IRSTI 28.23.27 https://doi.org/10.26577/JPCSIT.2023.v1.i2.04

N.M. Koishybayeva* , **M. Nurtas** , **A. Altaibek**
International Information Technology University, Kazakhstan, Almaty
*e-mail: nazerkekoishybaeva777@gmail.com

BUILDING A MODEL BASED ON MACHINE LEARNING METHODS FOR PREDICTING THE CREDITWORTHINESS OF CUSTOMERS

Abstract. A customer's credit rating is important for financial institutions, as lending can result in real and immediate losses. Scoring models are increasingly used in modern financial technologies and serve professionals to improve their efficiency. They are superior in their capabilities to the subjective assessments of people, as they are not subject to professional bias and cognitive distortions. In this article, we will focus on building machine learning models to predict customer creditworthiness. The main goal is to identify the most important factors that will help calculate the creditworthiness of customers, analyze customer characteristics. We will build Support Vector Machine(SVM), Decision Trees(DTs), Xgboost and Random Forest models and explore comparative analysis of their predictive accuracy.

Key words: creditworthiness, forecasting, scoring models, feature engineering, machine learning methods.

Introduction

Credit institutions, especially banks, have a significant impact on an economy that operates based on supply and demand [1]. One of the constant problems of credit organizations that require a timely solution is the issuance of a loan to an unreliable borrower and the refusal to issue a trustworthy one. A wrong decision will lead to losses and possible bankruptcy. One of the main reasons for poor cash flow management is lending to customers without assessing their creditworthiness. Small and medium businesses must have a strict credit check policy before attracting new customers to avoid financial problems [2]. A customer's creditworthiness is an important business principle that shows how creditworthy a customer is. A customer is creditworthy if the company believes it can repay the debt on time. It is determined by several factors. For example, income, payment history, credit score, outstanding obligations, etc [3].

Models are needed to produce numerical "scores" that help make better decisions and generalize consumer creditworthiness [4]. We will compare the available consumer credit risk model, which combines traditional credit factors such as debt-to-income ratios with consumer banking, which greatly enhances the ability of our model to make accurate forecasts. Analyzing these models and characteristics, we will find their patterns and how they affect the choice of solution. Numerous statistical models are used in the process of evaluating creditworthiness [5], this involves various statistical methods, such as linear regression, logistic regression, linear discriminant analysis, probit analysis, and naive bayes analysis [6].

However, when dealing with non-linear relationships, these strategies often have low or inadequate performance. With respect to actual application, it is difficult for them to satisfy these statistical hypotheses. Therefore, various artificial intelligence and machine learning methods were applied recently years for the credit rating, and these systems outperformed statistical analysis [7]. SVM [8], artificial neural networks (ANN) [9], and random forest [10] are some of these techniques. Credit scoring has recently paid a lot of attention to ensemble approaches like random forest, Xgboost [11]. Ensemble approaches are often more favorable when compared to other credit scoring algorithms, and these algorithms are currently accepted as industry standard.

Predicting loan defaults is generally of great interest, and numerous techniques and data sets have been investigated. The pertinent research evaluated "hard information" including income, age, and debt information as well as put a focus on "unofficial information" of borrowers in order to increase the effectiveness of the categorization model [12].

Overview of existing machine learning models for predicting customer creditworthiness

DTs. DTs are an example of supervised learning that is non-parametric and can be utilized for

tasks involving both classification and regression. The goal is to acquire easy-to-understand guidelines based on the data characteristics in order to construct a model that can estimate the value of a desired outcome. A tree can be modeled using a piecewise constant approach. DT classifiers have been extensively used for many different purposes. The tree flow view is akin to a progress chart, with cases organized in a hierarchical layout based on their characteristics. [13].

