MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
NATIONAL AVIATION UNIVERSITY
FACULTY OF AERONAVIGATIONS, ELECTRONICS AND TELECOMMUNICATIONS
DEPARTMENT OF TELECOMMUNICATION AND RADIO ENGINEERING SYSTEMS

ADMIT TO DEFENCE
Head of the Department

_____ R. Odarchenko
"_____"_____2022

DIPLOMA WORK

(EXPLANATORY NOTE)

BACHELOR'S DEGREE GRADUATE
BY SPECIALITY "TELECOMMUNICATIONS AND RADIO ENGINEERING"

Topic: «Data Mining algorithms for analyzing the customer base of a mobile operator»

Performer: _____ Yana Zhezhel
(signature)

Supervisor: _____ O. Holubnychyi
(signature)

N-controller: _____ D. Bakhtiyarov
(signature)

Kyiv 2022

NATIONAL AVIATION UNIVERSITY

Faculty of aeronavigations, electronics and telecommunications
Department of telecommunication and radio engineering systems
Speciality: 172 "Telecommunications and radio engineering"
Educational professional program: Telecommunication systems and networks

ADMIT TO DEFENCE
Head of the Department

_____ R. Odarchenko
"_____"_____2022

TASK
for execution of bachelor diploma work

Yana Zhezhel
(full name)

1.Topic of diploma work: «Data Mining algorithms for analyzing the customer base of a mobile operator»
approved by the order of the rector from « 25» April 2022 No433/ст

2. The term of the work: from 23 May 2022 to 17 June 2022.

3.Initial work data: SAS Enterprise Miner software

4. Explanatory note content: Introduction. The graphical package interface is a "specify and" interface With it, I completed all stages of the process Data Mining from the selection of data sources.

5. List of required illustrative material: SAS Enterprise Miner software is installed on my computer.

6. Work schedule

№ n/p	Task	Term implementation	Performance note
1.	Write a detailed content of diploma sections	23.05.2022-25.05.2022	Done
2.	Introduction	25.05.2022	Done
3.	Data Mining algorithms	26.05.2022-29.05.2022	Done
4.	Advanced algorithms	30.05.2022-02.06.2022	Done
5.	Stages and methods of finding new knowledge	03.06.2022-08.06.2022	Done
6.	Elimination of shortcomings and defense of the thesis	09.06.2022-17.06.2022	Done

7. Date of issue of the assignment: “20” May 2022.

Supervisor _____ O. Holubnychyi
(signature) (full name)

Accepted task for execution _____ Yana Zhezhel
(signature) (full name)

ABSTRACT

Graduate work on the topic “Data Mining alorithms for analyzing the customer base of a mobile operator ”. It contains 56 pages, 13 Figures, 6 sources.

The object of the study is a SAS Enterprise Miner software.

The purpose of the thesis is to complete all stages of the process Data Mining from the selection of data sources.

Research of the methods – the software which is installed on the computer.

Data Mining projects developed locally and in the client-server architecture. The package supports the implementation of all necessary procedures within a single integrated solution with the ability to work together and comes as distributed client-server application, which is especially convenient to implement data analysis on the scale of large organizations. SAS Enterprise Miner package designed for data analysts, marketing analysts, marketers, risk analysis specialists, fraud detection specialists.

Material of diploma work are recommended to be used in conducting scientific research, educational process and practical activity in the teaching of undergraduate disciplines.

CONTENTS

LIST OF ABBREVIATIONS

1) SAS Enterprise – Statistical Analysis System Enterprise
2) MCI – Media Control Interface
3) TASA – Telecommunication Access System Act
4) OR – Operating Room
5) BI – Bussiness Intelligence
6) PCA – Principal Component Analysis
7) MECE – Mutually Exclusive Collectively Exhaustive
8) XML – Extensible Markup Language
9) PMML – Predictive Model Markup Language
10) DTD – Document Type Declaration
11) HTML – Hypertext Markup Language
12) CHAID – Chi-square Automatic Interaction Detection
13) KXEN – Knowladge Extraction Engines
14) DBMS – Database Management System
15) IMS – Information Management System
16) IBM – International Business Machine
17) SQL – Structed Query Language
18) CRM – Customer Relationship Management
19) USD – United States Dollars
20) MSUA – Mobile Satellite Users Association
21) SEMMA – Sample , Explore , Modify, Model and Assess

INTRODUCTION

The rapid development of telecommunication technologies in the field of mobile communications is caused by increasing needs of users every time. Nowadays, the technologies cannot please all the needs of users, so next step in the development in the field of telecommunication systems is implementation of Data Mining algorithms. You may ask why there so many algorithms available. During the last 30 years the research has generated many variants of Data Mining algorithms that are suited to particular areas in the solution landscape. For better customer support service this process involves sifting information gathered from different sources. The patterns have found using Data Mining applications to help to recognize customer trends and be prepared to support them properly. Also it provides one of the best environments for involving business solutions. The smart systems are composed of many individual techniques designed to work to create powerful Data Mining models. Telecommunication systems have a lot of applications predicting rare events, for example the failure of network element or instance of phone fraud. The rarity is another issue that must be dealt with. The use of data mining technologies, as a rule, is based on the processing of large amounts of information accumulated in modern data warehouses. There are different concepts, technologies and practicals approaches to the construction of such repositories.

The purpose of the diploma work is research and analysis of algorithms of Data Mining for customer base of mobile operator.

The object of research is Data Mining algorithms which are used in the telecommunication systems.

The subject of research is the customer base of the mobile operator and service which it provides.

Research methods. To achieve the goals of diploma work it is necessary to perform such tasks:

– To install the SAS Enterprise Miner software

– To complete all stages of the process Data Mining from the selection of data sources
– To make conclusions

CHAPTER 1
DATA MINING ALGORITHMS

1.1. The overview

What does it mean? Data Mining algorithm or you can call it statistical it is mathematical expression of aspects of the patterns they find in data. Different algorithms provide perspectives on the complete nature of the pattern. The telecommunications generates and stores a lot of data. It includes calls detail data (which describes calls that traverse the network), the network data (which describes the state of hardware and software components), customer data (which describes customers). The amount of data is big that is why manual analysis is difficult sometimes it is not possible. The needs to manage with such a large volumes of data led to the development of knowledge-based systems. It performed functions such as identifying fraudulent phone calls and identifying network faults. The problem is that time consulting to obtain the knowledge from human experts and usually they do not have it. The appearing of Data Mining technology promised to solve this problems and that is why telecommunication systems was an early adopter of this technology. Data of telecommunications pose some interesting issues for Data Mining. Firstly, it concerns the scale since the databases may contain billions of records and mostly the largest in the world. Secondly, the raw data is not for Data Mining. Before you can use this data, summary features must be identified and data must be summarized by this features. Telecommunication systems have a lot of applications predicting rare events, for example the failure of network element or instance of phone fraud. The rarity is another issue that must be dealt with. Thirdly, the last problem is Data Mining concerns real-time performance: applications such as fraud detection require that any model/rules must be applied in real time. Each of this problems will be discussed in next topics in this work.

1.2. Telecommunication data and types

For now, in Data Mining you should understand what is Data. Without this the useful applications cannot be developed. So in this section I want to describe the three main types of telecommunication data. As I mentioned before the raw data is not suitable for Data Mining then we should use transformation steps. It is necessary to generate data that can be also described. Also I will show how we can use it for extraction useful information from data sets. Every time when you make a call in the telecommunication network, all information is saved as call detail record. It takes a lot of place (number of call, details record). For example, AT&T long distance customers generate over 300 million call detail records per day. Given some months of call detail data is kept online, it means tens of billions will need to be stored at any time. This records include sufficient information to describe important things of each call. For minimum, each call detail record will contain originating and terminating phone numbers, the date and time, duration of the call. It generated in real-time and available almost immediately for Data Mining, for example like a billing data which is available only once per month. The call detail records are used not only for Data Mining, since the goal is to extract knowledge at the customer level, but not at the level of individual phone calls. It associated with a customer must be summarized into single record that describes everything. The choice of summary features is critical in order to obtain a description from customer. I will write below the list of features that one might use when generating description of customer based on calls they originate and receive over time P:

- average call duration
- % no-answer calls
- % calls to/from a different area code
- % of weekday calls (Monday – Friday)
- % of daytime calls (9am – 5pm)
- average # calls received per day
- average # calls originated per day

- # unique area codes called during P

This eight features are used to make customer account. Such account has many applications, for example it can be used to distinguish between business and residential customers based on percentage of weekday and daytime calls. Most of this features were generated in straightforward method from the underlying data, but some of them, such as eight feature, is required more thought and creativity. Most people call only several area codes over short period of time, this one can help identify telemarketers, since they call to many different area codes. That example demonstrates that generating useful features is critical step within Data Mining. If bad features will be generated, it will be not be successful. But the construction of features must be guided by common sense, it should include data analysis. However, more detailed data analysis shown in Figure 1, indicates that period from 9 to 4 pm is actually more appropriate for purpose. Figure 1 it is plots, for each weekday hour, h, the business to call ratio, which is computed as: *% weekday business calls during h / % weekday residential calls during h*

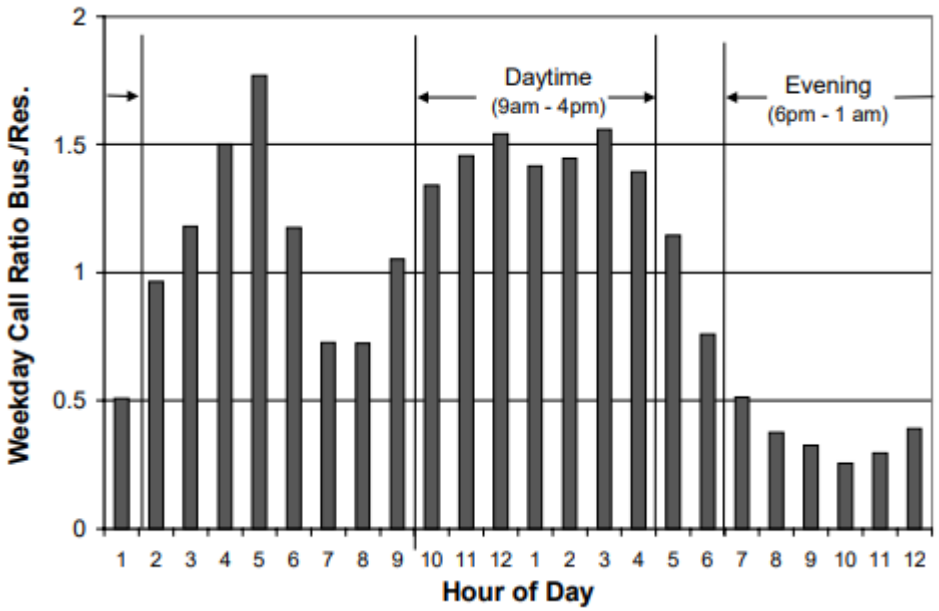

Fig.1.1. Comparison of Business and Residential Hourly Calling Patterns

This figure shows that during the period of 9 am to 4 pm, business place roughly 1.5 times as many of their total weekday calls as does a residence. At 5 pm the ratio is close to 1, indicating that the calls during this timeframe are not useful for distinguishing between a

business and a residence. The calls in the evening timeframe (6 pm – 1 am) are also useful in distinguishing between the two types of customers. For some applications, for example fraud detection, the summary descriptions, sometimes called signatures must be updated in real-time for millions of phone lines. This requires the use of fairly short and simple summary features that can be updated quickly.

