

1) select the range F4: I9 (4 columns according to the number of coefficients of the regression equation and 5 lines);

2) without removing selection, enter the formula = LINEST (D4: D14; A4: C14 ;; 1);

3) and then Press the keys <Ctrl + Shift + Enter>.

Additional statistical characteristics of regression for the given example are calculated in the field F4: I9.

The value of the determination coefficient is written in cell F6.

∦ Ci ∏h Ca	it opy ~ irmat Painter	Times New	Ro - 11	ĂĂ	= = =	≫- 8	Wrap Text		General	*	≠	A	Normal	Bad		Good	· ·	÷	× 🗊	∑ AutoSum ↓ Fill *	[™] ^A Z▼	H
📌 Fo	rmat Painter	ΒIU	• 🖽 • 🕹	<u>- A</u> -		€ # Ē	Merge & (Center 🝷	₽ •%,	00. 0. → 0.€ 00.	Conditional Formatting *	Format as Table *	Neutral	Calcu	ulation	Check Ce	T.	Insert De	elete Format	Clear -	Sort & F Filter * S	Find & Select *
Clipbo	ard 5		Font	5		Alignmen	t	5	Number	5				Styles				C	ells		Editing	
	- + E 🕻	× 🗸 .	∫x {=LIN	EST(D4:D1	4;A4:C14;;1)}																
Α	В	С	D	E	F	G	н	I	J	K	L	M	N	0	Р	Q	R	S	Т	U	V	W
alcula	tion of the d	leterminati	on coefficie	ent																		
	Producti on costs		Sales		Addition			eristics of														
	1 1005.8				-2 60053	the reg -0,91065		2301 52	3													
	2 1003,8					1,098388																
	3 1003.6					145,2995																
	4 889,8		1392,6		33,15457		#N/A															
	5 908,8	218,5			2099873	147783,7	#N/A	#N/A														
	6 992,8		1672,9																			
	7 988,3		1660,5																			
	8 975,8		2012,7																			
	9 893,9		2341,9																			
	10 865,1 11 879,6		2413,9 2477,5																			
	11 8/9,0	200,4	2477,5																			
	Коэф.дет	Порици	. Durret	0										4								

Fig. 4.1 Calculation of additional statistical characteristics of regression by the LINEST function.

4.2 PRACTICAL LESSION 4 and IWS 4

Subject: Calculation of the determination coefficient for the prediction problem.

Plan of the lesson.

1) Study the calculation scheme for calculating the determination coefficient for the prediction task, presented above.

2) Practice and task of the IWS. Apply the above described calculation scheme of statistical characteristics for its prediction task.

3) Analyze the results.

Literature.

1. Minko A.A. Prediction in business using Excel. M.: Eksmo, 2007, 208 p.

2. Larose D.T. Discovering knowledge in data - an introduction to data mining. Wiley-Interscience, Hoboken, New Jersey, 2005

3. Excel functions. https://support.office.com/ en-us/article/excel-functions.

TOPIC 5

CALCULATION OF CONFIDENCE INTERVALS FOR PREDICTED VALUES

5.1. Lecture material

Confidence interval for the predicted value is a random interval, which with a given probability α contains an unknown exact value of the prediction function F (X). The probability α is called a confidence level.

The confidence interval is calculated using the following formula /1, 2/:

$$y_p \pm t * s_p,$$

where $y_p = f(x_p)$ is the predicted value of the prediction function f; x_p is the value of the factor X for which it is necessary to calculate the value of the predicted variable Y;

t is the quantile of the order $(1 + \alpha) / 2$ of the Student's distribution with (n-k-1) degree of freedom (k is the degree of the polynomial of the regression function), α is the confidence level;

 s_p is the prediction error for the variable Y at the point y_p , defined by the formula /1, 2/:

$$s_p = s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}},$$

where

 s_{ε} - is the standard error of calculating the regression (calculated by the LINEST function;

 $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ - is the mean of $x_1, x_2, ..., x_n$ of the factor X variable;

 $SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$ - the sum of the squares of the deviations of the factor values from the mean.

The value of the quantile t of the Student's distribution depends on the given confidence level α , usually given as 90, 95 or 99%. Often the confidence level is set at 95%. The use of Student's quantile t imposes a number of conditions on the regression model such as independence and normal distribution of residuals with zero mathematical expectations and the same variances. Often in practice, since the confirmation of these conditions requires enough time resources, an empirical rule of "three sigma" is used.

This rule is not tied to the distribution of residues, and the number 3 is not much larger than the Student's t-test quantile for $\alpha = 95\%$. Then the upper and lower limits of the confidence interval can be calculated by the formula:

$$y_p \pm 3s_p$$
.

5.2 PRACTICAL LESSON 5 and IWS 5

Subject: Calculation of confidence intervals for the prediction task.

Plan of the lesson.

Study the scheme for calculating confidence intervals for the prediction task, presented above in the lecture material.

Below in the figure is an example for Excel, made in accordance with the described scheme of calculating confidence intervals for the task of forecasting sales volumes as a function of time periods, represented as a polynomial of the second degree $Y = b_0 + m_1 t + m_2 t^2$ [1].

	Copy	*				_		tor -	General	- -	Condit	≓ ≢ ional Fo	ermat as	Normal Neutral		Bad Calculation		Good Check Cell	•	Insert De	k 📰	∑ AutoSum ↓ Fill *	× AZ▼ Sort & F		
*	Forma board	at Painter B		Font	· • · = = =		g merge & cer		Number		TOTTIO	ting * T	able *		Style				Ŧ	*	• •	Clear *	Filter * S		
					ST(C2:C11;A2:B11																		,		
F		в	С		E		G	н	I	J		к	L	М		N	0	Р	Q	R	S	т	U	V	
Perio				Predict	Prediction period in square	Predict	Confidence	e interv	al																
	1	1	1282	11		2845,5917																			
	2	4		12		2825,3451																			
	3	9				2715,1517																			
	4	16 25		14	196	2551,1231	2226,9999	2875,2	246																
	5	25			m2	m1	ь0																		
-	7	49			16,60568182																				
	8	64			4,143533743																				
_	9	81			0,965632259		#N/A																		
	10	100			98,33968609																				
					1782930,834	63456,15	#N/A																		
					Average X	5,5																			
					Amount square																				
					St. Deviation 11																				
					St. Deviation 12 St. Deviation 13																				
					St. Deviation 13 St. Deviation 14																				
						100,01100																			
Þ		Коэф.дет	Дов.инт	Лист1	(+)										1 4										

Fig. 5.1 Screenshot of calculating confidence intervals for the task of forecasting sales volumes in relation to time periods.

2) Practice and task of the IWS. Apply the above confidence interval calculation scheme for your prediction task.

3) Analyze the results.

Literature.

1. Minko A.A. Forecasting in business using Excel. M.: Eksmo, 2007, 208 p.

2. Larose D.T. Discovering knowledge in data - an introduction to data mining. Wiley-Interscience, Hoboken, New Jersey, 2005

TOPIC 6

RECOVER MISSING DATA FOR THE PREDICTION TASK

6.1. Lecture material

Missing or missing data in the intellectual analysis is a problem for the data preparation phase. This lack of data can significantly affect the data analysis process. Lack of information is rarely useful. Other things being equal, more data is almost always better. Therefore, various approaches have been developed in this area to solve this problem or to circumvent it.

There are several approaches to solving this problem of missing data [1, 2]:

- replacement of missing data with a certain constant determined by the analyst;

- replacement of missing data by the average value of the field (for numeric variables) or by the mode (for categorical variables);

- replacement of missing data by a value obtained by using a function that approximates the available data.

One of the simple methods for restoring the missing values of the factor-variable is to replace the missing values with the average arithmetic values of the series of values of this factor. For example, if there is no i-th value x_i of the factor X, then $x_i = (x_{i-1} + x_{i+1})/2$ is assumed. This method of recovery has drawbacks. First, it does not work if there are no successive multiple factor values. Secondly, if the data are ordered in time, then the linear dependence of this factor on time is implicitly implied, which is not always true. Thirdly, if it is assumed the linear dependence of the factor on time, then the simple arithmetic mean of the nearest values is a very inaccurate method of approximation [1].

In the second method of recovering missing values, the regression function of the given (dependent) factor with missing values is first constructed based on one or more other factors (independent factors) also present in the data set. Then the missing values of the dependent factor are replaced by the calculated values of the constructed regression function for this factor. The main problem here is the choice of independent factors by which the regression function for the dependent factor will be built.

Independent factors are subject to certain requirements. First, they must correlate with the dependent factor for which the regression function is built. This means that there must be some dependence between them. Secondly, the values of the independent variables should be as deterministic as possible. Should not depend on random influences, especially on random influences that affect the predicted variable Y. However, the latter condition is difficult to verify in practice, therefore, the time factor is often taken as independent variables. But, by their nature, changes of the factor are not always determined only by a temporary factor. The way out of this situation is the construction of several data models, where the missing values are restored based on different independent variables [1].

Consider an example of restoring missing values.

In Figure 6.1. The table of initial data of production costs on the time periods is presented. There are missing values from the production cost data. It is necessary to restore the missing values. We can perform the reconstruction of the missing values using the second method described above, restoring the missing values, based on constructing a factor regression line with missing values for other factors in the original data.

In the figure 6.2. Represents the construction of the regression line of production costs from time, which allows you to calculate the missing values of production costs. To do this, you need to add the regression line formula to the missing value cells. This is shown in Figure 6.3.

See Figure 6.4. The table of initial data of expenses for advertising on the periods of time is presented. There are missing values from the cost of advertising. It is necessary to restore the missing values of advertising costs. Recover missing costs of advertising costs by using the above method of recovering missed values, based on building a line of regression costs for advertising cost factor production.

In Figure 6.5. The construction of the regression line of the dependence of advertising costs on production costs is presented, which allows calculating the missed costs from time to time. To do this, you need to add the regression line formula to the missing value cells. This is shown in Figure 6.6.

		Calibri	- 11	· A A =	= = =	89	🗃 Wrap Tex	rt	General			and a second sec	Normal	Bad		Good		÷	× 🖬	∑ AutoSum	¹ĭ <mark>A</mark> ₹	#
📲 Copy ؇ Forma	at Painter	ΒIU	- == - _	<u>> - A</u> - T		€E ¥E	📑 Merge &	Center 👻	s -%,	€.0 .00 .00 €.0	Conditional Formatting	Format as	Neutral	Calc	ulation	Check Ce	·	Insert De	elete Format	∑ AutoSum ↓ Fill * Clear *	Sort & Filter *	Find & Select *
Clipboard	5		Font	5		Alignm	ient	5	Numbe	r G				Styles					Cells		Editing	
	- 1 2	< 🗸 .	fx 890	,8																		
Α	В	С	D	E	F	G	н	I	J	К	L	М	N	0	Р	Q	R	S	Т	U	V	W
		Producti	Auverusi	Sales								_										
riod		on costs			_			Pro	d costs													
	Январь	890,8			_																	
	Февраль	002	210,5																			
	Март Апрель	903 888,8	208,6		900	•																
	Май	889,8	220,7	1292,0	890	•	****	•														
	Июнь	890,8	238,7		880			* +				!										
_	Июль	888,3	232,3		870			•	* *	•	Pri	od costs										
8	Август	885,8	240,1	2001,7					**	• •												
9	Сентябрь	883,9	238,1	2352,5	860 -						•											
	Октябрь	875,1	248,1		850 -		5	10	15		20											
	Ноябрь			2646,2		,	2	10	15		20											
	Декабрь	873,8		1982,7																		
	Январь	878,2	275,1																			
	Февраль Март	866,3 865,1	265,5 266,1																			
	март Апрель	876,6	282,6																			
	Апрель Май	862,5	282,0																			
	Июнь	865,9	286,3																			
19	Июль	857,5	284,3	2994,1																		
	Post pp at			07600	day 1.1	Kood sore	Die Die	1 Duc	2 Buer2	Duca	1 DucrE	Bucch	0	1								
	вост.пр.да	ан Ана.	пвыоорки	отоор.	фак М	соэф.дете	рм Лист	т Лист	и листэ	ЛИСТ	4 Лист5	Листо	\oplus	1				_	-	■ 町	_	- +

Fig.6.1. Chart of production costs from time.

| Painter | $ \begin{array}{ c c c c c } & & & & & & & & & & & & & & & & & & &$ | | | | |

 | | eneral | eral 🔹 🛃 | | |
 | Bad | | Good
 | | 🖶 🏹 🗒 | | ∑ AutoSum * A
▼ Fill * |
 | |
|---------|---|--|--|---|---
--
--|--
---|--|--|--
--
--|---|--|---|--|--
---|---|---|--|
| | RIU | • 🖽 • 💩 | - <u>A</u> - = | ==+ | e +e [| Merge 8

 | Center - | - % , | 00. 0.€
0.€ 00. | Conditional | Format as | Neutral
 | Calc | ulation | Check Ce
 | | Insert De | lete Format | | Sort &
 | Find & Select * |
| 5 | | Font | 6 | | Alignme |

 | | Number | | ronnatung - | Table |
 | Styles | |
 | | | ells | | Editing
 | Select - |
| | | fx 890,8 | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | 6 |

 | | | K | 1 | |
 | 0 | |
 | | c T | | | v w
 | |
| | | | | F | G | н

 | 1 | J | K | L | M | N
 | 0 | P | Q
 | к | 5 | 1 | 0 | V
 | W |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | 0,000 | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| арт | 903 | | | | |

 | | _ | | | |
 | | |
 | | | | |
 | |
| прель | | | | | |

 | | P | rod co | sts | y = -2,07 | 44x + 900,55
 | | |
 | | | | |
 | |
| ай | 889,8 | | 1667,3 | | | 910

 | | | | | R ² | 0,8529
 | | |
 | | | | |
 | |
| юнь | 890,8 | 238,7 | 1672,9 | | |

 | | | | | |
 | | |
 | | | | |
 | |
| юль | 888,3 | 232,3 | 1667,5 | | | 900

 | | | | | |
 | | |
 | | | | |
 | |
| вгуст | 885,8 | 240,1 | 2001,7 | | | 800

 | | | | | |
 | | |
 | | | | |
 | |
| ентябрь | 883,9 | 238,1 | 2352,5 | | |

 | •••• | ٠. | | | |
 | | |
 | | | | |
 | |
| ктябрь | 875,1 | 248,1 | 2468,5 | | | 880

 | | <u> </u> | • • | | Prod costs |
 | | |
 | | | | |
 | |
| оябрь | | | 2646,2 | | | 070

 | | | | | Linear (Prod costs) |
 | | |
 | | | | |
 | |
| екабрь | 873,8 | | 1982,7 | | | 8/0

 | | | | | |
 | | |
 | | | | |
 | |
| нварь | 878,2 | 275,1 | 1901,1 | | | 860

 | | | | | |
 | | |
 | | | | |
 | |
| евраль | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| арт | | | | | |

 | | 10 | 45 | 20 | |
 | | |
 | | | | |
 | |
| прель | | | | | | 0

 | 5 | 10 | 15 | 20 | |
 | | |
 | | | | |
 | |
| ай | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| юль | 857,5 | 284,3 | 2994,1 | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | | | | | |

 | | | | | |
 | | |
 | | | | |
 | |
| | В
варь
враль
враль
прт
рель
ий
онь
оль
тубрь
ябрь
ябрь
ябрь
варь
враль
враль
враль
враль | B C on costs app. saps. 890,8 sparb. 100,000,000,000,000,000,000,000,000,000 | B C D Producti Adversis 3 варь 80 058 200.8 варь 903 208.6 угг 903 208.6 угг 903 208.6 угг 903 208.6 угг 888.8 223.7 угс 885.8 240.1 инв 875.1 248.1 ябрь 877.2 275.1 варь 867.3 205.5 рогнь 886.3 232.5 угс 885.5 240.1 лабрь 873.2 275.1 варь 867.3 265.5 уго 85.6 245.5 уго 85.6 265.5 уго 865.6 282.6 уго 865.5 266.5 уго 865.5 265.5 уго 865.5 265.5 уго 865.5 265.5 уго 865.5 265.5 | B C D E Producti Advertisi Sales e.or.cost ng.cost volume baps. gc.ost volume baps. 210,5 1282,7 prir 903 208,6 1224,8 pens. 889,8 210,5 1223,1 wit 889,8 1667,3 167,3 prir 903 238,1 232,3 pins 875,1 248,1 2468,5 adops 875,2 275,1 1982,7 paps. 875,2 275,5 1971,6 pins 866,3 265,5 1971,6 pins 867,6 282,6 2138,2 pins 867,6 282,6 2138,2 pins 867,6 282,6 2138,1 pins 865,3 265,5 1971,6 pins 865,3 265,5 247,5 pins 865,6 285,6 2138,5 pins | B C D E F Producti Advertis Sales < | B C D E F G Producti Advertili Sales G G G Solution G G Solution G Solution Solution <td>B C D E F G H Producti Advertis Sales H</td> <td>B C D E F G H I Producti Advertii Sales no sol sol no sol sol</td> <td>B C D E F G H I J Producti Advertii Sales no. cost no. no.</td> <td>B C D E F G H I J K Producti Advertis Sales X</td> <td>B C D E F G H I J K L Producti Advertili Sales</td> <td>B C D E F G H I J K L M Producti Advertis Sales Image costs Volume Producti M Image costs M <th< td=""><td>B C D E F G H I J K L M N Producti Advertis Sales Olume Image Sales Image</td><td>B C D E F G H I J K L M N O Producti Adversiti Sales I J K L M N O encorts ng costs volume Image costs volume volume Image costs volume volume volume Image costs volume volume</td><td>B C D E F G H I J K L M O P Producti Adversiti Sales I J K L M O P B O Posteria Sales I J K L M O P B O O D E F G H I J K L M O P B O D <thd< th=""> <thd< th=""> <thd< th=""> D</thd<></thd<></thd<></td><td>B C D E F G H I J K L M N O P Q Producti Adversiti Sales I J K L M N O P Q B O Posts It J K L M N O P Q B O Dots Iteration Iteration</td><td>B C D E F G H I J K L M N O P Q R Producti Adversiti Sales I J K L M N O P Q R B O Poducti Adversiti Sales Image costs Image c</td><td>B C D E F G H I J K L M N O P Q R S Producti Advertis Sales O P Q R S Producti Advertis Sales O P Q R S Baps Bgosts Outside P P Q R S pane 2105 1282,7 P P P P Q R S with 888,8 232,7 1722,6 P P P Q R S with 888,8 232,3 1667,3 P Q R S with 888,3 232,3 1667,3 P Q R S Q P Q R M with 888,3 232,1 2352,5 P P Q P Q <t< td=""><td>B C D E F G H I J K L M N O P Q R S T Producti Advertis Sales Image official Sales Image official Image official Sales Image official <thimage official<="" th=""> <thimage official<="" th=""></thimage></thimage></td><td>B C D E F G H I J K L M N O P Q R S T U Producti Advertis Sales Image: Sales F G H I J K L M N O P Q R S T U Producti Advertis Sales Value Image: Sales Image: Sales</td><td>B C D E F G H I J K L M N O P Q R S T U V Producti Advertis Sales Value Image bits Value Value</td></t<></td></th<></td> | B C D E F G H Producti Advertis Sales H | B C D E F G H I Producti Advertii Sales no sol sol no sol sol | B C D E F G H I J Producti Advertii Sales no. cost no. no. | B C D E F G H I J K Producti Advertis Sales X | B C D E F G H I J K L Producti Advertili Sales | B C D E F G H I J K L M Producti Advertis Sales Image costs Volume Producti M Image costs M <th< td=""><td>B C D E F G H I J K L M N Producti Advertis Sales Olume Image Sales Image</td><td>B C D E F G H I J K L M N O Producti Adversiti Sales I J K L M N O encorts ng costs volume Image costs volume volume Image costs volume volume volume Image costs volume volume</td><td>B C D E F G H I J K L M O P Producti Adversiti Sales I J K L M O P B O Posteria Sales I J K L M O P B O O D E F G H I J K L M O P B O D <thd< th=""> <thd< th=""> <thd< th=""> D</thd<></thd<></thd<></td><td>B C D E F G H I J K L M N O P Q Producti Adversiti Sales I J K L M N O P Q B O Posts It J K L M N O P Q B O Dots Iteration Iteration</td><td>B C D E F G H I J K L M N O P Q R Producti Adversiti Sales I J K L M N O P Q R B O Poducti Adversiti Sales Image costs Image c</td><td>B C D E F G H I J K L M N O P Q R S Producti Advertis Sales O P Q R S Producti Advertis Sales O P Q R S Baps Bgosts Outside P P Q R S pane 2105 1282,7 P P P P Q R S with 888,8 232,7 1722,6 P P P Q R S with 888,8 232,3 1667,3 P Q R S with 888,3 232,3 1667,3 P Q R S Q P Q R M with 888,3 232,1 2352,5 P P Q P Q <t< td=""><td>B C D E F G H I J K L M N O P Q R S T Producti Advertis Sales Image official Sales Image official Image official Sales Image official <thimage official<="" th=""> <thimage official<="" th=""></thimage></thimage></td><td>B C D E F G H I J K L M N O P Q R S T U Producti Advertis Sales Image: Sales F G H I J K L M N O P Q R S T U Producti Advertis Sales Value Image: Sales Image: Sales</td><td>B C D E F G H I J K L M N O P Q R S T U V Producti Advertis Sales Value Image bits Value Value</td></t<></td></th<> | B C D E F G H I J K L M N Producti Advertis Sales Olume Image Sales Image | B C D E F G H I J K L M N O Producti Adversiti Sales I J K L M N O encorts ng costs volume Image costs volume volume Image costs volume volume volume Image costs volume volume | B C D E F G H I J K L M O P Producti Adversiti Sales I J K L M O P B O Posteria Sales I J K L M O P B O O D E F G H I J K L M O P B O D <thd< th=""> <thd< th=""> <thd< th=""> D</thd<></thd<></thd<> | B C D E F G H I J K L M N O P Q Producti Adversiti Sales I J K L M N O P Q B O Posts It J K L M N O P Q B O Dots Iteration Iteration | B C D E F G H I J K L M N O P Q R Producti Adversiti Sales I J K L M N O P Q R B O Poducti Adversiti Sales Image costs Image c | B C D E F G H I J K L M N O P Q R S Producti Advertis Sales O P Q R S Producti Advertis Sales O P Q R S Baps Bgosts Outside P P Q R S pane 2105 1282,7 P P P P Q R S with 888,8 232,7 1722,6 P P P Q R S with 888,8 232,3 1667,3 P Q R S with 888,3 232,3 1667,3 P Q R S Q P Q R M with 888,3 232,1 2352,5 P P Q P Q <t< td=""><td>B C D E F G H I J K L M N O P Q R S T Producti Advertis Sales Image official Sales Image official Image official Sales Image official <thimage official<="" th=""> <thimage official<="" th=""></thimage></thimage></td><td>B C D E F G H I J K L M N O P Q R S T U Producti Advertis Sales Image: Sales F G H I J K L M N O P Q R S T U Producti Advertis Sales Value Image: Sales Image: Sales</td><td>B C D E F G H I J K L M N O P Q R S T U V Producti Advertis Sales Value Image bits Value Value</td></t<> | B C D E F G H I J K L M N O P Q R S T Producti Advertis Sales Image official Sales Image official Image official Sales Image official Image official <thimage official<="" th=""> <thimage official<="" th=""></thimage></thimage> | B C D E F G H I J K L M N O P Q R S T U Producti Advertis Sales Image: Sales F G H I J K L M N O P Q R S T U Producti Advertis Sales Value Image: Sales Image: Sales | B C D E F G H I J K L M N O P Q R S T U V Producti Advertis Sales Value Image bits Value Value |

Fig. 6.2. The regression line of production costs from time.

D		Calibri	- 11	· A A	= = =	87 -	🔐 Wrap	lext.	Genera	al	-	≢		Normal	Bad		Good	*	÷ .	× 🗊	∑ AutoSur ↓ Fill *	° [™] <mark>A</mark> T	A l
I Cop I Cop	nat Painter	ΒIU	• 🖽 • 🕹	<u>- A</u> -	= = =	€E S E	📑 Merge	& Center 🝷	S • 9	% ,	00. 0. → 0.€ 00.	Conditional Formatting *	Format as Table *	Neutral	Calc	ulation	Check Co	ell 🗸	Insert D	elete Format	Clear •	Sort & Filter ≠ !	Find & Select *
Clipboar	d a		Font	5		Alignr	nent	5	- N	Number	r _i				Styles					Cells		Editing	
	* E 💈	$\langle \vee \rangle$	fx =-2,0	0744*A3+9	00,55																		
А	В	С	D	E	F	G	Н	I		J	К	L	M	N	0	Р	Q	R	S	т	U	V	w
			Advertisi	Sales																			
		on costs		volume																			
	Январь	890,8																					
	Февраль	896,4012	210,5	1282,7																			
	Март	903		1224,8						Pr	od co	osts	y = -2,0	743x + 900,55									
	Апрель	888,8		1292,6						•••			R ²	= 0,8681									
	Май	889,8		1667,3			910																
	Июнь	890,8		1672,9			900	•															
	Июль	888,3 885,8		1667,5 2001,7				*															
	Август	883,9		2352,5			890 -	· • • •	•														
	Сентябрь Октябрь	883,9	238,1	2352,5			880		~				Prod costs										
	Ноябрь	877,7316	240,1	2408,5			000		••••			•	Linear (Prod costs)										
	Декабрь	873.8		1982.7			870						Line	ar (Prod costs)	' <u> </u>								
	Январь	878,2		1902,7	-		850					• •											
	Февраль	866,3	265,5	1971,6						•													
	Март	865,1	266,1	1989,1	9,1		850					20											
16	Апрель	876,6	282,6	5 2138,2			0	5		10	15												
17	Май	862,5	286,5	2475,2	1		L																
18	Июнь	865,9	286,3	2395,6																			
19	Июль	857,5	284,3	2994,1																			
	Вост.пр.д	ан Ана	л.выборки	Отбор	о.фак К	Коэф.дет	ерм Л	ист1 Ли	ст2 Л	ист3	Лист4	Лист5	Лист6	+ :	4								
1																				m	🖹 🗉 🗕	_	-+

Fig. 6.3. Calculation of the missed values of production costs for the constructed line of regression of production costs from time.



