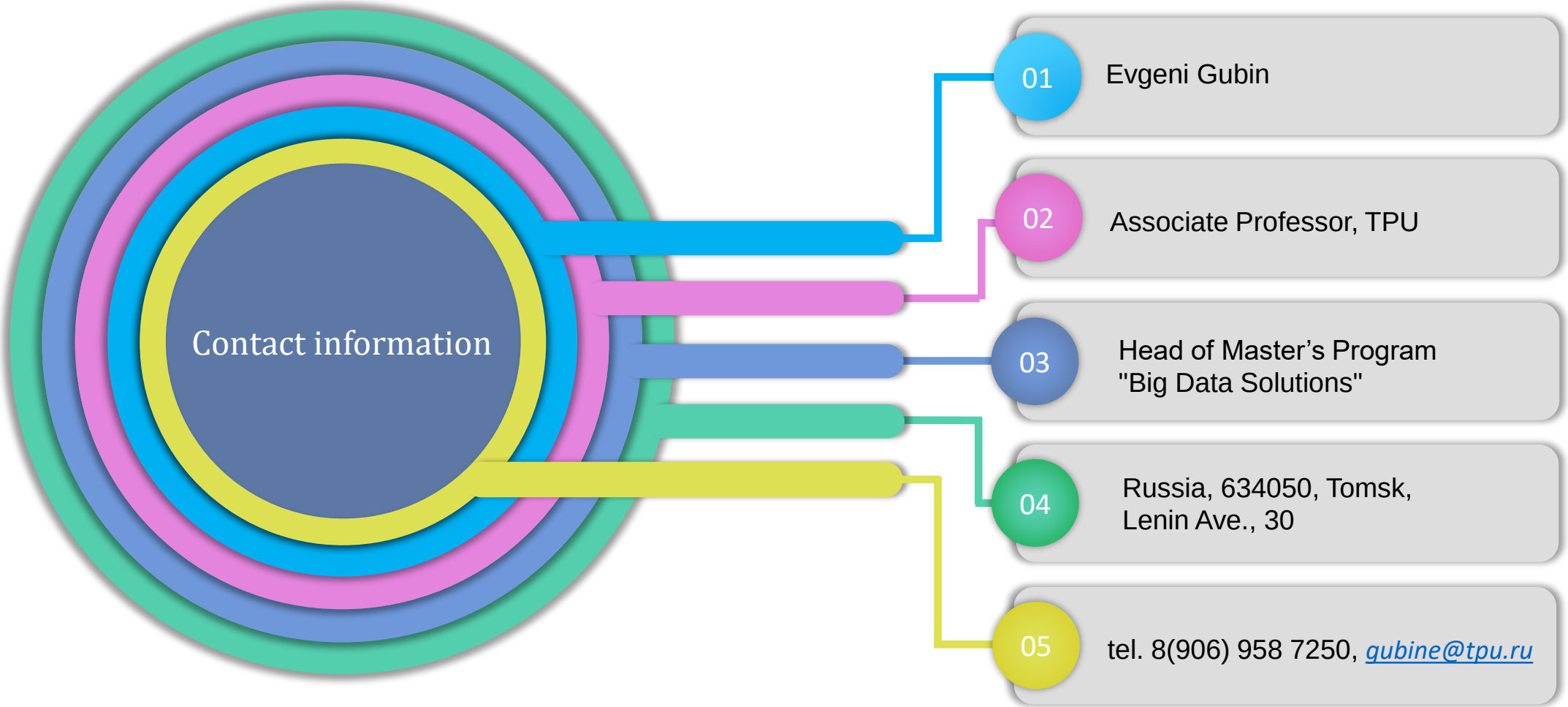
ТОМСКИЙ
ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ

reparation
ysis

Data

Gubin E., Ph.D

2019


Contact information
01
Evgeni Gubin
02
Associate Professor, TPU
03
Head of Master's Program
"Big Data Solutions"
04
Russia, 634050, Tomsk,
Lenin Ave., 30
05
tel. 8(906) 958 7250, gubine@tpu.ru

Purpose of course: *formation of basic skills of students in the preparation of initial data in the field of large amounts of data and use modern statistical instruments for data analysis (for example, the use of SAS technologies).*

- **About the course:** the process of collecting and preparing initial data is one of the most hard and complex stages in the analysis of large amounts of data, which sometimes takes up to 80% of the time. The use of statistical techniques and modern software can significantly reduce the time and cost at this stage and improve the efficiency and quality of the final results.
- **Audience:** Data Analysts and Data Engineers responsible for the processes of data analysis, collection, preparation and cleaning.
- **Required skills:** basic knowledge in programming, higher mathematics, statistics.

01

Credit risk modeling. Design and Application. Elizabeth Mays, Editor. Glen lake Publishing Company Ltd. 1998

02

Introduction to Scorecard for Model Builder. Fair Isaac Corporation. 2008 – 40pp.

03

Applying Data Mining Techniques Using Enterprise Miner Course Notes/ SAS Institute Inc. Cary, NC 27513, USA, 2003

04

The Little SAS Book: A Primer Second Edition. Cary, NC: SAS Institute

Links

- https://habr.com/ru/company/sas/blog/348168/ — SAS уроки
- https://drive.google.com/file/d/1j56iO9cRZL4OehOM8MqWpuPbxOdxWryF/view — Книга по SAS
- https://video.sas.com/detail/videos/how-to-tutorials — Обучение SAS Base

1. Introduction to Big Data

- Three sources and three components of Big Data
- Who works with Big Data
- Sources of Big data
- Data Mining and Methodology SEMMA and GRISP-DM
- Data format for Big Data

2. Descriptive statistics

- Definition of business goals and develop
- Explanatory variables and formulate the target function
- In addition, interactive statistical analysis of deviations is performed using data visualization

3. Data preparation of initial data

- Missing data
- Mistakes of data
- Outliers of data
- Duplicate cases(rows)
- Multicollinearity in the original data
- Digitalization of data

4. Partition initial data

- The formation of *training*, *validation* and *test samples*
- Basic principles of test sample formation (ratio of 60:20:20)

5. Predictive models for data analysis

- The concept of regression models and their strengths and weaknesses. Decision trees and their characteristic
- SAS code for log regression
- The results of the computational experiment to assess the accuracy of the predictive model

6. Final projects

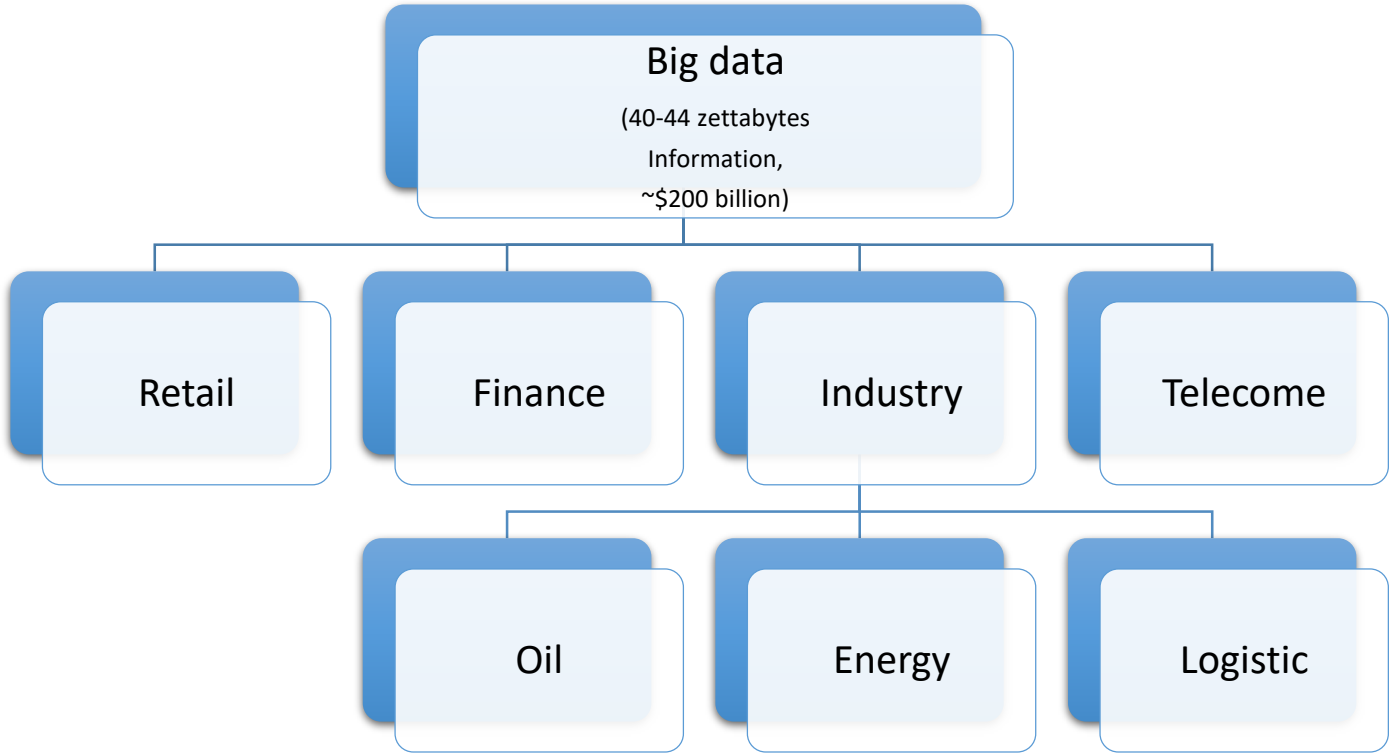
- Realization a full cycle of cleaning and preparing data on the example of the selected data set
- Formation of learning and test sample
- Conclusion

Big Data-three sources and three components – “V”

1. **Volume** — величина физического объёма
2. **Velocity** — скорость прироста и необходимости быстрой обработки данных для получения результатов.
3. **Variety** — возможность одновременно обрабатывать различные типы данных

Big Data – who works with Big data

Big Data is a horizontally scalable system that uses a set of techniques and
technologies that allow to process structured and unstructured information
and build relationships necessary to obtain uniquely interpreted human
data that has not lost relevance, and carrying the value of the goals
pursued by it



1. Retail
- *Walmart* (486 млрд.$),*Costco* (119 млрд.$),*The Kroger*(115 млрд.$) – US, X5 Retail Group (1287 млрд. Р),«Магнит» (1287 млрд. Р),«Лента»(365 млрд. Р)

2. Finance
- *HSBC* (GB, 2634 млрд. $), *BNP Paribas* (Франция, 2514 млрд. $),
- Sberbank (24 680 млрд. Р), ВТБ (24 680 млрд. Р), Газпромбанк (5 742 млрд. Р)

3. Industry
- *Exxon Mobil*(US),*Royal Dutch Shell* (GB),*British Petroleum*(GB),*General Electric*(US)
- Gazprom, Resent, Lukoil, Russia's railways