1.3. Network data

Every network element is capable of generating error and messages, which leads to amount of network data. This data must be stored and analyzed in order to support network management functions, for example fault isolation. For example, such message must indicate that “controller 7 experienced a loss of power for 40 seconds starting at 10:04 pm on Monday, March 13.” That is why expert systems have been developed to automatically analyze these messages and take actions, only involving a technician when a problem cannot be automatically resolved. As I described before, Data Mining technology is helping identify network problems by automatically extracting knowledge from the network data. This is accomplished by applying a time window to the data. For example, such summary must indicate that a hardware component experienced twelve instances of a power fluctuation in a 10 minutes period. The companies, like large businesses, have millions of customers. By the necessity this means maintaining a database of information on customers. This information will include name and address and must include other information such as service plan and contract information, credit score, family income and payment history. This information must be supplemented with data from sources, such as from credit reporting agency. The customer data maintained by telecommunication companies does not differ from that in other industries, the applications do not focus on this source of data. For example, customer data is used to supplement call detail data when trying to identify phone fraud.

1.4. Operations

The telecommunication assiduity was an early adopter of Data Mining technology and numerous operations live. I'll describe several operations in this section. Fraud it's serious problem for telecommunication companies, leading to billions of bones in lost each time. It can be divided into two orders subscription fraud and superimposition fraud. Subscription fraud occurs when client opens an account with the intention of noway paying for the charges. Superimposition fraud involves a licit account with some exertion, but also includes some exertion by a person other than the account holder. Superimposition fraud poses a big problem for the telecommunication assiduity and because why we concentrate on operations for relating this type of fraud. These operations should operate in real- time using the call detail records and when fraud is detected it should spark some action. This must be incontinently block the call and/ or kill the account, or involve opening an disquisition, which will affect in a call to the client to corroborate the legality of the account exertion. The system for relating fraud is to make account of client's calling geste and compare exertion against this geste. This Data Mining operation relies on detection.However, fraud can be linked after it occurs, If the call detail summaries are streamlined in real- time. Because geste doesn't indicate fraud, one fraud-discovery system augments this approach by comparing the new calling geste to biographies of general fraud — and only signals fraud if the geste matches one of these biographies. This introductory approach has been used to identify cellular cloning fraud, which occurs when the identification information associated with cell phone is covered and also programmed into a alternate phone. This Data Mining operation anatomized large quantities of cellular data in order to identify patterns of fraud. These patterns were used to induce observers, which watches a client's geste to one pattern of fraud. These observers were fed into a neural network, which determined when there's substantiation of fraud to raise an alert. Data Mining can also help descry fraud by relating and storing phone figures called when a phone is usedfraudulently.However, one should infer that the account is used fraudulently, If numerous calls appear from another phone to figures on this list. Fraud operations have characteristics that bear variations for standard Data Mining ways.

For illustration, the performance of a fraud discovery system should be reckoned at the clientlevel.However, this should count as only one alert; else the system may appear to perform better than it does, If client account generates 30 fraud cautions. Sophisticated costbased criteria can be used to estimate the system. This is important because misclassification costs for fraud are unstable and largely slanted. That's why, when erecting a classifier to identify fraud, one should know the relative cost of letting fraudulent call go through versus the cost of blocking a call from licit client. Another issue is since fraud is rare and the number of vindicated calls is low the fraud operation involves rare event where the underpinning class distribution is disposed. Data Mining algorithms have great dealing with disposed class distributions and prognosticating events. For illustration, if fraud makes up only 3 of all calls, Data Mining systems won't induce rules for fraud, since a dereliction rule, which noway predicts fraud, would be99.9 accurate. To deal with this issue, the training data is named to increase the proportion of fraudulent cases. Still, the use of anon-representative training set can be problematic because it doesn't give the Data Mining styles with accurate information about the class distribution.

1.5. Client profiling

Telecommunication companies have a great deal of data about guests. In addition to the general client data, telecommunication companies store call detail records, which describe geste of client. This information can be used to outline the guests and it can be used for marketing and soothsaying purposes. I begin with well- known marketing juggernauts in the telecommunication assiduity MCI’s musketeers and Family creation. It was launched in the United States in 1991 and was responsible for significant growth in client base. It offered to reduce calling freights when it's placed to others in one circle. This creation was began when request experimenters noticed subgraphs in the callgraph of network exertion which suggested the possibility of adding calling circles rather than approach of adding individual subscribers.MCI relied on customers to bring in members of calling circle, even it could have utilized its call detail data to generate a list of the people in each calling circle. The most reason for this is MCI did not want to anger customers by

using calling history. This demonstrates privacy concerns are an issue for Data Mining in the telecommunication industry, when call detail data is involved. The promotion relied on Data Mining to identify associations within data. Marketing application that relies on this technique is a Data Mining application for finding the set of non U.S. countries telecommunication customers. An issue with company is customer churn. It involves that customer is leaving one telecommunication company for another. Churn is a significant problem because of the associated loss of revenue and the cost of attracting new customers. The worst cases of churns occurred several years ago when competing long distance companies offered incentives, $60 or $100, for signing up with their company a practice which led to customers switching carriers in order to earn the incentives. Data Mining techniques permit companies the ability for historical data in order to predict when customer is likely to go. These techniques utilize billing data, call detail data, subscription information and customer information. Based on the induced model, the company can take action, if desired. For example, a wireless company may offer a free phone for extending their contract. Such effort utilized neural network to estimate the probability h(t) of cancellation at the given time t in the future. In the telecommunication industry, it is useful to profile customers based on patterns of phone usage, which can be extracted from the call detail data. Customer profiles can be used for marketing purposes, or better understand customer, which can lead to better forecasting models. The call detail data must be summarized to the customer level as described earlier here. Classifier induction program can be applied to a set of labeled training examples in order to build a classifier. This approach has been used to identify fax lines and to classify a phone line as belonging to a business or residence. Other applications have used approach to identify phone lines belonging to telemarketers and to classify a phone line as being used for voice or data. Two rules for classifying customer as being a business or residential customer using pseudo-code. These rules were generated by SAS Enterprise Miner, a sophisticated Data Mining package that supports multiple techniques. The rules were generated using a decision tree learner. Neural network was used to predict the probability of a customer being a business or residential customer, based on the distribution of calls by time of day. The probability estimate generated by the neural network was then used as an input to the

decision tree learner. Evaluation on a separate test set indicates that rule 1 is 89% accurate and rule 2 is 71% accurate.

- If < 43% of calls last 0-10 seconds and < 13.5% of calls occur during the weekend and neural network says that P > 0.58 based on time of day call distribution then business customer.
- If calls received over two month period from at most 3 unique area codes and <56.6% of calls last 0-10 seconds then residential customer.

It is noting because a telecommunication company generates a call detail record if the paying party is its customer, the company will also have a sample of calls for not customers. If a company has high market penetration, this sample may be large for Data Mining. Telecommunication companies have the technical ability to profile not customers as well as customers.

1.6. Fault isolation

Telecommunication networks are complex configurations of hardware and software. The network elements are capable of limited self-diagnosis and these elements may generate millions of status and alarm messages each month. In order to manage the network, alarms should be analyzed automatically in order to identify faults in a timely manner or before they occur and degrade network performance. The TASA is one tool that helps with the knowledge acquisition task for alarm correlation. This tool automatically discovers recurrent patterns of alarms within the network data along with their properties, using a specialized Data Mining algorithm. Network specialists use this information to construct a rule based alarm correlation system, which can be used in real time to identify faults. TASA is capable of finding rules that depend on temporal relationships between the alarms. Before standard classification can be applied to the problem of network fault isolation, the underlying time series data must be rerepresented as a set of examples. This summarization, process involves using a fixed time window and characterizing the behavior over this window. One may then label the example based on whether a fault

occurs within some other time frame, for example, within the following 4 minutes. Two time windows are required. Weiss & Hirsh view this task as an event prediction problem while Fawcett & Provost view it as an activity monitoring problem. Transforming time series data so standard classification tools can be used and has drawbacks. For example, using the scalar based representation all sequence information is lost. Timeweaver is a genetic algorithm based Data Mining system that is capable of operating on the raw network level timeseries data and making it unnecessary to represent the network level data. Given a sequence of timestamped events and a target event T, it will identify patterns that successfully predict T. The system is designed to perform well when the target event is rare, which is critical most network failures are rare. The target event is the failure of components in the 4ESS switching system.

1.7. Conclusion

I described how Data Mining is used in the telecommunication industry. Three main sources of telecommunication data were described, as were common Data Mining applications. One central issue is that telecommunication data is not in a form or at a level suitable for Data Mining. Other issues that were discussed include the large scale of telecommunication data sets, the need to identify very rare events. Data mining applications must consider privacy issues. This is correct in the telecommunications industry, since telecommunication companies maintain private information, such as whom each customer called. Most telecommunication companies utilize this information and privacy concerns have far been minimized. A more significant issue in the telecommunications industry relates to legal restrictions on how data may be used. In the USA, the information that telecommunications company acquires about their subscribers is referred to as Customer Proprietary Network Information and there are restrictions on how this data may be used. The Telecommunication Act of 1996, along with more clarifications from the Federal Communications Commission, generally prohibits the use of information without customer permission, even for the purpose of marketing the customers other services. In the case of customers who switch to other service providers,

the service provider is prohibited from using the information to try to get the customer back. Companies are prohibited from using data from one type of service in order to sell another service. The use of Data Mining is restricted in there are many instances in which knowledge extracted by process cannot be legally exploited. Much of the rationale for prohibitions relates to competition. For example, if large company can leverage the data associated with one service to sell another service, then companies provide fewer services would be at a competitive disadvantage. This has resulted in many successful applications. Given the fierce competition in the telecommunication industry, one can only expect the use to accelerate, as companies strive to operate more efficiently and gain a competitive advantage.

CHAPTER 2
ADVANCED ALGORITHMS

2.1.Introduction

You may ask why there are many algorithms exist. Research during the past years has generated many variants of Data Mining algorithms for areas in the solution landscape. Fig.1 shows specific algorithms fit into the solution landscape of business analytic problems areas: OR; forecasting; Data Mining; statistics; BI.