Fig. 6.4. Spot chart of advertising costs from production costs

$\begin{array}{c c} & & & \\ & & & & \\ & & & \\ & & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ &$				xt	General	÷			Normal	Normal Bad		Good	Good		i	Σ AutoSum * AZ▼		#				
🖺 O 💣 Fr	opy * ormat Painter	BIU									Conditional Formatting	Format as	Neutral	Cal	culation	Check Co	ell v	Insert D	elete Format	👽 Fill 👻	Sort & F	ind &
lipbo			Font	6		Align	ment	5	Number	r G		Table *		Styles				, T.,	Cells		Filter * Se Editing	elect *
	+ 1 3	< 🗸	fx 889.8																			
A	В	C	D	E	F	G	н	I	J	K	L	м	N	0	Р	Q	R	S	T	U	V	W
iod	Time	on costs	Advertisi																			
iou	1 Январь	890,8																				
	2 Февраль	896,4																				
	3 Март	903																				
	4 Апрель	888,8																				
-	5 Май	889,8																				
_	6 Июнь	890,8	238,7							Adv o	costs fror	n prod	costs									
	7 Июль	888,3	232,3					350														
	8 Август	885,8	240,1																			
	9 Сентябрь	883,9	238,1					300		• •												
	10 Октябрь	875,1	248,1					250			••											
	11 Ноябрь	877,73						200														
	12 Декабрь	873,8						150														
	13 Январь	878,2	275,1																			
	14 Февраль	866,3	265,5					100														
	15 Март	865,1	266,1					50														
	16 Апрель	876,6	282,6					0														
	17 Май	862,5	286,5					850	860	8	870 81	80	890	900	910							
	18 Июнь	865,9	286,3						 Adv cost 	ts from pro	nd costs —	— Linear (A	dv costs from	nrod costs)								
	19 Июль	857,5	284,3																			
	Вост.пр.д	ан Ана	л.выборки	Отбо	р.фак К	оэф.дет	ерм Лис	т1 Лис	т2 Лист3	Лист	4 Лист5	Лист6	(+)	1								
				_															Ħ	🖹 🗉 🗕 -		

Fig. 6.5. The regression line of the dependence of advertising costs on production costs.



Fig. 6.6. Calculation of the missed costs of advertising costs for the line of regression of advertising costs from production costs.

6.2 PRACTICAL LESSON 6 and IWS 6

Subject. Recover missing data for the prediction task.

Plan of the lesson.

1) Study the scheme for recovering missing data for the prediction task, presented above.

2) Practice and task of the IWS. Apply the above-described scheme for recovering missing data for its prediction task.

3) Analyze the results.

Literature.

1. Minko A.A. Forecasting in business using Excel. M.: Eksmo, 2007, 208 p.

2. Larose D.T. Discovering knowledge in data - an introduction to data mining. Wiley-Interscience, Hoboken, New Jersey, 2005
TOPIC 7

ANALYSIS OF DATA OUTLIERS FOR THE PREDICTION TASK

7.1. Lecture material

Data outliers are extreme values that lie near the limits of the data range or go against the trend of the remaining data. Identifying data outliers is important because they may represent errors in data entry. The task of the data outliers analysis is to find out whether the data outliers are values of "natural" origin, i.e. whether they are caused by the stochastic nature of the initial data or are conditioned by some unaccounted factors. If the data outliers are due to the presence of unknown factors, then it is necessary to deal with this situation, to identify the cause of the unknown factor. This can serve as an incentive for investigating the actual situation described by the initial data, and also for further improving the data model [1, 2].

There are several methods of data outliers detection.

One elementary robust method is to use the interquartile range [2]. The quartiles of a data set divide the data set into four parts, each containing 25% of the data:

-the first quartile (Q1) is the 25th percentile;

-the second quartile (Q2) is the 50th percentile, that is, the median;

-the third quartile (Q3) is the 75th percentile.

The interquartile range (IQR) is a measure of variability that is much more robust than the standard deviation.

The IQR is calculated as IQR = Q3 - Q1 and may be interpreted to represent the spread of the middle 50% of the data. A robust measure of data releases detection is therefore defined as follows. A data value is an outlier if:

1) It is located 1.5(IQR) or more below Q1, or

2) It is located 1.5(IQR) or more above Q3.

For example, suppose that for a set of test scores, the 25th percentile was Q1 = 165 and the 75th percentile was Q3 = 170, so that half of all the test scores fell between 165 and 170. Then the interquartile range, the difference between these quartiles, was IQR = 170 - 165 = 5.

A test score would be robustly identified as an data release if:

a. It is lower than Q1 – 1.5(IQR) = 165 – 1.5*(5) = 157,5, or

b. It is higher than Q3 + 1.5(IQR) = 170 + 1.5(5) = 177,5.

Other method of data outliers detection using regression function is described below [1].

The scheme for determining the data outliers includes the following steps:

1) the data outliers points x_i are removed from the initial data set;

2) the regression function f (X) is constructed from the remaining data and its statistical characteristics are calculated using the standard LINEST function (including the standard error s_{ε} of the constructed regression);

3) find the value of the regression function at the data outliers points f (X).

4) calculate the residuals at data releases points $e_i = y_i - f(X_i)$, where y_i - is the value of the i-th outlier.

5) find the normalized residues at the data outliers points $e_i = e_i / s_{\varepsilon}$..

6) If the absolute value of the normalized residue exceeds the number 3, it is considered that with a probability of 95% this data outliers is not accidental.

Let us consider an example of determining the randomness of data outliers.

The set of initial data is given, presented in Table 7.1 and in Figure 7.1.

Time	Period	Production	Advertising	Sales
		costs	costs	volume
January	1	890,8	200,8	1286,5
February	2	896,4	210,5	1282,7
March	3	903	208,6	1224,8
April	4	888,8	226,7	1292,6
May	5	889,8	230,62	1667,3
June	6	890,8	238,7	1672,9
July	7	888,3	232,3	1667,5
August	8	885,8	240,1	2001,7
September	9	883,9	238,1	2352,5
October	10	875,1	248,1	2468,5
November	11	877,73	253,58	2946,2
December	12	873,8	261,05	1982,7
January	13	878,2	275,1	1901,1
February	14	866,3	265,5	1971,6
March	15	865,1	266,1	1989,1
April	16	876,6	282,6	2138,2
May	17	862,5	286,5	2475,2
June	18	865,9	286,3	2395,6

Table 7.1. The set of initial data

July	19	857,5	284,3	2994,1

Cut							🔐 Wrap Tex		General				Normal	Bad		Good	· ·		× 📰	∑ AutoSum *	Ž Y H	
		ΒIU	• 🎛 • 🏅	<u>> - A</u> - I		€E 4E	🔛 Merge &	Center 🝷	₽ •%,	00, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	Formatting *	Format as Table *	Neutral		ulation	Check Co	zll v			Clear -		
Clipboard						Alig	nment	5	Number	5				Styles				Ce	lls	Ed	liting	
		X 🗸	<i>fx</i> 128	5,5																		
Α	В	С		E	F	G	Н	I	J	K	L	М	N	0	Р	Q	R	S 🔺	Form	at Shape		
-			Advertisi																FOIIII	lat shape		
Тіте Январь	Period	on costs																	3			
январь Февраль	2	890,8 896,4			_																	
Февраль Март	3							Sa	ales volum	е									> FILL			
Апрель	4		208,0		3	500													▲ LINE			
Май	5					000						_										
Июнь	6		238,7							\		1								lo line		
Июль	7	888,3	232,3	1667,5	2	500			1			\sim							0 5	olid line		
Август	8	885,8	240,1	2001,7	2	000				-									<u> </u>	radient line		
Сентябрь	9	883,9	238,1	2352,5	1	500																
Октябрь	10	875,1	248,1	2468,5	1	.000																
ноябрь	11					500																
Декабрь	12		261,05			0																
Январь	13		275,1	1901,1			2 3 4 5	6 7	8 9 10 11	12 13	14 15 16	17 18 19										
Февраль	14		265,5						Sales volum													
Март	15		266,1					-	 Sales volum 	e												
Апрель Май	10		282,6																			
Июнь	17		286,3																			
Июль	19																					
		00110	201/0	200 (/2																		
			ист3								1							•				

Fig. 7.1. Initial data with data outliers in the 11th period.

In the diagram of Fig. 7.1. One significant data outliers is seen in the 11th period.

1) In accordance with the scheme for determining the accidental data outliers, it is necessary to delete data outliers. For example, the outlier of period 11 is deleted and its data line is stored in free line 20 (see Figure 7.2.).

2) A regression function is constructed for the data of the predicted variable without a deleted data outliers (see Figure 7.2.). The statistical characteristics of the regression are calculated using the standard LINEST function, represented in cells G2: J6 (see Figure 7.2).

	Cut Copy *		Calibri					Wrap Text		General	•	ŧ	and the second s	Normal	Bad		Good	- •		× 🗊	∑ AutoSur ↓ Fill •			
*	Format P			• = - 4	<u>-</u> - <u>A</u> -	= = =			enter 👻	₽ -%,	FC	onditional prmatting *	Format as Table *	Neutral		lation	Check C	ell 🗸	Insert De	elete Format	Clear •		Find & Select *	
Clip	board	5		Font	5		Alignmer	t	5	Number	5				Styles				0	Cells		Editing		
	*	1 2	K 🗸	$f_x = \{=LIN\}$	IEST(E2:E1	9;B2:D19;;	1)}																	
F	A	В	С	D	E	F	G	н	Ι	J	К	L	M	N	0	Р	Q	R	S	Т	U	V	W	
				Advertisi																				
me		riod			volume	_																		
нва		1	890,8	200,8	1286,5					6 11526,03														
	аль	2	896,4	210,5	1282,7					4 12669,14														
lapı		3	903	208,6	1224,8			270,6237		#N/A														
npe		4	888,8	226,7	1292,6		14,05246			#N/A				Sale	s volum	ie								
ай		5	005/0	230,62 238,7	1667,3		308/48/	1025320	#N/A	#N/A	350	0				v =	79,973x + 11	171.6						
юн		0	890,8	238,7	1672,9		Malua faat	the 11th pe	de d		300					,-	13,3138 + 1	.,						
_		/	,			-	2012,158		riod							7								
вгус		8	885,8 883,9	240,1 238,1	2001,7			r the 11th p	and and		250	0		*	M	_								
ент Іктя	ябрь	10		238,1	2352,5		934.0422		period		200	0					Sales volun	ne						
	брь	10		246,1	1982,7	-			o for the	11th period	150	0	100											
нва		12		201,03	1982,7		3,451443		le for the	11th period	100	n 🔫	-				- Linear (Sale	es volume)						
	аль	13	,	265,5	1971,6		5,431445																	
Ларт		15		266,1	1989,1						50													
пре		16		282,6	2138,2							0 +				_								
ай	/10	17		286,5	2475,2							123	4567	8 9 10 11 12	131415161/	18								
юн	h	18		286,3	2395,6																			
юл	ь	19	857,5	284,3	2994,1																			
оя	БРЬ	11		253,58	2946,2																			
2	Ли	cr1	Лист2 Л	ист3	(+)										4									
1					0																			

Fig. 7.2. Calculation of the statistical characteristics of the regression without a deleted data outliers.

3) The values of the predicted variable for the deleted data outliers (see Figure 7.3) are calculated from the parameters of the regression: $Y = b_0 + b_1 t + b_2 X_1 + b_3 X_2$ (formula in Excel: =J2+I2*B20+H2*C20+G2*D20).

Cut	-	Calibri B I U		• A A =			Wrap Text Merge & Ci		eneral	v 00, 00	Condition	hal Format as	Normal Neutral	Bad Calcula	tion	Good Check C	ell v		elete Format	∑ AutoSun ↓ Fill *	Sort &	
Clipboard			Font	6		Alignmen		6	Number	6	Formattin	g * Table *		Styles				· .	cells		Filter * 1 Editing	Select *
	• E >	 V 	<i>fx</i> =J2+	12*B20+H2*	*C20+G2*	D20																
Α	В	С	D	E	F	G	н	Ι	J	К	L	M	N	0	Ρ	Q	R	S	Т	U	V	W
		Producti	Advertisi	Sales																		
Time	Period	on costs	ng costs	volume																		
нварь	1	890,8	200,8	1286,5		-10,1747	-9,17358	101,6506	11526,03													
Февраль	2	896,4	210,5	1282,7		11,78952	14,5223	71,10544	12669,14													
Ларт	3	903	208,6	1224,8		0,750701	270,6237	#N/A	#N/A									1 I				
Апрель	4	888,8	226,7	1292,6		14,05246	14	#N/A	#N/A				Sale	s volume	•							
Иай	5	889,8	230,62	1667,3		3087487	1025320	#N/A	#N/A													
1юнь	6	890,8	238,7	1672,9						35	00				y -	79,973x + 1	171,6					
1юль	7	888,3	232,3	1667,5		Value for t	he 11th per	iod		30	00			,	-							
Август	8	885,8	240,1	2001,7		2012,158				29	00			- t								
Сентябрь	9	883,9		2352,5		Residue fo	r the 11th p	eriod					1									
Октябрь	10	875,1	248,1	2468,5		934,0422				20	00				· -+	-Sales volur	ne					
Декабрь	12	873,8		1982,7		The norma		e for the 1	1th period	15	00					– Linear (Sale	(o mulum o					
Анварь	13	878,2	275,1	1901,1		3,451443	incear condu	e for the 1	2 mperiod	10	100				. –	- Linear (Said	es volume)					
Февраль	14	866,3	· · · ·	1971,6		3,431443																
иарт	15	865,1	265,5	1989,1							i00											
Апрель	15	876,6		2138.2							0 +				1							
апрель Иай	10	862,5		2475,2							1 2	34567	8 9 10 11 12	131415161718	5							
лан 1юнь	17	865,9		2395,6																		
1юль	18	857,5		2393,0																		
ЮЯБРЬ	19	877,73		2994,1																		
ЮЛБРБ	11	0/1,/3	255,56	2940,2																		
		ист2 Л		(+)										1								

Fig. 7.3. Calculation of the value of the predicted variable for remote emission by regression parameters.

4) The residuals at emission points $e_i = y_i - f(X_i)$, are calculated, where y_i - is the value of the i-th outlier (see Figure 7.4).



Fig. 7.4. Calculation of residues at data outliers points.

5) Calculation of normalized residues at emission points $e_i = e_i / s_{\varepsilon}$ for example (see Figure 7.5).

Cut		Calibri		· A A			Wrap Text		eneral	*	ŧ		Normal	Bad		Good	· ·	ÈE E	×	∑ AutoSun ↓ Fill *		
💞 Forma	t Painter	ΒIU	• 🖽 • 🟅	<u>> - A</u> - I		€ #2 🗄	Merge & Ci	enter 🍷 🔮	- % ,	€.0 .00 0.€ 00.	Condition: Formatting	al Formatas 1 * Table *	Neutral	Calcul	ation	Check C	el 🗸	Insert D	elete Format	Clear -	Sort & Filter ▼ :	
Clipboard	5		Font	6		Alignmen		5	Number	G.				Styles					Cells		Editing	
	2	K 🗸	<i>fx</i> =61	1/H4																		
Α	В	С	D	E	F	G	н	I	J	К	L	M	N	0	Ρ	Q	R	S	Т	U	V	W
		Producti	Advertisi	Sales																		
ime I	Period	on costs	ng costs	volume																		
нварь	1	890,8					-9,17358															
Ревраль	2		210,5				14,5223															
Ларт	3						270,6237	#N/A	#N/A									1				
прель	4	888,8	226,7			14,05246	14		#N/A				Sale	s volum	e							
Лай	5	00010				3087487	1025320	#N/A	#N/A	3	00					79,973x + 1	171.6					
юнь	6		238,7			Value for t		te d			00				y -	79,9758+1	171,0	-				
1юль	/ 8		232,3	2001,7		2012,158	ne 11th per	100						2								
вгуст ентябрь	8		240,1	2352,5		Residue fo	sho 11th e	aried			00		1	M								
ентяорь Эктябрь	10		238,1	2352,5		934,0422		enou		2	00					-Sales volur	ne					
lекабрь	10					The norma		e for the 1	1th period	1	00	-	5			- Linear (Sal						
нварь	13		275,1			3,451443	lizeu resiuu	e for the I	rinperiou	10	100	**				– Linear (Sai	es volume)					
евраль	13		265,5			5,152115					00											
Ларт	15		266,1																			
прель	16										0 +		8 0 101111	13141516171	0							
Лай	17	862,5	286,5	2475,2							1 2	34507	0 9 10 11 12	13141310171	0							
іюнь	18	865,9	286,3	2395,6																		
1юль	19	857,5	284,3	2994,1																		
ЮЯБРЬ	11	877,73	253,58	2946,2																		
	luces 1	Лист2 Л	lucr3	(+)										1								

Fig. 7.5. Calculation of normalized residues at data outliers points.

6) If the absolute value of the normalized residue exceeds the number 3, it is considered that with a probability of 95% this data outliers is not accidental.

In this example, the value of the predicted variable of the 11th period is not an accidental outlier, because the value of the normalized remainder is 3.45.

An analysis of the results of a randomized data outliers test includes a sufficiently in-depth analysis of the necessary actions in the event that data outliers are not accidental. Often this is a sign that there is an unaccounted factor.

If, as a result of the analysis, it turns out that these factors will not affect the predicted periods, such data outliers can be removed from the original data set.

If the data outliers are of a recurring nature, then they must be taken into account when allocating the seasonal component of the forecast.

If data outliers can appear in the future, but irregularly (for example, special promotions), these data releases can be taken into account by including additional factors in the data model.

7.2 PRACTICAL LESSON 7 and IWS 7

Subject: Analysis of data outliers for the predicting problem.

Plan of the lesson.

1) Examine the data outlier analysis scheme for the forecasting problem presented above.

2) Practice and task of the IWS. Apply the above described data outliers analysis scheme to your predicting task.

3) Analyze the results.

Literature.

1. Minko A.A. Predicting in business using Excel. Moscow: Eksmo, 2007, -208 p.

2. Larose D.T. Discovering knowledge in data - an introduction to data mining. Wiley-Interscience, Hoboken, New Jersey, 2005

TOPIC 8

SELECTION OF FACTORS FOR THE PREDICTION TASK

8.1 Lecture material

The task of selecting factors is as follows: it is necessary to select such factors in the data model that their inclusion increases the accuracy of the approximation of the initial data by the predicting function.

One of the most effective methods of selection of factors is the use of the adjusted coefficient of determination.

Earlier, we considered in detail the coefficient of determination, when the question of the accuracy of the approximation of the initial data by the predicting function was studied. The **coefficient of determination is convenient when estimating the accuracy of approximation by different prediction functions of the same initial data with the same set of factors**.

When it is necessary to compare the accuracy of approximation with the function of predicting the initial data with a different set of factors, and the task of selecting factors can be put in this way, then it is possible to use the adjusted coefficient of determination.

That is, you can first calculate the adjusted determination coefficient for a model with a given set of factors, and then remove one of the factors from the model and calculate the adjusted determination coefficient.

Then compare the obtained values of the adjusted determination coefficient: if elimination of the factor reduces the quality of the approximation, then this factor is significant for the model, and vice versa, if the quality of the approximation does not change or even increases, this means that this factor is not essential for the model.

The adjusted coefficient of determination $\overline{R_m^2}$ is calculated by the formula:

$$\overline{R_m^2} = \frac{(n-1)(1-R_m^2)}{n-m}$$
(8.1)

Where R_m^2 - "standard" coefficient of determination; *m*-number of factors; *n* is the number of data points in the source data set.

The adjusted coefficient of determination has the property that if k new factors are added to the model, then the new adjusted determination coefficient $\overline{R_{m+k}^2}$ can be either larger or smaller than the old adjusted coefficient of determination $\overline{R_m^2}$. If $\overline{R_{m+k}^2}$ is greater than $\overline{R_m^2}$, then it is considered that k new factors significantly affect the predicted variable Y, and these factors need to be added to the model. Otherwise (when $\overline{R_{m+k}^2}$ is less than or equal to $\overline{R_m^2}$, it is assumed that the new factors have little effect on the predicted variable, and therefore they can not be included in the model [1].

Below is an example of an assessment of the inclusion of a factor in the forecasting model. First, we consider a model with three factors:

$$Y = b_0 + b_1 t + b_2 X_1 + b_3 X_2 \tag{8.2.}$$

Then we add the factor of the time period square to this model, having received the model with four factors:

$$Y = b_0 + b_1 t + b_2 t^2 + b_3 X_1 + b_4 X_2$$
(8.3.)

Then, from this model, we remove the factor X_2 and get a model with three factors:

$$Y = b_0 + b_1 t + b_2 t^2 + b_3 X_1 \tag{8.4}$$

For each of these models, we calculate their statistical characteristics and the adjusted determination coefficient. Further, by comparing their values, it is

possible to draw conclusions about the importance of the included or excluded factors for the forecasting function of the problem under consideration.

Below in Tables 8.1, 8.2 and Figure 8.1, the initial data and the results of calculating the estimates for the model (8.2) are presented.

Time	Period	Production	Advertising	Sales
		costs	costs	volume
January	1	890,8	200,8	1286,5
February	2	896,4	210,5	1282,7
March	3	903	208,6	1224,8
April	4	888,8	226,7	1292,6
May	5	889,8	230,62	1667,3
June	6	890,8	238,7	1672,9
July	7	888,3	232,3	1667,5
August	8	885,8	240,1	2001,7
September	9	883,9	238,1	2352,5
October	10	875,1	248,1	2468,5
November	11	877,73	253,58	2012,2
December	12	873,8	261,05	1982,7
January	13	878,2	275,1	1901,1
February	14	866,3	265,5	1971,6
March	15	865,1	266,1	1989,1
April	16	876,6	282,6	2138,2
May	17	862,5	286,5	2475,2
June	18	865,9	286,3	2395,6
July	19	857,5	284,3	2994,1

Table 8.1. The initial data for the model (8.2)

Table 8.2. The results of calculating estimates from the model (8.2).

Number o	f factors -		3		
Regressior	n equation	Y=b0+b1*t	+b2*X1+b3	*X2	
-10,1748	-9,17357	101,6508	11526,03		
11,38783	14,02972	68,67564	12239,55		
0,751075	261,4473	#Н/Д	#Н/Д		
15,08639	15	#Н/Д	#Н/Д		
3093677	1025320	#Н/Д	#Н/Д		

The adjust	ed coefficie	ent of deter	mination	
0,71996				

	🕻 Cut 👌 Copy	-	Calibri	• 11 •	A A =	=	≫- ≣	Wrap Text		Seneral	•			Normal	Bad		Good	- -			∑ AutoSum ↓ Fill *	T AZ▼	H
, 1	🕈 Form	at Painter	ΒIU	- 🖽 - 🙅	• <mark>A</mark> • =	23	€ # []	Merge & C	enter -	* * % *	00, 0, 4 0,€ 00,	Formatting *	Table *	Neutral	Calc	ulation	Check C	Ŧ	*	* *	Clear 🕶	Filter *	Select *
								t	5	Number	5				Styles					iells		Editing	
				f_x {=LINE																			
				E	F	G	Н	I	J	K	L	M	N	0	Р	Q	R	S	Т	U	V	W	Х
			Advertisi ng costs																				
	1				N	lumbar of	factors			2													
	2							Y=b0+b1*t															
-	3	,						101,6508															
	4		,					68,67564															
	5	889,8	230,62				261,4473		#N/A														
	6	890,8	238,7	1672,9	1	15,08639	15	#N/A	#N/A														
	7	888,3	232,3	1667,5		3093677	1025320	#N/A	#N/A														
	8			2001,7																			
_	9	000/5		2352,5																			
_	10			2468,5			d coefficie	ent of deter	rmination														
	11					0,71996																	
_	12			1982,7 1901,1																			
-	15			1901,1																			
-	15			1989,1																			
-	16			2138,2																			
	17			2475,2																			
	18	865,9	286,3	2395,6																			
	19	857,5	284,3	2994,1																			
Þ		Лист1	Пист2 Л	ист3 Лист	т4 Листя	5 Лист	6 Дист								4								
		JINCLT		MCID TINC	IN TRICIS	5 MACI	Juch	' (±													■ ■		

Fig. 8.1. Screenshot of the calculation of the adjusted determination coefficient for the model (8.2).

As can be seen from Table 8.2. And figure 8.1 in regression model 8.2, the determination coefficient is quite high 0.751075, the adjusted determination coefficient is 0.71996.

Tables 8.3, 8.4 and figure 8.2 show the initial data and the results of calculating the estimates for the model (8.3), the four-factor model, in which the factor added is the time period squared factor.

Table 8.3. The initial data for the model (8.3)

Time	Period	Period	Production	Advertising	Sales
		in	costs	costs	volume
		square			
January	1	1	890,8	200,8	1286,5
February	2	4	896,4	210,5	1282,7
March	3	9	903	208,6	1224,8
April	4	16	888,8	226,7	1292,6
May	5	25	889,8	230,62	1667,3
June	6	36	890,8	238,7	1672,9
July	7	49	888,3	232,3	1667,5
August	8	64	885,8	240,1	2001,7
September	9	81	883,9	238,1	2352,5
October	10	100	875,1	248,1	2468,5
NOVEMBER	11	121	877,73	253,58	2012,2
December	12	144	873,8	261,05	1982,7
January	13	169	878,2	275,1	1901,1
February	14	196	866,3	265,5	1971,6
March	15	225	865,1	266,1	1989,1
April	16	256	876,6	282,6	2138,2
May	17	289	862,5	286,5	2475,2
June	18	324	865,9	286,3	2395,6
July	19	361	857,5	284,3	2994,1

Table 8.4. The results of calculating the estimates from the model (8.3).