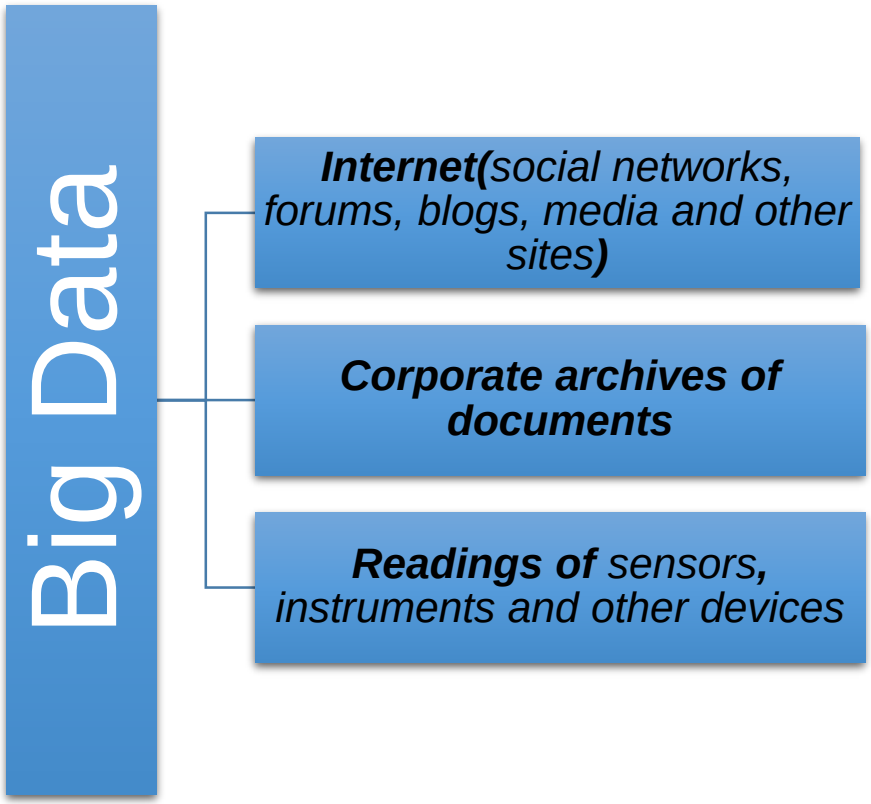
4. Telecom
- *Huawei* (China), *Google* (US), AT$T(US), *Sprint* (US), *T-Mobile* US, *Facebook*
- МТС, Мегафон, Yandex, «WhatsApp», «Одноклассники»

Structured data

- ✓**ETL** – data storage
- ✓Tables
- ✓Files
- ✓Internal data sources (CRM, ERP, etc.)
- ✓External data sources (social networks, Internet, etc.)

Unstructured data

- ✓Files without a predefined model
- ✓"Data lakes "- "the Idea of a data lake is to store initial data in its original format until it is needed."



Big Data Engineering включает проектирование и развертывание систем сбора и предварительной подготовки данных, для дальнейшего анализа

Big Data Analytics (**Scientist**) включает в себя расширенный анализ и построение прогнозной модели по исходным данным

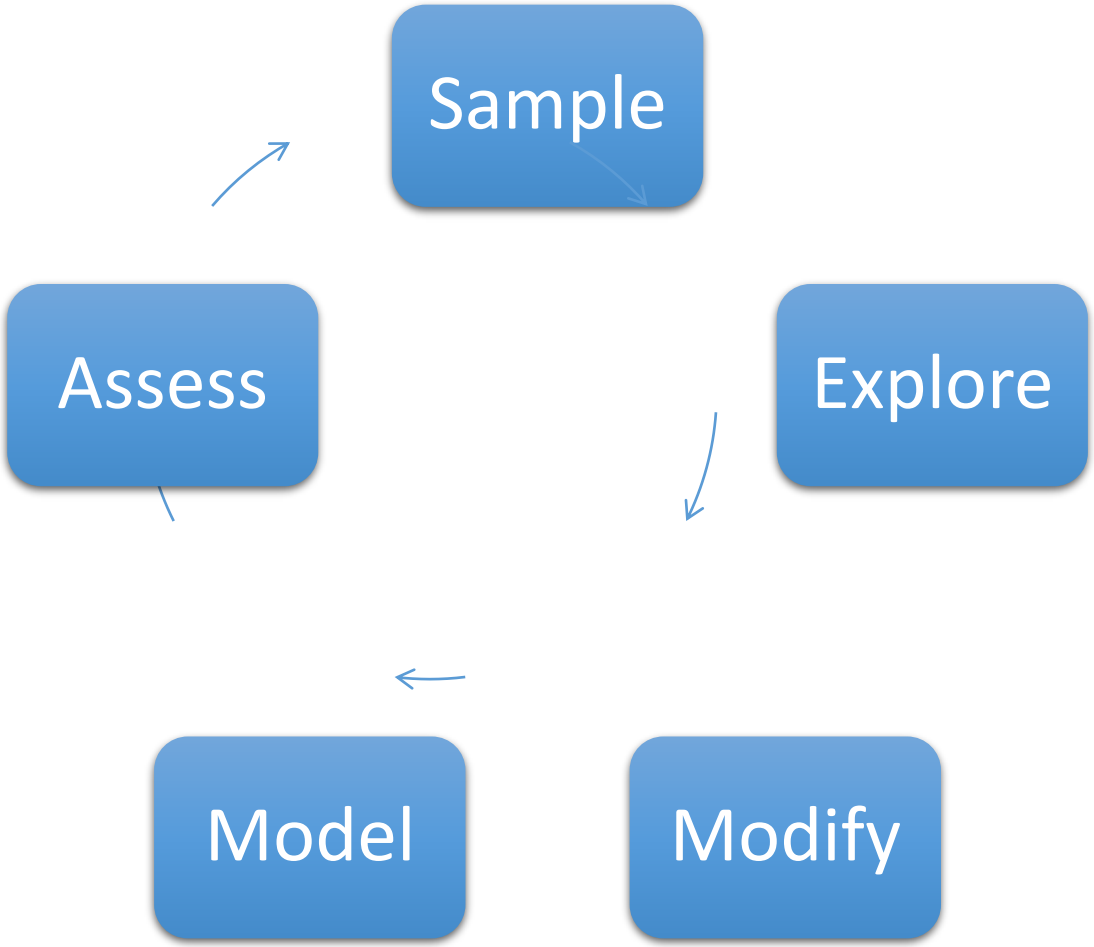
- Currently, Data Mining ("data mining") is defined as the process of finding secret mechanism in large amounts of data, which are often not structured and have a variety of formats (in the form of numbers, text, photos, etc.).

- You spend most of your data mining time preparing your data: cleaning, aggregating, transforming and modeling. Another problem is that statistical models are often built on data with a large number of observations or variables. Scalability requires careful selection and application of statistical methods.

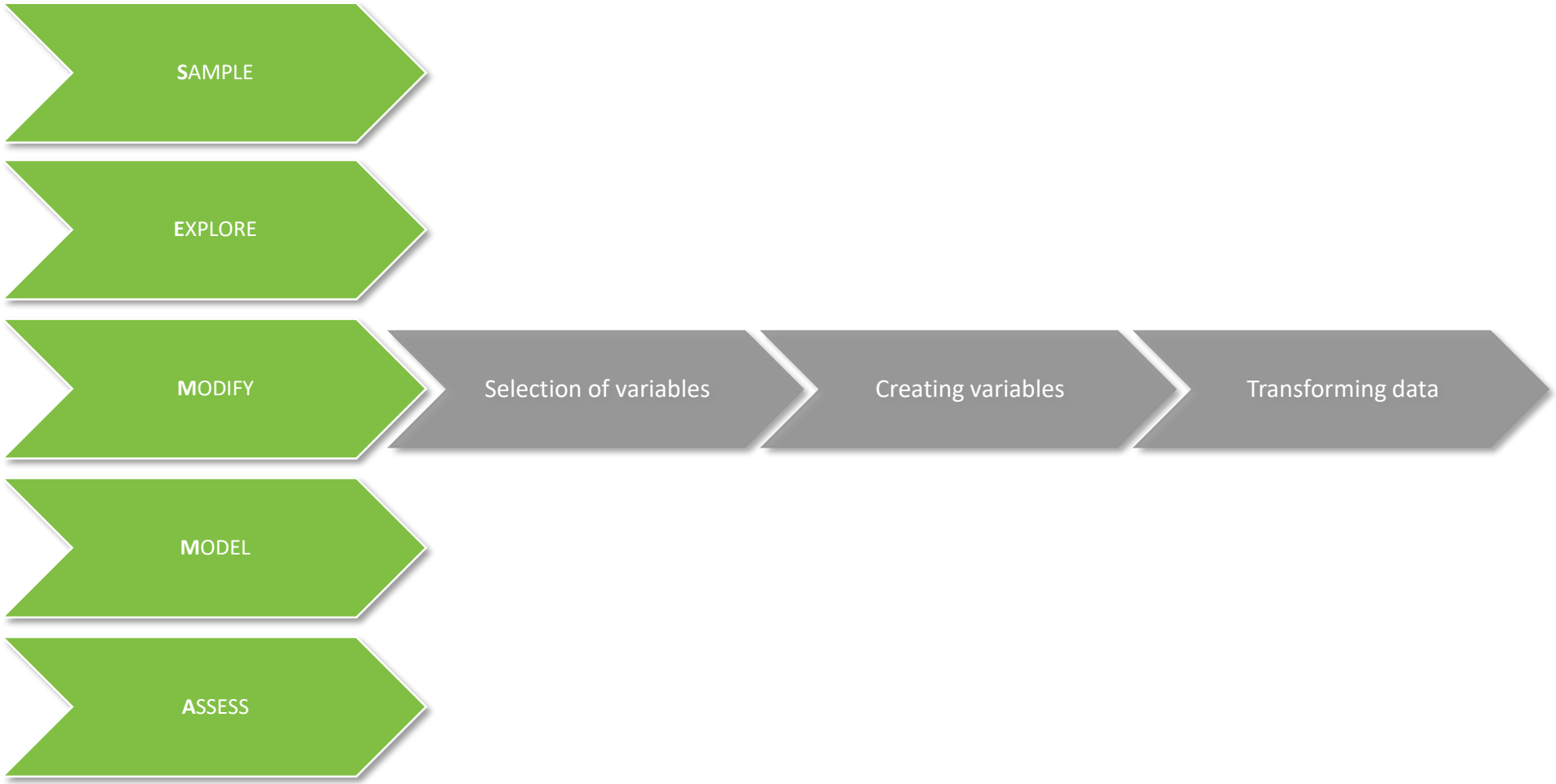
- The SAS Institute defines data mining as the process of **sampling**, **examining**, **modifying**, **modeling**, and **evaluating** (**SEMMA**) large amounts of data to identify previously unknown sample that can be used as a business advantage.