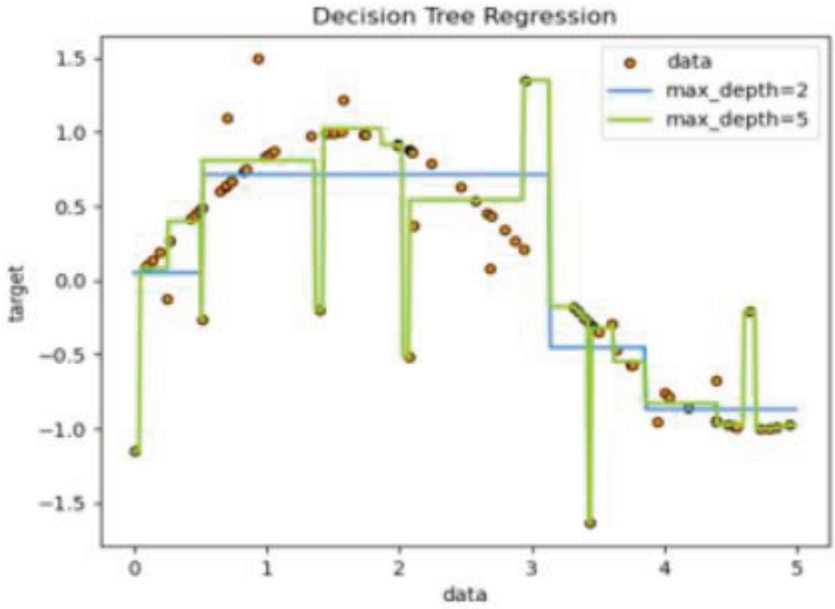

Figure 1 – An example shows how to work a DT regression

DTs are organized in a way that resembles an inverted tree, with data divided into several branches. The model consists of a sequence of logical decisions that are organized like a flowchart, with nodes denoting judgments that must be made regarding certain attributes. The decision-making process is reflected in the branches. Each branch's nodes stand in for classes and class distributions. The root node of a tree, which has the most information gain, is the largest node in the tree. After the initial node, one of the next nodes with the greatest increase in information is selected to be evaluated as a potential element for the subsequent node. Until all variables are compared or there are no more variables into which the samples may be separated, this process is repeated. The tree then finishes with nodes that depict the decision-making process when comparing classes or class distributions [14].

Random forest. The DT-based Random Forest methodology combines flexibility and strength into a single machine learning strategy. The approach can handle big data sets, where the alleged "curse of dimensionality" can lead to other models failing, and requires only a small random portion of the entire collection of observations [15]. This strategy adds variation to DT models by using the fundamentals of bagging—the random picking of features. The program then combines tree forecasts using a process depending on the amount of votes after generating a random forest. According to Breiman's description, Random Forest is a classifier made up of a group of structured classification trees h(x, k), k = 1,... where k are randomly independent and identically distributed vectors and each tree makes a single decision regarding the class that is most likely to match the input data x[16].