Number o	f factors -		4		
The regres	sion equati	on Y=b0+b2	1*t+b2*t^2	+b3*X1+b4	*X2
-13,0796	-10,1777	-2,10969	155,5671	12868,56	
11,92828	14,17594	2,378899	92,09033	12420,21	
0,764315	263,3282	#Н/Д	#Н/Д	#Н/Д	
11,35035	14	#Н/Д	#Н/Д	#Н/Д	
3148213	970784,6	#Н/Д	#Н/Д	#Н/Д	
The adjust	ed coefficie	ent of deter	mination		
0,717178					

h	🔏 Cut 🖺 Copy		Calibri	- 11	· A A	= = =	87 -	🔐 Wrap Text		General	*	≠		Normal Neutral	Bad		Good	-	÷ :	× 🖬	∑ AutoSun ↓ Fill * Clear *	° [▲] Z▼	A I
te	📌 Form	at Painter	ΒIU	• 🖽 • •	<u>ð</u> - <u>A</u> -	= = =	en 4n	🔛 Merge & (Center - C	· % •	€0 .00 (.00 →.0	Conditional F	ormat as Table *	Neutral	Calcu	ulation	Check Ce	ell 🗸	Insert De	elete Format	Clear -	Sort & Filter * S	Find & Select *
	Clipboard	5		Font	5		Alignm	ent	5	Number	5	onnocang			Styles					ells		Editing	
						0;B2:E20;;1)}																	
	А	в	С	D	E	F	G	н	I	J	К	L	м	N	0	Р	Q	R	S	т	U	v	W
			Period in	Producti	Advertisi	Sales																	
1	ime	Period	squaire	on costs	ng costs	volume																	
	нварь	1	l 1	890,8	200,8	1286,5		Number o	f factors	-	4												
	евраль	2										b3*X1+b4*	K2										
	1арт	3	-							7 -2,10969													
	прель	4								4 2,378899													
	1ай Іюнь	6							263,328		#N/A	#N/A											
	юнь юль	7						11,35035	970784,6		#N/A #N/A	#N/A #N/A											
	вгуст	/						5146213	970784,0	D HIN/A	#IN/A	#IN/A											
	ентябрь	9																					
	ктябрь	10						The adjust	ted coeffic	ient of dete	rmination												
	ОЯБРЬ	11						0,717178															
	екабрь	12						.,															
	нварь	13	3 169	878,2	275,1	1901,1																	
	евраль	14	1 196	866,3	265,5	1971,6																	
	1арт	15	5 225	865,1	266,1	1989,1																	
	прель	16																					
	Іай	17	-		· · ·																		
	юнь	18																					
	юль	19	361	857,5	284,3	2994,1																	
	»	Лист1	Лист2 Л	ист3 Л	ист4 Ли	ст5 Лист	6 ∣ Лис	17 🕀							4								
r		Листі	Лист2	истз Л	иста Ли	ст5 Лист	ь Лис	-T/ (+)						1	4						s 🗉		

Fig. 8.2. Screenshot of calculating the corrected determination coefficient for the model (8.3).

As can be seen from Table 8.4. And figure 8.2 in the regression model (8.3), the determination coefficient increased by 0.764315, and the adjusted determination coefficient decreased by 0.717178. This suggests that the added factor is not significant for the forecasting model, i.e. the addition of a "time period square" to the predicting model does not significantly affect the predicted variable.

Below in Tables 8.5, 8.6 and Figure 8.3, the initial data and the results of calculating the estimates for the model (8.4) are presented.

Time	Period	Period	Production	
		in	costs	Sales
		square		volume
			000.0	1206 5
January	1	1	890,8	1286,5
February	2	4	896,4	1282,7
March	3	9	903	1224,8
April	4	16	888,8	1292,6
May	5	25	889,8	1667,3

Table 8.5. The initial data for the model (8.4)

June	6	36	890,8	1672,9
July	7	49	888,3	1667,5
August	8	64	885,8	2001,7
September	9	81	883,9	2352,5
October	10	100	875,1	2468,5
NOVEMBER	11	121	877,73	2012,2
December	12	144	873,8	1982,7
January	13	169	878,2	1901,1
February	14	196	866,3	1971,6
March	15	225	865,1	1989,1
April	16	256	876,6	2138,2
May	17	289	862,5	2475,2
June	18	324	865,9	2395,6
July	19	361	857,5	2994,1

Table 8.6. The results of calculating the estimates from the model (8.4).

Number o	f factors -		3							
The regres	sion equati	on Y=b0+b2	1*t+b2*t^2	+b3*X1						
-13,7165	-1,39341	71,73884	13467,27							
13,89647	2,302831	51,68798	12491,62							
0,744074	265,0984	#Н/Д	#Н/Д							
14,53692	15	#Н/Д	#Н/Д							
3064840	1054158	#Н/Д	#Н/Д							
The adjust	The adjusted coefficient of determination									
0,712084										

b	👗 Cut 🖺 Copy		Calibri	- 11	· A A	= = =	89	🔐 Wrap Text		Seneral	×	≠		Normal	Bad	ulation	Good	-	÷ :	× 🗊	∑ AutoSum ↓ Fill *	ĽĂZŢ	AL.
		at Painter	ΒIU	• 🖽 •	<u>ð</u> - <u>A</u> -	= = =	€E 9E	🔛 Merge & (Center - C	🚰 • % ,	€.0 .00 0.€ 00.	Conditional Formatting *	Format as Table *	Neutral	Calc	ulation	Check Co	ell T	Insert De	elete Format	Clear •	Sort & I Filter * S	Find &
	Clipboard	6			6		Alignr	nent	5	Number	5	onnoccing			Styles					iells		Editing	
		- 1)	× 🗸	$f_x = \text{LII}$	NEST(E2:E2	0;B2:D20;;	1)}																
	А	в	С	D	E	F	G	н	I	J	K	L	м	N	0	Р	Q	R	S	т	U	V	w
			Period in	Producti	Sales																		
i	me	Period	squaire	on costs	volume																		
	варь	1	1	890,8	1286,5	5		Number o															
	евраль	2								Y=b0+b1*t													
	арт	3	9		· · ·					1 71,73884													
	рель	4								1 51,68798		2											
	ай	5	25						265,0984		#N/A												
	онь оль	7	36					14,53692	1054158		#N/A #N/A												
	оль густ	/ 8				-		3064840	1054150	s #N/A	#N/A												
	нтябрь	9		883,9																			
	тябрь	10						Скоррект	ипованин	й коэффиц		рминации											
	ОЯБРЬ	11				-		0,712084		ποσφφηά	дете	pininiaqrin											
	кабрь	12						-,															
	варь	13	169			l																	
	враль	14	196	866,3	1971,6	5																	
	арт	15	225	865,1	1989,1	L																	
	рель	16																					
	ай	17																					
	онь	18																					
	оль	19	361	857,5	2994,1	L																	
	►	Лист1	Лист2 Л	мст3 Л	ист4 Пи	ст5 Лии	т6 Ли	ст7 (+)							4								
	r	/merit	/INCIZ 1	nero n											•						🖹 🗉		+

Fig. 8.3. Screenshot of calculating the corrected determination coefficient for the model (8.4).

As can be seen from Table 8.6. and figure 8.3 in the regression model (8.4), the determination coefficient decreased by 0.744074, and the adjusted determination coefficient decreased also 0.712084. This suggests that the excluded factor "advertising costs" is essential for the forecasting model, i.e. it is necessary to leave the "advertising costs" factor in the forecasting model.

8.2 PRACTICAL LESSON 8 and IWS 8

Subject. Selection of factors for the prediction problem.

Plan of the lesson.

1) Examine the selection scheme for the factors for the prediction task, presented above.

2) Practical lesson and task of the IWS. Apply the above selection of factors for your prediction task.

3) Analyze the results.

Literature.

Minko A.A. Predicting in business using Excel. M.: Eksmo, 2007, -208 p.

TOPIC 9

ESTIMATION OF THE QUALITY OF THE DATA MODEL FOR THE PREDICTION TASK

9.1 Lecture material

Evaluation of the quality of the data model means to estimate the accuracy of the approximation of the input data by the prediction function $Y = f(t, X_1, X_2, ..., X_m; \varepsilon)$. There are various methods for estimating the accuracy of the approximation of the input data by the prediction function $f = f(b_1, b_2, ..., b_k; t; X_1, X_2, ..., X_m)$. The scheme of the effect of the random variable ε on the prediction function is adopted either additive $(f + \varepsilon)$, or multiplicative $(f * \varepsilon)$.

The quality of the data model is estimated by various indicators:

- The method of least squares (OLS);

- coefficient of determination;

- adjusted coefficient of determination;

- the average absolute deviation: $\frac{1}{n}\sum_{i=1}^{n}|e_{i}|;;$

- the average absolute error in percent: $\left(\frac{1}{n}\sum_{i=1}^{n} \left|\frac{e_{i}}{y_{i}}\right|\right) 100\%$.

The coefficient of determination and the adjusted coefficient of determination were previously considered in detail. The average absolute deviation and the mean absolute error in percent as well as the coefficient of determination are used when comparing data models built on the same data sets and containing the same set of factors.

The basic method of constructing the approximating function of the input data is the method of least squares. The main condition for the application of this method is the linear dependence of the approximating function on the unknown (sought) parameters.

The essence of the method of least squares is that the determination of the values of the unknown parameters $b_1, b_2, ..., b_k$ of the approximating function were based on the criterion of the minimum sum of squared deviations of the computed values

of the predicted variable by the approximating function from the observed values of the predicted (dependent) variable:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 \to min,$$

Where y_i are the observed values of the predicted variable at the i-th point of the initial data, $f(x_i)$ is the calculated values of the predicted variable from the approximating function at the i-th point.

Finding the "correct" values of the parameters $b_1, b_2, ..., b_k$ by the method of least squares is **due to a number of conditions**. Let ε_i be the random effects on the i-th point of the original data. Then the **conditions of least squares** have the following form [1]:

1) The mathematical expectation of random variables ε_i is equal to 0 for all data points.

2) The variances of the random variables ε_i must be the same for all data points. This condition is called the condition of data homoscedasticity, and its non-fulfillment is called heteroscedasticity of the data.

3) The random variables ε_i must be independent of each other. This condition is called the lack of autocorrelation.

4) The random variables ε_i must be independent of the factors.

5) There should be no strong linear relationship between the factors. This condition is called the absence of multicollinearity.

6) The random variables ε_i have a normal distribution with zero mathematical expectation and the same variance.

It should also be noted that the residuals e_i are taken as estimates of the random effects ε_i at the i-th point of the predicted variable. This is due to the fact that random effects can not be observed explicitly.

If these conditions are fulfilled, the values of the parameters $b_1, b_2, ..., b_k$ calculated by the least squares will be "correct": **unbiased**, **consistent** and **effective**. The **unbiased** property of the parameter values means that their mathematical expectations will be equal to the true values of these parameters. The property of **consistency** of the values of the parameters means that the variances of these values with an unlimited increase in the number n of data points will tend to zero. The **effectiveness** property of parameter values means that they have the smallest variance compared to any other estimates of these parameters.

When evaluating the quality of a data model, the feasibility of the method of least squares conditions is important. Thus, if the method of least squares conditions are not fulfilled, it is necessary to consider the correspondence of the chosen type of predicting function to the nature of the initial data. Often in the appearance of the picture of the original data, it is necessary to select a suitable type of prediction function (approximation).

One of the primary conditions of the method of least squares to be checked is the condition of homoscedasticity of data (the uniformity of the variance of residues for all data points) and the lack of autocorrelation of random effects ε_i (taking into account the use of residues - this will be the verification of the independence of the residues ε_i). Below in Figure 9.1, a graph of the homoscedasticity of the data is shown - the uniformity of the variance of the residues for all data points of factor X, and Figure 9.2 shows the graph of the hetaera.



Fig.9.1. Homoscedasticity of residues [2]



Fig.9.2. Heteroscedasticity of residues [2]

Consider the tests to verify these two conditions for the method of least squares. There are various developed tests to test each of these conditions. We show the tests of these conditions on one of them.

Verification of the homoscedasticity of data by the Spearman rank correlation test.

A hypothesis is advanced that there is no heteroscedasticity of the random variable e_i . It is assumed that the variance of the random variables e_i will either increase or decrease as the factor X increases, and therefore in the method of least squares regression the absolute values of the residuals e_i and the values of X will be correlated. Scheme of the test:

1) the data on X and the residuals e_i are ranked according to X and their ranks are determined; Ranks are defined as the location numbers of the values of X and e_i in the rows ordered in ascending order;

2) the Spearman rank correlation coefficient is determined by the formula

 $r = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$, where D_i is the difference between the ranks of X and e_i ;

3) The criterial statistics are calculated:

$$t_{x,e} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

4) The critical value t_{cr} is calculated as a quantile of the order of 1- α / 2 of the Student's distribution with the degree of freedom n-2.

If $t_{x,e} \le t_{cr}$, then this will indicate the absence of heteroscedasticity, or, conversely, the presence of homoscedasticity.

The following is an example of how to perform this test. The initial data of the example are presented in Table 9.1. The check on the homoscedasticity of the time factor and the residues in the predict function

$$Y = b_0 + b_1 t + b_2 t^2 + b_3 X_1 + b_4 X_2$$

Where X_1 - production costs, X_2 - advertising costs, Y - sales volumes, t - time.

Time	Perio	Perio	Producti	Adverti	Sales	Function	Residuals
	d	d in	on costs	sing	volume	forecast	
		squar		costs			
		е					
January	1	1	890,8	200,8	1286,5	1329,349	-42,8493
February	2	4	896,4	210,5	1282,7	1294,721	-12,0205
March	3	9	903	208,6	1224,8	1397,418	-172,618
April	4	16	888,8	226,7	1292,6	1446	-153,4
May	5	25	889 <i>,</i> 8	230,62	1667,3	1521,13	146,1696
June	6	36	890,8	238,7	1672,9	1537,63	135,2696
July	7	49	888,3	232,3	1667,5	1774,925	-107,425
August	8	64	885,8	240,1	2001,7	1822,27	179,4298
Septembe							
r	9	81	883,9	238,1	2352,5	1987,469	365,0307
October	10	100	875,1	248,1	2468,5	2061,72	406,7796
NOVEMB							
ER	11	121	877,73	253,58	2012,2	2074,541	-62,3407
December	12	144	873,8	261,05	1982,7	2123,879	-141,179
January	13	169	878,2	275,1	1901,1	1998,154	-97,0541
February	14	196	866,3	265,5	1971,6	2343,438	-371,838
March	15	225	865,1	266,1	1989,1	2442,189	-453,089
April	16	256	876,6	282,6	2138,2	2199,5	-61,2998
May	17	289	862,5	286,5	2475,2	2377,942	97,25768

Table 9.1. Initial data for the homoscedasticity test of residues

June	18	324	865,9	286,3	2395,6	2427,682	-32,082
July	19	361	857,5	284,3	2994,1	2616,842	377,2578

Table 9.2. Computation of the homoscedasticity test of residues

The regres	sion equati	ion Y=b0+b2	1*t+b2*t^2+b3	3*X1+b4*X2		
-13,0796	-10,1777	-2,10969	155,5671	12868,56		
11,92828	14,17594	2,378899	92,09033	12420,21		
0,764315	263,3282	#Н/Д	#Н/Д	#Н/Д		
11,35035	14	#Н/Д	#Н/Д	#Н/Д		
3148213	970784,6	#Н/Д	#Н/Д	#Н/Д		
Significanc	e level	0,05				
The sum o	f the squar	ed differend	ces of ranks			
1104		{=СУММК	BPA3H(B2:B20	;РАНГ(H2:H2O;	H2:H20;	L))}
Spearman	's correlatic	on coefficier	nt			
0,447078		{=1-6*J12,	/(20*(600-1))}			
Criteria sta	atistics					
1,943352		{=J14*KOF	РЕНЬ(20-2+КОІ	РЕНЬ(1-Ј14*Ј14	4))}	
Critical val	ue					
2,100922		{=СТЬЮДІ	РАСПОБР(L9;20	D-2)}		
There is ho	omoskedast	ticity				

	py 🏺 rmat Painter	Calibri B I L	• = -	<u>ð</u> - <u>A</u> -		सस ∰। %- ₽\		iter - 🧣		Cor	ditional Fo matting * T	ormat as N		Bad Calcu	lation	Good Check Cell	* * *	Insert De	lete Format	∑ AutoSun ↓ Fill *	Z Sort & Filter *	Find & Select *
Clipbo			Font	5		Alignment		5	Number	F2				Styles				(ells		Editing	
	* E	XV	f_X																			
A	В	C	D Producti	E	F	G Prediction	Н	I	J	ĸ	L	М	N	0	Р	Q	R	S	Т	U	V	W
lime	Period	squaire		ng costs	volume		Residue															
нварь		1	1 890,8	200,8	1286,5	1329,3493	-42,8493	Regressio	on equition	Y=b0+b1*	t+b2*t^2+b	3*X1+b4*	X2									
Феврал	ь	2	4 896,4			1294,7205							12868,56									
Ларт		5	9 903							14,17594												
прель		4 1				1446,0001	-153,4		0,764315		#N/A	#N/A	#N/A									
Лай 1юнь	_	5 2 6 3				1521,1304 1537,6304			11,35035	14 970784,6	#N/A #N/A	#N/A #N/A	#N/A #N/A									
юнь		7 4				1774,9248			5146215	970784,0	#IN/A	#19/24	#IN/A									
вгуст		8 6				1822,2702			Significanc	e level	0,05											
ентяб	ъ	9 8				1987,4693			-0													
жтябр	, 1	0 10	0 875,1	248,1	2468,5	2061,7204	406,7796		The sum of	f the square	ed differend	ces of rank	s									
оябры	1	1 12	1 877,73	253,58	2012,2	2074,5407	-62,3407		1104		{=CYMMK	BPA3H(B2:	В20;РАНГ(Н	2:H20;H2:	H20;1))}							
Цекабр						2123,8789			Spearman'													
нварь	1		,			1998,1541			0,447078		{=1-6*J12/	(20*(600-1	.))}									
еврал						2343,4378			Criteria sta													
1арт	1	5 22	,						1,943352 Critical val		{=J14*KOP	РЕНЬ(20-2+	КОРЕНЬ(1-	14*J14))}								
прель 1ай	1					2199,4998 2377,9423			2,100922		{=стьюдя		9-20-211									
юнь	1	-							There is ho			Merrobi (L	.5,20 2)]									
1юль		9 36				2616,8422					,											
	1 m	0	0	0																		
	ЛИСТІ	Лист2	ЛИСТЗ	\oplus										4								

Fig. 9.3. Screenshot of the initial data and calculation of the homoscedasticity test of the time factor and residues

Based on the results of the homoscedasticity test of the time factor and residues, it was found that $t_{x,e} = 1,943352$, $t_{cr} = 2,100922$. Accordingly, $t_{x,e} \le t_{cr}$, then there is a homoscedasticity of the time factor and residuals.

Verification of the condition of independence of the residual data by the Durbin-Watson test.

The verification of the independence of the residues, that is, the test for the **lack of autocorrelation** of the residues can be performed according to the Darbin-Watson test (criterion). The criterion is named after James Durbin and Jeffrey Watson [3]. The Durbin-Watson criterion is calculated by the following formula:

$$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} (e_i)^2}$$

In the absence of autocorrelation DW = 2, with positive autocorrelation, DW tends to zero, and for negative autocorrelation DW approaches 4. And according to the Durbin-Watson criterion, special tables are developed that allow one to determine the critical points of the observable statistics DW for a given number of observations n, the number of factors and a given level of significance. But in practice, the rule is often used:

- the interval from 0 to 4 is divided into several ranges: positive autocorrelation, negative autocorrelation and ranges of uncertainty;

- if DW \leq 0.5, then there is positive autocorrelation (the range of positive autocorrelation),

- if DW \geq 3.5, then there is negative autocorrelation (the range of negative autocorrelation),

- if $1.5 \le DW \le 2.5$, then there is no autocorrelation (the range of absence of autocorrelation).

- in other cases, these are uncertainty ranges when there is no confidence in the presence or absence of autocorrelation: 0.5 <DW <1.5 and 2.5 <DW <3.5.

If the value of the Durbin-Watson criterion falls within the autocorrelation ranges or uncertainty ranges when analyzing the given source data model, this should be the reason for further research on the selection of a more suitable prediction function.

Below in Table 9.3 and Figure 9.3, an example of the verification of the independence of residues according to the Durbin-Watson criterion is presented. The initial data is taken from Table 9.1.

Table 9.3 Calculation of the Durbin-Watson criterion for checking the independence of residues

Т	he	reg	ression		equation	
Y	′=b0+b1*t+	b2*t^2+b3*	*X1+b4*X2			
	-13,0796	-10,1777	-2,10969	155,5671	12868,56	
	11,92828	14,17594	2,378899	92,09033	12420,21	
	0,764315	263,3282	#Н/Д	#Н/Д	#Н/Д	
	11,35035	14	#Н/Д	#Н/Д	#Н/Д	
	3148213	970784,6	#Н/Д	#Н/Д	#Н/Д	
	the Durbir	-Watson cr	iterion			
		0,99674				

In this example, the value of the Durbin-Watson criterion is 0.99674, i.e. It falls within the uncertainty range of 0.5 <DW <1.5. This is the reason for choosing a more advanced forecasting function.

n P	Cut		Calibri	- 11	· A A	= = =	≫- 🖥	Wrap Text		General		1 de la compañía de		Normal	Bad	ulation	Good		H	×	∑ AutoSum ↓ Fill *	ĭ <mark>A</mark> Ţ	#
ste	 End Copy Form 	at Painter	ΒIU	• 🖽 • •	<u>ð</u> - <u>A</u> -	= = =	€ 40 🗄	Merge & C	ienter 🝷	General	€0 .00 C	ionditional F	ormat as	Veutral	Calcu	ulation	Check Ce	1 ·	Insert De	lete Format	Clear -	Sort & F Filter * S	ind &
	Clipboard	5		Font	6		Alignmen	t	5	Number	5		10010		Styles				0	ells	E	diting	
D		- 1 7	K 🗸	$f_x = su$	MXMY2(H3	:H20;H2:H1	9)/SUMSQ	(H2:H20)															
	А	в	C	D	E	F	G	н	I	J	к		м	N	0	Р	Q	R	s	т	U	V	W
		0					Predictio										~						
			Period in	Producti	Advertisi	Sales	n																
T	ime	Period	squaire	on costs	ng costs	volume	function	Residue															
	нварь	1	1	890,8	200,8	1286,5	1329,349	-42,8493	Regressi	ion equition	Y=b0+b1*t	+b2*t^2+b3	*X1+b4*X	2									
	евраль	2	4	896,4	210,5	1282,7	1294,721	-12,0205		-13,0796	-10,1777	-2,10969	155,5671	12868,56									
	Ларт	3	9	903	208,6	1224,8	1397,418	-172,618		11,92828	14,17594	2,378899	92,09033	12420,21									
	прель	4	16	888,8	226,7	1292,6	1446	-153,4		0,764315	263,3282	#N/A	#N/A	#N/A									
	1ай	5		889,8	230,62	1667,3		146,1696		11,35035			#N/A	#N/A									
1	юнь	6		890,8				135,2696		3148213	970784,6	#N/A	#N/A	#N/A									
	юль	7		,				-107,425															
	вгуст	8		885,8	240,1	2001,7	1822,27	179,4298		the Durbin	-Watson ci												
	ентябрь	9		883,9				365,0307			0,99674												
	ктябрь	10		875,1		2468,5		406,7796															
	ЮЯБРЬ	11		877,73				-62,3407															
	екабрь	12		873,8				-141,179															
	нварь	13		878,2				-97,0541															
	евраль	14		866,3				-371,838															
	1арт	15		865,1				-453,089															
	прель 1ай	16						-61,2998															
		17		862,5				97,25768															
	юнь	18		865,9																			
	юль	19	501	657,5	204,5	2994,1	2010,042	377,2578															
				2																			
	>	Листі	Лист2 Л	истз	(\pm)										4								

Fig. 9.4. A screenshot of the original data and the calculation of the independence test for residues according to the Durbin-Watson criterion

9.2 PRACTICAL LESSON 9 and IWS 9

Subject. Estimation of the quality of the data model for the prediction problem.

Plan of the lesson.

1) Study the scheme for assessing the quality of the data model for the prediction task, presented above.

2) Practical lesson and task of the IWS. Apply the above-described data quality model for your forecasting task.

3) Analyze the results.

Literature.

1. Minko A.A. Predicting in business using Excel. M.: Eksmo, 2007, -208 p.

2. http://metr-ekon.ru. Lectures on Econometrics with examples of solutions. The essence of heteroscedasticity.

3. https://ru.wikipedia.org/wiki/Robin_Watson_Charter_

TOPIC 10

DATA SMOOTHING FOR THE PREDICTION TASK

10.1. Lecture Material

Data smoothing is a procedure for preliminary processing of the initial data, which allows smoothing (decreasing) the values of random effects in the original data. Consider three methods of smoothing the data: the moving average, exponential smoothing, and the Holt method.

Moving average method.