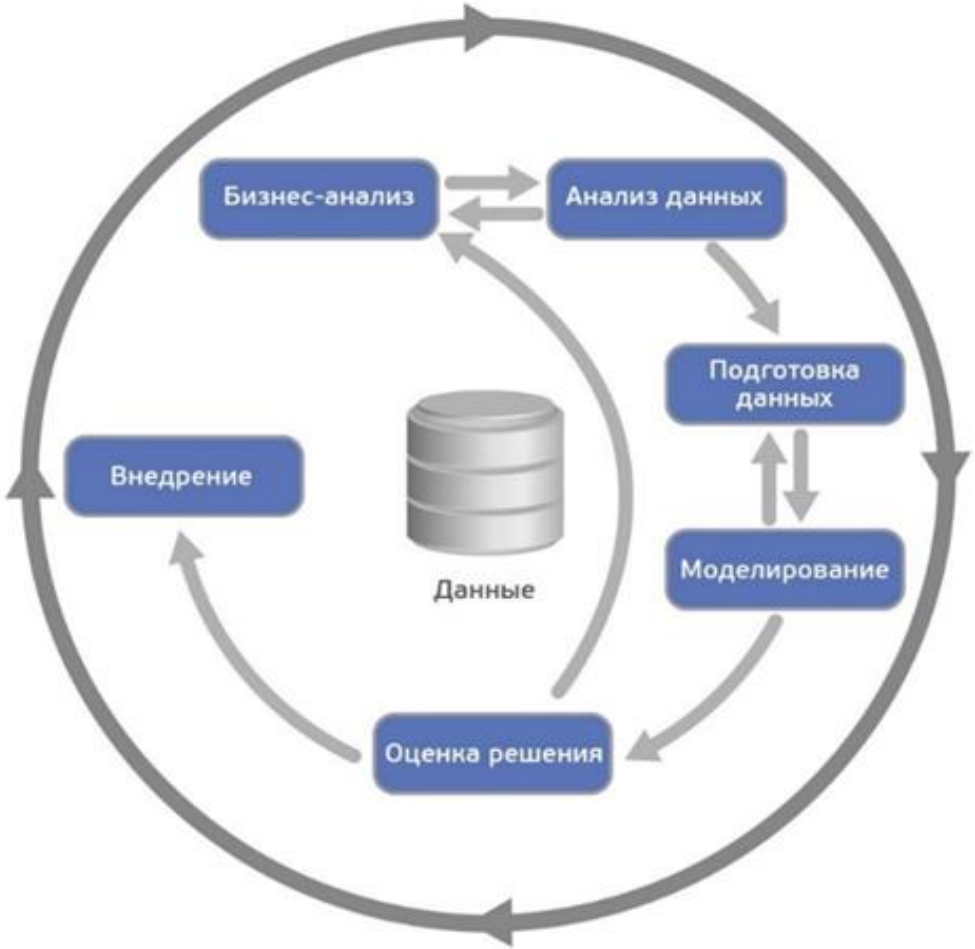
- **SAS Enterprise Miner** was designed to support the entire data mining process. SAS provides access to relational and heterogeneous data stores. The basic SAS language provides high power in data aggregation and transformation. Together, **SAS/STAT** and **Enterprise Miner** can support high practical implementation in numerical simulation of the business process under study. The functions of Enterprise Miner are organized into a well known logic algorithm called **SEMMA**:

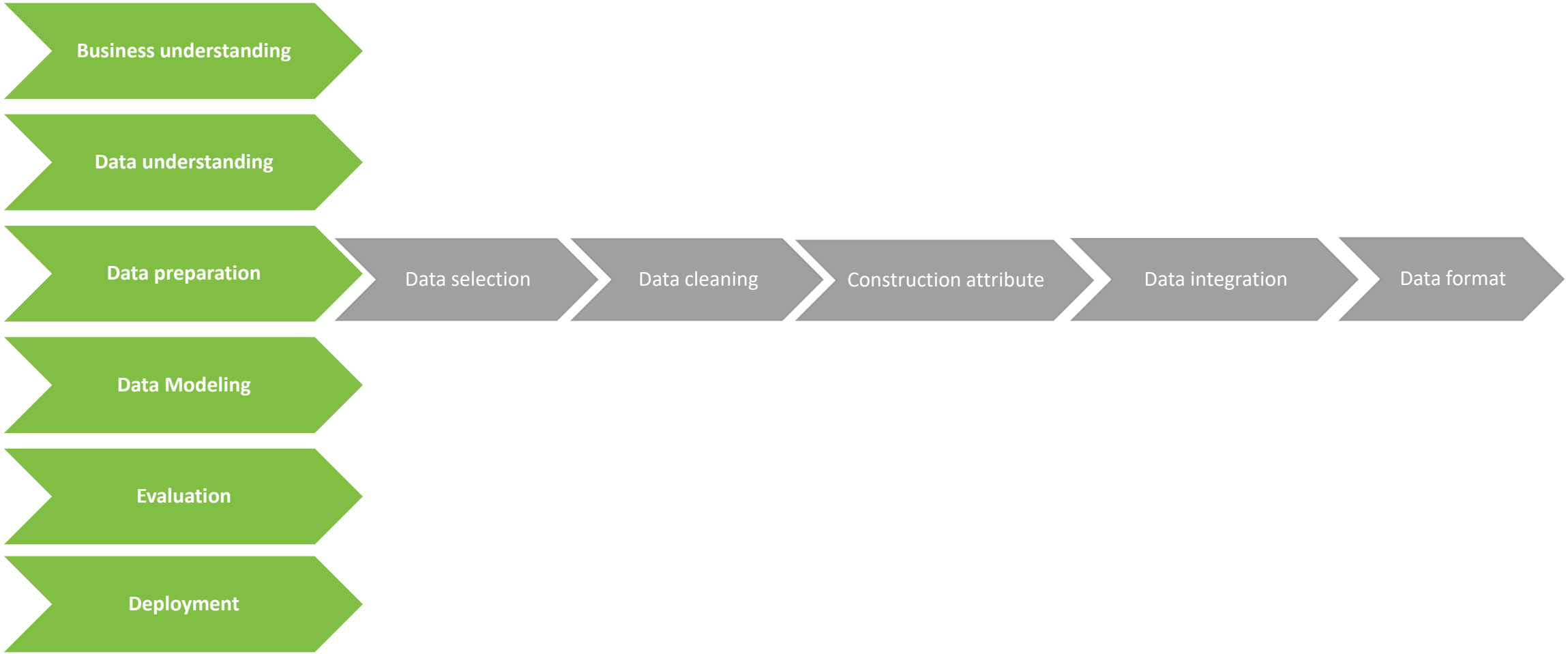
- **S**AMPLE - define input data, including sample from big data
- **E**XPLORE- exploring a dataset statistically and graphically
- **M**ODIFY- preparing data for analysis
- **M**ODEL- predictive model evaluation (regression model, decision trees, neural networks, etc.)
- **A**SSESS- comparison of competing predictive models (graphical comparison of respondents, yield graphs, etc.)

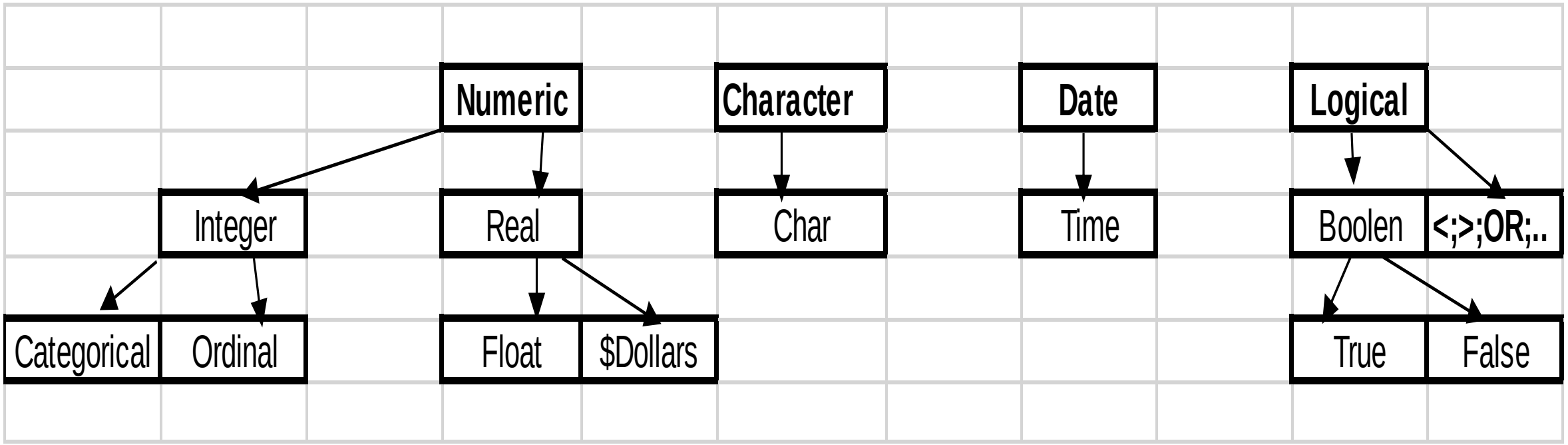

Sample
Explore
Modify
Model
Assess

Review of existing approaches. SEMMA




Бизнес-анализ
Анализ данных
Подготовка данных
Моделирование
Оценка решения
Внедрение
Данные


Business understanding
Data understanding
Data preparation
Data selection
Data cleaning
Construction attribute
Data integration
Data format
Data Modeling
Evaluation
Deployment


Numeric
Character
Date
Logical
Integer
Real
Char
Time
Boolen
<;>;OR;..
Categorical
Ordinal
Float
$Dollars
True
False

ID	Sex	DocType	Business	Education	CarStatus	AddIncome	Income	CreditSum	Age	DayOff	GB
1	Male	Driving licence	Industry	Secondary education	1	3318	16589	500000	33	90+	1
2	Female	other	Education	Secondary education	0	912	4562	100000	999	0	0
3	Female	other	Finance /Banking/Insuran	Higher education	0	919	4597	100000	30	0	0
4		other	Retail	Secondary education	0	4000	20000	280000	46	0	0
5	Male	Driving licence	Industry	Secondary education	1	12170	60850	500000	42	0	0
5	Male	Driving licence	Industry	Secondary education	1	12170	60850	500000	42	0	0
7	Male	Foreign passport	Industry	Higher education	1	728	3640	100000	25	60_90	0
8	Male	Drivine licence	Finance /Banking/Insuran	Secondary education	0	2468	12341	150000	43	0	0
9	Male	Military card	Industry	Secondary education	0	825	4123	270000	36	0	0
10	Female	other	Industry	Secondary education	0	468	2340	136000	31	0	0
11	Male	Drivine licence	Industry	Secondary education	1	430	2150		45	0	0
12	Female	Driving licence	Education	Secondary education	0	2468	12341	150000	22	0	0
13	Female	Driving licence	Industry	Secondary education	0	1170	5850	480000	31	30_90	0
14	Male	Drivine licence	other	Higher education	0	1063	5317	183000	28	0	0
15	Female	other	Education	Secondary education	0	1028	5139	214000	36	0	0
16	Female	other	Education	Secondary education	0	1025	5123	15000000	43	0	0

Example of formats of personal data of potential bank customers

Atribute	Field	Field Russian Name	Format/New	Format/Initial	Comment
ID	Identification number	Персональный номер	numeric	numerec	1234
SEX	Gender	Пол	numeric	char	1 - Male 2 - Female
DocType	Additional document	Дополнительный документ	numeric	char	1 - Military card 2 - Foreign passport 3 - Driving licence
Business	Employer scope of activity	Индустрия	numeric	char	1 - Finance / Banking / Insurance 2 - Industry 3 - Retail 4 - Education 5 - IT and Telecoms 6 - Other
Education	Education	Образование	numeric	char	1 -Graduate (Ученая степень) 2 - Higher education (Высшее) 3 - Secondary education (Среднее образование)
CarStatus	Owns car	Есть машина	numeric	char	1 - Yes 2 - No
AddIncome	Additional income	Дополопнительный доход	numeric	numeric	per month
Income	Income from mandatory work confirmed	доход по основному месту работы	numeric	numeric	per month
CreditSum	Loan amount	Сумма кредита	numeric	numerec	560000
Age	The age of the borrower	Возраст заемщика	numeric	numerec	43
DayOff	Overdue payment/days	Дни просрочки	char	char	90+/ 60_90 /30_60 / 0_30 / 0_0
GB/Target	Target function	Целевая функция	numeric	numeric	1- DayOff >= 90+ 0 - DayOff < 90

List of basic attributes of personal data

Problems of initial data set/ исходные («грязные») данные	**Format attribute/** формат переменной	**Comment/** комментарий
1.*Missing data*/ Отсутствующие данные	**Numeric**/числовой, **Char**/текст	1. **Add in** (average, median, frequency...) 2. **Delete this cases** (rows)
2.*Mistakes of data*/ Ошибки в данных	**Numeric**/числовой, **Char**/текст	1. **Add in** (average, median, frequency...) 2. **Delete** this cases (rows)
3.*Outliers of data* / Выбросы данных	**Numeric**/числовой	**Delete this cases** (rows)
4.*Duplicate cases*(rows) /Дублирующие наблюдения(строки)	**Duplicate ID** (observations)	**Remove one of the duplicate**
5.*Multicollinearity in the original data*/ Мультиколлинеарность	**Linear combination of variables** (attributes)	**Remove one of the attributes**
6.*Digitalization of data*/ Цифровизация данных	**Numeric**/числовой, **Char**/текст	**Converting to numeric format**
Selection of objective function, training and test samples/ Выбор целевой функции, обучающейся и тестовой выборок		

Selection of objective function, training and test samples/

Выбор целевой функции, обучающейся и тестовой выборок

Problems of initial data set/ исходные («грязные») данные	**Format attribute/** формат переменной	**Comment/** комментарий
Objective function/ Целевая функция	Binary (0,1)	“Bad” – 1; ”Good” – 0.
Training samples/ Обучающейся выборка	Sampling 70%-80%	Representative relative to the objective function (GB)/ Репрезентативная по GB
Testing samples/ Тестовая выборка	Sampling 30%-20%	Representative relative to the objective function (GB)/ Репрезентативная по GB

```
PROC IMPORT
   DATAFILE="C:\Users\gubine\Desktop\xls\Rezerv_11.xls"
    OUT=WORK.test1
     REPLACE
    DBMS=XLS;
   GETNAMES=YES;
RUN;
```

Before you begin to analyze the source data, you need to perform **descriptive statistics** data set for all incoming variables, including the target. This will help to give a common idea of the source data and define the input data and the target variable.