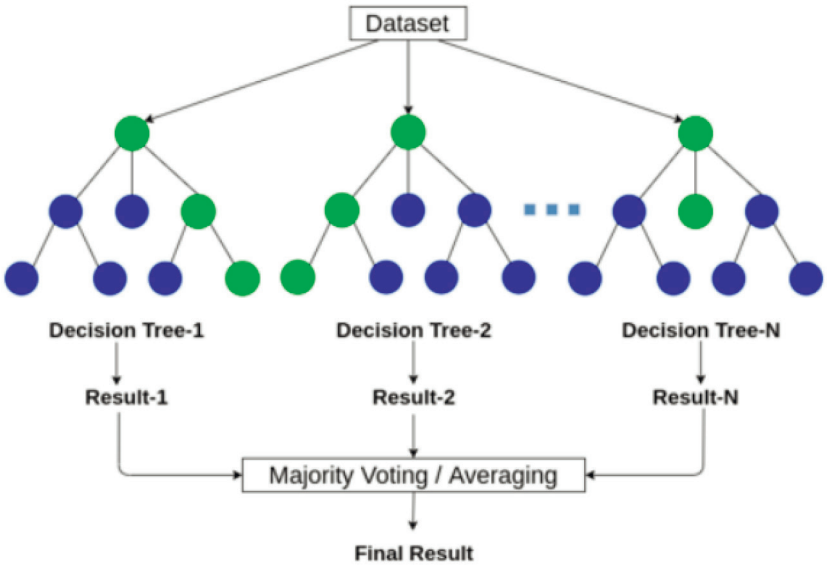

Figure 2 – An example shows of random forest work

SVMs. SVMs are a group of supervised learning techniques for classifying data, performing regression analysis, and identifying outliers [17]. Studies show that the SVM, the main idea of which is to move from the original feature space to a higher-dimensional (or even infinite-dimensional) space and search for a hyperplane in it that maximally separates classes, has proven to be very effective. classification method. Two important kernel parameters are C and γ. By applying the grid search method, we can find the best C and gamma values for the kernel [18].

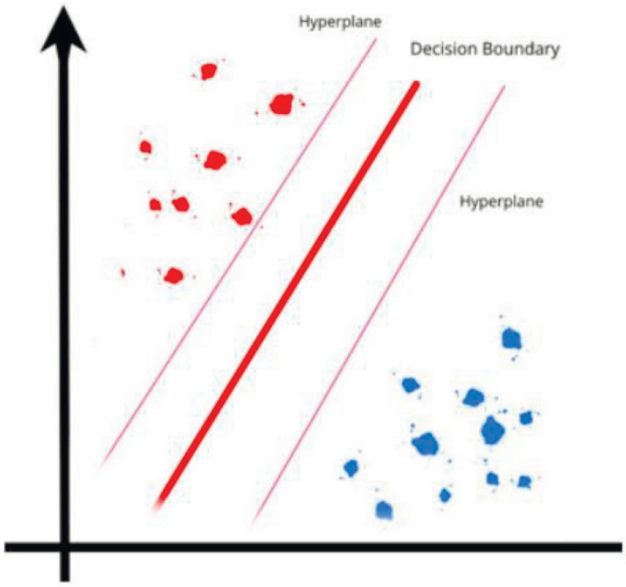

Figure 3 – An example of classifying customers using the SVMs

Xgboost**.** Xgboost is a commonly used algorithm for supervised machine learning that is capable of performing both classification and regression tasks. It is a version of the gradient boosting algorithm that is referred to as extreme gradient boosting. It is recognized for its rapidity and efficiency and is extensively employed in operational systems. Xgboost is a highly efficient and scalable implementation of gradient boosting and allows users to define custom optimization objectives and evaluation criteria. It has been used in many winning solutions of data science challenges [19].