Smoothing of data by the moving average method is carried out according to the formula [1]:

$$\widehat{y}_{i} = \frac{1}{2k+1} (y_{i-k} + y_{i-k+1} + \dots + y_{i-1} + y_{i} + y_{i+1} + \dots + y_{i+k-1} + y_{i+k}), \quad (10.1)$$

where k is a positive integer that is less than n. In practice, the value of k is set to 1, 2 or 3. That is, k values are taken to the left from the right and k values from the averaged value y_i , and also y_i itself.

The disadvantage of this method is the reduction of the smoothed data on the left and right by k values relative to the original data.

In the case of time series, the moving average method is calculated using the formula:

$$\widehat{y}_{i} = \frac{1}{k} (y_{i-k+1} + y_{i-k+2} + \dots + y_{i-1} + y_{i}).$$
(10.2)

In this case, we take on the left k-1 values from the averaged value y_i , and also y_i itself.

The disadvantage of this formula is the cut of the smoothed data on the left by k-1 values, i.e. a shift to the right of smoothed data relative to the original data.

Time	Period	Production	The smoothed
		costs	data (formula
			10.1)
February	2	1282,7	
March	3	1224,8	1266,7
April	4	1292,6	1394,9
May	5	1667,3	1544,267
June	6	1672,9	1669,233
July	7	1667,5	1780,7
August	8	2001,7	2007,233
September	9	2352,5	2274,233
October	10	2468,5	2277,733
NOVEMBER	11	2012,2	2154,467
December	12	1982,7	1965,333
January	13	1901,1	1951,8
February	14	1971,6	1953,933
March	15	1989,1	2032,967
April	16	2138,2	2200,833
May	17	2475,2	2336,333
June	18	2395,6	2621,633
July	19	2994,1	2694,85

Table 10.1. Initial data and results for the moving average method according to the formula (10.1)



Fig. 10.1 Sales volume charts: the initial (row 1) and smoothed (row 2) by formula (10.1)



Fig. 10.2. A screenshot of the initial data and smoothing using the moving average method according to the formula (10.1)

Table 10.2 Initial data and results for the moving average method using formula (10.2)

Time	Period	Production costs	The smoothed data (formula 10.2)
February	2	1282,7	
March	3	1224,8	1253,75
April	4	1292,6	1258,7
May	5	1667,3	1479,95
June	6	1672,9	1670,1
July	7	1667,5	1670,2
August	8	2001,7	1834,6
September	9	2352,5	2177,1
October	10	2468,5	2410,5
NOVEMBER	11	2012,2	2240,35
December	12	1982,7	1997,45
January	13	1901,1	1941,9
February	14	1971,6	1936,35
March	15	1989,1	1980,35
April	16	2138,2	2063,65
May	17	2475,2	2306,7

June	18	2395,6	2435,4
July	19	2994,1	2694,85





🔏 Cut		Calibri	- 11 -	ĂĂ	= = =	87 -	🔐 Wrap Te	xt	General	-			Normal	Bad		Good		÷	i	Σ AutoS	ium * Az	· #
Cop	nat Painter	BIU	-	- <u>A</u> -	= = =	€E 3E	📑 Merge &				Conditional Formatting *	Format as	Neutral	Calc	ulation	Check C		Insert D	elete Format	👽 Fill 🔻	Sort	& Find &
Clipboard									Number			Table -		Styles					Cells		Editing	- Select -
	+ + 5	< ./	∫x =AVE	RAGE(C1)	8-(19)																	
				E	F					K		м		0	Р	0	0	c	T	U	V	
A	В	С	D Сглажен	E	F	G	Н	I	J	K	L	M	N	0	P	Q	R	S	T	U	v	W
		Объемы																				Объемы
Время	Период		данные																		Период	продаж
Февраль	2																					
Март	3	1224,8	1253,75																			
Апрель	4	1292,6	1258,7		35	00																
Иай	5	1667,3	1479,95																			
Люнь	6	1672,9	1670,1		30	00				+												
1юль	7	1667,5	1670,2		25	00																
вгуст	8	2001,7	1834,6		25																	
ентябрь	9	2352,5	2177,1		20	00			Land Contract													
Октябрь	10	2468,5	2410,5					1			Sales	volume										
НОЯБРЬ	11		2240,35		15	00					Smoo	othed data										
Цекабрь	12		1997,45		10	00																
Январь	13	,	1941,9		10																	
Февраль	14				5	00																
Иарт	15		1980,35																			
Апрель	16		2063,65																			
Иай	17		2306,7			1 2	34567	8 9 10 1	1 12 13 14 15 1	61/1819	,											
1юнь	18																					
1юль	19	2 1, 1	2694,85																			
	0 1		2		-																	
>	Лист1	ист2 Л	ист3 Ли	CT4	÷									4								

Fig. 10.4. A screenshot of the initial data and smoothing using the moving average method according to the formula (10.2)

Exponential data smoothing.

Exponential data smoothing applies to data that depends on time, i.e. to time series. In this method, the smoothed values are calculated using the following formula:

$$\widehat{y_i} = (1 - \alpha)y_i + \alpha \widehat{y_{i-1}},\tag{10.3}$$

Where α is the smoothing coefficient; $\widehat{y_1}$ is taken as y_1 . The values of α are taken in the range (0.1), in practice - from 0.2 to 0.8.

The smoothing factor allows some balancing between the previous smoothed value $\widehat{y_{i-1}}$ and the current observed value y_i to get the current smoothed value $\widehat{y_i}$. The closer the smoothing coefficient is to zero, the less is the smoothing of the observed value, i.e. When $\alpha = 0$, there is no smoothing. Conversely, for $\alpha = 1$, the current smoothed value will coincide with the previous smoothed value, i.e. the graph of smoothed values will be a horizontal line (see Figure 10.6).

Table 10.3 Values of the α - smoothing coefficient and the smoothing formula with their values

	coeff smoothing 1	
0,2		{=(1-\$I\$2)*C3+\$I\$2*D2}
	coeff smoothing 2	
0,4		{=(1-\$I\$3)*C3+\$I\$3*E2}
	coeff smoothing 3	
0,6		{=(1-\$I\$4)*C3+\$I\$4*F2}
	coeff smoothing 4	
0,8		{=(1-\$I\$5)*3+\$I\$5*G2}
	coeff smoothing 5	
1		{=(1-\$I\$6)*C3+\$I\$6*H2}

Table 10.4 Initial data and results for the method of exponential smoothing by the
formula (10.3) for the given smoothing coefficients

Time	Period	Production	The	The	The	The	The
		costs	smoothed	smoothed	smoothed	smoothed	smoothed
			data 1	data 2	data 3	data 4	data 5
February	2	1282,7	1282,7	1282,7	1282,7	1282,7	1282,7
March	3	1224,8	1236,38	1247,96	1259,54	1271,12	1282,7
April	4	1292,6	1281,356	1274,744	1272,764	1273,167	1282,7
May	5	1667,3	1590,111	1510,278	1430,578	1336,556	1282,7
June	6	1672,9	1656,342	1607,851	1527,507	1400,513	1282,7
July	7	1667,5	1665,268	1643,64	1583,504	1453,464	1282,7
August	8	2001,7	1934,414	1858,476	1750,783	1549,654	1282,7
September	9	2352,5	2268,883	2154,89	1991,47	1693,5	1282,7
October	10	2468,5	2428,577	2343,056	2182,282	1840,515	1282,7

NOVEMBER	11	2012,2	2095,475	2144,542	2114,249	1891,507	1282,7
December	12	1982,7	2005,255	2047,437	2061,629	1914,257	1282,7
January	13	1901,1	1921,931	1959,635	1997,418	1915,792	1282,7
February	14	1971,6	1961,666	1966,814	1987,091	1924,967	1282,7
March	15	1989,1	1983,613	1980,186	1987,894	1936,696	1282,7
April	16	2138,2	2107,283	2074,994	2048,017	1970,813	1282,7
May	17	2475,2	2401,617	2315,118	2218,89	2056,974	1282,7
June	18	2395,6	2396,803	2363,407	2289,574	2124,94	1282,7
July	19	2994,1	2874,641	2741,823	2571,384	2274,88	1282,7



Fig. 10.5 Sales volume charts: the initial (row 1) and smoothed (row 2, row2, row3, row4, row5, row6) by formula (10.3)

	🔏 Cut 🖺 Copy	, -	Calibri		- A A					General	•	, _ , <i>∓</i> ,		Normal	Bad		Good	-		× 🗊	∑ AutoSu ↓ Fill *		
2	💞 Form	at Painter										Conditional Formatting *	Format as Table *	Neutral		ulation	Check C	ell 🔻	-	elete Format	Clear -	Filter *	Find & Select *
	Clipboard			Font	5		Alignmer	t	5	Numbe	ir 5				Styles				0	ells		Editing	
		* 1	XV	f _x																			
	Α	В	С	D	E	F	G	н	Ι	J	K	L	M	N	0	Р	Q	R	S	т	U	V	w
			Sales	Smoothe	Smoothe	Smoothe	Smoothe	Smoothe															
1	me	Period	volume	d data 1	d data 2	d data 3																	
	евраль		2 1282,),2 coeff sm		{=(1-\$I\$2											
	арт		3 1224,8			1259,54),4 coeff sm		{=(1-\$I\$3											
	прель	4	1292,6		5 1274,744),6 coeff sm		{=(1-\$I\$4											
	ай		5 1667,3		1510,278),8 coeff sm		{=(1-\$I\$5											
	юнь		5 1672,9		2 1607,851					1 coeff sm	ooth 5	{=(1-\$I\$6)*C3+\$I\$t	•H2}									
	юль		7 1667,9		1858,476	1583,504						3500 -						y = 1250,6	-0.0358+				
	згуст ентябрь		2352.5			1/50,/83						3500						y = 1250,64 R ² = 0.92					
	ктябрь	10	,		7 2343,056							3000						ales volume					
	ОЯБРЬ	1			5 2144,542							_											
	екабрь	1			5 2047,437							2500					s	moothed da	ta 1				
	нарь	1			1959,635							2000						moothed da	ta 2				
	евраль	14			5 1966,814								_					moouleu ua					
	арт	1	,		3 1980,186							1500						moothed da	ita 3				
	трель	10	5 2138,2		3 2074,994							1000	-										
	ай	1	2475,2	2 2401,617	2315,118	2218,89	2056,974	1282,7				1000						moothed da	ita 4				
	юнь	1	3 2395,6	2396,803	3 2363,407	2289,574	2124,94	1282,7				500						moothed da	ta 5				
	оль	1	2994,	2874,641	2741,823	2571,384	2274,88	1282,7										inoouncu uu	icu 5				
												0 +	1234	56789	10 11 12 13	14 15 16 17 18		xpon. (Smoo lata 4)	othed				
	Image: A start in the start	Лист1	Лист2 Ј	ЛистЗ Л	ист4 (()									1 4								

Fig. 10.6 A screenshot of the original data and exponential smoothing using formula (10.3)

Holt's smoothing method.

The method of smoothing Holt is based on adding a trend component to the data being smoothed. Calculated by the following formulas:

$$\widehat{y_{i}} = (1 - \alpha)y_{i} + \alpha(\widehat{y_{i-1}} + T_{i-1}); \quad T_{i} = (1 - \beta)(\widehat{y_{i}} - \widehat{y_{i-1}}) + \beta T_{i-1}, \quad (10.4)$$

Where α and β are smoothing coefficients, taking values from 0 to 1. The value $\widehat{y_1}$ is taken as y_1 , and the value calculated by the formula

$$T_1 = \frac{y_2 - y_1 + y_4 - y_3}{2}$$

is taken as the value of T_1 .

By Holt's method, it is possible to produce and predict for k periods ahead. Calculations are made by the formula:

$$\widehat{y_{n+k}} = \widehat{y_n} + kT_n \tag{10.5}$$

Table 10.5 Initial data and results for the Holt method according to the formula (10.4)

Time	Period	Production	The					
		costs	smoothed					
			data					
				Ті	Alfa =	0,6	Beta=	0,4
February	2	1282,7	1282,7	158,4		T _i {=(C3-C2	+C5-4)/2}	
March	3	1224,8	1164,5	-7,56		Ti sm {=(1-\$I	\$1)*(D3-2)-	+\$I\$1*E2}
April	4	1292,6	1220,276	30,4416				
May	5	1667,3	1380,821	108,5034				
June	6	1672,9	1432,55	74,43918				
July	7	1667,5	1481,867	59,36549				
August	8	2001,7	1654,181	127,1346				

September	9	2352,5	1857,228	172,682	
October	10	2468,5	1998,127	153,6126	
NOVEMBER	11	2012,2	1911,589	9,521932	
December	12	1982,7	1934,32	17,44755	
January	13	1901,1	1910,564	-7,27493	
February	14	1971,6	1939,343	14,35775	
March	15	1989,1	1950,631	12,51597	
April	16	2138,2	2018,149	45,51715	
May	17	2475,2	2173,659	111,5129	
June	18	2395,6	2195,528	57,72631	



Fig. 10.7 Sales volume graphs: the initial (row 1) and smoothed (row 2) sales volumes for the Holt method using formula (10.4)

In Fig. 10.7 also shows approximations of the original and smoothed sales volume data by predicting functions. The coefficient of determination of the prediction function for smoothed data is much higher.

Table 10.6. The results of calculating the predicting of the volume of sales by the Holt method using formula (10.5)

July								
	19	2994,1	2480,321	193,9664	Ti{=(1-\$I\$1	l)*(D19-D18	8)+\$I\$1*E18]	}
August								
	20		2674,287		прогноз {=	=\$D\$19+(B2	0-\$B\$19)*\$I	E\$19}
September								
	21		2868,254					
October								
	22		3062,22					

	🖌 Cut ከ Copy		Calibri			= = =		🔐 Wrap Te		General	-	≠	and the second s	Normal	Bad		Good	-	÷	× 🗊	∑ AutoSur	'ĭ <mark>A</mark> ₹	H
		at Painter	ΒIU	• 🆽 • 🕹	<u>></u> - <u>A</u> -	= = =	€E 3 E	📑 Merge 8	Center -	₽ -%,	00, 0, → 0,€ 00,	Conditional Formatting *	Format as Table *	Neutral	Calc	ulation	Check C	ell 🗸	Insert De	lete Format	Clear -	Sort & Filter * 1	
C	lipboard	5		Font	5		Alignm	hent	5	Number	r _a				Styles				c	ells		Editing	
		• :	× 🗸	fx =(1-	\$G\$1)*C4+	\$G\$1*(D3	-E3)																
	A	В	С	D	E	F	G	н	I	J	K	L	M	N	0	Р	Q	R	S	т	U	V	W
			Sales	Smoothe																			
m	e	Period	volume		Ti	Альфа=		,6 Бета		0,4													
e	враль		2 1282,7	1282,7	158,4	1		2+C5-C4)/2															
	рт	3	3 1224,8				{=(1-\$I\$	1)*(D3-D2)-	\$I\$1*E2}														
	ель	4	1292,6		30,4416						3500												
la		,	5 1667,3		108,5034						3500												
	нь	6	5 1672,9		74,43918						3000		122,7x ^{0,265}										
-	ль		7 1667,5		59,36549						-		= 0.757			Sale	s volume						
	уст		3 2001,7		127,1346						2500				N								
	тябрь		2352,5		172,682						2000					Smo	othed data I	Holt					
	ябрь	10			153,6126						2000		10			51110	ourea autor	-					
	ябрь	11			9,521932						1500					Pow	er (Sales vol	ume)					
	кабрь зарь	12			17,44755								y = -1,1	287x ² + 88,03 R ² = 0,9138		100	ci (buics voi						
	зарь зраль	14			14,35775						1000	-		R-=0,9136		Dob	. (Smoothed	data					
	рт рт	19	,		12,51597						500					Holt		uata					
	ель	10			45,51715						500												
la		17			111,5129						0	+											
	нь	18			57,72631						-	12345	6789	1011121314	15161718								
-	ль	19					Ti{=(1-\$	(D19-D	18)+\$I\$1	'E18}													
81	уст	20)	2674,287				{=\$D\$19+															
	тябрь	21		2868,254																			
К	ябрь	22	2	3062,22																			
,		Лист1	Лист2 / Л	ист3 Ли	аст4	(+)									4								
				_		•															n m -	_	- +

Fig. 10.8 The screenshot of the initial data, the results of smoothing and the forecast for the Holt method using formula (10.4) and (10.5)

10.2 PRACTICAL LESSON 10 and IWS 10

Subject. Smoothing data for the prediction task.

Plan of the lesson.

1) Study the data smoothing schemes for the prediction task presented above.

2) Practical lesson and task of the IWS. Apply the above described data smoothing schemes to your prediction task.

3) Analyze the results.

Literature.

1. Minko A.A. Forecasting in business using Excel. M.: Eksmo, 2007, -208 p.

TOPIC 11

TREND FUNCTIONS IN PREDICTING TASKS

11.1. Lecture material

A trend is the general tendency of changing the initial data. In time-dependent prediction functions, there are usually two components: the trend component, which is the main dependency in the predicting function, and the seasonal component, which may be absent. In this topic, it is consider schemes for building the trend component of the prediction function. The main way to identify the trend is the method of constructing the regression function. As previously shown, the scheme for constructing the regression function is:

- calculation of regression coefficients by the least squares method, which is performed in Excel by standard functions LINEST or LOGEST;

- writing a formula for the chosen regression function using the regression coefficients obtained at the previous stage;

- calculation of the necessary quality indicators of the constructed data model, for example, the adjusted determination coefficient, the mean absolute error or the average absolute error in percent;

- analysis of the quality of the constructed data model.

Below it is consider examples of constructing prediction functions that depend on one factor and prediction functions, depending on several factors.

11.1.1 Prediction functions depending on one factor

Polynomial prediction function.

The polynomial forecast function from one factor in the general form has the form of an algebraic polynomial [1]:

$$Y = b_0 + b_1 X + b_2 X^2 + \dots + b_m X^m,$$
(11.1)

Where the coefficients $b_0, b_1, b_2, \dots, b_m$ are determined on the basis of the initial data Y and X.

The following is an example of calculating quality metrics for a parabola and a cubic parabola for given source data:

 $Y = b_0 + b_1 t + b_2 t^2,$

$$Y = b_0 + b_1 t + b_2 t^2 + b_3 t^3.$$

Time	Period	Period in square	Period in cube	Production costs	parabola	cubic parabola
January	1	1	1	1286,5	1232,35	1014,37
February	2	4	8	1282,7	1322,12	1250,66
March	3	9	27	1224,8	1409,84	1441,64
April	4	16	54	1292,6	1495,51	1583,72
May	5	25	125	1667,3	1579,12	1709,36
June	6	36	216	1672,9	1660,68	1796,91
July	7	49	343	1667,5	1740,18	1860,82
August	8	64	512	2001,7	1817,63	1906,47
September	9	81	729	2352,5	1893,03	1939,3
October	10	100	1000	2468,5	1966,37	1964,72
NOVEMBER	11	121	1331	2012,2	2037,66	1988,12
December	12	144	1728	1982,7	2106,9	2014,93
January	13	169	2197	1901,1	2174,08	2050,56
February	14	195	2744	1971,6	2240,23	2128,48
March	15	225	3375	1989,1	2302,28	2169,91
April	16	256	4096	2138,2	2363,3	2264,46
May	17	289	4913	2475,2	2422,27	2389,47
June	18	324	5832	2395,6	2479,18	2550,36
July	19	361	6859	2994,1	2534,04	2752,54

Table 11.1 The initial data and the results of constructing regression functions for a parabola and a cubic parabola according to formula (11.1)

Table 11.2. The calculated characteristics of the regression functions of a parabola and a cubic parabola

Coeff of para								
-1,026883	92,853778	1140,524						
2,2723591	46,76701	203,22234						
0,7273164	264,95083	#н/д						
21,338032	16	#Н/Д						
2995814,5	1123183	#Н/Д						
Determinatio								
0,7273164								
Adjusted Determination coefficient of parabola								
0,7112762	{=1-(19-1)*(1-14)/(19-2)}							
Coefficients of cubic parabola								
Y=b0+b1*t+								
0,9019587	-28,06199	314,15832	727,3731					
0,4451412	13,503639	117,30559	275,9743					
0,7859135	242,46263	#Н/Д	#Н/Д					
18,355042	15	#н/д	#Н/Д					
3237175,6	881821,9	#н/д	#Н/Д					
Determinatio								
0,7859135								
Adjusted Determination coefficient for cubic parabola								
0,8897035	0,8897035 {=1-(19-1)*(1-I14)/(19-3)}							

Comparison of the adjusted coefficients of determination of the regression function of a parabola and a cubic parabola shows that the quality of the approximation of the initial data by a cubic parabola is higher than the parabola. This result can be visually observed by comparing the graphs of the original data and the constructed regression functions in Figure 11.1.


Fig. 11.1. The raw data graphs (row 1), the parabola regression function (row 2), and the cubic parabola regression function (row 3)

	ру т	Calibri B I U -	• 11 • . •			≻- ₽w === ඕм	frap Text	Gen		Conditi	onal Format a	Norma		Bad Calculatio		Good Check Cell	i fille	ert Delete	Format	∑ AutoSum	Sort & Find	1 &
Fc Clipbo	rmat Painter		Font			Alianment	erge a e	5	Number	Formatt	ing * Table *			Styles				* Cells	Ť	Clear *	Filter * Selec	ct *
	*	× √ f	έ =\$K\$2+	\$J\$2*B2+\$I	\$2*C2																	
A	В	С	D	E	F	G	н	I	J	К	L	м	N	0	Р	0	R	s	Т	U	V	W
		Period in	Period in	Sales	Parabol	parabol		Coeff.	of pararbola	Y=b0-												
ime	Period	square	cube	volume	а	a			b1*t+b2*t^2			35								_		
anuary	1	l 1		1 1286,5	1232,35			-1,026883	92,853778	1140,524		35										
ebruar	1 2		1		1322,12			2,2723591	46,76701			30	00									
larch	3				1409,84			0,7273164	264,95083	#N/A												
pril	4							21,338032	16			25	00				- A					
lay	5				1579,12			2995814,5	1123183	#N/A		20	- 00		/		<u> </u>	Sal	les volum			
une	6	5 36 7 49						Coeff. of det														
ly	8							0,7273164	termination			15	00						rabola			
ugust eptem	-							Adjusted det	ermination c	oeff_for.par	abola	10	m 🔰					-Cu	bic parab	iola		
ctobe	10								{=1-(19-1)*(:		abola	10										
lovem								Coeff. of cul		1 14//(15 2/)		5	00									
ecemi	-							Y=b0+b1*t+		^3												
anuary	13	169	219	7 1901,1	2174,08	3 2050,56		0,9019587	-28,06199	314,15832	727,3731		0 +									
ebruar	14	195	274	1971,6	2240,23	2128,48		0,4451412		117,30559			1.	23456/	891	0 11 12 13 14 15 1	6171819					
1arch	15	225	337	5 1989,1	2302,28	2169,91		0,7859135	242,46263	#N/A	#N/A											
pril	16			5 2138,2	2363,3	2264,46		18,355042	15	#N/A	#N/A											
1ay	17				2422,27			3237175,6	881821,9	#N/A	#N/A											
une	18																					
ıly	19	361	6859	9 2994,1	2534,04	2752,54		Coeff. of det	termination													
								0,7859135	1.1													
	-							Adjusted det		oeff. for cut 1-I14)/(19-3)												
								0,0097035	1-1-(19-1)-(1-1141/(19-3)												
	Полино		а, Гипер	б функц	Экспон	dynam i i	Стопон	функц S-н	nfinaa Pe	зульт.сгл.да	н Полин	own	0 1									
,	Полино	лот.функ	м пипер	олуулкц	SACTION.	wyrnal	creater.	wymaa 34	sopus Pe	зультеглада	ПОЛИН	o		•								-+

Fig. 11.2. Screenshot of the quality assessment of two models of the polynomial forecasting function: a parabola and a cubic parabola

Exponential predicting function.

The exponential function of prediction from one factor is calculated by the formula [1]:

$$Y = b_0 * (b_1)^X$$
(11.2)

To calculate the coefficients of the exponential function, we use the standard LGRF function.