This usually includes the following statistics:

- for each **numeric** input variable (attribute - "**Numeric**") - 1. total number of rows (observations) including missing ("**N**"), 2. the minimum value ("**Min**"), 3. .the maximum value ("**Ma**x"), 4. mean square deviation ("**St.div**"), 5. average ("**Mean**"), 6. the median ("**Median**"), etc.
- for a **text** variable (attribute – "**Char**") - 1. total number of rows (observations) including missing ("**N**"), 2. the frequency of incoming parameters ("**Frequency**") and their number ("**N**").

```
/* Descriptive Statistic for Data Analysis */

PROC MEANS DATA=test1 N  MIN MAX MEAN MEDIAN STD MAXDEC=2;
     TITLE 'Descriptive Statistics for numeric';
RUN;

PROC FREQ DATA=test1;
     TITLE ' Descriptive Statistics for char';
   TABLES SEX DocType Business Education CarStatus DayOff GB/
NOCUM NOPERCENT;
RUN;
```

The algorithm of methodology of preparation of initial data for the further analysis, as a rule, begins with check of presence of *missing* ("not filled" - missing) data. The reason for the presence of missing data in the source file may be: *technical errors*, *operator errors*, *omission of information* (purposeful or random), etc.

For correct further analysis, the missing values must either *be deleted* (the entire line-observation), or replaced by a specific algorithm. To make a correct decision it is necessary to remember that the amount of initial data should representatively reflect the studied General population. This means that there must be a sufficient number of observations (rows) for the appropriate predictive accuracy. Therefore before deleting or replacing missing data it is necessary to make a statistical assessment of future actions

As a base, experts suggest using the following recommendations to "*clean*" the source attributes in the presence of missing data:

- If the number of missing data is about *5%*, then keeping representativeness, these lines with "missing" can be deleted.
- If the amount of missing data *is more than 50%*, this attribute can be removed from further analysis.
- If the number of missing data is in the *range of 5% - 50%*, then for **numerical attributes** ("**numeric**") there are several options for replacing the missing values: *mean* ("**Mean**"), *median* ("**Median**"), "**Nearest neighbors**", etc. For **text attributes** ("**char**") it is possible to use the values: *the most common* or " *nearest neighbors*"

```
proc format;
  value $missfmt ' '='Missing' other='Not Missing';
  value missfmt  . ='Missing' other='Not Missing';
run;
proc freq data=rezerv_11; /* initial data */
  format _CHAR_ $missfmt.; /* apply format for the duration of this PROC */
    tables _CHAR_ / missing missprint nocum nopercent;
    format _NUMERIC_ missfmt.;
  tables _NUMERIC_ / missing missprint nocum nopercent;
run;
```

The next step in preparing the initial data for analysis is usually to check for **errors** in the data ("**mistakes data**"). The reason for this may be**: technical error**, **operator errors**, etc.

Correct further analysis in data errors is close to the missing values algorithm:

- or delete (entire line-observation),
- or replace by a specific algorithm.

To make a correct decision it is necessary to remember that the amount of initial data should representatively reflect the studied General population. This means that there must be a sufficient number of observations (rows) for the appropriate predictive accuracy. Therefore before deleting or replacing missing(errors) data it is necessary to make a statistical assessment of future actions.

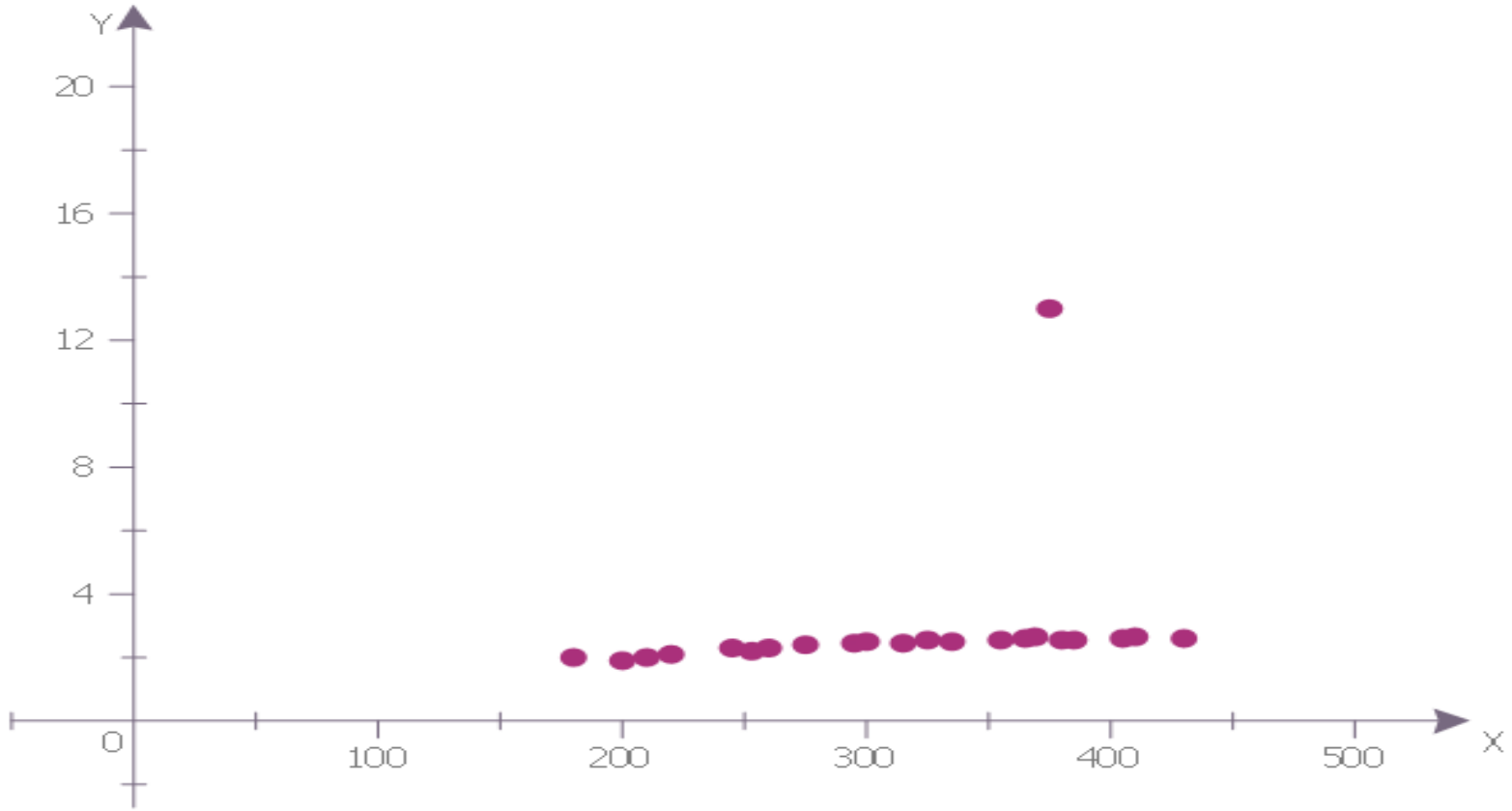
Outliers is an *unusually high* or *unusually low* value of a measured value (attribute). The concept of “Three Sigma" can be used for evaluation: if the range of the studied value goes **beyond ±3σ**, then it is possible to consider with a high degree of probability that it is an "outlier" that does not belong to the given population.

It is possible to define outliers by means of: *histograms*, *diagrams* (dot, box, scattering, etc.), *time series*.

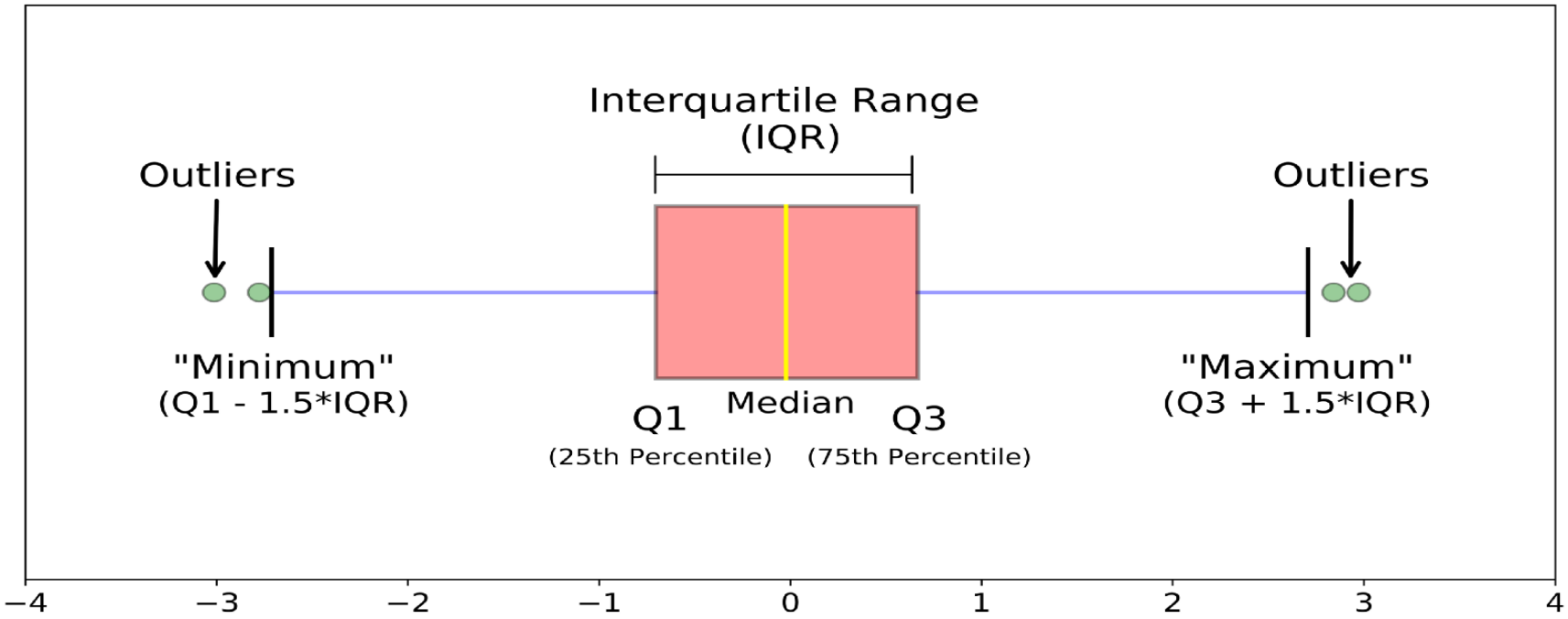
The reason of outliers: *data entry errors*, *small sampling*, or a *special factor* that requires *special study*.