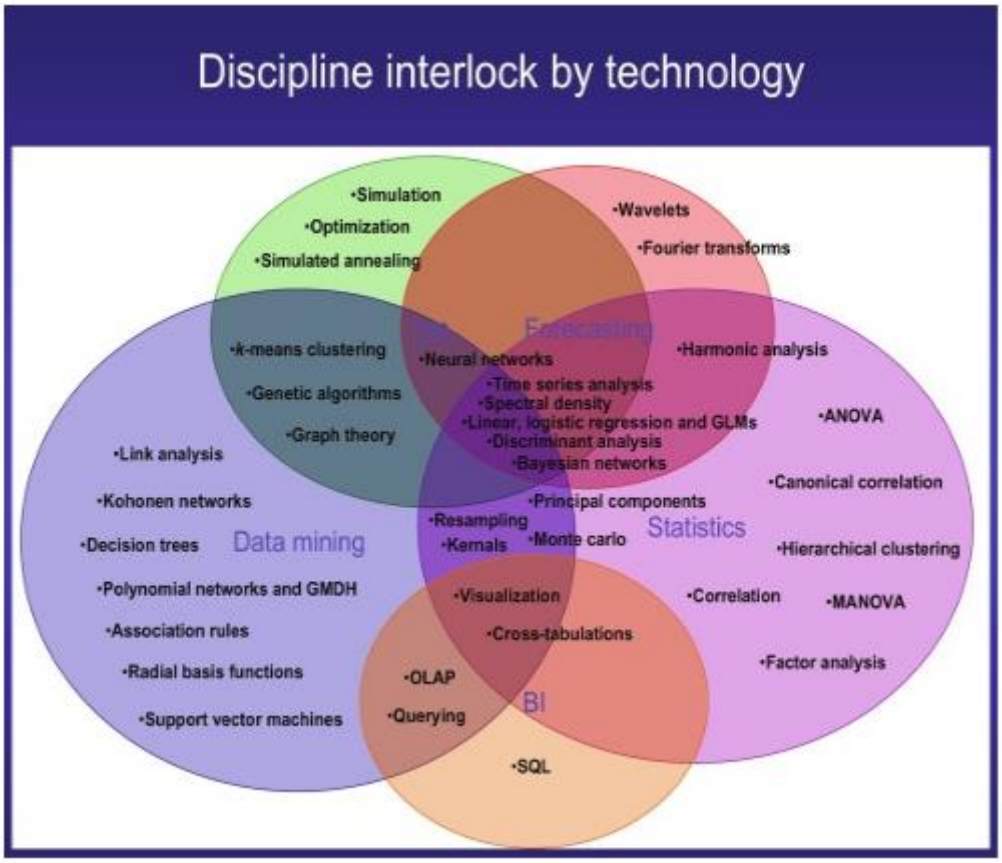

Fig.2.1. Specific algorithms and business analytic problem

It came from studies Elder Research on algorithms used in articles in different domains. You can see how field uses technique and what are suited to overlaps between areas. Data Mining includes a lot of techniques that are not considered in the world's

statistics. Operations research not only uses clustering, graph theory, neural networks and time series but also simulation and optimization. Forecasting overlaps Data Mining, statistics, and OR ,Fourier transforms and wavelets. For example, PCA is known in electrical engineering as the Karhunen-Loève transform and in statistics as the eigenvalue-eigenvector decomposition. One of the important products of higher education is that see the interconnections between ideas in different disciplines. Now, I will turn to the main job at hand in this chapter and look at advanced algorithms individually. Because it is implemented in different ways in each Data Mining or statistical package, I will consider how they are implemented in STATISTICA Data Miner.

2.2.Classification

Classification requires that you accept a number of assumptions. The fidelity of classes and ability will depend on how close your data set fits these assumptions. I stressed the importance of describing data set in terms of the nature of its variables, possible interactions with the target variable and with each other, and their underlying distributional pattern. In classification, I will try to satisfy assumptions as much as possible. Numerical Variables Operate Best Categorical variables can be used, but they should be decomposed into variables, if possible. Most data mining algorithms will eliminate cases with missing values in variables. Imputation of missing values it is one way to fix problem. Another way that some classification algorithm may fill missing values is to use surrogate variables. A surrogate variable has a similar splitting behavior to the variable with the missing value, and it can be used to replace a missing value in predictor variable. Variables are linear and independent in effects on the Target Variable. When we say that, we mean that there is a straight-line change in the target variable as each variable is varied over its range and that the effect of one variable is not related to effects of any other variable. Target variables must be independent also. Classification targets selected to define categories should be mutually exclusive and exhaustive. For example, data set used to classify shades of blue balls cannot contain any green balls. Mutually exclusive means only one target can be assigned to each case. If MECE is not

satisfied, assignment of some cases into categories may be arbitrary and not related to the predictor variables. Dean Abbott is an internationally recognized Data Mining and analytics expert with over two decades of experience applying advanced algorithms, data preparation techniques, and data visualization methods to world problems, including fraud detection, risk modeling, text mining, personality assessment, response modeling, survey analysis, planned giving, and predictive toxicology. He is also chief scientist of SmarterRemarketer, company which focus on behaviorally and data-driven marketing attribution and web analytics.

2.3.Meta data aspects

Currently, Data Mining algorithms require bringing all together data to be mined in a centralized data warehouse. A fundamental challenge is developing distributed versions of algorithms so that can be done while leaving some of the data in place. In addition, protocols, languages, and network services are required for mining distributed data to handle the mappings which are required. Such functionality is provided via metadata. XML is fast emerging on the World Wide Web. The XML files may be used to store metadata in representation to facilitate the mining of heterogeneous databases. PMML has been developed by the Data Mining community for the exchange of models between different data sites; these will be distributed over the Internet. This tools support interoperability between heterogeneous databases thus Distributed Data Mining.

XML is standard for representing data on the World Wide Web. Traditional Database Engines can be used to process XML documents conforming to DTDs. The XML files store metadata to facilitate the mining of multiple databases. PMML has been developed by the Data Mining community for the exchange of models between different data sites; it will be distributed over the Internet. Such tools support interoperability between databases thus Distributed Data Mining.

2.4.Web mining

This term is reserved for text that comes from web pages. It is not much different from text mining in written documents, but unstructured mining and analytics field has developed around it. Web mining is defined in the field as using traditional Data Mining algorithms and methods to discover patterns by using the web. It can be divided into three different types: web usage mining, web content mining, web structure mining. Figure shows these interrelated components of web mining.

Fig.2.2. Three components of web mining

Usage is the process of gathering information from server logs to understand the history on the web. Web content mining is the process of discovering what users are looking for there. Some users are looking at only textual data, when others are interested in multimedia data. This process tabulates information from text, image, audio, video data on the web. It sometimes is also called web text mining because text content is used often. It uses graphing methods to illustrate connection structures of websites. There are two types of web structure mining:

- Patterns from hyperlinks in the web page
- Treelike patterns of the web page structures that describe HTML or XML usage

The most important decision in the practice of Data Mining: selecting the right modeling algorithm to start with. Data Mining Algorithms and Procedural Analyses Basic Data Mining Algorithms contain:

1. Association Rules
2. Automated Neural Networks
3. Generalized Additive Models
4. General Classification/Regression Tree Models
5. General CHAID Models
6. Generalized k-Means Cluster Analysis Advanced Data Mining Algorithms

The first example is how STATISTICA Data Miner Recipe Interface packages all basic steps of Data mining project into an easy-to-use interface. The second example is KXEN. Both tools select the modeling algorithms, permit you to enter a few settings, and automatically generate model results. Use of either tool can be the best way for beginning data miners to build their first model.

2.5.STATISTICA Data miner recipe

After creation of your first model, you can feel a new sense of empowerment. It is amazing to see patterns in your data that you couldn't see before! The DMRecipe process will be described here shortly by myself. The DMRecipe Interface consists of several interactive steps that guide you through the process to create models: click on Data Mining and then Data Miner Recipe; select New on the recipe screen; click on Open/Connect Data File to select the input data set; click on Apply Data Transformations, if needed; click on Select Variables to select initial input variables; click on the downward-pointing triangle (▾) symbol to run the recipe. When model is completed, the results screen will display.

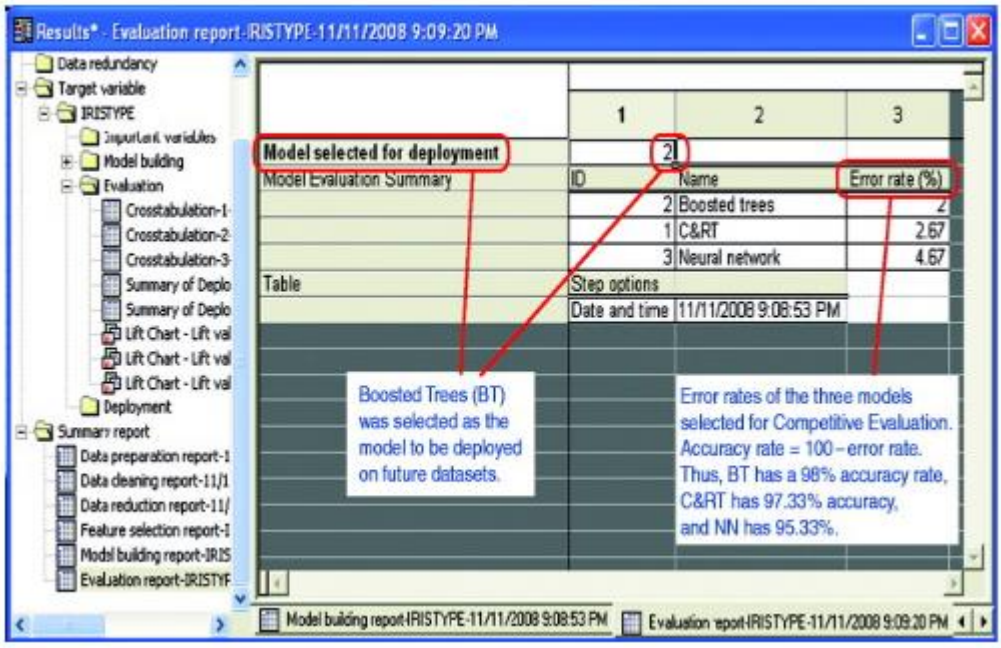

Fig.2.3. DMRecipe results screen

The DMRecipe Interface provides an automatic method for building Data Mining models. Figure 3 provides several reports to help the modeler evaluate the predictor power of the models. It shows that the Boosted Trees algorithm had the lowest error rate. The DMRecipe Interface selects modeling algorithms by default. As you can see, Boosted Trees had the lowest error rate among the algorithms trained, with an accuracy rate of 99%. Since Boosted Trees had the highest accuracy rate, it was selected as the model to be used for deployment to score future data sets. You can find that some data sets do not generate acceptable prediction accuracies with any model. STATISTICA Data Miner Recipe is a semi-automatic method for building complex analytical models for classification or numerical prediction. The DMRecipe Interface provides a step-by-step approach to data preparation, variable selection, and reduction, resulting in models trained with different algorithms. The first activity in the Data Mining process is preparation of the data set for modeling. Common data cleaning and transformation operations can be performed to provide data in the format suitable for the modeling algorithms. After the data set is prepared, you can conduct statistical analysis of the variables. Some variables can carry information similar to that of other variables, making them redundant. The DMRecipe tool provides measures of redundancy for continuous variables. You should let

it eliminate all but one from group of redundant variables. The resulting variable set will generate a much better model. That to eliminating redundant variables, you can reduce the number of variables even by eliminating variables highly correlated with the target variable. Multiple models are trained automatically. After building your Data Mining models, you can use it to score new data sets. KXEN An analogous series of steps are followed in the semi-automatic Modeling Assistant interface for processing all steps of simple Data Mining operation from model selection to model results. The screen permits you to select operations:

- Classification
- Clustering
- Time-series analysis
- Data exploration
- Data manipulation

The second screen shows you the input data file. The third allows you to view the data set or continue to analysis. The fourth provides a metadata report for each variable. The fifth provides a facility for variable selection. In KXEN, the variable selection prompts you for a list of variables to exclude rather than include. The counterintuitive response is puzzling at first, but it is good sense when viewed in context with the strengths of KXEN. This tool can accept variables that are redundant or collinear, and it automatically analyzes and excludes inappropriate variables from analysis. Variable exclusion is performed to permit you to avoid submitting variables to the modeling. The final screen provides report.