The following is an example of calculating the exponential forecasting performance indicators from a single factor for given source data:

$$Y = b_0 * (b_1)^t.$$

Table 11.3. The initial data and the results of constructing the exponential regression function from one factor using formula (11.2)

		Sales	exponential		
Time	Period	volume	regression	Coefficie	nts of function
January	1	1286,5	1320,98	1,04	1270,29
February	2	1282,7	1373,69	0,005	0,06267
March	3	1224,8	1428,5	0,749	0,13122
April	4	1292,6	1485,5	50,67	17
May	5	1667,3	1544,78	0,873	0,29274
June	6	1672,9	1606,42	{=LOGES	T(C2:C20;B2:B20;;1)}
July	7	1667 <i>,</i> 5	1670,52	Determi	nation coefficient
August	8	2001,7	1737,17	0,749	
September	9	2352,5	1806,49		
October	10	2468,5	1878,57		
NOVEMBER	11	2012,2	1953,53		
December	12	1982,7	2031,48		
January	13	1901,1	2112,54	{=\$G\$2*	POWER(\$F\$2;B14)}
February	14	1971,6	2196,83		
March	15	1989,1	2284,49		
April	16	2138,2	2375,64		
May	17	2475,2	2470,44		
June	18	2395,6	2569,01		
July	19	2994,1	2671,52		



Fig. 11.3. The graphs of the initial data (row 1), the exponential regression function (row 2) of one factor according to the formula (11.2

🔏 Cut		Calibri	- 11 -	ÂĂ	= =	≫-	Wrap Text		General	-	I ≢	a set	Normal	Bad		Good	-	ŧ i	× 🗊	∑ AutoSum ↓ Fill +	ĭ₽ZŢ	H.
✓ Forma		ΒIU	- 🖽 - 👌	• <u>A</u> •		€2 42 €	Merge & (Center 🝷	s -%,	€.0 .00 0.€ 00.	Conditional Formatting *	Format as Table *	Neutral	Calcu	lation	Check Ce	il v	Insert De	lete Format	Clear -	Sort & Filter *	Find &
Clipboard	5		Font	5					Numbe					Styles				0	ells		Editing	
,	• ±]	× 🗸	<i>fx</i> =\$G\$	2*POWER	(\$F\$2;B2)																	
А	В	С	D	E F	G	н	Ι	J	К	L	м	N	0	Р	Q	R	S	т	U	V	W	х
			Exponen tial regressio																			
ime	Period	volume	n	Coeffic	ients of fu	nction																
нварь	1	1286,5	1320,98		1270,29																	
евраль	2	1282,7	1373,69		0,06267																	
арт	3	1224,8			0,13122			F	3500						1							
прель	4	1292,6	1485,5	50,67					500													
ай	5	1667,3	1544,78		0,29274			- 3	3000				-									
юнь	6	1672,9	1606,42			;B2:B20;;1)}	_					L									
юль	7	1667,5	1670,52		nination co	efficient			2500		\wedge	1	I									
вгуст ентябрь	8	2001,7	1737,17 1806,49	0,749					2000					Sales volume								
ктябрь	10		1808,49							_												
оябрь	11		1953,53						1500	P				Exponential								
екабрь	12		2031,48						1000					regression								
нварь	13		2112.54	{=\$G\$2	*POWER(F\$2:B14)}			1000													
евраль	14	1971,6	2196,83						500													
Тарт	15	1989,1	2284,49																			
прель	16	2138,2	2375,64								9 11 13											
Іай	17	2475,2	2470,44						1 3	57	9 11 13	15 17	19									
юнь	18																					
юль	19	2994,1	2671,52																			
I	Полином	Лог.фу	нкц Гип	ерб.функі	Экспо	м.функц	Степен	.функц	S-образ	Резул	ьт.сгл.дан	Полино	M.F (+)	E 4								

Fig. 11.4. Screenshot of the construction and evaluation of the quality of the exponential regression function from one factor by the formula (11.2)

11.1.2. Forecasting functions, depending on several factors.

Exponential regression.

The exponential function of the regression from many factors in the general form is calculated by the following formula:

$$Y = b_0 * (b_1)^{X_1} * (b_2)^{X_2} * \dots * (b_m)^{X_m},$$
(11.3)

Where the coefficients $b_0, b_1, b_2, \dots, b_m$ are determined from the original data X_1, X_2, \dots, X_m and Y.

To calculate the coefficients of the exponential function of regression from many factors, the standard LGRFPRIBL function is used.

The following is an example of calculating the exponential forecasting performance indicators from three factors for given source data:

$$Y = b_0 * (b_1)^t * (b_2)^{X_1} * (b_3)^{X_2},$$
(11.4)

Time	Period	Production costs	Cost of	Sales	Trend
			advertising	volume	
January	1	905,8	199,8	1282	1303,2
February	2	902,5	211,5	1292,7	1346,58
March	3	903	206,6	1226,8	1375,57
April	4	889,8	225,7	1392,6	1585,13
May	5	889,8	219	1647,3	1636,04
June	6	892,8	236,7	1672,9	1540,28
July	7	888,3	231,3	1660,5	1678,17
August	8	875,8	241,1	2011,7	1956,56
September	9	883,9	238,1	2351,9	1807,64
October	10	875,1	248,1	2468,5	2010,12
NOVEMBER	11	871,6	256,9	2746,2	2095,7
December	12	873,8	251,9	1942,7	2095,67
January	13	868,2	273,1	1901,1	2183,13
February	14	866,3	264,5	1971,6	2317,85
March	15	862,1	267,1	1989,1	2471,69
April	16	866,6	282,9	2139,2	2292,83
May	17	862,5	287,5	2474,2	2431,14
June	18	863,9	286,3	2393,6	2435,33
July	19	858,5	285,3	2990,1	2657,8

Table 11.4 The initial data and the construction of the trend of the exponential regression function from three factors according to the formula (11.4)

The calculation of the trend of the exponential forecasting function from three factors: $Y = b0 * (b1) ^ t * (b2) ^ X1 * (b3) ^ X2$ is performed in Excel using the formula:

= \$ K \$ 3 * \$ J \$ 3 ^ B2 * \$ I \$ 3 ^ C2 * \$ H \$ 3 ^ D2.

Table 11.5. The results of calculating the characteristics of the exponential regression function from three factors using formula (11.4)

Incel costor	1 80 (81)	(02) /1	(00) 72
0,997792	0,987413	1,016948	191595799
0,00664	0,009145	0,035555	8,7102376
0,725445	0,150918	#Н/Д	#Н/Д
13,21128	15	#Н/Д	#Н/Д
0,902716	0,341646	#Н/Д	#Н/Д

Regression Y=b0*(b1)^t*(b2)^X1*(b3)^X2

Determination coefficient 0,725445



Fig. 11.5. The graphs of the initial data (row 1), the exponential regression function from the three factors (row 2) by the formula (11.4)

Image: Styles	🔏 Cut 🖺 Copy		Calibri	- 11	- A A	= = =	»- ~	Wrap Text	G	Seneral	*	≢	a see	Normal	Bad	l i	Good	· ·	E	× 🗊	∑ AutoSun ↓ Fill •		a a
Image F -sk52*tij2x402*tij2x422*sij2x422*sij2x422*sij2x422*sij2x422*sij2x422*sij2x422*sij2x422*sij2x422*sij2x422*sij2x42*			B <i>I</i> <u>U</u>	• 🖽 • •	<u>ð</u> - <u>A</u> -	= = =	€E 4E [Merge & Ci	enter - 🕻	- % ,	€.0 .00 C	onditional F prmatting *	ormat as Table *	Neutral	Calo	culation	Check Ce		Insert De	lete Format			
A B C D E G H I J K L M N O P Q R S T U V V Ima Period Image Image <thimage< th=""> Image Image <t< th=""><th>Clipboard</th><th>5</th><th></th><th>Font</th><th>5</th><th></th><th>Alignme</th><th>nt</th><th>5</th><th>Number</th><th>- Fa</th><th></th><th></th><th></th><th>Styles</th><th></th><th></th><th></th><th>C</th><th>ells</th><th></th><th>Editing</th><th></th></t<></thimage<>	Clipboard	5		Font	5		Alignme	nt	5	Number	- Fa				Styles				C	ells		Editing	
Ime Period Ime Freduction Engression Public (1)+(*)(2)/(1)+(*)(2)/(1)-(3)/(2) 1me Period on costs ng volume n anuary 1 905,8 199,8 1282 1303,197 0,997792 0,987413 1016948 10159579 0	*		× 🗸	<i>fx</i> =\$K	\$2*\$J\$2^B	2*\$I\$2^C2*	\$H\$2^D2																
Ima Pariod Oracut Single Single <td>А</td> <td>В</td> <td>С</td> <td>D</td> <td>E</td> <td>F</td> <td>G</td> <td>н</td> <td>Ι</td> <td>J</td> <td>К</td> <td>L</td> <td>М</td> <td>N</td> <td>0</td> <td>Р</td> <td>Q</td> <td>R</td> <td>S</td> <td>Т</td> <td>U</td> <td>v</td> <td>W</td>	А	В	С	D	E	F	G	н	Ι	J	К	L	М	N	0	Р	Q	R	S	Т	U	v	W
me period oppose n Regression >-b0*01*1**tic2/9x1*15/0729 i <th< th=""><th></th><th></th><th>Producti</th><th></th><th>Salar</th><th>three factors</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></th<>			Producti		Salar	three factors																	
nuary 1 905.8 1998 1282 1201.9772 0,98743 1,016948 1915599 branary 2 902,5 211,5 122,7 1345,581 0,0064 0,009145 0,035555 8,7102376 arch 3 003 205,6 1226,8 1355,573 0,72544 0,00914 8 N/A 8N/A ne 6 802,8 236,7 157,9 1340,282 HN/A 18N/A 8N/A ne 6 802,8 236,7 157,9 1340,282 HN/A 18N/A 8N/A sgsta 8 875,8 241,1 2011,7 1956,56 ptember 9 883,9 233,1 251,9 1307,611 Determination coefficient maary 13 868,2 273,1 1901,1 2131,314 branary 14 866,3 264,5 1270,9 1340,281	me P	Period				-		Regression	Y=b0*(b1	\^t*(h2)^X1	1*(h3)^X2												
braney 2 902,5 211,5 1292,7 1342,6 1375,573 0,725445 0,150918 M/A		1										9											
arch 3 903 2066 1226,8 1375,73 0,725445 0,150918 M/A																							
ay 5 889,8 219 1647,3 1650,4 0.902216 0,34164 #N/A #N/A <td>arch</td> <td>3</td> <td>3 903</td> <td>206,6</td> <td></td>	arch	3	3 903	206,6																			
ne 6 802,8 236,7 197,9 190,22 940,4 91	oril	4	889,8	225,7	1392,6	1585,128		13,21128	1	5 #N/A	#N/A												
y 7 8883 231,3 1660,5 1078,174 9 ygust 8 875,8 241,1 2017,7 1956,56 9 ytobe 9 883,3 238,1 231,9 1007,641 Determination coefficient 900 tober 10 875,1 248,1 2017,7 1956,56 9 6 9 tober 10 875,1 248,1 2017,2 0,725445 900	ay	5	5 889,8	219	1647,3	1636,04		0,902716	0,34164	5 #N/A	#N/A												
agest 88 875,8 241,1 2011,7 1955,66 additional state ptember 9 883,9 238,1 2351,9 1807,641 Determination coefficient 0 875,1 245,1 268,0 2010,122 0,725445 0 875,1 255,9 2746,2 2095,697 200 nuary 13 886,2 273,1 1901,1 213,851 200 arch 15 882,1 257,1 1891,2 2137,851 200 arch 15 882,2 223,2 2293,62 2293,62 2393,6 yil 16 866,6 282,9 2139,2 2292,83 200 200 ne 18 863,9 285,3 299,0 267,805 200 200 y 17 858,5 285,3 299,0 267,805 200 200 action 12 3 4 5 6 7 8 9 10111213141516171819 200 200 200	ne	6	5 892,8	236,7	1672,9	1540,282		#N/A	#N/A	#N/A	#N/A												
agast 8 875,8 241,1 2011,7 195,56 1 <td>ily</td> <td></td> <td>7 888,3</td> <td>231,3</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>3500</td> <td></td>	ily		7 888,3	231,3								3500											
10 875,1 248,1 2465,5 2010,122 0,725445 1 0 0 0VEMBE 11 871,6 256,9 246,2 2095,697 0 <td></td> <td>0000</td> <td></td>												0000											
OverMele 11 871.6 256.9 2246.2 2095.697 200 December 12 873.8 251.9 194.7 2095.674 200 Darrow 13 886.2 273.1 1901.1 1218.134 200 5ales volume 5ales volume <td< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>tion coeff</td><td>icient</td><td></td><td>3000</td><td></td><td></td><td></td><td>- /</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></td<>									tion coeff	icient		3000				- /							
eccember 12 873.8 251.9 1942.7 2095.674				,				0,725445							Λ								
nuary 13 888,2 273,1 1901,1 2183,134 bruary 14 866,3 266,5 1971,6 2317,851 arch 15 882,1 267,1 1989,1 2471,686 sile solume regression three factors 100 1 2 3 4 5 6 7 8 9 10111213141516171819 1 2 3 4 5 6 7 8 9 10111213141516171819												2500				\sim							
mbm/ary 10 866,5 264,5 1971,6 2317,851 arch 15 862,1 267,1 1989,1 2471,866 arki 16 866,6 282,0 2139,2 2292,83 ay 17 862,5 287,5 2474,2 2431,139 ne 18 883,9 286,3 2393,6 2435,335 y/ 19 858,5 285,3 2990,1 2657,805												2000			λ	1							
arch 15 862,1 267,1 1989,1 2471,666 1999,1 2471,666 1999,1 2471,666 100												2000		\sim				volume					
vrii 16 866,6 282,9 2139,2 2292,83 av 17 862,5 287,5 2474,2 243,139 ne 18 863,9 285,3 2990,1 2657,805 by 19 858,5 285,3 2990,1 2657,805												1500	1	\checkmark				ssion three fa	ctors				
av 17 862,5 287,5 2474,2 2431,139 ne 18 863,9 286,3 2393,6 2435,335 y 19 858,5 285,3 2990,1 2657,805			,									-	<				0						
ne 18 8639 2863 29336 245335 ly 19 8585 2853 2990,1 2657,805 0 1 2 3 4 5 6 7 8 9 10111213141516171819												1000											
19 858,5 285,3 2990,1 2657,805 1 2 4 5 6 7 9 10111213141516171819				,																			
												500											
	,	-	,.	,		,						0											
													2345	678910	111213141	516171819							
Descuer Backway Funnah Auron Backway Funnana these fasters Connou duran Conference Desc. Co																							
Designed Designed Lange Company three factors Company three factors																							
			l llor du	LINE L EM	non6 duur	- Avene	u dunun	Exponen	three fa	ctore CT	anou duur	L S.ofin	aa Dr	av ()									

Fig. 11.6 Screenshot of the construction and evaluation of the quality of the exponential function of the regression from three factors by the formula (11.4)

In Excel, there are a number of implemented standard functions that allow you to approximate the original data. And a number of functions, such as: linear,

exponential, logarithmic, polynomial, power functions allow you to build trend lines, determine the coefficient of determination, make a prediction for a specified number of periods forward (backward). Often it is necessary to build the forecasting function using the standard LINEST and LOGEST functions, calculating the coefficients and statistical characteristics of the approximating regressions. This scheme is used in the examples discussed above.

11.2 PRACTICAL LESSON 11 and IWS 11

Subject. Highlighting the trend function for the prediction task.

Plan of the lesson.

1) Study the trend allocation scheme for the prediction task presented above.

2) Practical lesson and task of the IWS. Apply the above described schemes to the selection of the trend function for its prediction task.

3) Analyze the results.

Literature.

1. Minko A.A. Predicting in business using Excel. M.: Eksmo, 2007, -208 p.

TOPIC 12

SEASONAL COMPONENT IN PREDICTION FUNCTIONS

12.1. Lecture material.

The seasonal component in the prediction functions determines periodic changes in the values of the prediction function. The type (model) of interaction between the trend component and the seasonal component of the prediction function can be additive or multiplicative. In the additive model of the prediction function, seasonal changes are added to the trend, i.e. F(t) = T(t) + S(t), where T and S denote, respectively, the trend and seasonal components of the function f, and in the multiplicative model, the trend and seasonal variations are multiplied; F(t) = T(t) * S(t).

The scheme for allocating the seasonal component from the initial data and including it in the prediction function has the following stages:

1) Based on the input data, the trend of the predicted variable (trend of the prediction function) is constructed T. The previous chapter shows how this is done. This is usually done by constructing a regression function.

2) Seasonal components s_i are calculated for each regression data point using the formulas:

- for the additive model $s_i = y_i - T_i$, where i = 1, ..., n;

- for the multiplicative model $s_i = y_i/T_i$, where i = 1, ..., n,

Where y_i are the observed values of the predicted variable at the i-th data point; T_i - the value of the trend of the predicted variable in the i-th data point; Total points of input data n.

3) The seasonal components $s_1, s_2, ..., s_n$ are divided into groups according to k seasons, for example, 12 months of the year; For each season there is an average \overline{S}_l

4) The seasonal coefficients S_1 , S_2 , ..., S_k are calculated:

- for the additive model $S_i = \overline{S_i} - m$, $i = \overline{1 \dots k}$, where m is the correction factor for computing each seasonal coefficient: $m = \frac{1}{k}(\overline{S_1} + \overline{S_2} + \dots + \overline{S_k})$

- for the multiplicative model $S_i = \overline{S_i} * m$, $i = \overline{1 \dots k}$, where m is the correction factor for computing each seasonal coefficient: $m = \frac{k}{\overline{S_1} + \overline{S_2} + \dots + \overline{S_k}}$.

5) The prediction function with seasonal coefficients is calculated:

- for the additive model $f_i = T_i + S_p$;;

- for the multiplicative model $f_i = T_i * S_p$.

Where S_p is the seasonal coefficient for the p-th season, which includes T_i -the trend function at the i-th data point,

6) The residuals (prediction errors) are calculated $e_i = y_i - f_i$.

7) The remainder of the constructed data model is analyzed and the model quality indicators are computed.

If the residue analysis and the quality indicators of the constructed model are satisfactory, i.e. the constructed model is adequate to the initial data, then a prediction is made for the constructed model. If analysis of the residuals shows unsatisfactory results, then a change in the trend function approximating the initial data is necessary.

8) Calculation of the prediction. If it is necessary to predict the depended variable at time t_{n+j} , then the value of the prediction function f_{n+j} is calculated using the formulas:

- for the additive model $f_{n+j} = T_{n+j} + S_p$,

- for the multiplicative model $f_{n+j} = T_{n+j} * S_p$.

As can be seen from the formulas, in order to make a prediction, it is necessary to calculate the value of the trend T_{n+j} at the time of the forecast. The trend value of the prediction function T_{n+j} at time t_{n+j} in the case when the predicted variable depends only on time is not difficult: it is calculated by the regression formula of this prediction function for the required time moment.

The situation is more complicated if the prediction function depends on several factors. In this case, to calculate the predicted values of the trend prediction function, it is necessary to know (calculate) the predicted values of all factors. In this case, it is need first calculate the predicted values of the factors, and then calculate the predicted values of the function. That is, to predict the values of

factors, it is need follow the same steps as for the prediction function. For this it is necessary for each factor to construct a function of regression by another factor, usually in this capacity the time factor is taken. And the regression function for this factor is to calculate the predicted values of this factor at the necessary instants of time.

Below is an example of constructing a prediction function, taking into account the allocation and use of the seasonal component in the prediction function.

1) Based on the input data, the trend of the predicted variable (trend of the prediction function) is constructed T. In the example, this is done by constructing the regression function. In the example, the regression equation is taken

 $Y = b_0 + b_1 t + b_2 t^2 + b_3 X_1 + b_4 X_2$, where t - time, X_1 - production costs, X_2 -advertising costs.

Period	Period	Production	Advertising	Sales	Regression
	in	costs	costs	volume	function
	square				
1	1	890,8	200,8	1286,5	1329,349
2	4	896,4	210,5	1282,7	1294,721
3	9	903	208,6	1224,8	1397,418
4	16	888,8	226,7	1292,6	1446
5	25	889,8	230,62	1667,3	1521,13
6	36	890,8	238,7	1672,9	1537,63
7	49	888,3	232,3	1667,5	1774,925
8	64	885,8	240,1	2001,7	1822,27
9	81	883,9	238,1	2352,5	1987,469
10	100	875,1	248,1	2468,5	2061,72
11	121	877,73	253,58	2012,2	2074,541
12	144	873,8	261,05	1982,7	2123,879
13	169	878,2	275,1	1901,1	1998,154
14	196	866,3	265,5	1971,6	2343,438
15	225	865,1	266,1	1989,1	2442,189
16	256	876,6	282,6	2138,2	2199,5

Table 12.1 Initial data and constructed regression function.

17	289	862,5	286,5	2475,2	2377,942
18	324	865,9	286,3	2395,6	2427,682
19	361	857,5	284,3	2994,1	2616,842

Calculation of regression coefficients by the standard LINEST function.

The regression equation is $Y = b0 + b1 * t + b2 * t ^ 2 + b3 * X1 + b4 * X2$

2) Seasonal components s_i are calculated for each regression data point using the formulas:

- for the additive model $s_i = y_i - T_i$, where i = 1, ..., n;

Where y_i are the observed values of the predicted variable in the i-th data point, T_i is the trend value of the predicted variable in the i-th data point.

3) The seasonal components $s_1, s_2, ..., s_n$ are divided into groups according to k seasons; for each season there is an average $\overline{S_i}$.

4) The seasonal coefficients S_1 , S_2 , ..., S_k are calculated:

- for the additive model $S_i = \overline{S_i} - m$, $i = \overline{1 \dots k}$, where m is the correction factor for computing each seasonal coefficient: $m = \frac{1}{k}(\overline{S_1} + \overline{S_2} + \dots + \overline{S_k})$

For the example in question, Table 12.2 shows the values of the calculated seasonal components, the mean seasonal component, seasonal coefficients.

Table 12.2 Calculation of seasonal components, mean seasonal component, seasonal coefficients

Sales	Regression	Residuals	Average	Seasonal
volume	function		season	coefficient
1286,5	1329,349	-42,8493	-69,9517	-101,107
1282,7	1294,721	-12,0205	-191,929	-223,084
1224,8	1397,418	-172,618	-312,853	-344,008
1292,6	1446	-153,4	-107,35	-138,505
1667,3	1521,13	146,1696	121,7136	90,55862
1672,9	1537,63	135,2696	51,59384	20,43882
1667,5	1774,925	-107,425	134,9165	103,7615

1822,27	179,4298	179,4298	148,2747
1987,469	365,0307	365,0307	333,8757
2061,72	406,7796	406,7796	375,6246
2074,541	-62,3407	-62,3407	-93,4957
2123,879	-141,179	-141,179	-172,334
1998,154	-97,0541	-97,0541	
2343,438	-371,838	-371,838	
2442,189	-453,089	-453,089	
2199,5	-61,2998	-61,2998	
2377,942	97,25768	97,25768	
2427,682	-32,082	-32,082	
2616,842	377,2578	377,2578	
	1987,469 2061,72 2074,541 2123,879 1998,154 2343,438 2442,189 2199,5 2377,942 2427,682	1987,469365,03072061,72406,77962074,541-62,34072123,879-141,1791998,154-97,05412343,438-371,8382442,189-453,0892199,5-61,29982377,94297,257682427,682-32,082	1987,469365,0307365,03072061,72406,7796406,77962074,541-62,3407-62,34072123,879-141,179-141,1791998,154-97,0541-97,05412343,438-371,838-371,8382442,189-453,089-453,0892199,5-61,2998-61,29982377,94297,2576897,257682427,682-32,082-32,082

Excel formulas for computed columns (on the first line, column names and line numbering, see Figure 12.1):

The regression function is = \$ R \$ 2 + \$ Q \$ 2 * A2 + \$ P \$ 2 * B2 + \$ O \$ 2 * C2 + \$ N \$ 2 * D2;

Residuals of regression function = E2-F2;

Mean \overline{S}_{l} - = AVERAGE (G2; OFFSET (G2; 12; 0));

The seasonal coefficients S_i - = H2-AVERAGE (\$ H \$ 2: \$ H \$ 13).

5) The forecasting function with seasonal coefficients is calculated:

- for the additive model $f_i = T_i + S_p$,

Where S_p is the seasonal coefficient for the p-th season, which includes T_i - the trend function at the i-th data point.

For the example in question, Table 12.3 shows the values of the prediction function.

Table 12.3 Construction of the prediction function as the sum of the trend and the calculated seasonal coefficients

Sales volume	0	Remaining regression function	Average season	Seasonal coefficient	Prediction function
1286,5	1329,349	-42,8493	-69,9517	-101,107	1228,243
1282,7	1294,721	-12,0205	-191,929	-223,084	1071,636

-					
1224,8	1397,418	-172,618	-312,853	-344,008	1053,409
1292,6	1446	-153,4	-107,35	-138,505	1307,495
1667,3	1521,13	146,1696	121,7136	90,55862	1611,689
1672,9	1537,63	135,2696	51,59384	20,43882	1558,069
1667,5	1774,925	-107,425	134,9165	103,7615	1878,686
2001,7	1822,27	179,4298	179,4298	148,2747	1970,545
2352,5	1987,469	365,0307	365,0307	333,8757	2321,345
2468,5	2061,72	406,7796	406,7796	375,6246	2437,345
2012,2	2074,541	-62,3407	-62,3407	-93,4957	1981,045
1982,7	2123,879	-141,179	-141,179	-172,334	1951,545
1901,1	1998,154	-97,0541	-97,0541		1897,047
1971,6	2343,438	-371,838	-371,838		2120,354
1989,1	2442,189	-453,089	-453,089		2098,181
2138,2	2199,5	-61,2998	-61,2998		2060,995
2475,2	2377,942	97,25768	97,25768		2468,501
2395,6	2427,682	-32,082	-32,082		2448,121
2994,1	2616,842	377,2578	377,2578		2720,604

The Excel formula for the prediction function (on the first line, column names and line numbering, see Figure 12.1): = F2 + I2.

6) Calculation of residuals (prediction errors) $e_i = y_i - f_i$.

For the example in question, Table 12.4 shows the values of the prediction residuals (errors).

Sales	Regressio	Remainin	Average	Seasonal	Predictio	The
volum	n function	g	season	coefficien	n	predictio
e		regressio		t	function	n
		n function				residuals (errors)
1286,5	1329,349	-42,8493	-69,9517	-101,107	1228,243	58,25742
1282,7	1294,721	-12,0205	-191,929	-223,084	1071,636	211,0636
1224,8	1397,418	-172,618	-312,853	-344,008	1053,409	171,3909
1292,6	1446	-153,4	-107,35	-138,505	1307,495	-14,8951
			121,713			
1667,3	1521,13	146,1696	6	90,55862	1611,689	55,61099
			51,5938			
1672,9	1537,63	135,2696	4	20,43882	1558,069	114,8308

			134,916			
1667,5	1774,925	-107,425	5	103,7615	1878,686	-211,186
			179,429			
2001,7	1822,27	179,4298	8	148,2747	1970,545	31,15502
			365,030			
2352,5	1987,469	365,0307	7	333,8757	2321,345	31,15502
			406,779			
2468,5	2061,72	406,7796	6	375,6246	2437,345	31,15502
2012,2	2074,541	-62,3407	-62,3407	-93,4957	1981,045	31,15502
1982,7	2123,879	-141,179	-141,179	-172,334	1951,545	31,15502
1901,1	1998,154	-97,0541	-97,0541		1897,047	4,052622
1971,6	2343,438	-371,838	-371,838		2120,354	-148,754
1989,1	2442,189	-453,089	-453,089		2098,181	-109,081
2138,2	2199,5	-61,2998	-61,2998		2060,995	77,20515
			97,2576			
2475,2	2377,942	97,25768	8		2468,501	6,699056
2395,6	2427,682	-32,082	-32,082		2448,121	-52,5208
			377,257			
2994,1	2616,842	377,2578	8		2720,604	273,4963

The Excel formula for calculating the remainders of the prediction function (for the first line, column names and line numbers, see Figure 12.1): = E2-J2.