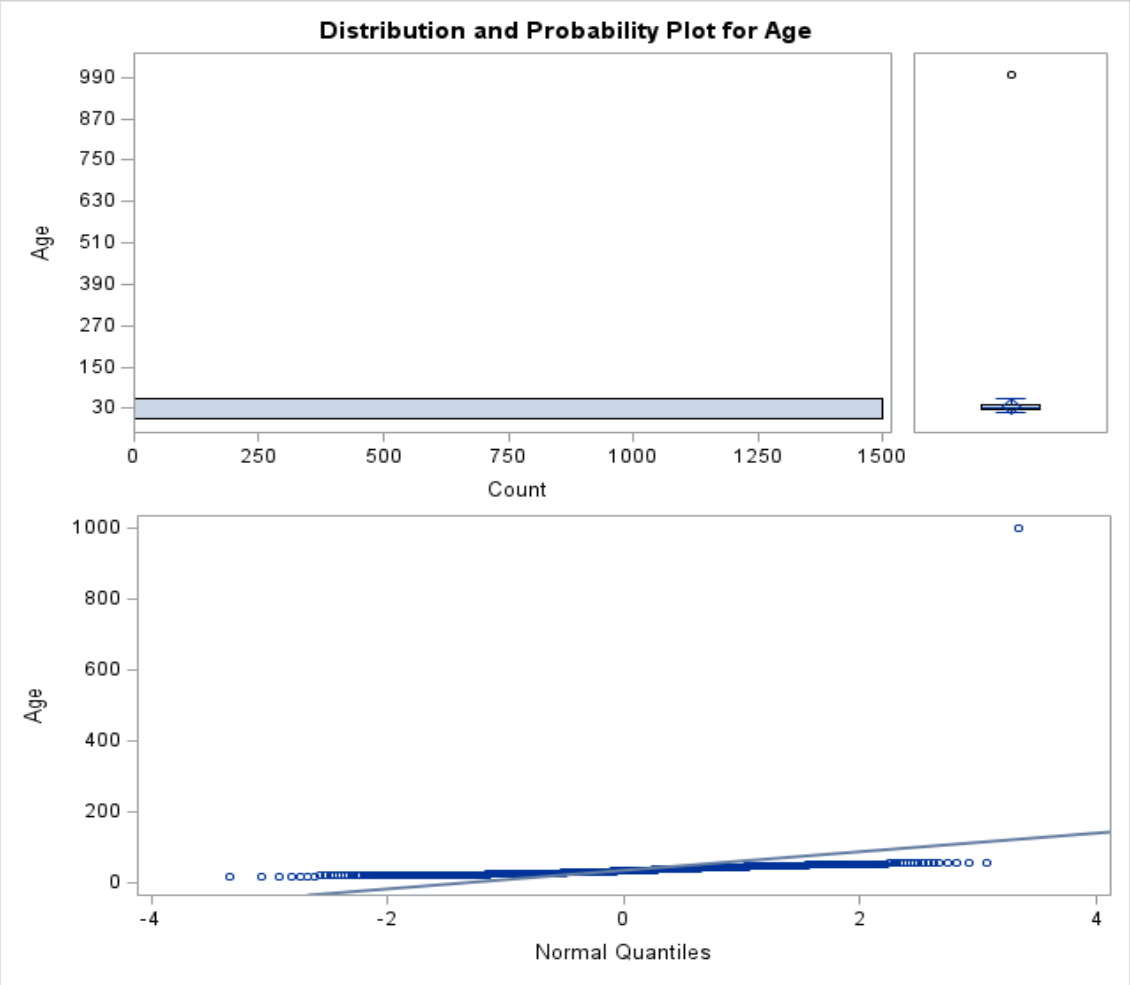
Depending on the nature of the outliers, the following steps can be taken:

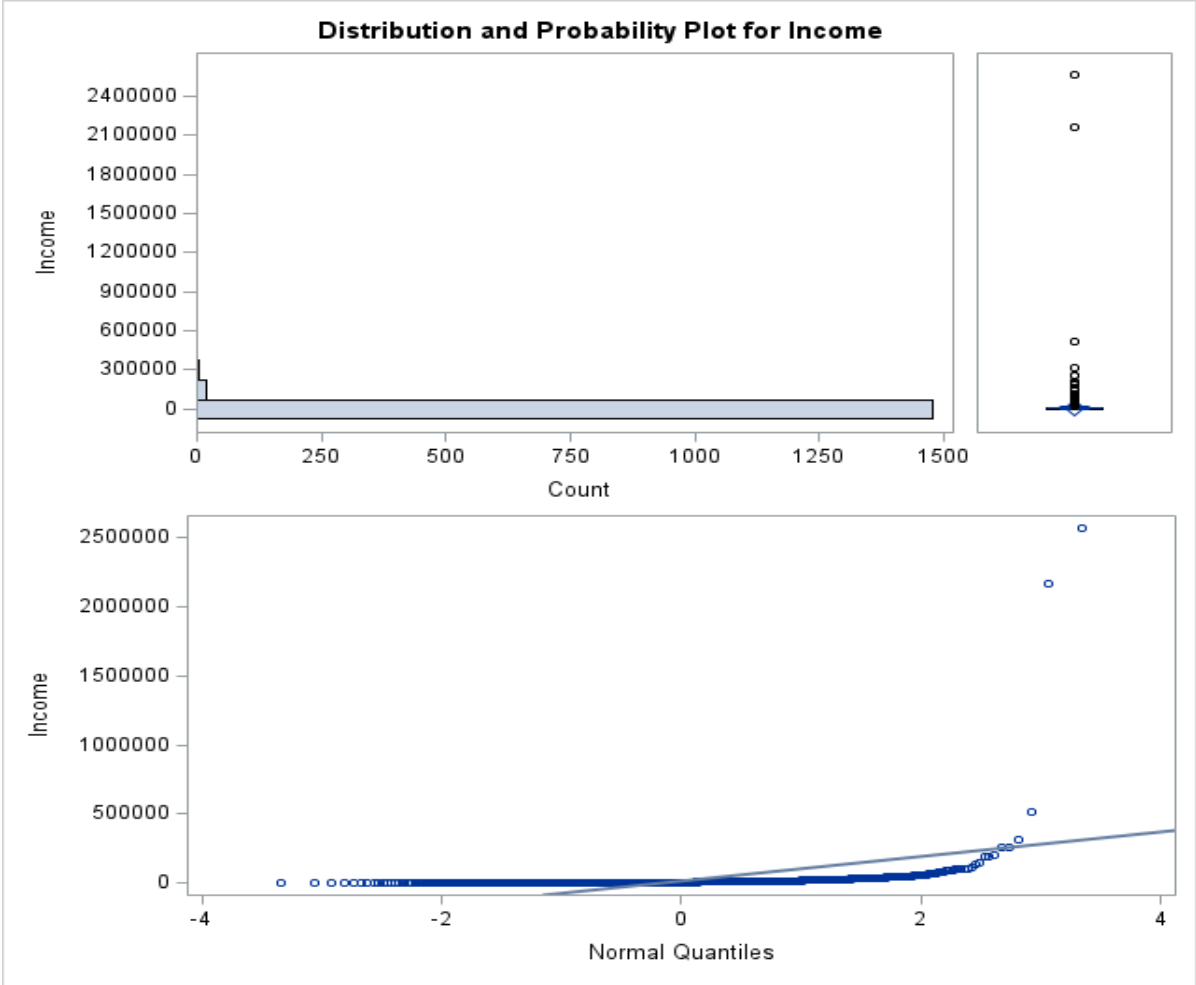
1. *exclude* from consideration; 2. *to increase* the sample size, 3. *take outliers* as a follow-up analysis.


Y
20
16
12
8
4
0
100
200
300
400
500
X

The easiest way to detect outliers is to consider all the observations behind interquartile range as outliers. This can be done visually with a box plot (whiskers plot). The box plot shows mean or median, interquartile range, minimal and maximal values and outliers. The box plot example is shown in fig. A.2.




Distribution and Probability Plot for Age
Age
990
870
750
630
510
390
270
150
30
0
250
500
750
1000
1250
1500
Count
Age
1000
800
600
400
200
0
-4
-2
0
2
4
Normal Quantiles


Distribution and Probability Plot for Income
Income
2400000
2100000
1800000
1500000
1200000
900000
600000
300000
0
0
250
500
750
1000
1250
1500
Count
Income
2500000
2000000
1500000
1000000
500000
0
-4
-2
0
2
4
Normal Quantiles

Влияние выбросов на прогнозные результаты

Временная шкала	Значения
1	2
2	4
3	4
4	6
5	7
6	9
7	10
8	9
9	12
10	10
11	11
12	12

Выброс в точке 10

Влияние выбросов на прогнозные результаты

Временная шкала	Значения
1	2
2	4
3	4
4	6
5	7
6	9
7	10
8	9
9	12
10	100
11	11
12	12

Выброс	36.35 (-11.97;84.67)	прогнозное значение
Нет_выброса	12.88 (10.46; 15.30)	прогнозное значение

Влияние отсутствующих данных (missing) на прогнозные результаты

Временная шкала	Значения
1	2
2	4
3	4
4	6
5	7
6	9
7	10
8	9
9	12
10	10
11	11
12	12

Прогноз	Граница нижняя	Граница верхняя	Исх. данные
12,83	**9,31**	**16,35**	Средняя **(8)**
12,88	**10,46**	**15,3**	Искомая **(10)**

```
proc univariate data=test1 robustscale plot; /* outliers */
 var  income age;
run;
```

Income

Extreme Observations

Lowest		Highest	
Value	**Obs**	**Value**	**Obs**
0	1492	263040	*1310*
0	1477	312750	*719*
0	1471	519500	*1031*
0	1465	2163580	*480*
0	1457	2570370	*881*

Age

Extreme Observations

Lowest		Highest	
Value	**Obs**	**Value**	**Obs**
19	1170	57.5589	*1101*
19	183	57.5808	*1427*
20	1445	59.1041	*1103*
20	1426	59.3973	*896*
20	1208	999.0000	*2*

```
proc sort data=test1/  ** not duplicate  **/
  nodupkey out=NotDuplicate; by id;
run;
```

The same observations influence the regression coefficients, increasing the variance of the model, so duplicate observations must be found and removed from the analysis

The presence of multicollenarity in the initial variables (attributes) leads to inaccuracy of the predictive model and the final result will depend heavily on different samples.

Therefore, it is difficult to estimate the effect of explanatory variables on the target function.