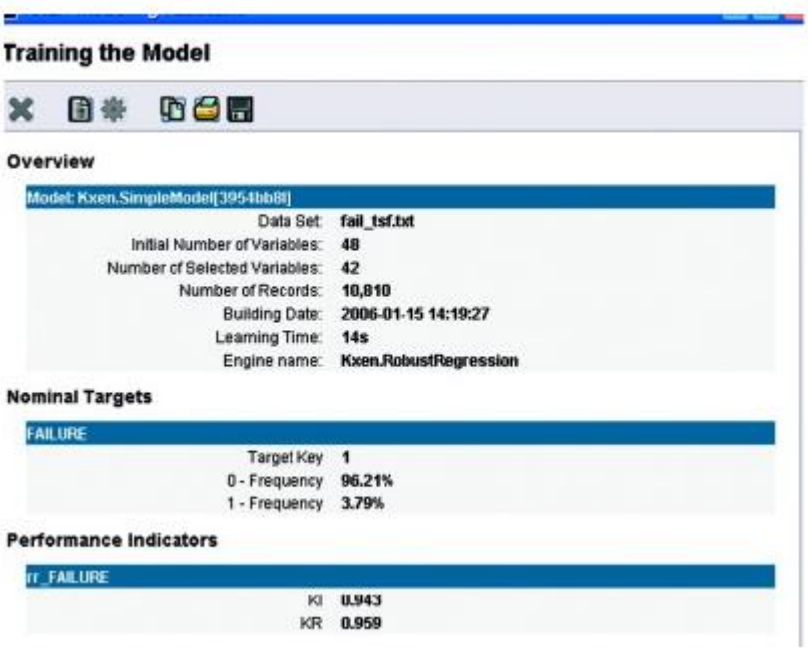

Fig.2.4. Results

A number of reports are available in KXEN similar to in the DMRecipe. One of the most powerful is the operation of the Consistent Coder. KXEN derives automatically a number of new variables as combinations and transformations of existing variables. This automated reduction of unnecessary dimensionality can produce very powerful models without significant danger of multicollinearity. It is very efficient for modeling problems in which models must be developed. Separate sales forecasting models should be created for each category of sales merchandise in retail operations. KXEN can create numbers of models very quickly. These tools provide analogy to the ideal, in which Data Mining is as easy to use as the automobile interface.

2.6.The essence of analytical technologies

Analytical technologies have been used by mankind for a long time. A simple example of analytical technology is the Pythagorean theorem, which allows determine the length of the hypotenuse having known leg lengths. Another example of analytical technology is algorithm for processing information by the human brain. Even a child's brain can perform tasks beyond the control of modern computers, for example,

recognizing familiar faces in a crowd or effectively managing multiple dozens of muscles while playing football. The uniqueness of the brain is the ability to solving new tasks - playing chess, driving a car, etc. But at the same time, the brain is poorly adapted to process large amounts of numerical information - a person cannot multiply two multi-digit numbers without using calculator or calculation algorithm in a column. Real tasks with numbers, much more difficult than multiplication and man to solve such tasks require additional techniques and tools. Under analytical technology we will understand the techniques that are based certain models, algorithms, mathematical theorems allow for known data to evaluate the values of unknown characteristics and parameters. Analytical technologies are needed primarily by approving people important decisions - managers, analysts, experts, consultants. Income the company is largely determined by the quality of these decisions - the accuracy of forecasts, the optimality of the chosen strategies. And development depends on the quality of these decisions companies. Analytical technologies can be used to solve problems forecasting, for example, exchange rates, commodity prices, demand, income company, unemployment rate and optimization, such as procurement plan, plan investment, development strategies. It should also be noted that for real tasks business and production, there are no clear algorithms for solving them. Managers and experts find solutions to such problems only on the basis personal experience. Often classical methods are ineffective for many practical tasks, as it is impossible to accurately describe reality with a small number of model parameters, or calculation. The model takes a lot of time and computing resources. Analytical technologies allow to create models that significantly increase efficiency of decisions. Among the classical approaches to data analysis in practice the most deterministic technologies and probabilistic technologies became common. Deterministic analytical technologies such as the Pythagorean theorem used by man for many centuries. During this time it was created a huge number of formulas, theorems and algorithms for solving classical tasks - determination of volumes, finding solutions of systems of linear equations, finding the roots of polynomials. Complex and effective methods of analysis have been developed problems of optimal control, solution of differential equations, etc. To apply the algorithm, it is necessary that the task is complete described by a certain deterministic model (some set of known

functions and parameters). In this case, the algorithm gives the exact answer. Example, to apply the Pythagorean theorem you need to check that the triangle - rectangular. In practice, observation tasks are common random variables, for example, the problem of forecasting the share price. For such problems cannot be constructed by deterministic models, therefore a fundamentally different, probabilistic approach is used. Parameters of probabilistic models are distributions of random variables, their average values, dispersions, etc. As a rule, these parameters are unknown in advance, but for their evaluation statistical methods used in the fixed samples are used values (historical data). Such methods suggest that some probabilistic model is known tasks. For example, in the problem of forecasting the course we can assume that tomorrow's share price depends only on the price for the last 2 days (autoregressive model). If this is true, then observe the course throughout several months allow to estimate rather precisely coefficients of this dependence and predict the course in the future. Unfortunately, classical methods are ineffective for many practical tasks. This is due to the fact that the impossible is complete enough describe reality using a small number of model parameters, or calculating the model requires a lot of time and computing resources. In particular, consider the problems that arise when finding a solution to the problem optimal distribution of investments.

- in the real problem, none of the functions is known exactly, only known approximate or expected values of profit. To get rid of from uncertainty, we are forced to lock functions, losing at this in the accuracy of the description of the problem;
- a deterministic algorithm for finding the optimal solution can be applicable only if all given functions are linear. IN real business tasks this condition is not met. Although these functions can be approximated by linear, the solution in this case will be far from the optimal;
- if one of the functions is nonlinear, then the simplex method is unsuitable, and there are two traditional ways to find a solution to this problem. First path - use the gradient descent method to search maximum profit. In the case where the function definition area profit has a complex form, and the function itself - several local maxima, the gradient method can lead to suboptimal decision. The second way is to

conduct a full list of investment options. If each of the 10 functions is set to 100 points, you will have to check about 1020 options, which will require at least a few months of work modern computer.

Probabilistic technologies also have significant disadvantages solving practical problems. We illustrated the work of the probabilistic approach on the example of a simple linear autoregressive model, but the dependence that occur in practice, often nonlinear. Even if there is a simple one dependence, its appearance is unknown in advance. If we want to consider for forecasting the share price of several interrelated factors (for example, number of transactions, dollar exchange rate, etc.), you will have to resort to construction multidimensional statistical model. However, such models either suggest Gaussian distribution of observations (which is not performed in practice) or not substantiated theoretically. In multidimensional statistics for lack of the best often use unsubstantiated heuristic methods, which are inherent very close to neural network technology. In recent years, there has been a rapid development of analytical systems new type. They are based on artificial intelligence technologies that mimic natural ones processes, such as the activity of brain neurons or the process of natural selection. When developing modern analytical technologies, their ability is taken into account:

- understanding the task, the overall process and knowledge of the capabilities of other systems and people involved in the interaction;
- communication with users through understanding of natural language, drawings, images, and signs;
- knowledge based on common sense; coordinating decision-making, planning and action;
- learning from previous experiences and adaptation of behavior

There is an understanding of these capabilities in people and their implementation in program development central in the creation of the latest analytical technologies capable of acquiring and use knowledge. From the growth of capacity to conduct information analysis, decision making, flexible design and production depends on national competitiveness. When adding intelligence to computer systems removes many limitations in the solution real tasks.

2.7.The concept of Data Mining

Most organizations accumulate huge amounts during their activities volumes of data, but the main thing they want to get from them is useful information. How to find out from the data what is more profitable for customers organizations how to allocate resources efficiently or how to minimize losses? The latest technologies are designed to solve these problems mining analysis used to find models and relationships hidden in the data environment - models that can not be found by conventional methods. A model, like a map, is an abstract representation of reality. Map can point the way from the airport to the house, but it can't show an accident that created a traffic jam, or repair work that is currently underway and require a detour. As long as the model does not correspond to existing real relationship, it is impossible to get a favorable result. There are two types models: predictive and descriptive. The former use one set of data from known results for building models that explicitly predict results for other sets, and the second describes the dependencies in existing data, which in turn are used to make decisions or actions. Of course, a company that has long been on the market and knows its customers enjoys many models. Intelligence technology can not only confirm these empirical observations, but also find new, previously unknown models. At first, this may give the user only a slight advantage, but if you combine it for each product and each customer, gives a large separation from those who do not use such technologies. On the other hand, for using methods of intellectual analysis can find a model that will lead to a radical improvement in the financial and market situation companies. The term Data Mining got its name from two concepts: the search for valuable information in a large database and mining. Both processes require either sifting a huge amount of "raw" material, or reasonable research and search for useful values. Most often, Data Mining is translated as data mining, extraction information, data excavation, data mining, search tools regularities, extraction of knowledge, analysis of patterns, "extraction of grains of knowledge from the mountains data ", excavation of knowledge in databases, information penetration of data, "washing" the data. The concept of "knowledge discovery in databases" (Knowledge