7) Analysis of the constructed data model for quality indicators.

Seasonal coefficient	Prediction function	the prediction residuals (errors)		b4	b3	b2	b1	b0
-101,107	1228,243	58,25742		-13,0796	-10,1777	-2,10969	155,5671	12868,56
-223,084	1071,636	211,0636	Regressior	n equation Y	=b0+b1*t+b	2*t^2+b3*>	(1+b4*X2	
-344,008	1053,409	171,3909						
-138,505	1307,495	-14,8951						
90,55862	1611,689	55,61099	Determina	tion coeff re	gression equ	ation		
20,43882	1558,069	114,8308		0,764315				
103,7615	1878,686	-211,186						
148,2747	1970,545	31,15502	Determina	tion coeff of	prediction f	unction		
333,8757	2321,345	31,15502		0,936665				
375,6246	2437,345	31,15502						
-93,4957	1981,045	31,15502						
-172,334	1951,545	31,15502						

1897,047	4,052622			
2120,354	-148,754			
2098,181	-109,081			
2060,995	77,20515			
2468,501	6,699056			
2448,121	-52,5208			
2720,604	273,4963			

Excel formula for calculating the coefficient of determination of the regression function (see column 12.1 for column names and line numbering):

```
= 1- (SUMSQ (G2: G20) / SUMSQ (E2: E20-AVERAGE (E2: E20))).
```

The Excel formula for calculating the determination coefficient of the prediction function (column names and line numbering, see Figure 12.1): = 1- (SUMSQ (K2: K20) / SUMSQ (E2: E20-AVERAGE (E2: E20))).

The value of the coefficient of determination of the prediction function is 0.936665, which is much better than the value of the coefficient of determination of the regression function 0, 764315, which indicates that the quality of the model with seasonal component is more qualitative than the model without taking it into account.

λα ΩΩ		Ci	alibri	- 11	· A A	= = =	≫	Wrap Text	G	General	~	₽	1	Normal	Bad		Good	· ·	ŧ	i 🏹	∑ AutoSum ↓ Fill *	ĭŽŢ	H	
	rmat Paint	er B	ΒIU	• == •	<u>ð</u> - <u>A</u> -	553	€ 12 €	Merge & C	enter - C	≩ •% ,	€0 .00 C	onditional ormatting *	Format as Table *	Neutral	Calcu	lation	Check Ce	T. T.	Insert *	Delete Forma	t 🧶 Clear -	Sort & Filter *		
Clipbo	ard	5		Font	5		Alignmer	t	5	Number	G ₂				Styles					Cells	E	diting		
	* 1	\times	\sim	fx {=1-((SUMSQ(K2	:K20)/SUM	5Q(E2:E20-	AVERAGE(E	2:E20)))}															
А	В		С	D	E	F	G	н	I	J	К	L	м	N	0	Р	Q	R	S	Т	U	V	w	
							Residuals				Residuals													
						L .	of				of													
				Cost of		Regressio			-	Predictio	predictio													
riod				advertisi		n	n	of season		n	n													
rioa	squa	1	on costs 890,8	ng 200,8	volume	function			coeff	function 7 1228,243			-12 079	6 -10,1777	-2 10060	155 5671	12969 56							
	2	4	896,4	210,5			-12,0205							Y=b0+b1*t										
	3	0	903	208,6						3 1053,409		-	requation	1 00.01 0		5 ALIST A								
	3	16	888,8	208,0	1224,0			,		5 1307,495														
	5	25	889,8	230,62									oeff of reg	ression funct	tion									
	6	36	890,8	238,7						2 1558,069			0,76431											
	7	49	888,3	232,3						5 1878,686												_		
	8	64	885,8	240,1	2001,7	1822,27	179,4298	179,4298	148,2747	7 1970,545	31,15502	Determ of	oeff of pre	diction funct	ic 3500									
	9	81	883,9	238,1	2352,5	1987,469	365,0307	365,0307	333,8757	7 2321,345	31,15502		0,93666	5	3000 -				1	-				
		100	875,1	248,1		2061,72	406,7796	406,7796	375,6246	5 2437,345	31,15502				2500				K	7				
		121	877,73	253,58				-62,3407		7 1981,045					_			\sim		-	ales volume			
		144	873,8	261,05					-172,334	1951,545					2000									
		169	878,2	275,1				-97,0541			4,052622				1500		P				egression functio			
		196 225	866,3 865,1	265,5 266,1				-371,838 -453,089			-148,754				1000 -	\mathbf{e}				p	rediction functio	n		
		256	876,6	282,6				-61,2998			77,20515				-					— p	rediction errors			
		289	862,5	282,0				97,25768			6,699056				500	~	~		,	s	eason coeff			
		324	865,9	286,3							-52,5208				0 -	\frown		\sim	1					
		361	857,5	284,3				377,2578			273,4963				-500	1 😼 5	7 9 11	13 15 17	19 21	23				
	20	400	859,064	296,762	2	2511,216				2659,491														
	21	441 8	856,9897	301,5146	5	2539,235				2873,111					-1000									
		484 8	854,9154	306,2672	2	2563,035				2938,66														
	23	529 8	852,8411	311,0198	3	2582,616				2489,12														
	Адд м	одел	lb Ce30	мем но	мультипл	мод Пр	оиз затр-	прогноз	Прогноз	з зат рекл	Прогно	з объема	а продаж	(+) :	4									

Fig. 12.1 A screenshot of the construction of the prediction function taking into account the seasonal component.

8) Calculation of the prediction.

The example uses the regression equation

 $Y = b_0 + b_1 t + b_2 t^2 + b_3 X_1 + b_4 X_2$, where t - time, X_1 - production costs, X_2 - advertising costs.

Since the prediction function depends on several factors, to calculate the predict from the predicted variable, it is necessary to make a prediction for all factors of the regression equation. The prediction by time and square of time is not difficult. It is necessary to construct regression functions for X_1 - production costs, X_2 - advertising costs.

The prediction of values of the factor "production costs" is presented below.

Period	Production	Prediction	Prediction
	costs	period	of
			production
			costs
1	890,8	20	859,064
2	896,4	21	856,9897
3	903	22	854,9154
4	888,8	23	852,8411
5	889,8		
6	890,8		
7	888,3		
8	885,8		
9	883,9		
10	875,1		
11	877,73		
12	873,8		
13	878,2		
14	866,3		
15	865,1		
16	876,6		
17	862,5		
18	865,9		

Table 12.6. Input and prediction for the "production cost" factor





Fig. 12.2. Graph of the regression function with the prediction for the "production cost" factor

	🔏 Cut		Calibri	- 11	· A Ă	= = =	87 -	🗃 Wrap Te	đ	General	-			Normal	Bad		Good		ŧ	i	∑ AutoSum ↓ Fill *	Ž AT	#
	ስ Copy 💕 Form	at Painter	ΒIU	- 🎛 - 🕹	<u>></u> - <u>A</u> -	= = =	62 3 2	🔄 Merge 8	Center 🝷	\$ -%,			Format as	Neutral	Calcu	ulation	Check Co	ell 🔻	Insert D	elete Format	A Classe	Sort & Fi Filter * Se	ind &
c	lipboard			Font	5		Alignr	nent	5	Number	5	Tormatang	Tuble		Styles					Cells		Editing	neer
		- = >	< 🗸	fx =-2,	0743*C2+9	900,55																	
	A	В	С	D	E	F	G	н	I	J	К	L	м	N	0	Р	Q	R	S	т	U	V	w
				Predictio n of																			
		Producti																					
er	iod 1	on costs 890,8		on costs 859,064																			
_	2	896,4		856,9897																			
	3			854,9154																			
	4		23	852,8411																			
_	5									Produ	ction	costs											
_	6												/=-2,0743	+ 900,55									
	/ 8												$R^2 = 0,$	681									
	9							30	_														
	10							80															
	11							70		M	Δ		produ	tion costs									
_	12							50		9	M												
_	13						8	50					 Linear costs) 	production									
_	14 15						8	40					coscoj										
	16						8	30															
	17						8	20															
	18							1234	5678	9 101 11 21 31 41	516171819	920212223											
	19	857,5																					
>		Адд моде	ль сезо	мем на	мультипл	мод Пр	оиз зат	р-прогноз	Прог	юз зат рекл	Про	гноз объема	а продаж	(+) :	4								
									_												圖 町		-+

Fig. 12.3. Screenshot of the prediction for the factor "production costs"

Below is a prediction of the values of the factor "advertising costs"

Table 12.6. Baseline data and prediction for the "advertising costs" factor

Period	Advertising costs	Prediction period	Prediction of advertising costs	
1	200,8	20	296,762	
2	210,5	21	301,5146	
3	208,6	22	306,2672	
4	226,7	23	311,0198	
5	230,62			
6	238,7			
7	232,3			
8	240,1			
9	238,1			
10	248,1			
11	253,58			
12	261,05			
13	275,1			
14	265,5			
15	266,1			
16	282,6			
17	286,5			
18	286,3			
19	284,3			



Fig. 12.4. Graph of the regression function with the prediction for the "advertising costs" factor

🔏 Cut		Calibri	- 11	· A A	= = =	87 -	🔐 Wra	ap Text		General	-			Normal	Bad		Good		÷	× 🖬	∑ AutoSur ↓ Fill *	° č <mark>∕</mark> z▼	#
🖺 Copy ؇ Forma	at Painter	ΒIU	• = •	<u>ð</u> - <u>A</u> -	5 5 5	€E 3 E	🔁 Mer	rge & C	enter -	- %,	00, 0, → 0,€ 00,	Conditional Formatting *	Format as	Neutral	Calcu	ulation	Check Ce		Insert De	elete Format	Clear •	Sort & Filter * S	Find &
Clipboard			Font	6		Align				Numbe					Styles					Cells		Editing	
	- E 2	< 🗸 .	<i>fx</i> =4,	7526*C2+2	01,71																		
A	В	С	D Predictio	E	F	G		н	I	J	К	L	М	N	0	Р	Q	R	S	Т	U	V	W
			n of																				
		Predictio																					
riod 1	ng costs 200,8	n period	ng costs 296,762	1																			
2	210,8		301,5146																				
3	208,6		306,2672																				
4	226,7	23	311,0198	3						Adv	ertisii	ng costs											
5	230,62						350						y = 4,7526	× + 201 71									
6	238,7						300					_	R ² = 0										
7	232,3																						
8	240,1						250		~														
10	238,1						200	~					Adv	ertising costs									
11	253,58						150						Line	ar (Advertising									
12	261,05						100						cost										
13	275,1						50																
14	265,5																						
15	266,1						0 -	1.1.2	4567	8.0.10111212	44 64 64 74	91000010000											
16 17	282,6 286,5							123	4307	8 9 10111213	141310171	81920212225											
17	286,3																						
19	280,3																						
	, <u>.</u>																						
									_														
>	Адд моде	ль сезо	мем но	мультипл	мод П	роиз зат	гр-прог	HO3	Прогн	оз зат рек	Про	гноз объема	продаж	+ :	4								

Fig. 12.5. Screenshot of the prediction for the factor "advertising costs"

The prediction for the prediction variable (sales volume) is presented below.

Table 12.6. Initial data and prediction for the prediction function for time periods 20-23

Пе	Произв	Затраты						
рио	одстве	на	Объем	Функция	Остаток			
д	нные	реклам	прода	регресси	функции	Среднее	Сезонны	Функция
	затраты	у	ж	и	регрессии	сезона Si	е коэф	прогноз
1								
	890,8	200,8	1286,5	1329,349	-42,8493	-69,9517	-101,107	1228,243
2								
	896,4	210,5	1282,7	1294,721	-12,0205	-191,929	-223,084	1071,636
3								
	903	208,6	1224,8	1397,418	-172,618	-312,853	-344,008	1053,409
4								
	888,8	226,7	1292,6	1446	-153,4	-107,35	-138,505	1307,495
5								
	889,8	230,62	1667,3	1521,13	146,1696	121,7136	90,55862	1611,689
6								
	890,8	238,7	1672,9	1537,63	135,2696	51,59384	20,43882	1558,069
7								
	888,3	232,3	1667,5	1774,925	-107,425	134,9165	103,7615	1878,686
8								
	885,8	240,1	2001,7	1822,27	179,4298	179,4298	148,2747	1970,545
9								
	883,9	238,1	2352,5	1987,469	365,0307	365,0307	333,8757	2321,345
10								
	875,1	248,1	2468,5	2061,72	406,7796	406,7796	375,6246	2437,345

23	852,841	311,019		2582,616				2489,12
22	854,915	306,267		2563,035				2938,66
21	856,989	301,514		2539,235				2873,111
20	859,064	296,762		2511,216				2659,491
19	857,5	284,3	2994,1	2616,842	377,2578	377,2578		2720,604
18	865,9	286,3	2395,6	2427,682	-32,082	-32,082		2448,121
17	862,5	286,5	2475,2	2377,942	97,25768	97,25768		2468,501
16	876,6	282,6	2138,2	2199,5	-61,2998	-61,2998		2060,995
15	865,1	266,1	1989,1	2442,189	-453,089	-453,089		2098,181
14	866,3	265,5	1971,6	2343,438	-371,838	-371,838		2120,354
13	878,2	275,1	1901,1	1998,154	-97,0541	-97,0541		1897,047
12	873,8	261,05	1982,7	2123,879	-141,179	-141,179	-172,334	1951,545
11	877,73	253,58	2012,2	2074,541	-62,3407	-62,3407	-93,4957	1981,045



Fig. 12.6. The prediction graphic for the prediction function, the seasonal coefficients graphic, the prediction function error graphic.

12.2 PRACTICAL LESSON 12 and IWS 12

Subject. Segregation of the seasonal component of the data model for the prediction task.

Plan of the lesson.

5) Study the scheme for allocating the seasonal component of the data model for the prediction problem presented above.

6) Task of the IWS. Apply the above chart of the seasonal component of the data model for its prediction task.

7) Analyze the results.

Literature.

Minko A.A. Predicting in business using Excel. M.: Eksmo, 2007, -208 p.

TOPIC 13

DECISION TREES

13.1. Lecture material

The work of decision trees is based **on the process of recursive partitioning of the initial set of observations or objects into subsets** associated with **classes defined by the selected attributes of objects** on each cycle of the recursive partition.

The partition is performed using the decision rules, in which the values of the attributes are checked according to the specified condition. Let's consider the main idea of algorithms for constructing decision trees using the example.

The scheme for using decision trees is as follows. Let's consider the scheme of construction and use of decision trees on **the standard example of issuing a loan to the client** [1].

The database on which to base the forecasting, for example, contains the following initial data on the bank's customers that are attributes of this database:

- age,

- availability of real estate,
- education,
- the average monthly income,
- return a loan by the client on time.

The task is to predict, based on the data listed above (except of the last attribute), it is worth whether giving out a loan to a new client.

The task is solved in two stages:

- construction of a classification model (training stage);

- use of the constructed model for making decisions on new clients.

At the stage of building the classification model, a classification tree is constructed or a set of classification rules is created.

At the stage of using the model, the built decision tree, namely, the path from its root to one of the vertices, which is a set of rules for a particular client, is used to answer the question «Whether to issue the loan? "

The rule is a logical construction, represented in the form

"If: <condition> then: <operator>".

Let's assume for this example there are statistics presented in Table 13.1.

Table 13.1. Customer input data

	Age	property	education	average monthly	loan
		availabili		income	repaymen
		ty			t
1	40(age>30)	yes	Higher	high	yes
			education		
2	50(age >30)	yes	Higher	high	yes
			education		
3	65(age >60)	no	specialized	average	no
			secondary	_	
4	32(age >30)	yes	Higher	average	yes
	_	-	education	_	
5	$25(age \leq 30)$	no	secondary	low	no
			education		
6	$20(age \leq 30)$	no	secondary	low	no
			education		
7	$28(age \le 30)$	yes	secondary	average	yes
			education	_	

In Fig. 13.1. an example of a classification tree, which solves the problem of "Whether to issue the loan?".



Fig.13.1. A variant of the decision tree for the task of "issuing a loan to a client" (1-issuing a loan, 0-not issuing).

As can be seen from the decision tree of this example, the internal nodes of the tree (age, availability of real estate, income and education) are attributes of the database described above. These attributes are called *splitting attributes*. The final nodes of a tree, or sheets, are called *class labels*, which are the values of the dependent categorical variable "to give out" or "not to issue" a loan.

At each internal node, it is necessary to specify a validation condition (*splitting predicate*) that would split the set associated with this node into subsets. For this check, one of the database attributes must be selected, except for the dependent variable. This question of selecting the attribute of splitting is the main point in the automated construction of the decision tree.

The general rule for selecting an attribute can be formulated as follows: **the selected attribute must split the set so that the resulting subsets consist of objects belonging to the same class** or are as close to it as possible, i. e. the number of objects from other classes ("impurities") in each of these sets was as small as possible.

A number of methods for constructing decision trees have been developed, including: Classification and regression trees (CART) method, C4.5 method [1].

Algorithm C4.5 uses the concept of **information gain or entropy reduction** to select the optimal splitting (split) in each node of the decision tree. **Because the decrease in entropy leads to an increase in information and vice versa.**

Let us first consider the concept of entropy. Suppose that there is a variable X whose k values have probabilities p_1, p_2, \ldots, p_k . How much information is needed to transmit a stream of symbols representing the values of the observed X? The answer to this question is the entropy X and is defined as

$$H(X) = -\sum_j p_j log_2(p_j).$$

Algorithm C4.5 uses the concept of entropy as follows.

Suppose that there is a splitting attribute s that divides the learning set of data T into several subsets, T_1 , T_2 , ..., T_k . The average amount of information on these subsets

can be calculated as the weighted sum of the entropies of the individual subsets, as follows:

$$H_s(T,D) = \sum_{i=1}^s P_i H_s(T_i,d),$$

Where P_i represents the fraction of records in the subset i, $H_s(T,D)$ is the entropy of the splitting of the data set T by the splitting attribute S with the dependent variable D, $H_s(T_i, D)$ - is the entropy of the splitting subset T_i from the splitting attribute S with the dependent variable D, *s* is the number of values of the splitting attribute S (determines the number of subsets (classes) of splitting).

Then, information gain or entropy reduction can be determined by splitting the initial set T into subsets of the splitting attribute s as:

 $Gain(T,S) = H(T,D) - H_s(T,D),$

Where H (T, D) is the entropy of the initial data set T by the attribute of the D-dependent variable.

 $H(T,D) = \sum_{i=1}^{d} P_i H(T_i,D),$

Where d is the number of values of the attribute D-dependent variable.

Thus, the information is incremented by dividing the initial training data set T in accordance with the splitting attribute S. At each node, C4.5 selects the optimal splitting, which has the highest gain of information Gain (T, S).

We calculate the entropy and increment of information by the definition of the optimal attribute for an example.

We calculate the entropy of the initial data set H (T).

H (T, Credit) = $-\frac{4}{7}log_2\frac{4}{7} - \frac{3}{7}log_2\frac{3}{7} \approx 0,9852$. Here, Credit is a dependent variable.

We calculate the increments of information for various attributes.

The increase in information on the splitting attribute "Income":

$$Gain (T, Income) = H(T, Credit) - \frac{2}{7}H(T_{income=high}, Credit) - \frac{3}{7}H(T_{income=average}, Credit) - \frac{2}{7}H(T_{income=low}, Credit)$$

$$\approx 0.9852 - \frac{2}{7}\left(-\frac{2}{2}log_{2}\frac{2}{2} - \frac{0}{2}log_{2}\frac{0}{2}\right) - \frac{3}{7}\left(-\frac{2}{3}log_{2}\frac{2}{3} - \frac{1}{3}log_{2}\frac{1}{3}\right) - \frac{2}{7}\left(-\frac{0}{2}log_{2}\frac{0}{2} - \frac{2}{2}log_{2}\frac{2}{2}\right) \approx 0.9181$$

The increase in information on the splitting attribute "Age":

$$Gain (T, age) = H(T, Credit) - \frac{3}{7}H(T_{age>30}, Credit) - \frac{1}{7}H(T_{age>60}, Credit) - \frac{3}{7}H(T_{age<30}, Credit)$$

$$\approx 0.9852 - \frac{3}{7}\left(-\frac{3}{3}log_2\frac{3}{3} - \frac{0}{3}log_2\frac{0}{3}\right) - \frac{1}{7}\left(-\frac{0}{1}log_2\frac{0}{1} - \frac{1}{1}log_2\frac{1}{1}\right) - \frac{3}{7}\left(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}\right) \approx 0.9181$$

Increase in information on the attribute of splitting " Property ":

$$\begin{aligned} Gain (T, \text{property}) &= H(T, \text{Кредит}) - \frac{4}{7} H \Big(T_{property=yes}, \text{Кредит} \Big) \\ &- \frac{3}{7} H \Big(T_{property=no}, \text{Кредит} \Big) \approx 0.9852 - \frac{4}{7} \Big(-\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} \Big) \\ &- \frac{3}{7} \Big(-\frac{0}{3} \log_2 \frac{0}{3} - \frac{3}{3} \log_2 \frac{3}{3} \Big) \approx 0.9852 \end{aligned}$$

The increase in information on the splitting attribute "Education":

$$\begin{aligned} Gain (T, \text{education}) &= H(T, \text{Credit}) - \frac{3}{7} H \left(T_{education=high}, \text{Credit} \right) - \\ \frac{1}{7} H \left(T_{education=spec \ secondary}, \text{Credit} \right) - \frac{3}{7} H \left(T_{education=secondary}, \text{Credit} \right) \\ &\approx 0.9852 - \frac{3}{7} \left(-\frac{3}{3} \log_2 \frac{3}{3} - \frac{0}{3} \log_2 \frac{0}{3} \right) - \frac{1}{7} \left(-\frac{0}{1} \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1} \right) \\ &- \frac{3}{7} \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) \approx 0.9181 \end{aligned}$$

Thus, calculations of the information gain show that the optimal at the first level is the verification of the attribute "Property".

Thus, the calculation of the information gain shows that the optimal at the first level is the verification of the attribute "Property", in which the gain of information Gain (T, property) = 0.9852. With this in mind, the decision tree of this example will look like Figure 13.2.



Fig. 13.2 Optimal solution tree for example

Select the data set for the right branch from the source data set (Table 13.2.). This is necessary for the recursive application of the algorithm for constructing a decision tree to determine the splitting attribute for the next level. To do this, the set of possible attributes includes: age, education, income. The attribute "property" is used for splitting on the first level.

Table 13.2 Data set for selecting the second splitting attribute.

	Age	property	education	average monthly	loan
		availabili		income	repaymen
		ty			t
1	40(age>30)	yes	Higher	high	yes
			education		
2	50(age >30)	yes	Higher	high	yes
			education		
4	32(age >30)	yes	Higher	average	yes
			education		
7	$28(age \le 30)$	yes	secondary	average	yes
			education		

We calculate the entropy of the initial data set H (T) for the next level of splitting.

 $H(T, Credit) = -\frac{4}{4}\log_{2}\frac{4}{4} - \frac{0}{4}\log_{2}\frac{0}{4} \approx 0$

Since the entropy of a given set for the next level is zero, this means that the process of building the decision tree for this example is complete. The resulting decision tree for the given source data set of the statistics of the example is optimal (Figure 13.2). However, one should bear in mind that the original sample may not cover all possible cases in reality, so the final decision on the structure of the decision tree remains with the user.

Below is represented the solution of the above example using data mining tool Rapidminer.

In the Fig.13.3 screenshot of source data for the standard example of issuing a loan to the client is presented.

In the Fig.13.4 screenshot of the developing process structure for the standard example of issuing a loan to the client is presented. Input and output points of process operators (mod = Model, unl = Unlabelled data, lab = Labelled data, exa = Example set, ori = Original data).

In the Fig.13.5-13.8. screenshot of set role operator parameters for the standard example of issuing a loan to the client are presented.

Parameters of the operator "Set role" [2]:

1) The parameter 'attribute_name', which role should be changed. The name can be selected from the dropdown menu or manual typed.

Range of the parameter 'attribute_name':

- target_role;

- set of additional role.

The target role of the selected Attribute is the new role assigned to it. As possible target roles are used:

- regular: attributes without a special role. Regular Attributes are used as input variables for learning tasks;

- label: this is a special role; an attribute with the label role acts as a target Attribute for learning Operators; the label is also often called 'target variable' or 'class';

- prediction: this is a special role; an Attribute with the prediction role is the result of an application of a learning model.

More detail description of the parameters is in the Rapidminer Manual [2].

And in the Fig.13.9. screenshot of results for the standard example of issuing a loan to the client is presented.