To assess multicollinearity is correlation matrix of attributes. If the value of the *pair correlation coefficients is higher than 0.7 – 0.8*, this indicates possible problems with the quality of the future predictive model. Therefore, it is necessary to change the original variables so that they (attributes) are not so strongly correlated.

```
proc princomp data=test1 /* correlation */
        outstat=test1_stat noprint;
run;
```

TYPE	_NAME_	ID	CarStatus	AddIncome	Income	Age	GB
MEAN		750,9993338	0,326449034	3050,689141	15253,4457	36,15862849	0,131912059
STD		433,4468372	0,469069983	17917,40377	89587,01887	26,47480476	0,338507915
N		1501	1501	1501	1501	1501	1501
CORR	ID	1	0,017117208	-0,000380217	-0,000380217	-0,058908699	-0,003125426
CORR	CarStatus	0,017117208	1	0,007317044	0,007317044	0,010531656	-0,120234308
CORR	AddIncome	-0,000380217	0,007317044	1	1	0,018290776	-0,034779742
CORR	Income	-0,000380217	0,007317044	1	1	0,018290776	-0,034779742
CORR	Age	-0,058908699	0,010531656	0,018290776	0,018290776	1	-0,077438074
CORR	GB	-0,003125426	-0,120234308	-0,034779742	-0,034779742	-0,077438074	1

Initial data very often have a different format (*numeric*, *char*, *Boolean*, etc.), which does not give them a good analysis.

For qualitative analysis of big data, including neural networks, logistic regressions, machine learning, visualization, etc. this is a significant inconvenience, and sometimes it is impossible to correctly carry out the required predictive analysis.

To minimize these shortcomings, it is proposed to translate all the original *data into a digit format*. Then the initial file will take the form of data set, where *all attributes will be in the format numeric* (categorical, ordinal,..).

It should be noted that during the digitization process, the initial file variables (attributes) should be changed to new variables (attributes) with the appropriate format, and the old formats (attributes) should be removed.

The main recommendations for digitizing the initial file are as follows:

1. **Text variable** (*Char*) to digital - ***Numeric categorical***. For example, *profession*, *gender*, *second document*, etc.
2. **Text ordinal variables** (*Char*) to digital - ***Numeric ordinal***. For example, *education (class)*, *salary*.
3. **Numeric variables** (*Numeric real*, *numeric integer*) should be converted to digital - ***Numeric ordinal***. For example, *age*, *loan amount*, etc.

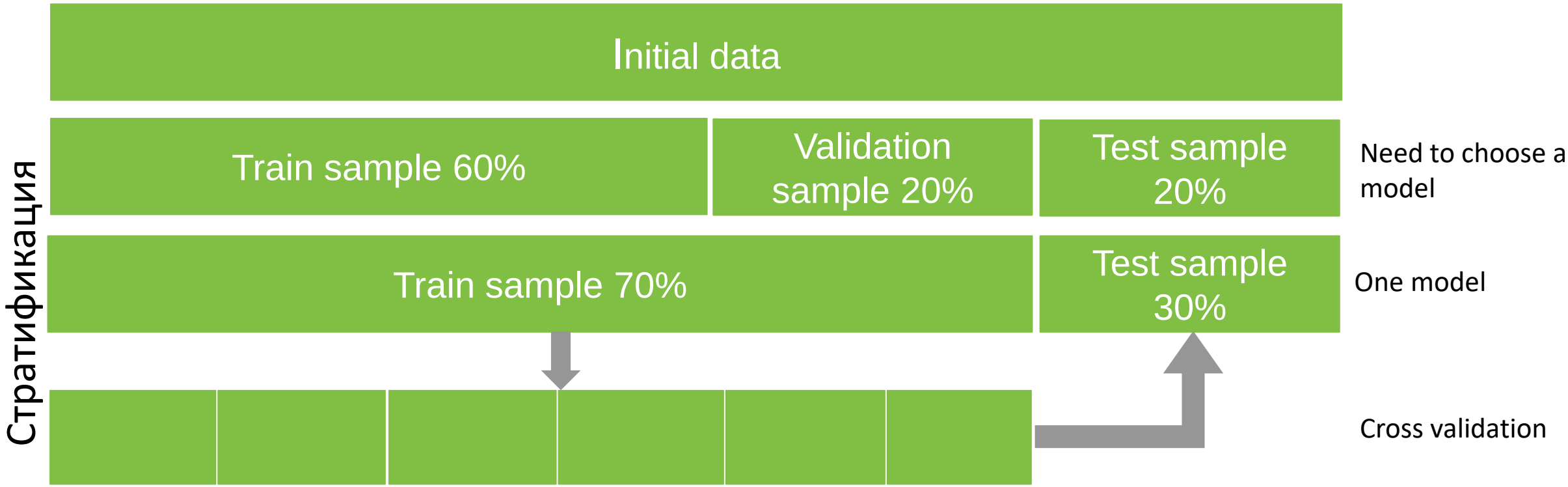
```
/**** Digitalization Data ****/

Data test2;
 set test1;

 if sex="Male"                        then sex_=1;
 if sex="Female"                      then sex_=2;

  if DocType="Military card"          then DocType_=1;
  if DocType="Foreign passport"       then DocType_=2;
  if DocType="Drivine licence"        then DocType_=3;
  if DocType="Driving licence"        then DocType_=3;
  if DocType="other"                  then DocType_=4;
```

```
if Business="Finance /Banking/Insurance" then Business_=1;
   if Business="Industry"                    then Business_=2;
   if Business="Retail"                      then Business_=3;
   if Business="Education"                   then Business_=4;
   if Business="IT and Telecoms"             then Business_=5;
   if Business="other"                       then Business_=6;

    if Education="Graduate"                then Education_=1;
    if Education="Higher education"        then Education_=2;
    if Education="Secondary education"     then Education_=3;

 drop sex DocType Business Education DayOff;/* delete Char variables */

run;
```

ID	CarStatus	AddIncome	Income	CreditSum	Age	GB	sex_	DocType_	Business_	Education_
1	1	3317,8	16589	500000	32,9	1	1	3	2	3
2	0	912,4	4562	100000	999	0	2	4	4	3
3	0	919,4	4597	100000	29,8	0	2	4	1	2
4	0	4000	20000	280000	46,5	0		4	3	3
5	1	12170	60850	500000	42	0	1	3	2	3
5	1	12170	60850	500000	42	0	1	3	2	3
7	1	728	3640	100000	25,4	0	1	2	2	2
8	0	2468,2	12341	150000	43,1	0	1	3	1	3
9	0	824,6	4123	270000	36,3	0	1	1	2	3
10	0	468	2340	136000	30,5	0	2	4	2	3
11	1	430	2150		44,7	0	1	3	2	3
12	0	2468,2	12341	150000	21,9	0	2	3	4	3
13	0	1170	5850	480000	30,8	0	2	3	2	3
14	0	1063,4	5317	183000	27,9	0	1	3	6	2
15	0	1027,8	5139	214000	36	0	2	4	4	3
16	0	1024,6	5123	15000000	43,1	0	2	4	4	3

Digitalization of data (finally)

ID	CarStatus	AddIncome	Income	CreditSum	Age	GB	sex_	DocType_	Business_	Education_
1	1	3317,8	16589	500000	32,9	1	1	3	2	3
2	0	912,4	4562	100000	999	0	2	4	4	3
3	0	919,4	4597	100000	29,8	0	2	4	1	2
4	0	4000	20000	280000	46,5	0		4	3	3
5	1	12170	60850	500000	42	0	1	3	2	3
5	1	12170	60850	500000	42	0	1	3	2	3
7	1	728	3640	100000	25,4	0	1	2	2	2
8	0	2468,2	12341	150000	43,1	0	1	3	1	3
9	0	824,6	4123	270000	36,3	0	1	1	2	3
10	0	468	2340	136000	30,5	0	2	4	2	3
11	1	430	2150		44,7	0	1	3	2	3
12	0	2468,2	12341	150000	21,9	0	2	3	4	3
13	0	1170	5850	480000	30,8	0	2	3	2	3
14	0	1063,4	5317	183000	27,9	0	1	3	6	2
15	0	1027,8	5139	214000	36	0	2	4	4	3
16	0	1024,6	5123	15000000	43,1	0	2	4	4	3


Initial data
Train sample 60%
Validation sample 20%
Test sample 20%
Need to choose a model
Train sample 70%
Test sample 30%
One model
Стратификация
Cross validation

```
/*The program divides file (ONE) into two (Train and Test)
randomly*/
DATA One;
 SET test1;
 LABEL x = 'Random number';
 x=RANUNI(int(time()));
RUN;
```

There were 1499 observations read from the data set WORK. ONE.
The data set WORK.ONE has 1499 observations and 11 variables.

```
DATA Test;
 SET One;
 IF x>0.75;
RUN;
```

The data set WORK.TEST has 366 observations and 11 variables.

```
DATA Train;
 SET One;
 IF x<=0.75;
RUN;
```

The data set WORK.TRAIN has 1133 observations and 11 variables.

```
/*The program divides file (ONE) into two (Train and Test)
randomly*/
DATA One;
 SET test1;
 LABEL x = 'Random number';
 x=RANUNI(int(time()));
RUN;
```

There were 1499 observations read from the data set WORK. ONE.
The data set WORK.ONE has 1499 observations and 11 variables.

```
DATA Test;
 SET One;
 IF x>0.75;
RUN;
```

The data set WORK.TEST has 366 observations and 11 variables.

```
DATA Train;
 SET One;
 IF x<=0.75;
RUN;
```

The data set WORK.TRAIN has 1133 observations and 11 variables.

The results of the computational experiment to assess the accuracy of the predictive model

Comment	Python	SAS	SAS Enterprise Miner
No data preparation ("dirty»)	59%	59%	57%
With the use of the proposed method data preparation	69%	73%	75%
Increase the accuracy of a predictive model	+10%	+14%	+18%

Initial data for the log. regressions

Atribute	Field	Field Russian Name	Format/New	Format/Initial	Comment
ID	Identification number	Персональный номер	numeric	numerec	1234
SEX	Gender	Пол	numeric	char	1 - Male 2 - Female
DocType	Additional document	Дополнительный документ	numeric	char	1 - Military card 2 - Foreign passport 3 - Driving licence
Business	Employer scope of activity	Индустрия	numeric	char	1 - Finance / Banking / Insurance 2 - Industry 3 - Retail 4 - Education 5 - IT and Telecoms 6 - Other
Education	Education	Образование	numeric	char	1 -Graduate (Ученая степень) 2 - Higher education (Высшее) 3 - Secondary education (Среднее образование)
CarStatus	Owns car	Есть машина	numeric	char	1 - Yes 2 - No
AddIncome	Additional income	Дополнительный доход	numeric	numeric	per month
Income	Income from mandatory work confirmed	доход по основному месту работы	numeric	numeric	per month
CreditSum	Loan amount	Сумма кредита	numeric	numerec	560000
Age	The age of the borrower	Возраст заемщика	numeric	numerec	43
DayOff	Overdue payment/days	Дни просрочки	char	char	90+/ 60_90 /30_60 / 0_30 / 0_0
GB/Target	Target function	Целевая функция	numeric	numeric	1- DayOff >= 90+ 0 - DayOff < 90

Logistic regression model

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{1,1} + \ldots + + b_k x_{i,j} + \varepsilon_i$$

where p_i— the probability of default on the loan for i-th borrower;

x_{ij} — value of j-th independent variable ;