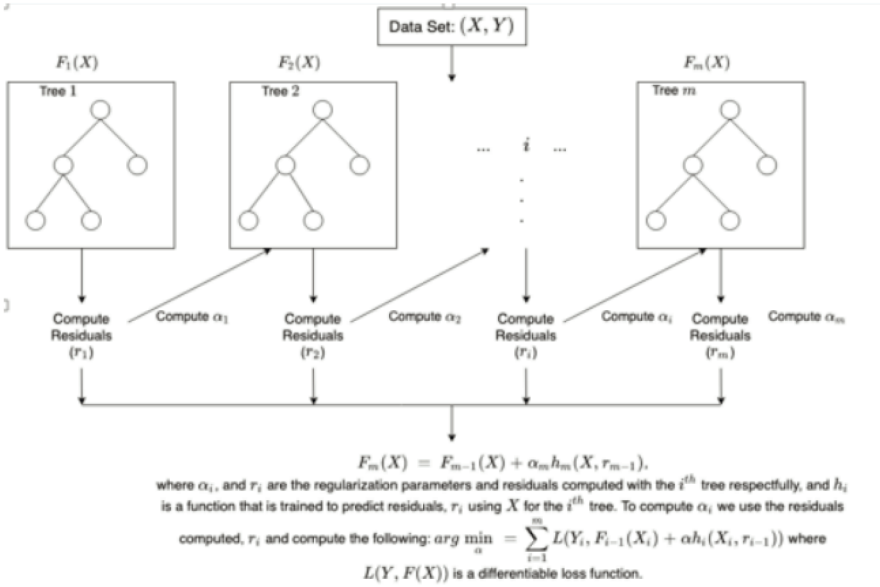

Figure 4 – An example of how Xgboost works [20]

Data preprocessing

We sourced pre-made data from the platform Kaggle, a data science and machine learning competition platform and a community of over 2 million data scientists and machine learning experts. Predicting whether a customer will repay a loan or hit a hardship is a critical business need and there is enough data collected in this platform to help us build models and analyze them. We will get the opportunity to use real data based on the experiments received.

To determine a person's creditworthiness, the following data is typically considered:

1) Employment history: This includes information about a person's current and previous employment, salary, and job stability.

2) Credit history: This includes information about a person's credit accounts, such as credit cards, loans, and mortgage payments, and whether they have made payments on time.

3) Income: Information about a person's income, including salary, investment income, and any other sources of income, is also considered.

4) Debt: The amount of debt a person has, including credit card debt, student loans, and other types of debt, is considered.

5) Demographic information: This might encompass the individual's age, place of residence, and other personal details that may be put to use to recognize them.

This information is usually obtained from credit bureaus, which maintain credit reports on individuals and businesses. Lenders use this information to evaluate a person's creditworthiness and determine whether they are likely to repay a loan [21].

According to some criteria's above, we found data in the platform "Kaggle" that would help us to explore them [22]:

Table 1 – Research data options of applicant

Name of parameters	Type
ID	ID of applicant
CODE_GENDER	Man and woman
FLAG_OWN_CAR	Having a car or not
FLAG_OWN_REALTY	Having house reality or not
CNT_CHILDREN	Children Numbers
AMT_INCOME_TOTAL	Annual Income
NAME_INCOME_TYPE	Income Type
NAME_EDUCATION_TYPE	Education
NAME_FAMILY_STATUS	Marriage Condition
NAME_HOUSING_TYPE	House Type
DAYS_BIRTH	Age
DAYS_EMPLOYED	Working Years
FLAG_MOBIL	Having a phone or not
FLAG_WORK_PHONE	Having a Work Phone or not
FLAG_PHONE	Having a phone or not
FLAG_EMAIL	Having an email or not
OCCUPATION_TYPE	Occupation Type
CNT_FAM_MEMBERS	Famliy Size

Table 2 – Research data options of applicant

Name of parameters	Type
ID	ID of applicant
MONTHS_BALANCE	Applications account open month
STATUS	Applicant status

Feature engineering

Feature engineering is the process of creating, selecting, and transforming data into features that is usable to train a machine learning model. This process involves extracting relevant features from raw data, selecting the right features to create a model, and transforming the data into the desired format. Feature engineering is an important step in the machine learning process, as it can significantly improve the performance of a model by providing it with more relevant, meaningful, and useful features. We have done some transformations with the data in order to properly use the function [23].

Gender

```
new_data['Gender'] = new_data['Gender'].replace(['F','M'],[0,1])
print(new_data['Gender'].value_counts())
iv, data = calc_iv(new_data,'Gender','target')
ivtable.loc[ivtable['variable']=='Gender','IV']=iv
data.head()
```

```
0    10061
1     5874
Name: Gender, dtype: int64
This variable's IV is: 0.04386069618605326
0    10061
1     5874
Name: Gender, dtype: int64
```

	Variable	Value	All	Good	Bad	Share	Bad Rate	Distribution Good	Distribution Bad
0	Gender	0	10061	9994	67	0.631377	0.006659	0.630815	0.728261
1	Gender	1	5874	5849	25	0.368623	0.004256	0.369185	0.271739

Figure 5 – Some example to feature engineering

IV, WOE: Concept and Application

IV, WOE calculations play a crucial part in machine learning. It is employed to ascertain the optimal data split points by optimizing the information value of the split. It also helps to identify the key factors influencing the prognosis. In addition, the calculation is used to assess the effectiveness of the model and identify areas for improvement [24].