Discovery in Databases (KDD) can be considered synonymous with Data Mining. The concept of Data Mining first appeared in 1978 and gained high popularity in modern interpretation since about the first half of the 1990s years. To date, data processing and analysis has been carried out within the application statistics, while mainly solving the problem of processing small databases. The concept of modern Data Mining technology is the basis templates (patterns) that reflect fragments of multifaceted relationships in data. These templates are patterns inherent in subsets of data that can be compactly expressed in a human-readable form. Search for templates is carried out by methods that are not limited by a priori assumptions about the structure of the sample and the type of distribution of values of the analyzed indicators. An important provision of Data Mining - the non-triviality of the wanted templates. This means that the found templates should reflect non-obvious, unexpected regularities in the data, such as the so-called hidden Knowledge. The society came to understand that poor data (raw Data) contain a deep layer of knowledge, when properly excavated which can be found real nuggets. In general, Data Mining technology quite accurately Gregory Piatetsky-Shapiro is one of the founders of this directly: "Data Mining is a process of detecting" raw "data before unknown, non-trivial, practically useful and accessible interpretation of knowledge, necessary for decision-making in various spheres of human activity ".The essence and purpose of Data Mining technology can be described as follows: it technology that is designed to search large amounts of non-obvious data, objective and useful in practice patterns. Unobvious - this means that the found patterns are not detected by standard processing methods information or expertly. Objective - this means that detected regularities will fully correspond to reality, in contrast to the expert thought, which is always subjective. Practically useful - it means that conclusions have a specific meaning that can be found practical application. Data Mining is the process of identifying, researching and modeling large volumes of data to identify previously unknown structures (patterns) for the purpose Achieving Business Advantage (SAS Institute Definition). Data Mining is a process that aims to identify new significant correlations, samples and trends resulting from the screening of large amounts of data that stored using sample recognition techniques plus application of statistical and mathematical methods

(Gartner Group definition). Data Mining is a multidisciplinary field that has emerged and developed on the basis of such sciences as applied statistics, pattern recognition, artificial intelligence, database theory and more. Consider the essence of some disciplines at the junction of which appeared Data Mining technology:

- statistics - is the science of methods of data collection, processing and analysis for identification of patterns inherent in the phenomenon under study. Statistics are available a set of methods for planning an experiment, data collection, presentation and generalization, as well as analysis and conclusions based on these data. It operates on data obtained from observations or experiments;
- machine learning can be described as a process of obtaining program of new knowledge. Mitchell in 1996 gave the following definition: "Machine learning is a science that studies computer algorithms that automatically improve at work.
- artificial intelligence - a scientific field in which and tasks of hardware or software modeling of species are solved human activities that are traditionally considered intellectual. Term intelligence comes from the Latin intellectus, meaning mind, human mental abilities. Accordingly, artificial intelligence (AI, Artificial Intelligence) is interpreted as the property of automatic systems to take over certain functions of human intelligence. Artificial intelligence is called a property intelligent systems perform creative functions that are traditionally considered human prerogative.

The concept of Data Mining is also closely related to database technologies and the concept of data. Research on the historical aspect of this problem allows identify the main points associated with the emergence of intelligent systems data analysis. So in 1968 the first industrial was put into operation DBMS IMS system from IBM. Already in 1975 the first standard appeared Association for Data Processing Languages - Conference on Data System Languages (CODASYL), which defined a number of fundamental concepts in the theory of database systems data that are still fundamental to the network data model. Further development of database theory is associated with the American name mathematics E.F. Kodda, who is the creator of the relational data model. During the 80's years, many researchers have experimented with a new approach in areas structuring

databases and providing access to them. The purpose of this search was to obtain relational prototypes for easier data modeling. In as a result, in 1985 a language called SQL was created. For today day almost all DBMS provide this interface. About this time specific types of data appeared - "graphic image", "document", "sound", "map", data types for time, time intervals, double-byte character strings character representations have been added to the SQL language. The results of all these efforts made possible the emergence of Data Mining technology, data warehouses, multimedia databases and web-databases. To successfully carry out the process of finding new knowledge a prerequisite is the availability of a data warehouse. A data warehouse is a subject-oriented, integrated, time-bound, constant collection of data for supporting the decision-making process. Subject orientation means that data grouped and stored according to the areas they describe, and not to the applications that use them. Integration means data meet the requirements of the whole enterprise, not one function of the business. Hereby, the data warehouse ensures that the same reports are generated for different ones analysts, will contain the same results. Binding to time means that the repository can be considered as a set of "historical" data, ie it is possible restore the picture at any time. The attribute of time is always obvious present in data warehouse structures. Invariability means getting hit once in the repository, the data is stored there and does not change. In the data warehouse only added. For the organization and operation of information storage specialized software is created that provides effective user interaction. The key opportunity to use the latest technologies has become a huge drop in prices over the last few years on storage devices information from tens of dollars for storing megabytes of information, up to tens cents. This has significantly reduced the cost of collection and increased storage of large amounts of information. Falling prices for processors with simultaneous increasing their speed contributed to the development of technologies related to processing huge arrays of information. As a result, many were overcome barriers encountered in finding new knowledge in storage information. Client-server architecture is also a necessary attribute data mining technologies. This approach provides opportunities perform the most time consuming data processing procedures on high-performance server for both project developers and users. On on the same server can be stored, and at the request of clients,

executed corporate projects. Data mining technologies are an important part of the market modern information technologies. Gartner Group Agency analysis of information technology markets, in the 1980s introduced the term "Business Intelligence" (BI), business intelligence or business intelligence. This term proposed to describe various concepts and methods that improve business decisions through the use of decision support systems. In 1996 year, the agency clarified the definition of this term: "Business Intelligence - software operating within the enterprise and providing functions access and analysis of information stored in the data warehouse, and what ensure the adoption of correct and sound management decisions. The concept of BI combines various tools and technologies for data analysis and processing scale of the enterprise. On the basis of these means BI-systems, the purpose are created which - to improve the quality of information for management decisions. BI systems are also known as Decision Support Systems (DSS, DSS, Decision Support System). These systems convert data into information on on the basis of which decisions can be made, ie that supports decision-making. Gartner Group defines the composition of the Business Intelligence systems market as a set software products of the following classes:

- Data Warehousing;
- operational analytical processing systems;
- information and analytical systems;
- Data Mining tools;
- Query and Reporting tools.

Gartner's classification is based on the method of functional tasks, where software products of each class perform a specific set of functions or operations with using special technologies. Here are some brief quotes from the most influential members of business organizations who are experts in this relatively new technology. Guide with Purchasing Data Mining Products (Enterprise Data Mining Buying Guide) Aberdeen Group: "Data Mining is a useful mining technology information from databases. However, due to significant differences between tools, experience and financial condition of product suppliers, businesses need to carefully evaluate prospective developers Data Mining and Partners. To make the most of power Commercial level data mining tools, the company

must choose clear and convert data, sometimes integrating information retrieved from external sources, and set up a special environment for Data Mining algorithms. Data Mining results largely depend on the level data preparation, not from the "great features" of any algorithm or set of algorithms. About 75% of the work on Data Mining is data collection, which is carried out before the tools themselves are launched. Illiterate by using some tools, the company can waste their own potential and sometimes millions of dollars".Opinion of Herb Edelstein, a world-renowned expert in Data Mining, Data Warehousing and CRM: "A recent study by Two Crows has shown that Data Mining is still in its infancy. Many organizations are interested in this technology, but only a few are active implement such projects. We managed to find out another important point: the process of implementing Data Mining in practice is more complex than expected. IT teams have become fascinated with the myth that Data Mining tools are simple in use. It is assumed that it is enough to run such a tool on terabyte database, and useful information will appear instantly. Actually, a successful Data Mining project requires an understanding of the nature of the activity, knowledge of the data and tools as well as the data analysis process".Care must be taken before using Data Mining technology analyze its problems, limitations and critical issues related to it, and also understand what this technology cannot. In particular, Data Mining cannot replace the analyst. Also, technology cannot answer the questions that were not specified. It cannot replace the analyst, it only gives him a powerful tool to facilitate and improve its work. Because it is given technology is a multidisciplinary field for application development, including Data Mining, it is necessary to involve specialists from different fields, and also ensure their quality interaction. Different Data Mining tools have different degrees of "friendliness" interface and require certain user skills. Therefore software the software must correspond to the level of training of the user. The use of Data Mining must be inextricably linked to promotion user qualifications. However, experts in Data Mining, who would be well versed in business, so far little. Successful analysis requires high-quality data processing. According to the statement analysts and database users, the reprocessing process can take up to 80% percent of the entire Data Mining process. So that technology works for itself, it will take a lot of effort and time to go to the preliminary data analysis of choice model and its adjustment.

With Data Mining you can find really valuable ones information that will soon give big dividends in the form of financial and competitive advantage. However, Data Mining often does a lot erroneous discoveries that do not make sense. Many experts say that Data Mining tools can produce huge numbers statistically unreliable results. To avoid this, a check is needed adequacy of the obtained models on test data. It should also be noted that a quality Data Mining program can cost money expensive enough for the company. An option is to buy ready-made solutions with pre-verification of its use, such as a demo with a small sample of data. Data Mining Tools, Unlike statistical, theoretically do not require a strictly defined number retrospective data. This feature can cause detection unreliable, erroneous models and, as a result, adoption based on them wrong decisions. Statistical significance should be monitored discovered knowledge. Data Mining has quite significant differences from other methods of analysis data. Traditional methods of data analysis (statistical methods) and OLAP mainly focused on testing pre-formulated hypotheses (verification-driven Data Mining) and on "rough" intelligence analysis that is the basis of operational analytical data processing (Online Analytical Processing, OLAP), while one of the main provisions of Data Mining - search unobvious patterns. Data Mining tools can find the following regularities independently and also independently build hypotheses about relationships. Because the very formulation of the hypothesis of dependencies is the most difficult task, the advantage of Data Mining over others methods of analysis is obvious. Most statistical methods for detection relationships in the data use the concept of sampling averaging, which leads to operations on non-existent quantities, while Data Mining operates with real values. OLAP is more suitable for understanding retrospective data, Data Mining relies on retrospective data for getting answers to questions about the future. The potential of Data Mining gives the "green light" to expand boundaries application of technology. Regarding the prospects of Data Mining, the following are possible directions of development:

- selection of types of subject areas with corresponding heuristics, formalization of which will facilitate the solution of relevant data mining problems that belong to these areas;

- creation of formal languages and logical means by which it will be formalized reasoning and automation of which will be a tool solving Data Mining problems in specific visual areas;
- creation of Data Mining methods that can not only extract from the data regularities, but also to form some theories based on empirical data;
- overcoming the significant backlog of tools Data Mining from theoretical advances in this field.

If you look at the future of Data Mining in the short term, it is obvious that the development of this technology is most focused on areas business related. Data Mining products can be the same common and necessary, like e-mail, and, for example, used by users to find the lowest prices for a particular product or the cheapest tickets. In the long run, the future of Data Mining is real fascinating - it can be a search by intelligent agents as a new species treatment of various diseases and a new understanding of the nature of the universe. However, Data Mining also poses a potential danger - it is growing the amount of information becomes available through the world wide web, including information of a private nature, and more and more knowledge can be extracted from it. Not so long ago the largest online store "Amazon" was in the center scandal over his patent "Methods and systems of assistance users when buying goods ", which is nothing more than another product Data Mining, designed to collect personal information about visitors store. The new technique allows you to predict future requests based on facts of purchases, as well as draw conclusions about their purpose. The purpose of this methods - what was said above - getting as much as possible customer information, including private (status, age, benefits, etc.). Thus, data on the privacy of buyers are collected shop, as well as members of their families, including children. The latter is prohibited legislation of many countries - the collection of information about minors is possible there only with parental permission. Research notes that there are both successful solutions that use Data Mining, and a bad experience with this technology. Areas where the use of Data Mining technology is likely will be successful, have the following features:

- require knowledge-based solutions;
- have a changing environment;

- have available, sufficient and relevant data;
- provide high dividends from the right decisions.

