Ministry of Science and Higher Education of the Russian Federation

Federal State Budgetary Educational Institution

of Higher Education

"PETROZAVODSK STATE UNIVERSITY"

Institute of Mathematics and Information Technology

**Report on the subject "Data Processing Technologies"**

Completed by a student of group 22503 Masaeva Olga

Course leader: Sedov

Alexey Vladimirovich,

Candidate of Technical Sciences, Associate Professor

Petrozavodsk 2022

*Problem Statement*: The main source of the problem is the competition produced on the Kaggle platform [1] “Santander Customer Transaction Prediction”. Quote from the competition description: “Our mission in Santander is to help people and businesses thrive. In this assignment, we invite you to help us determine which customers will make a particular transaction in the future, regardless of the amount of the transaction. The data provided for this contest has the same structure as the real data we have to solve this problem.”

Data Description: Quoted from [1] “You are provided with an anonymous dataset containing numeric feature variables, a binary target column, and an ID\_code string column. The task is to predict the value of the target column in the test set.”

*Purpose:* Optimal solution of the problem of binary classification of transaction data.

To achieve this goal, the following tasks were set:

1. Primary study of data (sizes, statistical characteristics)
2. Plot illustration of characteristics
3. Application of correlation and factor analysis of features
4. Selection of a model and implementation of a model for classification
5. Analysis of model metrics
6. Drawing conclusions

Data analysis was implemented in the language python programming, using the Jupyter notebook software environment, as well as the numpy, pandas, matplotlib, sklearn, lightgbm, scipy libraries

# Primary data exploration

The data was located in the csv file train.csv, converted into a data frame, a head of which is presented below:

ID\_code target var\_0 var\_1 var\_2 var\_3 var\_4 var\_5 var\_6 \

0 train\_0 0 8.255 -6.7863 11.9081 5.0930 11.4607 -9.2834 5.1187

1 train\_1 0 11.5006 -4.1473 13.8588 5.3890 12.3622 7.0433 5.6208

2 train\_2 0 8.6093 -2.7457 12.0805 7.8928 10.5825 -9.0837 6.9427

3 train\_3 0 11.0604 -2.1518 8.9522 7.1957 12.5846 -1.8361 5.8428

4 train\_4 0 9.8 369 -1.4834 12.8746 6.6375 12.2772 2.4486 5.9405

var\_7 ... var\_190 var\_191 var\_192 var\_193 var\_194 var\_195 \

0 18.6266 ... 4.4354 3.9642 3.1364 1.6910 18.5227 -2.3978

1 16.5338 ... 7.6421 7.7214 2.5837 10.9516 15.4305 2.0339

2 14.6155 ... 2.9057 9.7905 1.6704 1.6858 21.6042 3.1417

3 14.9250 ... 4.4666 4.7433 0.7178 1.4214 23.0347 -1.2706

4 19.2514 ... -1.4905 9.5214 -0.1508 9.1942 13.2876 -1.5121

var\_196 var\_197 var\_198 var\_199

0 7.8784 8.5635 12.7803 -1.0914

1 8.1267 8.7889 18.3560 1.9518

2 -6.5213 8.2675 14.7222 0.3965

3 -2.9275 10.2922 17.9697 -8.9996

4 3.9267 9.5031 17.9974 -8.8104

In the training sample there are 200 real features (var\_0-199) and one target binary feature (target) - there is a binary classification problem. In total, there are 200,000 records in the sample.

Figure 1. shows a histogram of the target feature count on the available data. It can be seen that the sample is unbalanced: the number of records with a value of 0 is about 9 times more than with a value of 1. Statistics confirm this assumption: the proportion of zeros in the target feature is 0.89951, while the fraction of units is 0.10049.