Weight of Evidence(WoE) is a statistical technique used to assess the strength of evidence for or against a hypothesis. It combines data from multiple independent sources to create a measure of the total evidence that is more powerful than any single source. The WoE measures how strongly the evidence supports or refutes a hypothesis, and it is applicable for evaluating the precision of a decision or forecast [25].

$$woe_i = \ln\frac{P_{yi}}{P_{ni}} = \ln\frac{y_i/y_s}{n_i/n_s} \quad (1)$$

where P_{yi} refers to the likelihood of an event taking place given the evidence, and P_{ni} is the probability of the event occurring without the evidence.

Information Value (IV) in ML is a measure of the predictive power of a given feature or set of features. It measures how much additional information is revealed by a given feature, relative to what is already known. Information Value can help identify which features are most important for making predictions and optimize models.

$$IV_i = (P_{yi} - P_{ni}) * woe_i \quad (2)$$

The total IV value of a variable can be calculated by summing the difference between the conditional positive rate (CPR) and the conditional negative rate (CNR) for each category of the variable, weighted by the proportion of observations in each category [26]:

$$IV = \sum_i^n IV_i \quad (3)$$

Table 3 – This table shows relationship between IV value and predictive power.

IV	Ability to predict
<0.02	Almost no predictive power
0.02~0.1	weak predictive power
0.1~0.3	Moderate predictive power
0.3~0.5	Strong predictive power
>0.5	Predictive power is too strong, need to check variables

Results of predicting

As a result of the research, the most suitable model for making predictions the creditworthiness of the client should be proposed.

Table 4 – This table shows results of the models.

Name of model	Accuracy score
Xgboost	0.94
SVM	0.59
Random forest	0.89
DT	0.83

In the table above, you can see what results the models showed when predicting. The answer to the question of which model is best for predicting a client's creditworthiness depends on several factors, like the scale and intricacy of the dataset, the type of data used (numerical, categorical, etc.) and the desired accuracy of the forecasts.

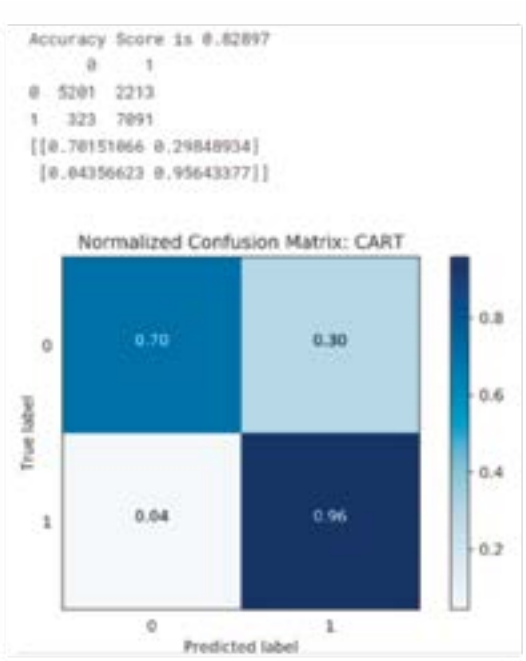

Figure 6 – Results of "DT".