The scope of Data Mining is not limited - it is everywhere any data. But first of all Data Mining methods today, to put it mildly, intrigued commercial enterprises that deploy projects based on Data Warehousing. There are many experiences enterprises shows that the benefits of using Data Mining can reach 1000%. For example, there are reports of economic effects, 10-70 times which exceeded the initial costs from 350 to 750 thousand dollars; about the project in 20 million USD, which paid off in just 4 months. Another example is the annual savings of 700 thousand. USD due to the introduction of Data Mining in the network of supermarkets in Great Britain. Data Mining is of great value to executives and analysts in their daily activities. Business people realized that for using Data Mining techniques they can get tangible benefits in competition.

2.8.Conclusion

Today we are witnessing the active development of intelligent technologies data analysis (Data Mining), the emergence of which is associated primarily with the need for analytical processing of large amounts of information that accumulate in modern databases. Most companies accumulate under huge amounts of data during their time, but the main thing they want from to get them is useful information. How can you find out from the data that is more profitable for the company's customers how to allocate resources efficiently or how to minimize losses? To solve these problems and designed the latest mining technologies used to find models and relationships hidden in the data environment - models that can not be found by conventional methods.

CHAPTER 3
STAGES AND METHODS OF FINDING NEW KNOWLEDGE

3.1.The overview

It is important to understand that building a Data Mining model is part of a larger process that includes all stages starting with defining the basic problem that the model will solve, to deployment of the model in the working environment. This process can be specified using the following six basic steps. Creating a data mining model is available dynamic iterative process. After reviewing the data, the user can find that the existing data is not enough to create the necessary models data mining, which, accordingly, leads to the need for search additional data. You can develop several models and understand that they are not solve the formulated problem. Therefore, a change in characteristics is required task. You may need to update already deployed models account of new data received. Therefore, it is important to understand that Creating a data mining model is a process and that is every step of the way such a process can be repeated as many times as necessary for creating an effective model. Consider in more detail each of these stages: Problem definition. In order to make full use of everything the benefits of intelligent technology need to clearly present the goal future analysis. The construction of the model is carried out depending on the purpose. If it is necessary to increase the profit of the trading organization, then for the purposes of: "increase sales "and" increase the effectiveness of advertising "must be built different models. At the same stage, methods for evaluating the results are determined future project and possible costs for its implementation. Data preparation and review. This is the longest possible stage occupy from 50% to 85% of the time the whole process of finding new knowledge. On at this stage it is necessary to determine the sources of data. It can be data collected by the organization itself or external data from the public sources (weather information or census) or private sources (various archival data, databases of notary offices, etc.). Data evaluation. When building a model you need to remember one thing rule concerning the correctness of the input data: "If the

input of the problem "garbage" arrives, then the result will also be "garbage". Input data can be in one base or in several. Before "loading" data into storage should take into account that different data sources may be designed for specific tasks and, accordingly, there are problems associated with data aggregation: different formats of numerical data representation (e.g. goals or material); different data encoding (for example, different date format); different ways of storing data; different units of measurement (inches and centimeters); as well as the frequency of data collection and the date of the last update. Even, if the data is in the same database, you still need to pay attention to missed values and values of unrealistic magnitude, so-called "emissions". The analyst must always know how, where and under what conditions the data is collected, and be make sure that all the data used for the analysis is measured in the same way. Data aggregation and purification. At this stage, construction is underway data warehouses for further processing, ie filling storage or adding to it the data selected in the previous stages. In this one at the same time there is a cleaning, ie correction of all detected errors. There are various aspects to data cleansing. They are all aimed at finding and correction of errors made at the stage of information collection. Error in data can be considered: missed value, impossible event (incorrect typed value - "emission"). Correction is based on common sense, use of rules or with the involvement of an expert well versed in subject area. The record in the database that contains the error must be corrected or, in controversial cases, excluded from further consideration. After verifying the data, they are converted and formatted accordingly evaluation results. This is done for greater convenience of observation data. Discrete event data is converted to a specially designed or standard form, which reflects the time and description of events. If users are easy will understand this form, they will be able to quickly learn the events that took place based on the construction of this form. It may seem that this step duplicates the collection phase data, but in fact these are two completely different stages. On the first of them it happens data collection to speed up machine processing of information without loss of quality, on the second the data are reduced to a form convenient for visual control user. The person conducting the analysis can more fully imagine the input data. This is necessary for various types of reports when required briefly describe the input data used for analysis. Data selection. If the repository

is formed and the types of models that are defined will be built to solve problems, there is a selection of necessary data just for these models. This means not only reducing the number of records in base under a certain condition, but also a change in the number of fields, merging different tables in one, or, conversely, the creation of several on the basis of one table. That is, the transformation takes place in "three dimensions": by the number of records, by number of fields and structure. Data conversion. Serves to enrich the resulting base, ie adding different relationships based on existing fields "not just" price "and "quantity", and their product - "total amount", not debt and income, and the ratio of debt to income), adding intervals by the number of the month you can put the number quarter, and the percentage of implementation of the plan can be supplemented by characteristics "good", "satisfactory"), adding critical values (maximum, average, minimum). Building a model is an iterative process, that is, you need to build a series models to find the one that best meets your goals. Models can be divided into two groups. Once the model type is defined, the construction algorithm must be selected models or technology of finding knowledge. The essence of the process of building a controlled model is to finding dependencies on one piece of data ("learning model") and checking these dependencies on the rest of the data (accuracy assessment). Model is considered built if the cycle of "training" and inspections is completed. If the accuracy of the model during the next iterations does not improve, it says so on the completion of the model. Because "training" and test data are in the same database, there is often a need for a third set data - control, which is selected from such data that do not intersect with "training" and test. It is needed for independent evaluation accuracy model. Typically, all three datasets belong to a set of data, necessary for the implementation of a particular project. The most well-known test method is called simple estimation. IN in this case, the division of data into two sets is random. The ratio of the amount of test data to the amount of data on which it occurs construction of the model should be in the range from 5% to 33%. After building the model, it is used to predict the values on the test set. To a lesser extent accuracy of the model is considered the ratio of the number of successful results to the total the number of examples in the test set (you can use a variable such as degree of inaccuracy equal to 1 - "measure of accuracy"). If a not very large base is used to build the model data, the so-

called cross-estimation of accuracy is applied. In this In this case, the data are randomly divided into two approximately equal parts. After this model will be based on one of them and the other is used for determination of accuracy. Then parts of the database change roles. Received two independent estimates of accuracy are combined (as an arithmetic mean or other way) to best assess the accuracy of the model built on the whole base. For even smaller databases, several thousand records use cross-estimation accuracy. In this case, the base is divided by n approximately equal disparate groups. Then the first of these groups becomes a test set, and the others groups are united, and on their basis the model is built. Received the model is used to predict the values for the test set and thus the first value of accuracy is obtained. Similarly acquire all n independent values of accuracy. The middle of them is the accuracy of the whole models. Another method is used to find accuracy in small ones databases. In this case, the model will be based on the data of the whole bases. After that, a set is randomly created from the database records test kits (at least 200, and sometimes even more than 1000). One record can be present in various test kits. For any of them is determined precision. The average of these is the accuracy of the whole model. Once the construction of the model is complete, you can adjust the model, using other parameters or even change the algorithm to build the model, because you can never say which algorithm, which technology to find knowledge will give better results. One cannot be sure that a certain technology will work best. It is often necessary to build a large number models and evaluate each to find the best one. In addition, for different models require different data preparation and the inevitable repetition of steps. All that increases the time of finding the best model, so it is necessary to apply parallel computing technologies.

3.2.Basic models of intelligent computing

Consider the main types of models used for finding new knowledge based on information storage data. The purpose Intelligent technology is finding new knowledge that the user can continue to use to improve the results of their activities. Result modeling

is the identification of relationships in data. Neural networks are systems with architecture that conditionally simulate the work of neurons. The mathematical model of a neuron is somewhat universal nonlinear element with the ability to widely change and customize it characteristics. Neural networks are a set of interconnected layers neurons that receive input data, process them and generate on output result. Between the nodes of the visible input and output layers can find a certain number of hidden layers. Neural networks are implemented opaque process. This means that the built model usually does not have clear interpretation. Many packages implement neural algorithms networks used in the processing of commercial information, when pattern recognition, handwriting interpretation, interpretation cardiogram. Hardware or software implementations of neural network algorithms called a neurocomputer.

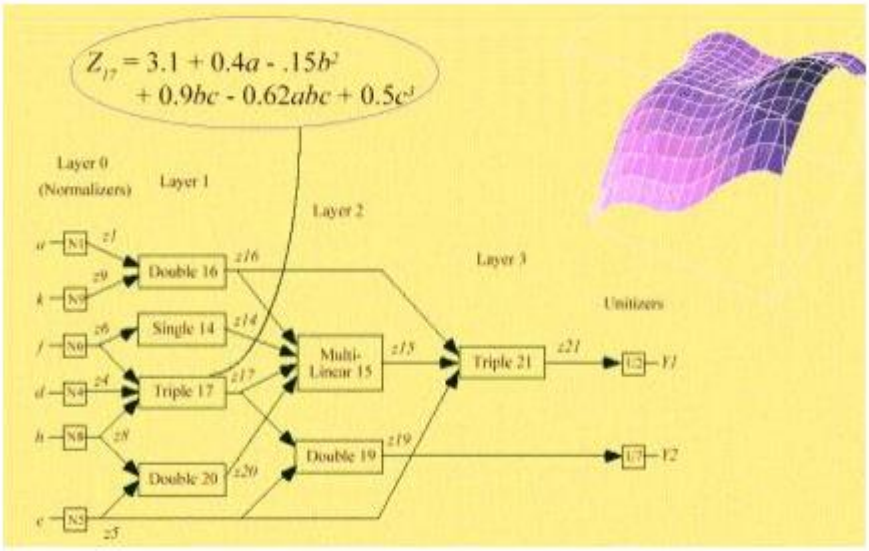

Fig.3.1. Polynomial neural network

Decision trees are a method widely used in finance and business numerical forecast problems are more common. As a result of application of this method, a hierarchical structure is created for the training data sample rules of classification of the type, "IF ... THEN ...", having the form of a tree. For, to decide to which class to assign an object or situation, it is necessary answer the question that stands in the nodes of this tree, starting

with it root. Or of the form "The value of the variable B belongs to the subset of signs C?". So, in the end, you can get to one of the end nodes, where defined object class. This method guarantees a substantive representation of the rules and his easy to understand. Today there is a rise in interest in products that apply decision trees. This is mainly due to the fact that most commercial problems are solved by them faster than by algorithms neutron networks, they are simpler and clearer to users. At the same time it cannot be said that decision trees always work flawlessly: for certain data types they may be unacceptable.