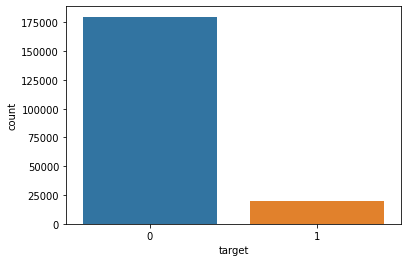
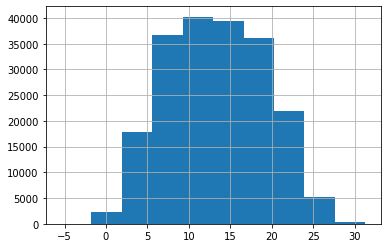
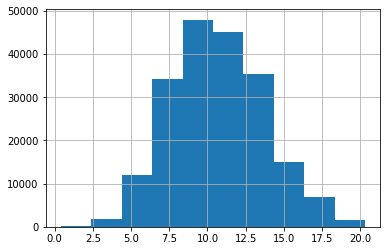
Fig 1. Target feature count histogram.

Figure 2 shows histograms of individual features of the sample data. Visually, the graphs resemble a normal distribution, but when applied to them the Shapiro-Wilk test for normality (the scipy.stats.shapiro function [2]), the p-values ​​were significantly small (of the order of 1.1724418070891532e- 36), which rejects the hypothesis of normality of distributions. When approaching, the presence of more than one peak in the histogram is noticeable, the distribution may be a mixture of normal distributions.

Figure 3 also shows histograms of the 1st feature, divided by the value of the target variable. There is also a visual similarity of the histograms, but this fact also requires statistical verification. Despite the external similarity of the distributions, they failed the Kolmogorov-Smirnov homogeneity test (function ks\_2samp [3]), since the pvalue is below 1%.

Figure 4 shows the graphs of the distribution of the mathematical expectation and the standard deviation of features.



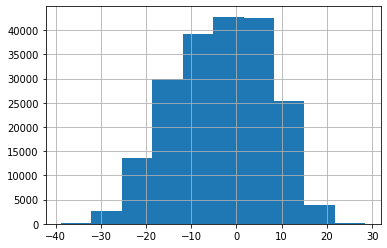
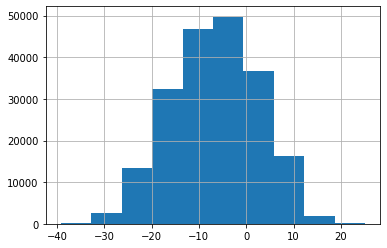


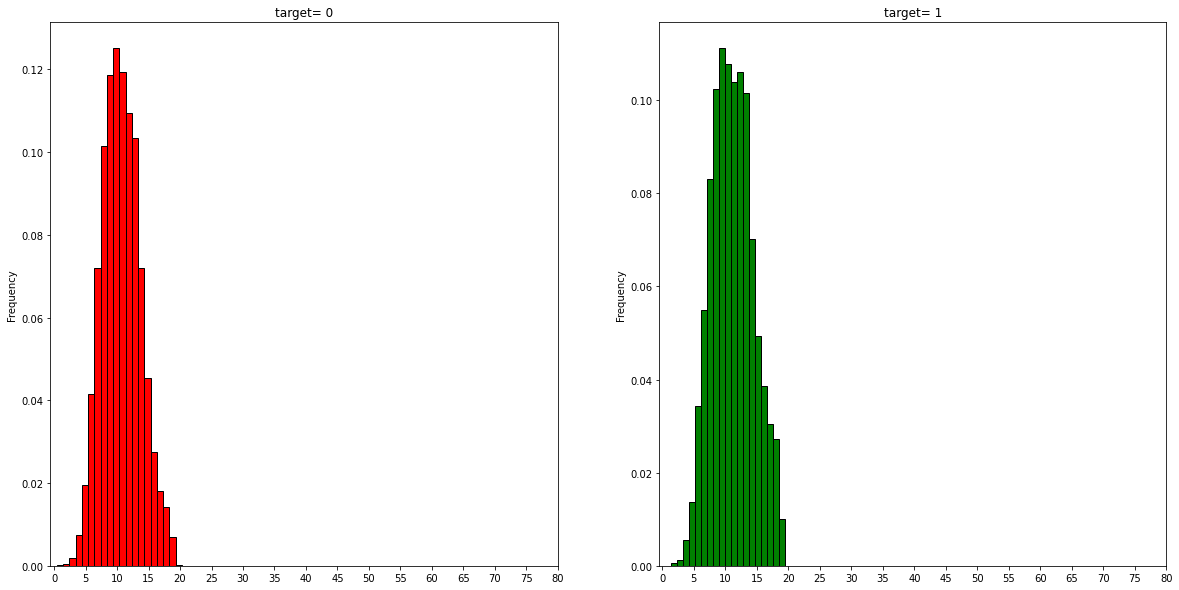
Fig 2. Graphs of parameter distribution histograms 1(upper left), 20(upper right), 100(lower left), 199 (lower right)

Fig 3. histograms of feature var\_0 divided by target value.