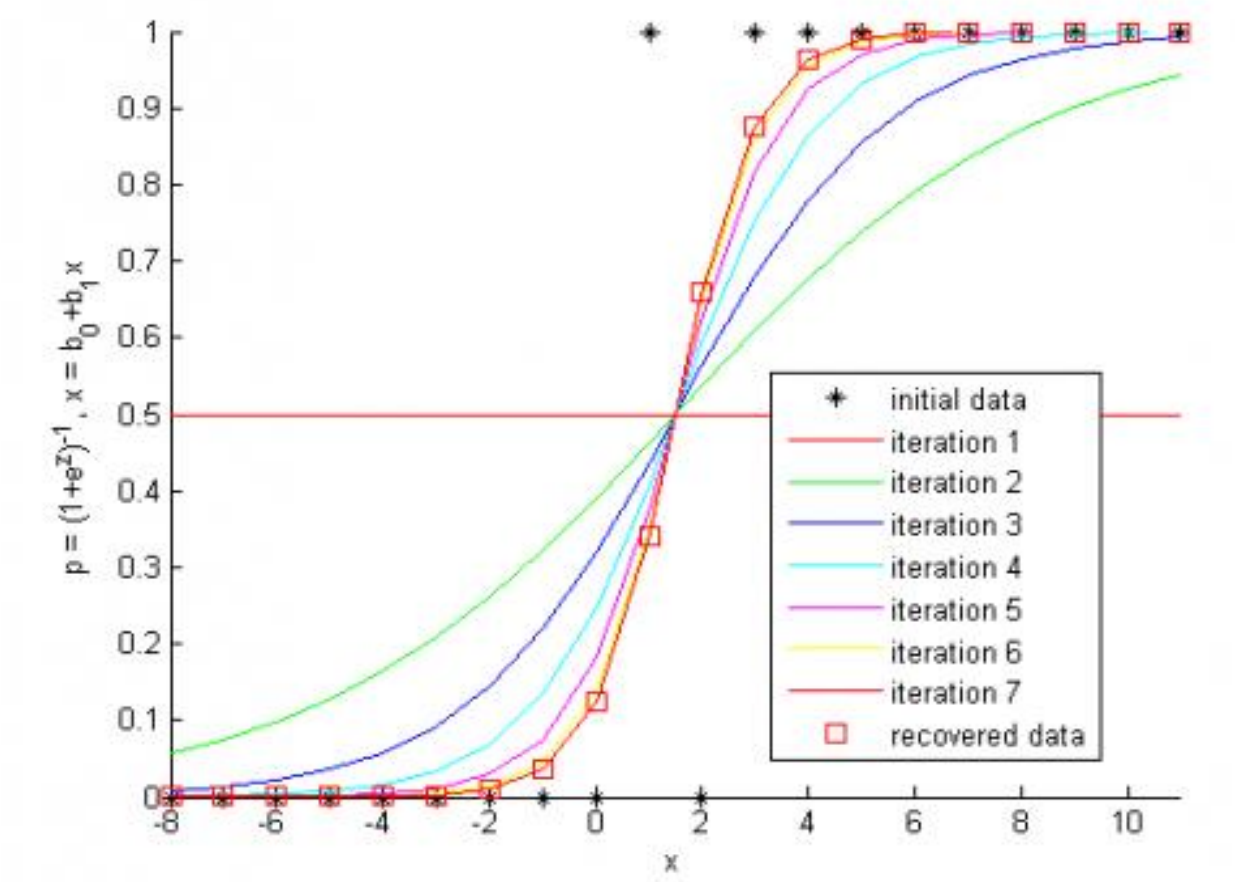
b_0 — independent model constant,

b_j — model parameters;

ε_i — random error component.

```
ods graphics on;
proc logistic Data = train descending;
class TITLE EC_CARD REGN / param = ref;
model GB = AGE CarStatus CreditSum sex_ DocType_
Education_ Income_ Business_ / selection = stepwise
slstay=0.15 slentry=0.15 stb;
score data=train out = Logit_Training   fitstat outroc=troc;
score data=test  out = Logit_Validation fitstat outroc=vroc;
Run;
ods graphics on;
```

The iterative calculation of the parameters of the logistic regression



Logistic regression model

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{1,1} + \ldots + + b_k x_{i,j} + \varepsilon_i$$

where p_i— the probability of default on the loan for i-th borrower;

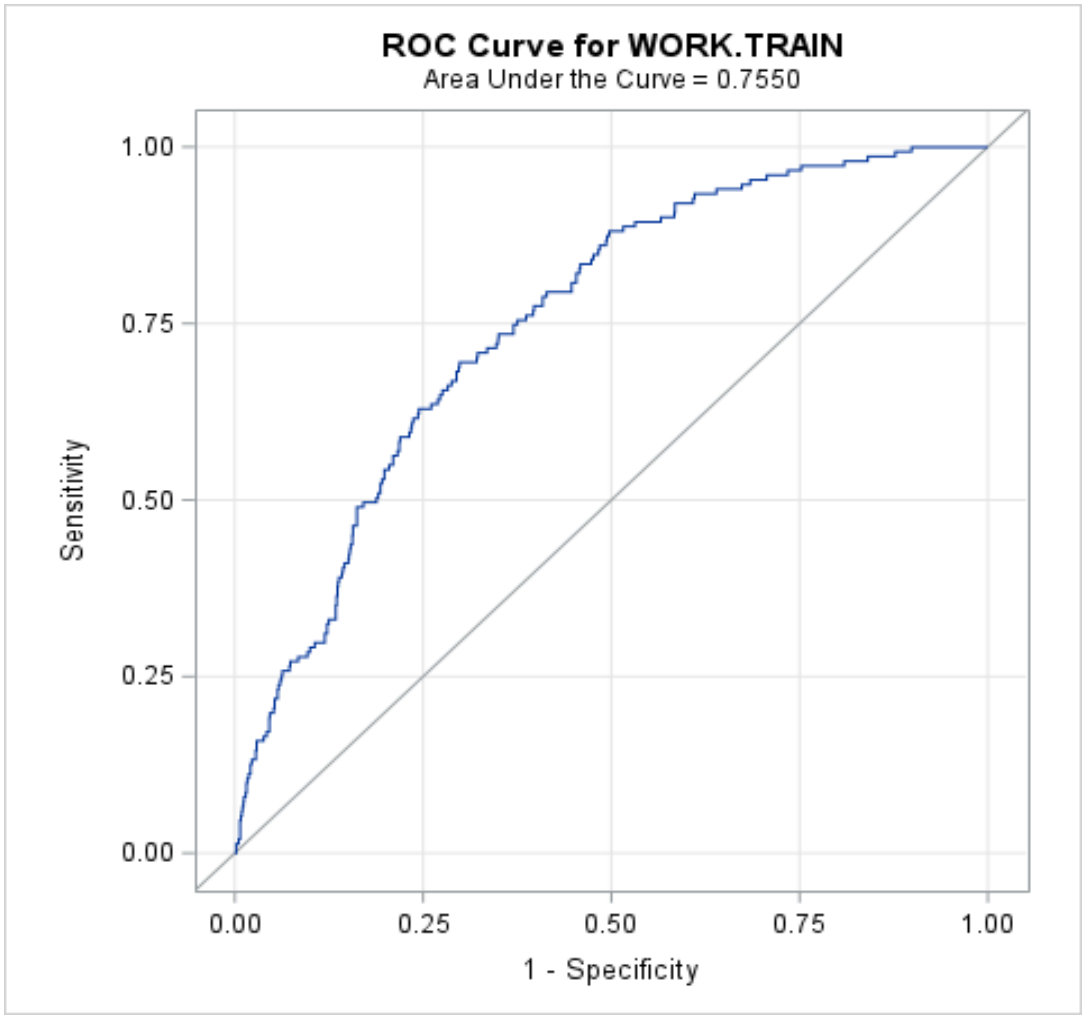
x_{ij} — value of j-th independent variable ;

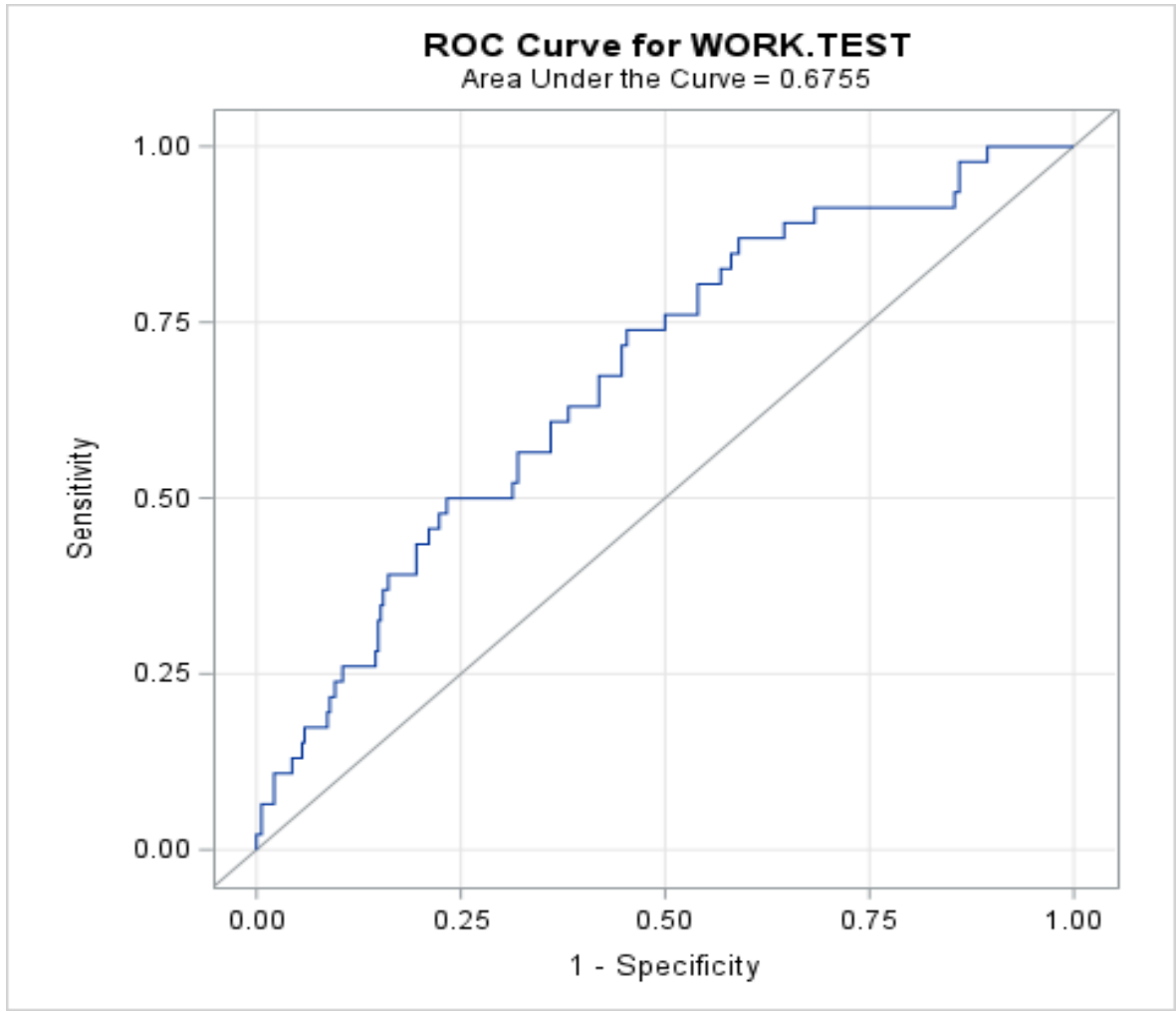
b_0 — independent model constant,

b_j — model parameters;