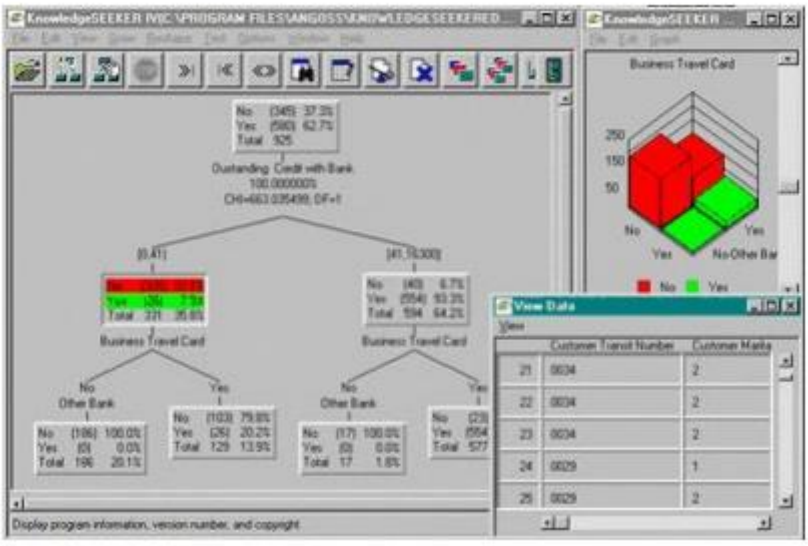

Fig.3.2. The system processes banking information

The fact is that the individual nodes on each branch is given less number of data records - the tree can segment data into a large number individual cases. The more such individual cases, the less educational examples fall into each such case, and their classification becomes less reliable. If the tree is very "branched" - consists of unjustifiably a large number of small branches - it will not give statistically sound answers. As practice shows, most systems use decision trees, this problem does not find a satisfactory solution. Reflection systems based on similar cases. The idea of the algorithm simple. To predict the future or choose the right one solutions, systems find in the past

close analogues of the current situation and choose the same answer that was right for them. Therefore, this method is still called the "nearest neighbor" method. Reflection systems based on similar cases give good results in various tasks. Chief of them the downside is that they don't create any models or rules at all, summarizing previous experience. In choosing a solution, they are based on the whole array of available historical data, so it is impossible to say on the basis what specific factors do these systems build their responses. Algorithms for detecting associations find rules for individual subjects, which appear together in one transaction, for example in one purchase. Sequence is also an association, but it depends on time. The association is recorded as A-> B, where A is called a prerequisite, B - a consequence. Frequency of occurrence of each individual item or group of items is determined very simply - the number of occurrences of this item in all events (purchases) is calculated and divided by the total number of events. This value is measured as a percentage and is called "prevalence". Low prevalence (less than one thousand percent) indicates the insignificance of the association. To determine the importance of each resulting associative rule it is necessary to obtain a value called "trust A to B" (relationship A and B). This value shows how often with the appearance of A appears B and calculated as the ratio of the frequency of occurrence (prevalence) of A and B together to prevalence of A. That is, if the confidence in A to B is 20%, it means that when buying product A in every fifth case and buy product B. If the prevalence of A is not equal to the prevalence of B, then the confidence of A to B is not equal trust B to A. In fact, buying a computer often leads to buying floppy disk than buying a floppy disk before buying a computer. Another important characteristic of the association is the strength of the association. The greater the power, the stronger the effect that the appearance of A has on the appearance of B. Power is calculated by the formula: (trust A to B) / (prevalence B). Some association search algorithms first sort the data and only then this determines the relationship and prevalence. The only difference is such algorithms are the speed or efficiency of finding associations. This is important in due to the huge number of combinations that need to be sorted out for finding more meaningful rules. Association search algorithms can create your own databases of prevalence, trust and power to which can be requested upon request. For example: "Find all associations in which for goods X

with a confidence of more than 50% and a prevalence of at least 2.5% " finding sequences adds a time variable that allows you to work with a series of events to find successive associations over a period of time.

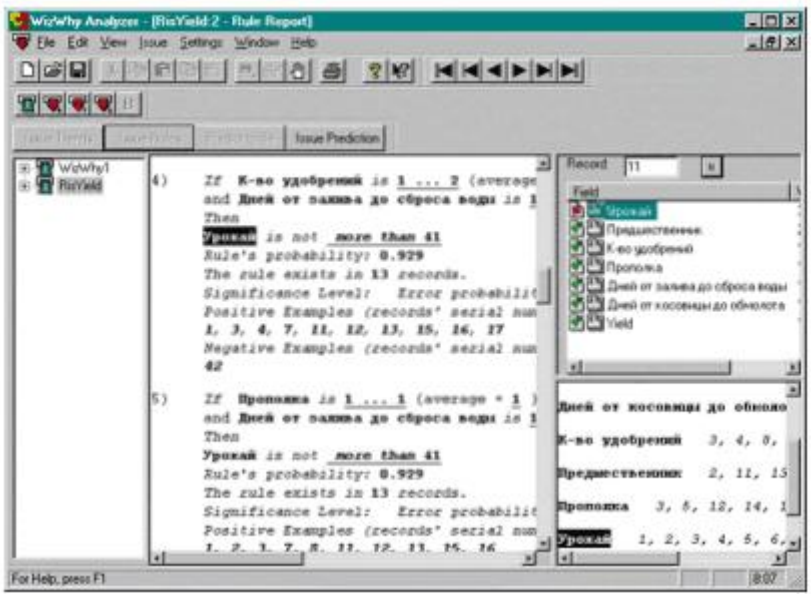

Fig.3.3. The system has found rules that explain low yields some crops

Summing up this method of analysis, we must say that by chance there may be a situation when the goods in the supermarket will be grouped by found models, but this, instead of expected profit, will have the opposite effect. This may be due to the fact that the customer is long will not go to the store in search of the desired product, while buying more something that catches his eye and something he never planned to buy. Fuzzy logic is applied to data sets where affiliation data to any group is a probability in the range from 0 to 1. Clear logic manipulates results that can be either true or false. Fuzzy logic is applied in cases where there is a "can be" in addition to "yes" or "no".In Japan, this area is experiencing a boom. The program of the organization is to create closer to human computing devices. LIFE unites 48 companies, including: Hitachi, Mitsubishi, NEC, Sharp, Sony, Honda, Mazda, Toyota. From foreign LIFE participants It is possible to allocate: IBM, Fuji Xerox, to activity of LIFE also shows interest NASA. Power and

intuitive simplicity of fuzzy logic as a methodology problem solving ensures its successful use in embedded systems control and analysis of information. At the same time there is a human connection intuition and experience of the operator. Unlike traditional math, which requires at each step of modeling accurate and unambiguous formulations regularities, fuzzy logic offers a completely different level of thinking, due to which the creative modeling process takes place at a high level abstractions, in which only a minimal set of patterns is postulated. Genetic algorithms are a powerful tool for solving various combinatorial problems and optimization problems. However, genetic algorithms are now included in the standard toolkit of intelligent methods calculations. This method is so named because it mimics the process to some extent natural selection in nature. Let us need to find solutions to problems, the most optimal in terms of some criterion, where each solution is fully described a certain set of numbers or quantities of non-numerical nature. Let's say if we it is necessary to choose a set of a fixed number of market parameters, which is essential affect its dynamics, it will be a set of names of these parameters. About this set can be said as a set of chromosomes that determine the qualities of the individual - this solution to the problem. Values of defining parameters solutions are called genes. The search for the optimal solution is similar on the evolution of the population of individuals represented by sets of chromosomes. There are three mechanisms in evolution: first, the selection of the strongest - sets of chromosomes that correspond to the most optimal solutions; second, crossbreeding - the production of new individuals by mixing chromosome sets of selected individuals; and, thirdly, mutations are random changes genes in some individuals in the population. As a result of generational change is produced the solution of the problem, which can no longer be further improved. Genetic algorithms have two weaknesses. First, the staging task does not allow to analyze the statistical significance of the obtained with their help solutions and, secondly, to effectively formulate tasks, determine the criterion for the selection of chromosomes by force only a specialist. Because of these factors, genetic algorithms should be considered rather as a tool research than a data analysis tool for practical application in business and finance. Evolutionary programming is the youngest area of intelligence calculations. Hypotheses about the type of dependence of the target variable on other variables

formulated by the system in the form of programs in some internal language programming. If it is a universal language, then theoretically it is possible express the dependence of any kind. The process of building such programs is built as an evolution in the world of programs (this method is a bit like genetic algorithms). If the system finds a program that accurately expresses the dependency, which is sought, she begins to make small modifications to it and selects among the child programs built in this way are those that increase accuracy. The system "grows" several genetic lines of programs that compete between itself in the accuracy of finding the desired dependence. Special broadcast module, translates the found dependencies from the internal language of the system on user-friendly language (mathematical formulas, tables, etc.), making them achievable. In order to make the results clearer for non-mathematical user, there is a large arsenal of various tools visualization of identified dependencies. The search for the dependence of target variables on others is carried out in the form functions of any particular kind. For example, in one of the most successful algorithms of this type - the method of group accounting of arguments (MSUA) dependence is sought in the form of polynomials. And complex polynomials are replaced by a few simple ones that take into account only some features (groups arguments). Paired combinations of features are usually used. This method does not have big advantages in comparison with neural networks with ready a set of standard nonlinear functions, but, the dependence formulas are obtained, c principle, amenable to analysis and interpretation (although in practice it is still difficult). Data visualization programs in a sense are not a means of analysis information because they only present it to the user. But, visual representation of, say, four variables clearly summarizes huge amounts of data.

3.3.SAS Enterprise miner software

It is an integrated component of the SAS system designed specifically for discovery in huge data sets of information necessary for adoption decisions. Designed to search and analyze the deep hidden regularities in SAS data, Enterprise Miner includes statistical

methods analysis, the relevant methodology for implementing Data Mining (SEMMA) projects and graphical user interface. An important feature of SAS Enterprise Miner is its full integration with the software product SAS Warehouse Administrator, designed for the development and operation of information repositories, and others components of the SAS system. Data Mining projects can be developed both locally and in the client-server architecture. The package supports the implementation of all necessary procedures within a single integrated solution with the ability to work together and comes as distributed client-server application, which is especially convenient to implement data analysis on the scale of large organizations. SAS Enterprise Miner package designed for data analysts, marketing analysts, marketers, risk analysis specialists, fraud detection specialists, as well as engineers and scientists responsible for making key decisions in business or research activities.