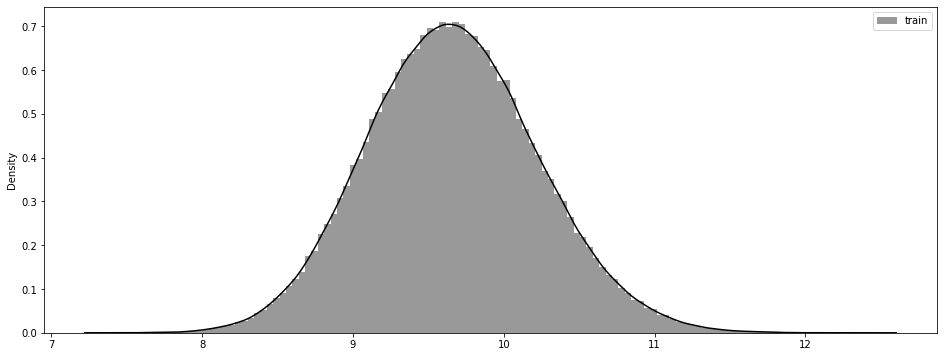
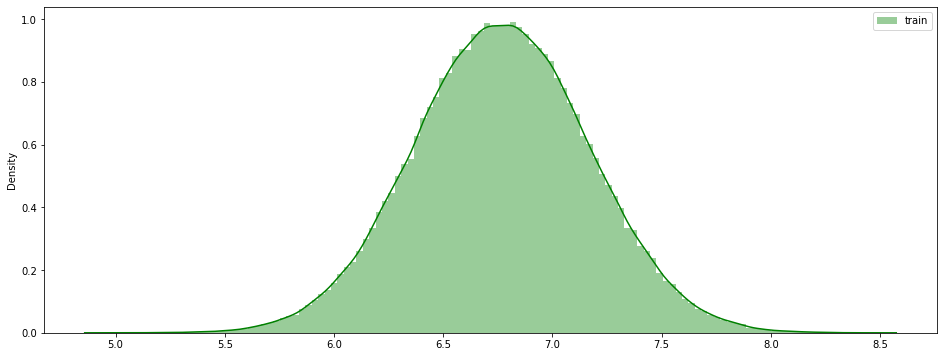


Fig. 4. Graphs of distributions of mathematical expectations and standard deviations of features

# Correlation and Factor analysis

As a correlation analysis, Pearson's paired correlation coefficient was calculated. Without identical pairs, the lowest correlation value is observed between the pair var\_75 and var\_191 cor = 2.703975e-08, the highest correlation value is observed between the pair var\_139 and var\_26 cor = 0.009844. Given the fact that the largest pairwise correlation is an extremely small value, we can conclude that the features are independent.

Also, given the large amount of data, the use of the principal component method (PCA algorithm [4] with standard parameters and n\_components=2 ) was considered to reduce the dimension of the feature vector. The result of data compression to two components is the variance vector “Results of variance in 2 columns used : [0.005654 0.00514221]” , which is 0.010796214605994834 in total. Since the data loss is negligible - 0.01 , the principal component method takes place



5. Graph of accumulation of sampling points in the new coordinate system

# Using models

As models, 2 models were taken - logistic regression and LGBM. Both models showed good results and worked fast enough even without dimensionality reduction.

In the competition [1], ROC AUC was used as a quality metric.