Image: Second part of the Case Deposition y data Blook 2) Image: Second part of the Case Deposition	2	H •	 • 				Views: D	Find data, operators .etc 🖇	All Studio
Remplete (8 samplete. 9 special attitudes. 5 regular at			et (//Local Repor	sitory/data/Book3)					
No. Age property ave delaction average mo. base regetma 0 1 40 ys Higher educa. high ys 0 30 ys Higher educa. high ys n 0 3 65 no specialistas average ys n 0 3 65 no specialistas average ys n no 0 3 0 specialistas average ys no no no 0 30 no secondary ed. tow no no no			examples 0 sne	rial attributes. 5 regula	r affrihetes)		xample set y/Loc		
1 40 ys Hydre 40ca. hydr. ys 2 50 ys Hydre 40ca. hydr. ys 3 40 50 no specialization average no 6 30 no specindary ed. hydr. no 6 30 no specindary ed. hydr. no Advinsion hydr. no						average mo	loan repaym	ritter (ur o examples). (aii	,
3 45 no secales s. average no 4 32 us Hyber 64ca average us 5 25 no secondary ed. tow no 0 30 no secondary ed. tow no 0 30 no secondary ed. tow no Achaeced secondary ed. tow no no	Data	1				high			
Statistics 4 32 yes Higher educa. average yes Solar 25 no secondary ed. tow no 6 30 no secondary ed. tow no Aburded 30 no secondary ed. tow no Aburded 30 no secondary ed. tow no	100	2	50	yes	Higher educa	high	yes		
A A2 ys High Holds Alends y ys Chains A A A No No		3	65	no	specialized s	average	no		
6 30 no secondary ed. low no	Statistics								
Chails Activations Annotations									
Anototisms Acontractions		6	30	no	secondary ed	low	no		
	Charts								

Fig.13.3. Screenshot of source data for the standard example of issuing a loan to the client



Fig.13.4. Screenshot of the developing process structure for the standard example of issuing a loan to the client.

Elle Edit Process View Connections Cloud Setting	s chausious Darb	Views: Design R	esuits Turbo Prep Auto Model			ita, operators .elc 🔎 All Studio 👻
	Process Process Process		100% 🔑	P P 🔒 🍹 🖬	Parameters ×	Ioan repayment 💌 🛈
• ddd (123) Bookl (123 - vt.) Artistist 11.88 AM - 458 Symm Bookl (123 - vt.) Artistist 3.38 FM - 458 Symm Bookl (123 - vt.) Artistist 10.22 FM - 58 Symm Bookl (123 - vt.) Artistist 10.22 FM - 58 Symm Bookl (123 - vt.) Artistist 10.22 FM - 58 Symm Bookl (123 - vt.) Artistist 10.27 FM - 58 Symm Bookl (123 - vt.) Artistist 10.27 FM - 58 Symm	Retrieve Book1	Decision Tree		183. 195. (target role set additional roles	label 🔹 🐨
Boold (12) - v1, Entries 30 PM - 36 pmu - 37 puese PM REF (12) - v1, Entries 100 PM - 37 puese Y Operators X set rol X Bending (1) Altobalas (1)	Set Role	4 4				
Names & Role (1)					Change compatibility (8.0.00 Help X regular: Attributes withou are used as input variable	e possoe A
					 id: This is a special role. An identifier for the Examples Examples. Different Blend 	Attribute with the id role acts as an It should be unique for all
	Leverage the Wisdom of Crowds to get operat	for recommendations based on your proce				An Attribute with the label role acts arning Operators. The label is also

Fig.13.5. Screenshot of set role operator parameters (target attribute "loan repayment") for the standard example of issuing a loan to the client

Eile Edit Dicess Mer Connections Cloug getings Egensions Heig News Design Results Turbo Prep Auto Model Find data, operators, edit Auto Model Sign in Image: Sign in the second	Корзина	ew process*> – RapidMiner Studio Trial 9.0	001 @ 123-RK			<u>,</u>		
Vision Vision <td>2018 грант</td> <td>Repository ×</td> <td>Process Process</td> <td>Views Design Results T</td> <td></td> <td>E Set Role</td> <td>×</td> <td>All Studio •</td>	2018 грант	Repository ×	Process Process	Views Design Results T		E Set Role	×	All Studio •
Image: Set Role Image: Set Rol	Actilo5 20:	Cata (123) Book1 (123 4, Briteria 11 128 AM - 438 byen Book2 (123 4, Briteria 11.02 AM - 438 byen Book2 (123 4, Briteria 12.02 FM - 198 byen Book2 (123 4, Briteria 12.02 FM - 198 byen Titanic-Data (123 4, Briteria 13.01 FM - 138	Retrieve Book1	to mod		101		• ①
Ling over 1 Ling over 1 </td <td></td> <td>Book2 (123 - v1, 5/16/18 50 / M4 - 56 bytes) PRIT1 (123 - v1, 5/16/18 11 50 / M4 - 575 bytes) Operators X</td> <td>4 m 1 m</td> <td></td> <td></td> <td></td> <td></td> <td></td>		Book2 (123 - v1, 5/16/18 50 / M4 - 56 bytes) PRIT1 (123 - v1, 5/16/18 11 50 / M4 - 575 bytes) Operators X	4 m 1 m					
BS/SEG Rej Kurson Classic Reg Kurson Reg Classic Reg Kurson Reg Classic Reg Kurson Reg Signifier Reg Kurson Reg Optimizer Reg Reg Reg Reg Reg Reg Reg Reg Reg Reg Reg Reg Reg Reg	altyn corr	Blending (1) Bending (1) Attributes (1) Bending & Roles (1)				Channe com	anim/ (0.000)	
USS9 3932/001 Attribute. The default role is regular, other roles are classified as special. An Attribute. The default role is angular, other roles are classified as special. An Attribute. The default role is angular, other roles are classified as special. An Attribute. The default role is angular, other roles are classified as special. An Attribute. The default role is angular, other roles are classified as special. An Attribute. The default role is angular, other roles are classified as special role. Attribute. The default role is angular, other roles are classified as special role. The default role is angular, other roles are classified as special role. The default role is angular, other roles are classified as special role. The default role is angular, other roles are classified as special role. The default role is angular, other roles are classified as special role. The default role is angular, other roles are classified as special role. The default role is angular, other special role is angular, other s	BENGO P					Help X tomp to rotorial		
Differentiation v	BRASMUS:		Leverane the Wisdom of Crowds to net operator record	rmentations based on your process design!		Attribute. The de special. An Exam special role can o than one Attribu last Attribute. Th	fault role is regular, other roles are class pleSet can have many special Attributes nly appear once. If a special role is assig et, all roles will be changed to regular es e different types of roles are explained i	sified as s. but each gned to more kcept for the
	Committee				_			

Fig.13.6. Screenshot of set role operator parameters (regular attribute "age") for the standard example of issuing a loan to the client

🗋 ڬ 🖬 • 🕨 • 🔳	igs Egtensions Help	Views: Design F	Results Turbo Prep Au	ito Model		ata, operatora, elc 🔎 All Studio 🕶
Repository X import Data v import	Process Process Process Retrieve Book1 pr	Decision Tree		100% 🖉 🔍 🔍 🧟 🦉 🕅 100% 100% 100 100 100 100 100 100 100 10	attribute name target role	average monthly income
Bool2 (122)	Set Balo				set additional roles	Edit List (1)
Blending (1) B						
The Attributes (1)					Attribute. The default role is re special. An ExampleSet can han special role can only appear or than one Attribute. all roles wil	11) bes how other Operators handle this guiar, other roles are classified as which not if a special role as sagned to more the changed to regular except for the pes of roles are explained below in the

Fig.13.7. Screenshot of set role operator parameters (regular attribute "average monthly income") for the standard example of issuing a loan to the client

Корзина	ew process*> – RapidMiner Studio Trial 9.0.						
2018 грант	Pile Edit Process yew Connections Cloug Setti Repository × Compository = Local Repository (2)	Process Process Process	Views: Design Results Turb	9 Prep Auto Model 100% 🖉 🖉 👂 🥃 🦉 🖉	Find data Parameters × For the second secon	coorations .etc 🔎 All Studio ¥	Sign in
ACTIONS 201	data (133) Bookt (123 - + 4 antenia + 136 AM + 135 symm Bookt (123 - + 4 antenia + 136 AM + 135 symm Bookt (123 - + 4 antenia + 132 PM - 186 symm Tank-0 Cata (123 - + 4, antenia + 132 PM - 196 symm Tank-0 Cata (123 - + 4, antenia + 132 PM - 196 Poccesses (123)	Retrieve Book1	Decision Tree		target role set additional roles	regular 🔹 🗇	× in
ACIIDS 20: REVIEW	Bool2 (123 - vt. 615/16 835 / 94 - 36 5yka) QP INT3 (123 - vt. 615/16 835 / 94 - 36 5yka) V Operators satrol	Set Bale					×
	Bending (1) Bending (1) Set Role Set Role				✓ Chance compatibility (9.0.001) Help ×		* 300%
P. CLASS ERASMUST					Attribute. The default role is regu- special. An ExampleSet can have special role can only appear once than one Attribute, all roles will b	many special Attributes, but each e. If a special role is assigned to more be changed to regular except for the	×
Diariesez Czerwinski	No results were found.	Leverage the Wisdom of Crowds to get operator reco	mmendations based on your process design!		last Attribute. The different type: parameter section.	s of roles are explained below in the	
0						EN 🔺 🏲 🛋 🗊 🐗 🏟	23:06 18.08:2018

Fig.13.7. Screenshot of set role operator parameters (regular attribute "education") for the standard example of issuing a loan to the client

Repository ×	Process					
impon Data = *	Process		100% 🖉 🔎 🔎		arameters ×	
T Local Repository (123)			100 10 10 10 10 10		tribute name	property availability •
 Cocal Repository (123) data (123) 	Retrieve Book1	Decision Tree			get role	regular 💌 🛛
Book1 (122 - v1, 8/18/18 11:55 AM - 528 bytes)	Diep C nt	tra mod		res tal	getrole	regular • 0
Book2 (123 - v1, 8/18/18 9 38 PM - 96 bytes) Book3 (123 - v1, 8/18/18 10 22 PM - 99 bytes)	1 A	eri D		res se	t additional roles	🔂 Edit List (1)
Titanic+Data (123 - v1, 8/15/18 11/21 PM - 18						
processes (123)						
Book2 (123 - v1, 8/18/18 9:36 PM - 98 bytes)	Set Role					
<						
Operators ×		ſ				
set rol X						
Blending (1)						
- Attributes (1)						
Vames & Roles (1)						
Set Role				н	Change compatibility (9.0.001) elp × mp to rotorian Process	()
				P	escription)
				T	ne role of an Attribute describe	es how other Operators handle this
				sp	becial. An ExampleSet can have	ular, other roles are classified as many special Attributes, but each
						e. If a special role is assigned to more
				th	ian one Attribute, all roles will l	be changed to regular except for the
	Version Bellinder of Arrisk Is adversely	or recommendations based on your process design		la		be changed to regular except for the s of roles are explained below in the

Fig.13.8. Screenshot of set role operator parameters (regular attribute "property availability") for the standard example of issuing a loan to the client

			Views: Design	Results Turbo Prep	Auto Model	Find data, operators .elc	🔎 All Studio 🔻
P	ExampleSet (//Local Repository Result History		Decision Tree) ×	ocal Repository/data/Book2)	ExampleSet (//Local Repositor	pleSet (//Local Repository/Book2) × //data/Book1) ×	ory
Clin Clin	-						= • • 10
Graph	Q Q	property availability					-
Grap	Tree *	no yes					-
2	Vode Labels	no yes					
Descript	on 📝 Edge Labels						
Annotati	ns						
							۰. ا
							2

Fig.13.9. Screenshot of results for the standard example of issuing a loan to the client

13.2 PRACTICAL LESSON 13 and IWS13

Subject. The construction and use of decision trees for the classification problem.

Plan of the lesson.

1) Study the scheme for constructing and using decision trees for the classification problem, presented above.

2) The task of the IWS. Apply the above scheme of constructing and using the decision tree for your classification problem.

3) Analyze the results.

Literature

1. Larose D.T. Discovering knowledge in data - an introduction to data mining. Wiley-Interscience, Hoboken, New Jersey, 2005

2. Rapidminer Studio, manual. https://docs.rapidminer.com/latest/studio/

TOPIC 14

CLUSTERING

14.1. Lecture material

Clustering refers to the grouping of observation records into classes of similar objects.

A cluster is a collection of records that are similar to each other, and are not similar to records in other clusters.

Clustering differs from classification in that there is no target variable for clustering. The clustering task does not attempt to classify, evaluate, or predict the value of the target variable. Instead, clustering algorithms tend to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of records in a cluster is maximized, and the similarity of records outside this cluster is minimized [1].

Examples of clustering tasks in business and research include the following:

• Targeted marketing of niche products for small businesses that does not have a large marketing budget.

• For the purpose of accounting audit, for the segmentation of financial behavior in benign and suspicious categories.

• As a measurement tool to reduce, when a data set has hundreds of attributes.

Clustering is often performed as a preliminary step in the data mining process, with clusters being used as additional inputs to other downstream techniques, such as neural networks.

14.1.1 Similarity of measures.

Cluster analysis is confronted with many of the same issues that are dealt with in the classification problem. For example, the following issues should be addressed:

- How to measure the similarity?

- How to standardize or normalize numeric variables?

One measure of the similarity of records is the Euclidean distance between records:

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i} (x_i - y_i)^2}$$

Where $x = x_1, x_2, ..., x_m$, and $y = y_1, y_2, ..., y_m$ represent the m attribute values of the two records.

Of course, there are many other indicators, such as the distance of Manhattan:

$$d_{Manhatten}(x, y) = \sum_{i} |x_i - y_i|$$

The Minkowski distance is the general case of the above metrics for the total exponent r:

$$d_{Minkowski}(x,y) = \sqrt[r]{\sum_{i} |x_i - y_i|^r}$$

A particular case of the Minkowski distance is the Manhattan distance corresponding to r = 1. For r = 2, the usual Euclidean distance is obtained.

14.1.2. Transformation of data.

Variables, as a rule, have ranges that differ significantly from each other. Therefore, for data analysis programs, it is necessary to normalize numerical variables, standardize the scale of the effect of each variable on the results. There are several methods for normalization, two of the most common methods: min-max normalization and Z-score standardization. Let X refer to the initial value of the field and X * refers to the normalized value of the field.

The Min-Max normalization method scales by the difference in range:

$$X^* = \frac{X - \min(X)}{range(X)} = \frac{X - \min(X)}{\max(X) - \min(X)},$$

Where X is the value of the normalized field, $\min(X)$ is the minimum value of the normalized field, $\max(X)$ is the maximum value of the normalized field.

For example, consider the variable "time to 60" from the car data set, which determines the time (in seconds) how much each car spends to reach 60 km per hour. Let's calculate the min-max normalization for three cars (Car1, Car2, Car3), having "time to 60" 8, 15 and 25 seconds respectively.

Then for Car1 the normalized value of the field "time to 60" will be equal to:

$$X^* = \frac{X - \min(X)}{\max(X) - \min(x)} = \frac{8 - 8}{25 - 8} = 0$$

For Car2, the normalized field value "time to 60" will be:

$$X^* = \frac{X - \min(X)}{\max(X) - \min(x)} = \frac{15 - 8}{25 - 8} = 0,41$$

For Car3, the normalized field value "time to 60" will be:

$$X^* = \frac{X - \min(X)}{\max(X) - \min(x)} = \frac{25 - 8}{25 - 8} = 1.$$

Thus, min-max values of normalization will fluctuate from zero to one if new data values do not occur that lie outside the original range.

Z-score standardization is very widespread in the world of statistical analysis, it works as a difference between the field value and the mean value of the field and the scaling of this difference by the standard deviation (SD) values of the field. I.e,

$$X^* = \frac{X - \operatorname{mean}(X)}{SD(X)}.$$

Then for Car1 the normalized value of the field "time to 60" will be equal to:

$$X^* = \frac{X - \text{mean}(X)}{SD(X)} = \frac{8 - 15}{2,9} = -2,4$$

For Car2, the normalized field value "time to 60" will be:

$$X^* = \frac{X - \operatorname{mean}(X)}{SD(X)} = \frac{15 - 15}{2,9} = 0$$

For Car3, the normalized field value "time to 60" will be:

$$X^* = \frac{X - \text{mean}(X)}{SD(X)} = \frac{25 - 15}{2,9} = 3,4.$$

Summarizing, the standardization values of the Z-count are generally in the range of -4 to 4, with an average value having Z-account of zero standardization.

14.1.3 Clustering algorithms. Hierarchical cluster method.

In this method, a hierarchical agglomerate algorithm is realized, which consists in uniting smaller clusters into larger clusters.

Before starting clustering, all objects are considered separate clusters.

In the course of the algorithm, clusters are combined. **First, select a pair of nearby clusters, which are combined into one cluster.** As a result, the number of clusters becomes N-1.

The procedure is repeated until all the classes are combined. At any stage, the union can be interrupted by obtaining the required number of clusters.

The closeness of clusters is determined by the distance between them.

You can use several criteria to determine the distance between clusters A and B:

- The "nearest neighbor" method, based on the minimum distance between any record of cluster A and any record of cluster B.

- The method of " distant neighbor", based on the maximum distance between any record of cluster A and any record of cluster B. Moreover, cluster clusters that are smaller in size are the priority for the association.

- The method of "**the least mean connection between the records of cluster A and records of cluster B**".

The "nearest neighbor" method.

For example, take a one-dimensional data set [1]:

1 4 9 15 16 18 25 32 32 44

It is assumed that each element is a separate cluster. The process of combining clusters and constructing a hierarchy of clusters for an example is shown in Fig. 14.1. Consider the sequence of algorithm steps for this example.

Step 1. Since singleton clusters are considered, the minimum distance between any record of cluster A and any record of cluster B will be searched between the
elements. These are clusters with 32,32 elements, the distance between them is 0.

Step 2. The search of variants of clusters shows that clusters with **elements 15,16** with a distance equal to 1 should unite.

Step 3. At this step it is necessary to combine the cluster (15,16) with the cluster (18) with the distance equal to 2.

Step 4. At this step, clusters (1) and (4) are combined with a distance of 3.

Step 5. Clusters (1,4) and (9) are united with a minimum distance of 5 between elements 4 and 9.

Step 6. Clusters (1,4,9) and (15,16,18) are united with a minimum distance equal to 6 between cluster element 9 (1,4,9) and cluster element 15 (15,16,18).

Step 7. Clusters (1,4,9, 15,16,18) and (25) are combined with a minimum distance of 7 between the cluster element 18 (1,4,9,15,16,18) and the cluster element 25 (25).

Step 8. Clusters (1,4,9, 15, 16, 18, 25) and (32, 32) are united with a minimum distance of 7 between cluster element 25 (1,4,9,15,16,18,25) And cluster element 32 (32,32).

Step 9. Clusters (1,4,9, 15,16,18,25,32,32) and (44) are combined with a minimum distance of 12 between the cluster element 32 (1,4,9,15,16,18, 25,32,32) and the cluster element 44 (44).

At each step, the variants are sorted and there is a variant of combining clusters with the minimum distance.

steps	1	4	9	15	16	18	25	32	32	44
1								32,32	d=0	
2				15,16	d=1					
3				15,16,	18					
				d=2(16	5,18)					
4	1,4	d=3								
5	1,4,9	d=5((4,9)							
6	1,4,9,	15,16,1	8 d=6((9,15)						
7	1,4,9,	15,16,1	8,25 d	=7(18,2	5)					
8	1,4,9,	15,16,1	8,25,32	,32 d=	7(25,32	()				
9	1,4,9,	15,16,1	8,25,32	,32,44	d=12(3	2,44)				

Fig. 14.1. Steps for clustering the data set using the "nearest neighbor" method.

The method of "distant neighbor".

For example, the "distant neighbor" method is shown in Fig. 14.2.

The difference in the distant neighbor algorithm for this example begins with step 4, where the distance is taken as the maximum between the elements. The situation is similar in step 5.

The structural difference in the hierarchy begins with step 6, where, unlike the previous method, other clusters are combined.

шаги	1	4	9	15	16	18	25	32	32	44
1								32,32	d=0	
2				15,16	d=1					
3	1,4	d=3								
4				15,16,1	18					
				d=3(15	5,18)					
5	1,4,9	d=8((1,9)							
6							25,32,	32		
							d=7(25	5,32)		
7	1,4,9,	15,16,1	8 d=1′	7(1,18)						
8							25,32,	32,44	d=19(2	5,44)
9	1,4,9,	15,16,1	8,25,32	,32,44	d=43(1	,44)				

Fig. 14.2. Steps for clustering a data set using the far-neighbor method.

The method of the least mean connection between the records of cluster A and records of cluster B.

"The method of the least mean connection between records of cluster A and records of cluster B" is shown for example in Fig. 14.3. As you can see from the comparison of the clustering patterns in three methods, the result of the clustering of the "least mean communication method between the records of cluster A and records of cluster B" coincides with the result of the "distant neighbor" method.

steps	1	4	9	15	16	18	25	32	32	44
1								32,32	d=0	
2				15,16	d=1					
3	1,4	d=3								

4			15,16,18						
			d=2,5(1)	5,18=3;1	6,18=2)				
5	1,4,9								
	d=6,5(1,9=8;4	l,9=5)							
6						25,32	,32 d	=7	
7	1,4,9, 15,16,1	8 d=11	,66						
8						25,32	,32,44	d=14	1,33
9	1,4,9, 15,16,1	8,25,32	,32,44 d	=23.91					

Fig. 14.3. Steps for clustering a data set using the method of least average communication between cluster records.

14.1.4 Algorithms for clustering. Method K-means (k-means).

The k-means method is a simple and effective algorithm for building data clusters. The algorithm includes the following steps [1]:

1) The user sets the number of clusters to which the original data set should be divided.

2) Randomly assign records as cluster centers.

3) For each record is the closest cluster to it.

4) For each cluster there is a centroid of the cluster and the center of each cluster changes according to the centroid value.

5) Steps 3-5 are repeated **until there is a minimum of the sums of the quadratic deviation of cluster points from the center of the cluster or centroid** of the clusters will cease to change.

The criterion of "closeness" in step 3 usually uses the Euclidean distance

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i}(x_i - y_i)^2}.$$

The cluster centroid in step 4 is as follows. Suppose that there are n data points (a_1,b_1,c_1) , (a_2,b_2,c_2) ,..., (a_n,b_n,c_n) . The centroid of these points is at the point $(\sum a_i/n, \sum b_i/n, \sum c_i/n)$. If the centroids of the clusters cease to change, the algorithm terminates. The completion of the algorithm can also be determined by minimizing the sums of quadratic deviation of cluster points (sum of squared errors-SSE) from the center of the cluster:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2,$$

Where $p \in C_i$ is the point C_i of the cluster, m_i is the centroid of the cluster, $d(p, m_i)$ is the deviation (error) of the points from the center of the cluster.

Also, to assess the rationality of clustering, the ratio of the variations between clusters and the sums of the quadratic deviation of cluster points is used:

 $\frac{d(m_1,m_2)}{SSE}.$

This indicator is growing for a more rational version of clustering.

Example. Let there be given 5 points of a two-dimensional space:

а	b	с	d	e	
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	

The number of clusters is set to -2.

The centroid for cluster C1 will be (1,2), and the centroid for cluster C2 will be (5,3). This is usually determined randomly.

	m ₁	m ₂	Cluster
А	1	4	C1
В	2.24	2	C2
С	3.16	1	C2
D	4.12	0	C2
Е	0	4.12	C1

Table 14.1. Calculation of the distance of Euclid for example.

Cluster C1 includes (a, e), and cluster C2 includes (b, c, d).

Calculating the sums of the quadratic deviation of cluster points (sum of squared errors-SSE) from the center of the cluster for an example:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2 = 1^2 + 2^2 + 1^2 + 0^2 + 0^2 = 6.$$

Calculating the ratio:

$$\frac{d(m_1,m_2)}{SSE} = \frac{4,12}{6} = 0,68.$$

In the next iteration, new values of cluster centroids are determined.

For cluster C1, the centroid value is: $\left(\frac{1+1}{2}, \frac{3+2}{2}\right) = (1; 2.5)$

For cluster C2, the centroid value is: $\left(\frac{3+4+5}{3}, \frac{3+3+3}{3}\right) = (4; 3).$

Now, for the new values of cluster centroids, the distances from the center of the clusters for each point are calculated.

Table 14.2. The calculation of the Euclidean distance for an example is the second
iteration

	m ₁	m ₂	Claster
a	0.5	3	C1
b	2.06	2	C2
С	3.04	1	C2
d	4.03	1	C2
e	0.5	3.16	C1

Calculating the sums of the squared deviation of cluster points (sum of squared errors-SSE) from the cluster center for the second iteration:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2 = 0.5^2 + 2^2 + 1^2 + 1^2 + 0.5^2 = 6.5$$

Calculating the ratio:

 $\frac{d(m_1,m_2)}{SSE} = \frac{3,04}{6.5} = 0,46.$

As SSE increases, and the evaluation of the rationality of clustering decreases, the previous version is considered the most rational.

Below is presented an example of solving the above problem using the Rapidminer software tool [2].

In the Fig. 14.1 the screenshot of source data for the a one-dimensional data set is presented.

In the Fig. 14.2 the screenshot of process structure of clustering task by k-means method is presented.

In the Fig. 14.3 the screenshot of clustering task results by k-means method is presented.