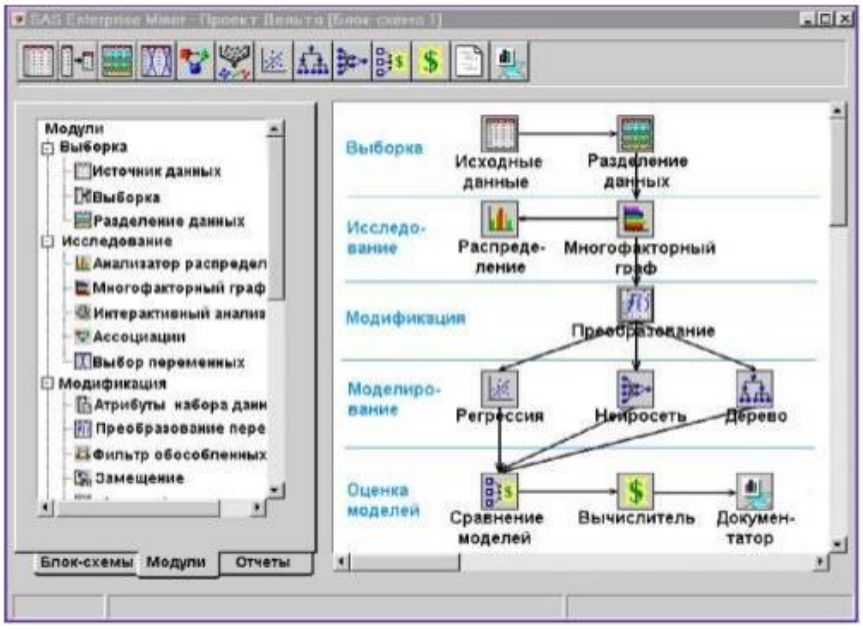

Fig.3.4. SAS Enterprise Miner main screen

A data source links SAS Enterprise Miner to an existing analysis table. To specify a data source, you need to define a SAS library and know the name of the table that you will link to SAS Enterprise Miner. I followed such steps to specify a data source.1.Select File

√New √Data Source from the main menu. The Data Source Wizard –Step 1. It tells SAS Enterprise Miner where to look for initial metadata values. The default and typical choice is the SAS table that I link to in the next step. Select Next to use a SAS table (the common choice) as the source for the metadata.

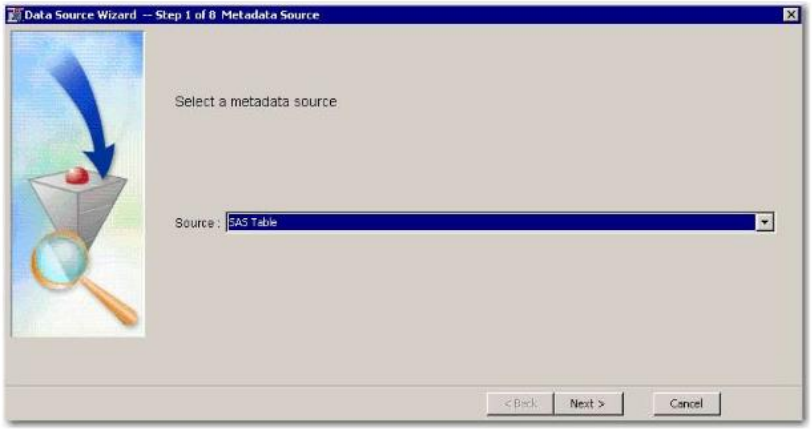

Fig.3.5 Step 1

The Data Source Wizard continues to Step 2. Select a SAS Table. In this step, I selected the SAS table that I want to make available to SAS Enterprise Miner. I typed the library name and SAS table name .

Fig.3.6 Step 2

3 step. Select Browseto choose a SAS table from the libraries that are visible to the SAS Foundation Server. The Select a SAS Table window appears. 4.One of the libraries listed is named AAEM, which is the library name defined in the Library Wizard. Double-click Aaem. The panel on the right is shows the contents of the library. 5.Select the Pva97nk SAS table. 6.Select OK. The Selected a SAS Table window closes and the selected table appears in the Table field.

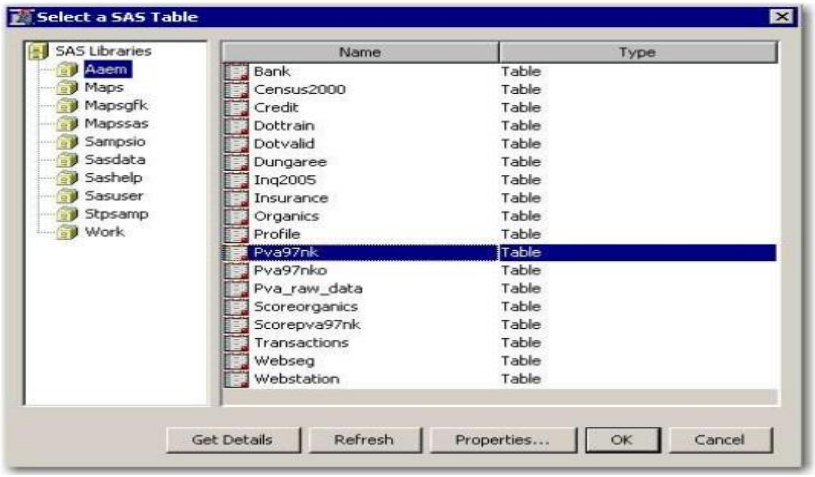

Fig.3.7 Step 3-6

7.Select Next. The Data Source Wizard proceeds to Step3 Table Information. This step of the Data Source Wizard provides basic information about the selected table. The SAS table PVA97NK is used to demonstrate the predictive modeling tools of SAS Enterprise Miner. The table contains 9,686 cases and 28 variables.

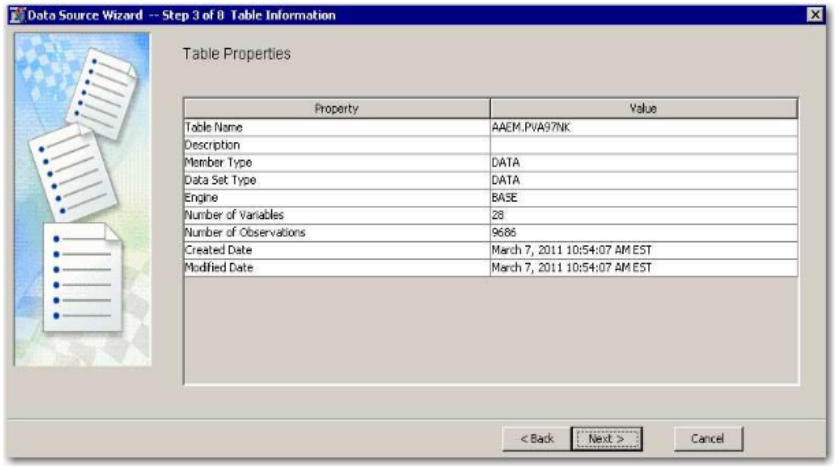

Fig.3.8 Step 7

3.4.Conclusion

The essence and purpose of Data Mining technologies can be described as follows: this technologies that are designed to search large amounts of non-obvious data, objective and useful in practice patterns. Scope of application Data Mining is not limited by anything - it is everywhere where there is any data. But first of all Data Mining methods today intrigued companies deploying projects based on modern information technologies. The experience of many such companies shows that the return on the use of Data Mining can reach 1000%. Data Mining technologies are needed first specialists who make important decisions - managers, analysts, experts, consultants. The company's income is largely determined by quality these decisions - the accuracy of forecasts, the optimality of the chosen strategies. And from the quality of these decisions depends on the development of the company. It should also be noted that for real business and production problems there are no clear solution algorithms. Ago managers and experts solve such problems only on a personal basis experience. Often, classic techniques are ineffective for many practical tasks, because it is impossible to accurately describe the reality of using a small number of model parameters, or the calculation of the model takes a lot of time and computing resources. Analytical technologies allow create models that significantly increase the effectiveness of solutions. This system draws its conclusions from the data

already accumulated bank in the process of working in the retail lending market. Thus in During the implementation process, the system is configured exactly for the data set for which targeted specific bank. In other words, the dm-Score system is ready work with the data that is available and does not require fixation on any a specific rigidly set questionnaire. In the process of analyzing data on borrowers and loans apply various mathematical methods that are found in them factors and their combinations that affect the creditworthiness of borrowers, and the strength of their influence. The identified dependencies form the basis for decision-making in the corresponding block. The analysis unit should be used periodically for analysis of new data of the bank (new borrowers come, current ones payments) to ensure the relevance of the system and the adequacy of the approved her decisions.

CONCLUSION

The use of data mining technologies, as a rule, is based on the processing of large amounts of information accumulated in modern data warehouses. There are different concepts, technologies and practicals approaches to the construction of such repositories. Also developed quite powerful technology of their use - OLAP - a technology that has a variety architecture, features and wide practical possibilities. In the activities of many companies often have to decide tasks, the formulation of which is informal, and the solution is ambiguous. Even if a strict algorithmic approach is impossible, and to get an accurate solution in principle it is impossible, there are other effective ways of the decision. An important place these include neurocomputer technology and neural networks. IN materials of the third section consider the essence, architecture, practice of construction and software tools for the implementation of neurocomputer technology, as well as Hopfield and Kohonen neural networks and modern practice and promising directions of their application. Evolutionary theory has proven its effectiveness as a solution complex formalizable problems of clustering, associative search, and when solving time-consuming problems of optimization, approximation, intellectual data processing. Concepts of evolutionary computation include genetic algorithms, genetic programming, evolutionary strategies and evolutionary programming. Evolutionary technologies of intellectual analysis today successfully used to address a number of large and economically significant tasks in business and other important projects. An important direction in the development of data mining is widespread application of fuzzy computing theory, which in the modern world is seen as a consortium of computational methodologies that collectively provide the basis for understanding, designing and developing intellectual systems, in particular data mining systems. Presented materials on fuzzy method software and modern application practice such methods.

REFERENCES

1. C. Baek, T. Doleck, “Educational data mining: a bibliometric analysis of an emerging field,” *IEEE Access*, vol. 10, pp. 31289–31296, 2022, DOI: 10.1109/ACCESS.2022.3160457.

2. O. Chernyak, P. Zakharchenko, “Data Mining,” *IEEE Access,* vol.837, pp. 6-237, 2010, DOI: 49.2206/ACCESS.2010.7754191.

3. M. Kantardzic, “Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition,” *IEEE Access,* vol.534, pp. 510-528, 2011, DOI: 10.1002/9781118029145.

4. K. Tsiptsis, A. Chorianopoulos, “Data Mining Techniques in CRM: Inside Customer Segmentation,” *IEEE Access,* vol.357, pp. 333-348, 2010, DOI: 10.1002/9780470685815.

5. R. Kothari, “Wiley Encyclopedia of Engineering,” *IEEE Access,* vol.507, pp. 432-448, 2006, DOI: 10.1002/9780471740360.

6. P. Ponniah, “Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals,” *IEEE Access,* vol.509, pp. 493-495, 2001, DOI: 10.1002/0471221627.