*Logistic regression*:

Logistic regression or logit model (English logit model) is a statistical model used to predict the probability of an event occurring by comparing it with a logistic curve. This regression returns the answer as the probability of a binary event (1 or 0). This model is often used for classification problems. For our data model, the LogisticRegression function ([5]) with default parameters was used, and the full set of the dataset was used as the data. The results of the metrics on the validation set are presented below:

Confusion\_matrix:

[[35625 435]

[ 3033 907]]

Accuracy: 0.9133

Precision: 0.6759

Recall: 0.2302

F1: 0.3535

AUC : 0.6091

Since the class sizes are not balanced, this metric should not be used to judge the performance of the algorithm. Accuracy has an average value, and recall is completely low, due to which the F metric also has a low value. The AUC value is about 0.6, which means that the quality of the algorithm is better than random class selection, but not much.

For a compressed principal component logistic regression model, the results of metrics are presented below:

Confusion\_matrix:

[[36013 40]

[ 3888 59]]

Accuracy: 0.9018

Precision: 0.5960

Recall: 0.0149

F1: 0.0292

AUC: 0.5069

It can be seen that a much smaller number of elements were classified in the second class in this model than in the previous model. Characteristic values ​​have decreased, especially completeness. The value of the area under the ROC curve indicates that this choice is not much better than a random choice.

*Light Gradient Boosted Machine*

In machine learning, the LightGBM classifier is part of the Boosting family and is the most common classification model in the machine learning community today. LightGBM is a powerful machine learning model that you can shape based on the problem you're working on. Light GBM is a fast, distributed, high performance decision tree supported gradient forcing algorithm used for ranking, classification and other machine learning problems.

For the current task, the full data set and the lightgbm[6] model were used with parameters:

params = {

'device' : 'cpu',

'n\_estimators': 7000,

'num\_leaves': 20,

'max\_depth': -1,

'min\_data\_in\_leaf': 80,

'learning\_rate': 0.008,

'boosting': 'gbdt',

'objective': 'binary',

'metric': 'auc',

'n\_jobs': -1

}

The training time was several minutes. On the validation set, the results of the metrics were as follows:

Confusion\_matrix:

[[17779 133]

[ 1463 625]]

Accuracy: 0.9202

Precision: 0.8245

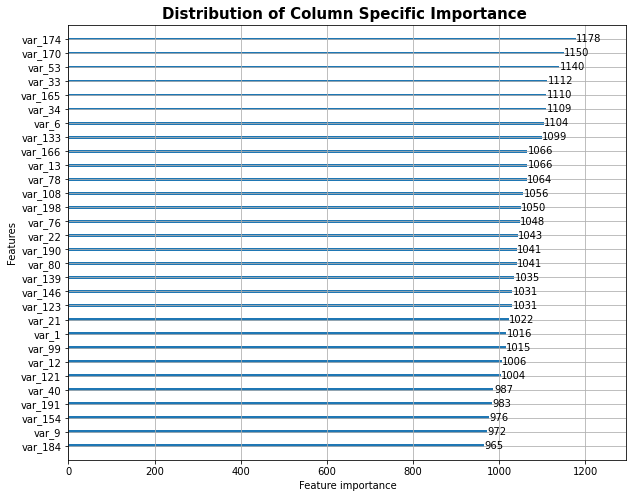
Recall: 0.2993

F1: 0.4392

AUC: 0.6460

The results of the metrics are better than the logistic regression model. High accuracy, with low recall gives an average estimate of the F-metric. The results of the model on hidden test data when sent to the competition [1] are 0.89 , a very good result.

Also, as a demonstration of the model, you can bring the graph in Fig. 6.

fig. 6 graph illustrating the importance of each variable in the constructed model.

# Conclusion

As a result of the work, an analysis of transaction data was made and, on its basis, a binary classification model was selected.the example of compression by the method of principal components and logistic regression models

Used materials:

1. Web resource [link: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html> ]
2. Web resource [link: [https://docs.scipy.org/doc/scipy /reference/generated/scipy.stats.ks\_2samp.html](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html) ]
3. Web resource [link: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html> ]
4. Web resource [ link: [https ://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html) ]
5. Web resource [link: <https://lightgbm.readthedocs.io/en/latest/Python-API.html>]