Fig. 14.1. Screenshot of source data for the a one-dimensional data set

🗋 🐸 🔚 • 🕨 🕨 🔳		Views: Design	Results Turbo Prep	Auto Model		it, operators, etc 🔎 All Studio
Repository ×	Process				Parameters ×	
O import Data	Process			100% 🔎 🔎 📮 🧸 💣 🗄	Process	
	Process				logverbosity	init *
DB	Retrieve Book2 Clustering					
Local Repository (123) data (123)	Diep ent ens du	0		res	logfile	
Book1 (123 - v1, 8/18/18 11:55 AM - 528 bytes)	VA the			185		
Book2 (123 - +1, #/18/18 B 38 PM - 98 bytes)	×					
Titanic+Data (123 - v1. 8/15/18 11:21 PM - 100						
Processes (123)						
Book2 (123 - v1. 8/18/18 8:35 PM - 36 bytes)						
< ***** I >*						
Operators ×						
Data Access (50)						
 Blending (77) Cleansing (25) 						
Modeling (143)					Show advanced parameters	
Predictive (61)					No. of Concession, Name	
T Segmentation (14)					Change compatibility (9.0.001	1
k-Means						
i k-Means (Kernel)					Help ×	
k-Means (fast)					Process	
X-Means					RapidMiner Studio Core	
BBSCAN					Synopsis	
Expectation Maximization Clustering						outer most operator of every process
Support Vector Clustering						
Random Clustering					Description	
Agglomerative Clustering	Leverage the Wisdom of Crowds to get operator	recommendations based on your pro	cess design!			ctly one operator of this class, and it ne process. This operator provides a
					set of parameters that are of gl	

Fig. 14.2. Screenshot of process structure of clustering task by k-means method



Fig. 14.3. Screenshot of clustering task results by k-means method

14.2 PRACTICAL LESSON 14 and IWS 14

Subject. Building and using clustering methods.

Plan of the lesson.

1) Study the scheme for constructing and using the clustering methods presented above.

2) The task of the IWS. Apply the above scheme of building and using clustering methods for your clustering task.

3) Analyze the results.

Literature.

1. Larose D.T. Discovering knowledge in data - an introduction to data mining. Wiley-Interscience, Hoboken, New Jersey, 2005

2. Rapidminer Studio, manual. https://docs.rapidminer.com/latest/studio

TOPIC 15

ASSOCIATION RULES

15.1. The lecture material

The purpose of the search for association rules is to find patterns between related events in databases. For the first time, the association rule mining problem was proposed to find typical shopping patterns made in supermarkets, so sometimes it is also called market basket analysis. Very often buyers buy more than one product. In most cases, there is a relationship between these goods. A market basket is a set of goods purchased by a buyer within the framework of a single transaction. Transactions are quite typical operations, for example, they can describe the results of visits to the store buyer. A transaction is a set of events that occurred simultaneously. Each such transaction is a set of goods purchased by the buyer for one visit. By registering all business transactions during the whole period of their activity, trading companies accumulate transaction collections, formed as transaction database.

Transaction database is a two-dimensional table, which consists of the transaction number (TID) and the list of purchases purchased by the buyer during this transaction. Each transaction corresponds to the purchase of an individual buyer. An example of such a table is shown in Table 15.1.

TID	The list of purchases
001	Bread, milk, tea, sugar
002	Milk, cheese, sour cream
003	Milk, bread, cheese, tea
004	Sausage, cheese, bread, sugar
005	Bread, tea, milk, sour cream
006	Sugar, tea, butter
	-

Table 15.1. Transaction database example

Let D be the set of transactions presented in Table 15.1, where each transaction T in D represents a set of elements contained in a variety of possible goods, the set denoted as I (bread, milk, tea, sugar, cheese, sour cream, sausage, butter). Suppose that there are a certain set of elements A (for example, bread and milk) and another set of objects B (for example, tea). Then the association rule takes the form, if A, then B (i.e,), where *antecedent* A and *consequent* B are the proper subsets of I. For example, trivial rules such as: *"bread and milk*, and then *tea"*.

As a result of this type of analysis, one can establish a regularity of the following kind: "If a set of goods (or a set of elements) A has been encountered in the transaction, then one can do the conclusion that in the same transaction a set of elements B should appear. Establishment such regularities enable us to find very simple and understandable rules, called *associative*.

The method of associative rules uses two parameters that characterize this method: *support* and *confidence* of the rule. *Support* for a specific association rule is the proportion of transactions in D which contain both A and B. That is,

 $Support = \frac{Number \ of \ transactions \ containing \ both \ A \ and \ B}{Total \ number \ of \ transactions}$.

The confidence of the association rule is a measure of the accuracy of the rule,

the percentage of transactions containing A that also contain B.

 $Confidence = \frac{Number of transactions containing both A and B}{Number of transactions containing A} .$

A set of elements is a set of elements contained in I, and k-itemset is a set of elements containing k elements. For example, {bread, milk} is a set of 2 items, and {sugar, tea, butter} is a 3-set of items, each of the set I. The frequency of the elements of a set is simply the number of transactions that contain a particular set of elements. A frequent set of items is a set of elements that arises at least a certain minimum number of times, with a set frequency $\geq \varphi$. For example, suppose that we set $\varphi = 4$. Then the sets that occur more than four times are called often. We denote the set of frequent sets of k-objects as F_k .

The mining of association rules from large databases is a two-step process [1]:

1. Find all frequent itemsets; that is, find all itemsets with frequency $\geq \phi$.

2. From the frequent itemsets, generate association rules satisfying the minimum support and confidence conditions.

The *a priori algorithm* takes advantage of the a priori property to shrink the search space. The *a priori property* states that if an itemset Z is not frequent, then adding another item A to the itemset Z will not make Z more frequent. That is, if Z is not frequent, will not be frequent. In fact, no *superset* of Z (itemset containing Z) will be frequent. This helpful property reduces significantly the search space for the a priori algorithm.

Below is an example of mining associative rules using the Rapidminer program for the data presented in Table 15.1 (Fig. 15.1-15.6) [2].

In the Fig. 15.1 is presented screenshot of source data with items in separate columns. In the Fig. 15.2-15.4 is presented screenshot of process structure with define of characteristics of process operators.

In Rapidminer there are possibility to write a list of items in one column. For above example we prepare source data as list of items in one column, rename items by short name (Table 15.2).

TID	The list of purchases	The list of purchases in one
		column
001	Bread, milk, tea, sugar	a b c g
002	Milk, cheese, sour cream	b d f
003	Milk, bread, cheese, tea	b a d c
004	Sausage, cheese, bread, sugar	eldlalg
005	Bread, tea, milk, sour cream	a c b f
006	Sugar, tea, butter	g c h
	_	-

Table 15.2. Transaction database example with a list of items in one column

In the Fig. 15.5 is presented screenshot of source data with list items in one column.

In the Fig. 15.6 is presented screenshot of process structure with define of characteristics of process operators FP- growth with list items in one column.

In the Fig. 15.7 is presented screenshot of frequent items of association rules with items on separate columns.

In the Fig. 15.8 is presented screenshot of association rules with items on separate columns.

In the Fig. 15.9 is presented screenshot of frequent items of association rules with list items in one column.

In the Fig. 15.10 is presented screenshot of association rules with list items in one column.



Fig. 15.1. Screenshot of source data with items in separate columns

) 늘 📕 • 🕨 • 🔳		Views: Design Result	s Turbo Prep Auto Model		Find data, operatorsetc 🔎 All Stud
pository ×	Process			Parameters ×	
G Import Data	Process		100% 🔎 🔎 📮 🥁 📹	Process	
	Process			logverbosity	init
🗖 data (123)	Retrieve Book1 ass	Create Association			
Book1 (123 - v1, 8/18/18 11:55 AM - 528 bytes)				logfile	
Book1 assoc (123 - v1, 8/30/18 2:19 PM - 369 byte:					
Book1 assoc1 (123 - v1, 8/30/18 5:42 PM - 452 bys	V A			res 👌	
Book1 assoc 2 (123 - v1, 8/30/18 6:12 PM - 161 by					
Book2 (123 - v1, 8/18/18 9:38 PM - 98 bytes)					
Book3 (123 - v1, 8/18/18 10:22 PM - 99 bytes)					
BookTrend (123 - v1, 8/19/18 9.58 AM - 815 bytes)	FP-Growth				
BookTrend1 (123 - v1, 8/19/18 10:38 AM - 1 kB)		fre			
Titanic+Data (123., ut. 8/18/18.15.21 PM., 108.18)					
rators ×					
rch for Operators					
Data Access (50)					
Blending (77)					
Cleansing (25)					
Modeling (143)					
Predictive (61)				Show advanced parameters	
Segmentation (14)				Change compatibility (9.0.001)	
Associations (6)					
IFP-Growth					
Treate Association Rules				Help ×	
I Apply Association Rules				Process	
🛒 Generalized Sequential Patterns				RapidMiner Studio Core	
Item Sets to Data				Synopsis	
🛒 Unity item Sets				The root operator which is the out	er most operator of every process.
Correlations (8)					
Similarities (4)				Description	
Feature Weights (17)					one operator of this class, and it must be the
Optimization (20)					operator provides a set of parameters that are logging and initialization parameters of the
Time Series (13)	Leverage the Wisdom of Crowds to get open	ator recommendations based on your process design!		random number generator.	roading and inclanation parameters of the
Scoring (12) V Set more operators from the Marketplace		Activate Wisdom of Crowds			
ser more spendiors from the Marketplace		•		Baramotore	

Fig. 15.2. Screenshot of process structure of association rules defining

🗋 늘 🖩 • 🕨 • 🔳		Views: De	sign Results	Turbo Prep Au	to Model		Find data: operatora: etc 💫 All Studio
Repository ×	Process					Parameters ×	
🕒 Import Data 📃	• Process •				100% 🖉 🧣 🖓 🖓 🕅 🔀	FP-Growth	
	A Process					input format	items in separate columns
 data (123) Book1 (123 - v1. 8/18/18 11.00 AM - 828 bytes) 	Retrieve Book1 ass	Create Association			195.4	🕑 trimitem names	
Book1 assoc (123 - v1, 8/30/18 2 19 PM - 369 6 Book1 assoc1 (123 - v1, 8/30/18 5 42 PM - 452					m ns (min requirement	support
Book1 assoc 2 (125 - v1, 8/30/18 8/12 PM - 161 Book2 (123 - v1, 8/18/18 9:38 PM - 58 bytes)	fP	Growth				min support	0.5
Book3 (123 - v1, 8/18/18 18:22 PM - 99 bytes) BookTrend (123 - v1, 8/19/18 9:58 AM - 815 byte	a					min items per itemset	1
BookTrend1 (122 - v1, 8/19/18 18:36 AM - 1 (8) Titasic+Flata (22 - v1, 8/19/18 13.31 PH - 106)		_				maxitems per itemset	3
Operators X						max number of itemsets	1000000
Search for Operators						I find min number of itemsets	
Data Access (50)	^					min number of itemsets	100
Blending (77)							
Cleansing (25)						must contain list	Edit Enumeration (0)
Modeling (143)							
Predictive (61)							
Segmentation (14)						Show advanced parameters	
Associations (6)							
FP-Growth						Help 🔀	
Treate Association Rules							
Apply Association Rules						FP-Growth	
Generalized Sequential Patterns						Concurrency	
Item Sets to Data						Tags: Associations, Market, Baske Itemset, Itemset, Item set, Mining	t, Upselling, Up-selling, Crossselling, Cross-selling, Francent Patterns
Inity item Sets							CALIFORNIA ANTINALA
Correlations (8)						Synopsis	
Similarities (4)						ExampleSet, using the FP-tree d	es all frequently-occurring itemsets in an ata structure.
Feature Weights (17)						Jump to Tutorial Process	
 Optimization (20) 							
Time Series (13)		get operator recommendations based on your pro	cess design!			Description	
Scoring (12)	× 1	🖌 Actua	te Wisdom of Crowds				comptimer get a suggestion of the following form:
Get more operators from the Marketplace		·					sometimes get a suggestion of the following form:

Fig. 15.3. Screenshot of process structure with define of characteristics of process operators FP- growth with items on separate columns

🗋 늘 🖩 • 🕨 • 🔳		Views Design	Results Turbo Prep Auto Model			Find data, operatora, etc. 🔗	All Studio •
Repository ×	Process				Parameters ×		
🕒 Import Data 🛛 🚍	• Process		100%	🗵 👟 🗸 🐛 🔍 🔍	F Create Association Rules		
	A Process				criterion	confidence	• D
T 📑 data (123)	Retrieve Book1 ass	Create Association Rules					
Book1 (123 - v1. 8/18/18 11.55 AM - 628 bytes)	Diep at D	(ne nd)		195	min confidence	0.5	Ð
Book1 assoc (123 - v1, 8/35/18 2-19 PM - 366 6		ane na		m			
Book1 assoc1 (123 - v1, 8/30/18/5/42 PM - 452				ns 🤇			
Book1 assoc 2 (123 - v1, 8/30/18 8/12 PM - 151	Les						
Book2 (123 - v1, 8/18/18 9/38 PM - 88 bytes) Book3 (123 - v1, 8/18/18 10/22 PM - 98 bytes)							
BookTrend (122 - v1, 8/19/18 10/22 PM - 59 5(98))		Growth					
BookTrend1 (122 - v1, 8/19/18 10:36 AM - 1 x8)		era)					
Titanic+Data (122 - 47, 875) (112 - 57, 714) Titanic+Data (122 - 47, 875) (112 - 57, 714)		The Design of th					
Operators ×							
Search for Operators							
Data Access (50)	~						
Blending (77)							
Cleansing (25)							
Modeling (143)							
Predictive (61)							
Segmentation (14)					E Show advanced parameters		
Associations (6)					es successions		
IFP-Growth							
Treate Association Rules					Help ×		
Apply Association Rules					Create Association R	ules	<u>^</u>
Generalized Sequential Patterns					RapidMiner Studio Core		
Item Sets to Data					Tags: Associations, Market, Basket, Upp- Itemset, Item-set, Item set, Mining, Frequ		selling
Inity item Sets						Cellis LassesLine	
 Correlations (8) 					Synopsis This operator generates a set of assoc	intian rules from the stree set of t	
Similarities (4)					itemsets.	action rules from the given sec of i	request
Feature Weights (17)					Jump to Tutorial Process		
 Optimization (20) Time Series (13) 							
 Scering (12) 	Leverage the Wisdom of Crowds to	get operator recommendations based on your process de	sign!		Description		
Get more operators from the Marketplace		🛹 Activate Wisdo	om of Crowds		Association rules are if/then statemen	ts that help uncover relationships	s between
Click to select, drag to move.					enantical consultated data the account		

Fig. 15.4. Screenshot of process structure with define of characteristics of process operator "Create associative rules" with items on separate columns



Fig. 15.5. Screenshot of source data with list items in one column

👌 📒 🔹 🕨 🔹 🔳		Views Desi	gn Results	Turbo Prep	Auto Model			Find data operatora, etc.	🔎 Ali Studi
epository X	Process						Parameters ×		
🕒 Import Data 🛛 🗉 👻	Process				100% 🔎 🔎 🔎	12 🔹 🖉 🛙	FP-Growth		
							input format	item list in a column	•
🛅 data (123)	Process Retrieve Book1 ass	Create Association							
Book1 (123 - v1. 8/18/18 11:55 AM - 528 bytes)	Dirp at D	provide a second second second second				195 /	item separators	1	
Book1 assoc (123 - v1, s/35/18 2 19 PM - 369 byte		(te ni)				195			
Book1 assoc1 (123 - y1, 8/30/18 5/42 PM - 452 by	V.1.					ns (use quotes		
Book1 assoc 2 (123 - v1, 8/30/18 8 12 PM - 161 by							🕑 trim item names		
Book2 (123 - v1, 8/18/18 9:38 PM - 58 bytes)							C Dum Kern Hames		
Book3 (123 - v1, 8/18/18 10.22 PM - 99 bytes)							min requirement	support	•
BookTrend (123 - v1, 8/19/18 9:58 AM - 815 bytes) BookTrend1 (123 - v1, 8/19/18 10:38 AM - 1 x8)		rowth							
Titanic+Data (122 - v1. 8/19/18 10 38 AM - 1 48)		era p					min support	0.5	
1							min items per itemset	1	
rators ×									
							max items per itemset	3	
Data Access (50)							max number of itemsets	1000000	
Blending (77)									
Cleansing (25)							Ind min number of itemsets		
Modeling (143)							min number of itemsets	100	
Predictive (61)							min number of itemsets	100	
Segmentation (14)							Show advanced parameters		
Associations (6)									
FP-Growth							Help ×		
Treate Association Rules									
Apply Association Rules							FP-Growth		
F Generalized Sequential Patterns							concurrency		
Item Sets to Data							Tags: Associations, Market, Basket Itemset, Itemset, Itemset, Mining,		g. cross-seiling.
Correlations (8)							Synopsis		
Similarities (4)							This Operator efficiently calculate	es all frequently-occurring items	ets in an
Feature Weights (17)							ExampleSet, using the FP-tree da		
Optimization (20)							Jump to Tutorial Process		
Time Series (13)	Leverage the Wilsdom of Crowds to get	operator recommendations based on your proc	ss design						
Scoring (12)							Description		
Get more operators from the Marketplace		Adivate	Wisdom of Crowds				When online shopping, you will s	ometimes get a suggestion of t	ne following form:

Fig. 15.6. Screenshot of process structure with define of characteristics of process operators FP- growth with list items in one column

ult History	🛒 Frequent	titemSets (FP-Growth) 🛛 🛛	🛒 AssociationRules (Create Association	Rules) ×			Repository ×
-	No. of Sets: 47	Size	Support	Item 1	Item 2	Item 3	🕒 Import Data 🖉 🖛
Data	Total Max. Size: 3	1	0.600	cheese			Training Resources (connected)
	Min. Size: 1	1	0.600	tea			Samples Samples (connected)
	Max. Size: 3	1	0.400	Mik			 DB
notations	Contains item:	1	0.400	bread			Local Repository (123)
I Sarona		1	0.400	sour cream			 Cloud Repository (disconnected)
	Update View	1	0.200	Bread			
		1	0.200	Sausage			
		1	0.200	Sugar			
		1	0.200	butter			
		1	0.200	milk			
		1	0.200	sugar			
		2	0.200	cheese	tea		
		2	0.400	cheese	Milk		
		2	0.400	cheese	bread		
		2	0.200	cheese	sour cream		
		2	0.200	cheese	Sausage		
		2	0.200	cheese	sugar		
		2	0.200	tea	Milk		
		2	0.200	tea	bread		
		2	0.200	tea	sour cream		
		2	0.200	tea	Bread		
		2	0.200	tea	Sugar		
		2	0.200	tea	butter		
		2	0.200	tea	milk		
		2	0.200	Mik	bread		
		2	0.200	1010	entir meam		~

Fig. 15.7. Screenshot of frequent items of association rules with items on separate columns

	· · ·			Views: Design F	tesults	Turbo Prep	Auto Model							tata, operatora, etc. 🤌 All Stud
ult History	🛒 FrequentitemSets (FP	Growth)	🛛 🛒 AssociationRules (C	reate Association Rules)										Repository ×
	Show rules matching	No.	Premises	Conclusion		Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti		G Import Data
Data	all of these conclusions:	34	cheese	Milk		0.400	0.667	0.875	-0.800	0.160	1.667	1.800	^	F Training Resources (connected)
	cheese tea	35	cheese	bread		0.400	0.667	0.875	-0.800	0.160	1.667	1.800		Samples Samples Samples (connected)
Z	Milk	36	MIIK	cheese		0.400	1	1	-0.400	0.160	1.667			 I DB
Graph	bread sour cream	37	bread	cheese		0.400	1	1	-0.400	0.160	1.667	00		Eccal Repository (123)
or agen	Bread Sausage	38	Sausage	cheese		0.200	1	1	-0.200	0.080	1.667			Cloud Repository (disconnected)
-	Sugar butter	39	sugar	cheese		0.200	1	1	-0.200	0.080	1.667			
	milk	40	Bread	tea		0.200	1	1	-0.200	0.080	1.667	80		
scription	undat.	41	Sugar	tea		0.200	1	1	-0.200	0.080	1.667			
		42	butter	tea		0.200	1	1	-0.200	0.080	1.667			
1		43	milk	tea		0.200	1	1	-0.200	0.080	1.667	00	- 81	
otations		44	Sausage	bread		0.200	1	1	-0.200	0.120	2.500			
		45	sugar	bread		0.200	1	1	-0.200	0.120	2.500	80		
		46	Bread	sour cream		0.200	1	1	-0.200	0.120	2.500	00		
		47	milk	sour cream		0.200	1	1	-0.200	0.120	2.500			
		48	Bread	milk		0.200	1	1	-0.200	0.160	5			
		49	milk	Bread		0.200	1	1	-0.200	0.160	5	00		
		50	Sausage	sugar		0.200	1	1	-0.200	0.160	5			
		51	sugar	Sausage		0.200	1	1	-0.200	0.160	5			
		52	Sugar	butter		0.200	1	1	-0.200	0.160	5			
		53	butter	Sugar		0.200	1	1	-0.200	0.160	5	00		
		54	cheese, tea	Milk		0.200	1	1	-0.200	0.120	2.500			
		55	tea, Milk	cheese		0.200	1	1	-0.200	0.080	1.667			
	Min. Criterion:	56	cheese, tea	bread		0.200	1	1	-0.200	0.120	2.500			
	confidence *	57	tea, bread	cheese		0.200	1	1	-0.200	0.080	1.667			
	Min. Criterion Value:	58	Milk, bread	cheese		0.200	1	1	-0.200	0.080	1.667			
		60	chases environaam	sale		0.200			JN 200	0 120	2 500	~	\sim	

Fig. 15.8. Screenshot of association rules with items on separate columns

ult History	🛒 Frequen	attemSets (FP-Growth) \times	🛒 AssociationRules (Create Association	Rules) ×			Repository ×
_	No. of Sets: 39	Size	Support	Item 1	Item 2	Item 3	🕒 Import Data 🛛 🗉 🗣
Data	Total Max. Size: 3	1	0.600	a			 Training Resources (connected)
	Min. Size: 1	1	0.600	b			 Samples Community Samples (connected)
	Max. Size: 3	1	0.600	c			 BB
iotations	Contains Item:	1	0.600	d			Local Repository (123)
TO SERVITE		1	0.400	r			Cloud Repository (disconnected)
	Update View	1	0.400	9			
		1	0.200	8			
		1	0.200	h			
		2	0.400	а	5		
		2	0.400	a	c		
		2	0.400	8	d		
		2	0.200	8	č.		
		2	0.200	3	9		
		2	0.200	а	e		
		2	0.400	b	c		
		2	0.400	b	d		
		2	0.400	b	t:		
		2	0.200	c	d		
		2	0.200	c	r		
		2	0.200	c	g		
		2	0.200	c	h		
		2	0.200	d	r		
		2	0.200	d	g		
		2	0.200	d	6		
		2	0.200	0	e		
		2	0.200				~

Fig. 15.9. Screenshot of frequent items of association rules with list items in one column

	- · ·			Views: Design	Results	Turbo Prep	Auto Model							data, operatora, etc. 🤌 All Stu
ult History	FrequentitemSets (FP-1	Growth)	🕺 🛒 AssociationRules (Cr	eate Association Rules) $ imes$										Repository ×
-	Show rules matching	No.	Premises	Conclusion		Support	Confidence	LaPlace	Gain	p-9	Lift	Convicti		G Import Data
Data	all of these conclusions:	33	а	b		0.400	0.667	0.875	-0.800	0.040	1.111	1.200	^	F Training Resources (connected
	a	34	b	a		0.400	0.667	0.875	-0.800	0.040	1.111	1.200		Samples Gommunity Samples iconnecte
-	c d	35	а	c		0.400	0.667	0.875	-0.800	0.040	1.111	1.200		> 📕 DB
Graph	t	36	c	1		0.400	0.667	0.875	-0.800	0.040	1.111	1.200		Local Repository (123)
	9	37		d		0.400	0.667	0.875	-0.800	0.040	1.111	1.200		Cloud Repository (disconnected)
-	h	38	d	а		0.400	0.667	0.875	-0.800	0.040	1.111	1.200		
scription		39	b	c		0.400	0.667	0.875	-0.800	0.040	1.111	1.200		
acalpatri		40	c	b		0.400	0.667	0.875	-0.800	0.040	1.111	1.200	1	
		41	b	d		0.400	0.667	0.875	-0.800	0.040	1.111	1.200		
		42	d	b		0.400	0.667	0.875	-0.800	0.040	1.111	1.200	- 11	
notations		43	b	1.		0.400	0.667	0.875	-0.800	0.160	1.667	1.800		
		44	a	b, c		0.400	0.667	0.875	-0.800	0.160	1.667	1.800		
		45	b	a, c		0.400	0.667	0.875	-0.800	0.160	1.667	1.800		
		46	c	a, b		0.400	0.667	0.875	-0.800	0.160	1.667	1.800	- 11	
		47		a		0.200	1	1	-0.200	0.080	1.667	80		
		48	1	b		0.400	1	1	-0.400	0.160	1.667	80		
		49	h	c		0.200	1	1	-0.200	0.080	1.667			
		50		d		0.200	1	1	-0.200	0.080	1.667			
		51	e	g		0.200	1	1	-0.200	0.120	2.500	00		
		52	h	g		0.200	1	1	-0.200	0.120	2.500	~		
		53	a, b	c		0.400	:10	1	-0.400	0.160	1.667			
		54	a, c	b		0.400	1	1	-0.400	0.160	1.667			
	Min. Criterion:	55	b, c	а		0.400	1	1	-0.400	0.160	1.667	80		
	confidence *	56	a,f	b		0.200	1	1	-0.200	0.080	1.667			
	Min. Criterion Value:	57	c, d	а		0.200	1	1	-0.200	0.080	1.667			
	-1	6.0	. 1	~		0 200	4		-n 200	0.090	1 667	~	\sim	

Fig. 15.10. Screenshot of association rules with list items in one column

15.2 PRACTICAL LESSON 15 and IWS 15

Subject. Building and using association rules methods.

Plan of the lesson.

1) Study the scheme for constructing and using the association rules methods presented above.

2) The task of the IWS. Apply the above scheme of building and using association rules methods for your clustering task.

3) Analyze the results.

Literature.

1. Larose D.T. Discovering knowledge in data - an introduction to data mining. Wiley-Interscience, Hoboken, New Jersey, 2005

2. Rapidminer Studio, manual. https://docs.rapidminer.com/latest/studio