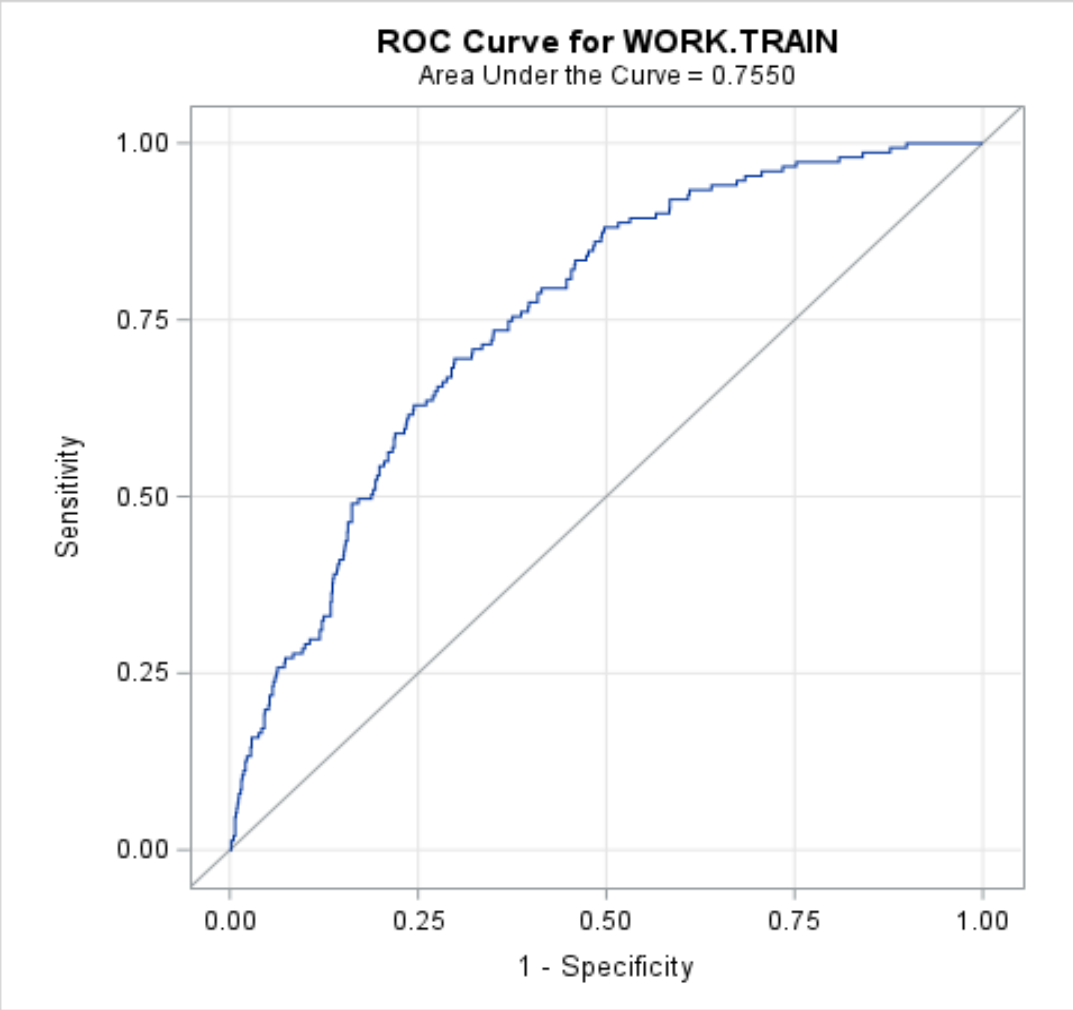
ε_i — random error component.

Classification results for training and test samples



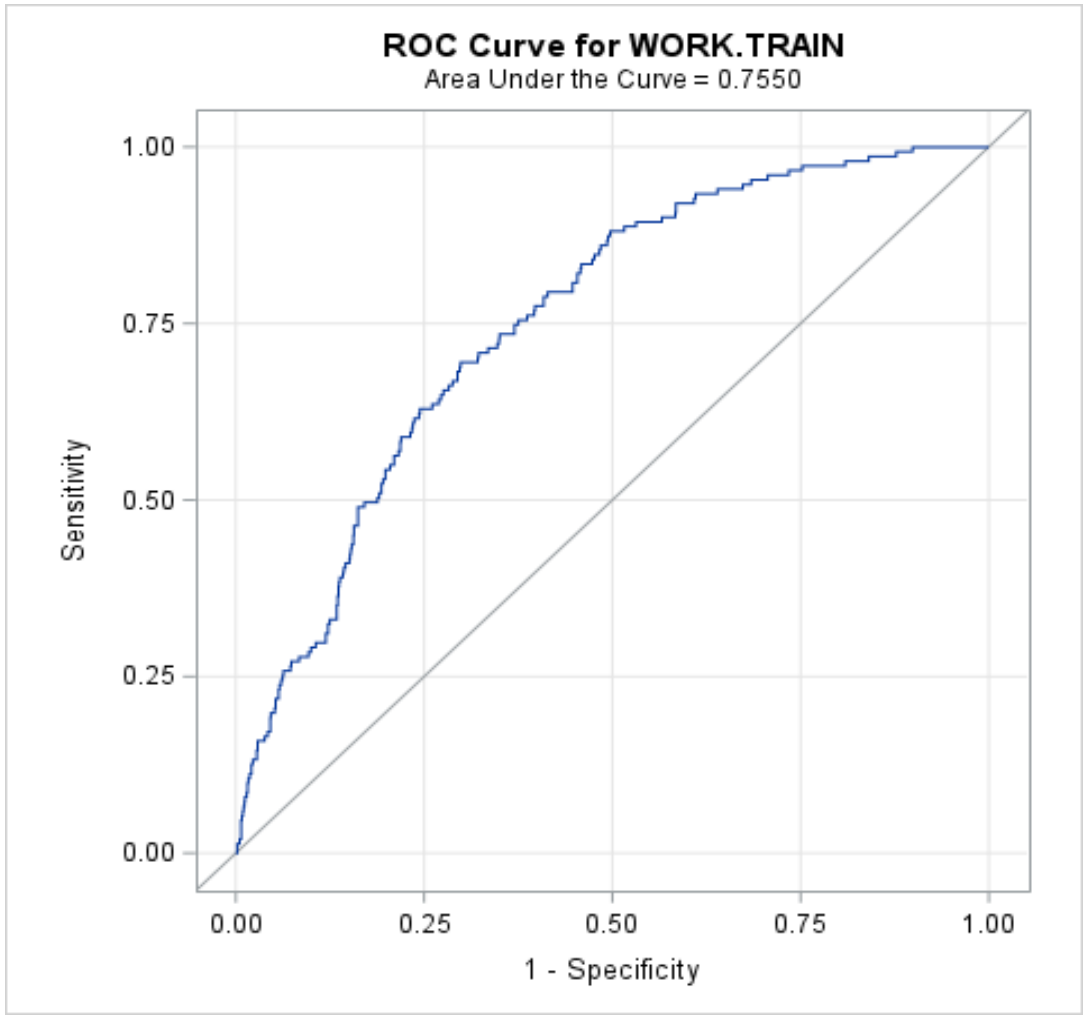


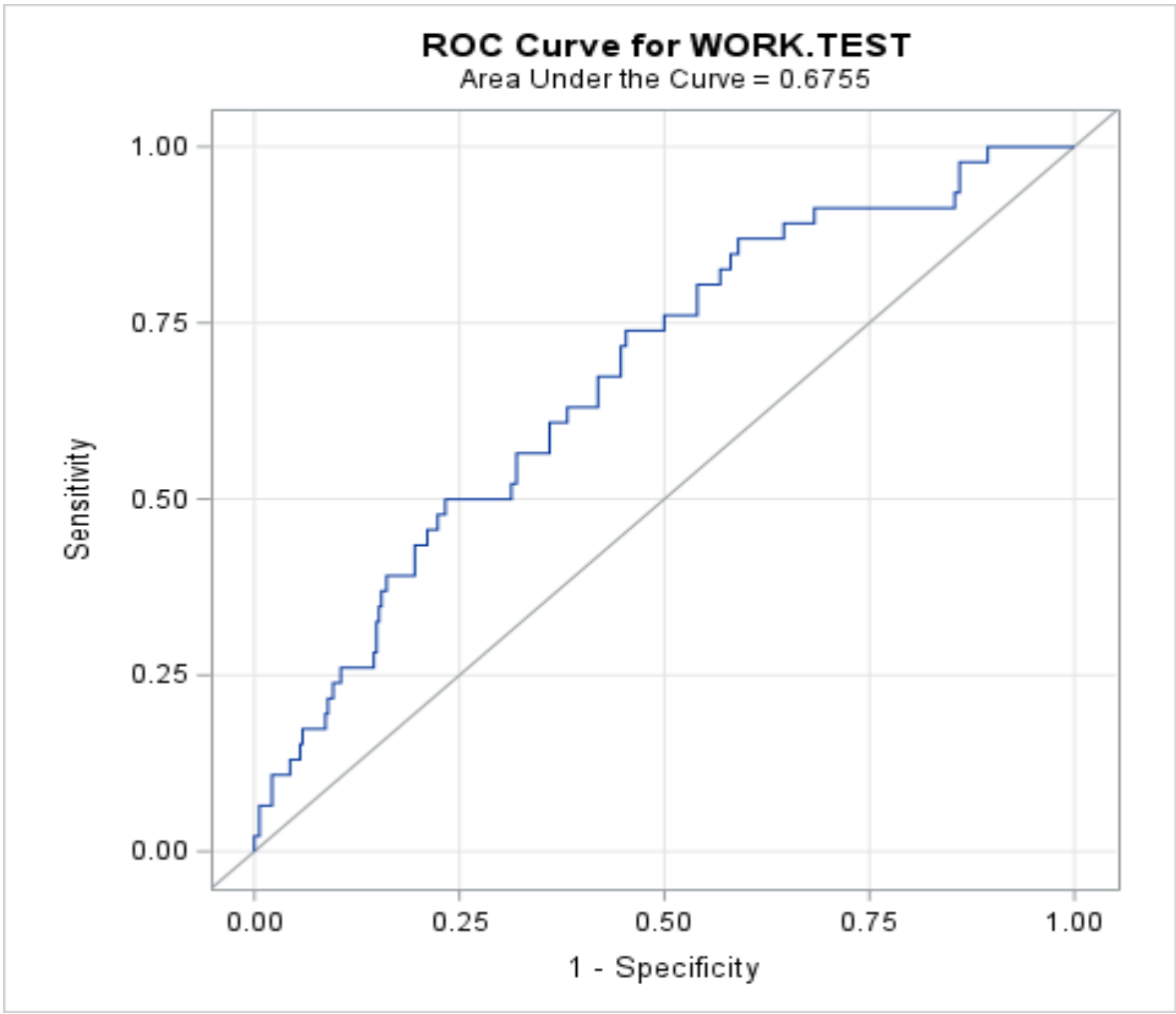
ROC logistic regression curve



		Истинные значения	
		0	1
Предсказанные	0	True Negative (TN)	False Negative (FN)
	1	False Positive (FP)	True Positive (TP)

The iterative calculation of the parameters of the logistic regression





Export converted data from SAS (test1)format to rezerv1 xlsx.

```
proc export data=test1
outfile="C:\Users\gubine\Desktop\xls\rezerv1.xlsx"
dbms=xlsx
replace;
run;
```

In this paper, a method of data preparation for the construction of predictive models of classification is proposed.

Data preparation steps include the following steps: 1. verification of initial data for errors (misprints), 2. on the absence of data ("missing"), 3. on data emissions ("outliers"), 4. on the presence of duplicate lines (observations), 5. to check the original explanatory variables (attributes) for multicollinearity and 6. transformation of source data into digital format ("digitalization") and 7. select the target variable.

The obtained technique is implemented in Python, SAS, SAS Enterprise Miner software packages. *Comparison of the accuracy of the results obtained without data preparation and using the proposed data preparation technique showed an increase in the predictive power of the predict model by almost 20%.The highest accuracy (75%) is demonstrated by the solution obtained with the help of SAS Enterprise Miner.*

TOMSK
POLYTECHNIC
UNIVERSITY

Tomsk Polytechnic University
Institute of Cybernetics

Thanks for your attention

And…

.. *future preparation of data for analysis according to* ***the proposed method***

Tomsk-2021