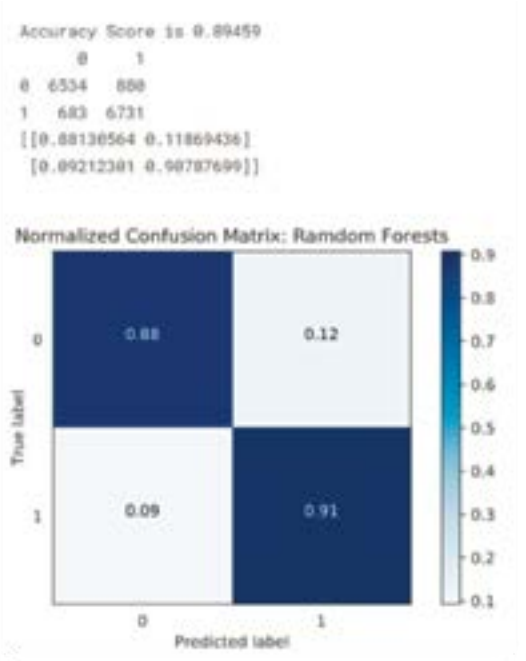

Figure 7 – Results of "random forest"

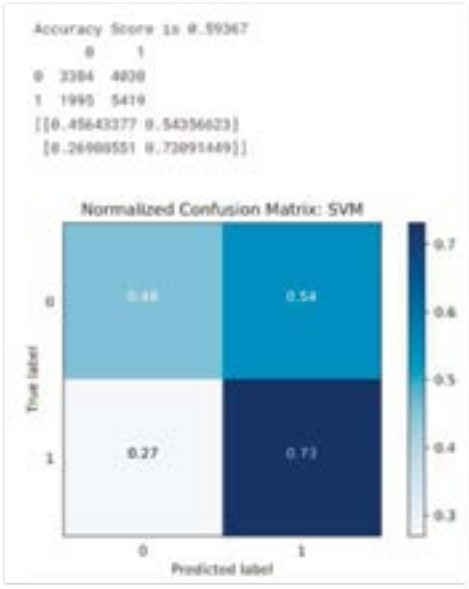

Figure 8 – Results of "SVM".

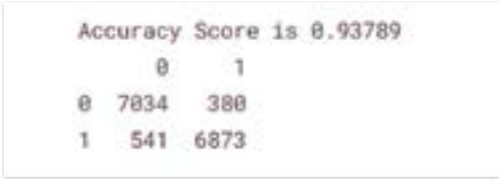

Figure 9 – Results of "Xgboost"

Conclusion

The result of this research is: understanding the patterns that will help credit institutions predict the creditworthiness of a client; and build an accurate predictive model. For comparison, we have chosen such models as Random forest, DTs, SVM, Xgboost. The best results were obtained for the optimal Xgboost model. But this is not, of course, the point of inquiry. Since, depending on the choice of character-

istics, the result may change, it is important to choose the right characteristics and build the model correctly.

Obtaining good outcomes is feasible by utilizing the models derived from the aforementioned experiments and possessing appropriate data. The credit strategy department of a bank or other lending institution may use such a model for decision speed. This will ultimately reduce the potential loss of revenue due to non-performing loans, which can be caused by clients who are unable to fully repay their debt repayments.

We believe that our results are not the end in this research, for future study more powerful models of consumer behavior will be built that can be developed with machine learning methods, and are exploring further refinements and larger datasets in ongoing research.

References

1. Dr.S.Sheeba Gladis, Dr.M.Josephine Rani. A study on the role of banks in development of the district, 2022. URL: https://www.journalppw.com/index.php/jpsp/article/view/7902/5147

2. Kozlova E.P., Galantna E.N. Accounting in commercial banks. – M .: Finance and statistics, 2000 .-- 640 p

3. HighRadius Corporation. How to Check the Creditworthiness of a New Customer? URL: https://www.highradius.com/resources/Blog/how-to- check-the-creditworthiness-of-a-new-customer/

4. Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer Credit Risk Models via Machine-Learning Algorithms. URL: https://core.ac.uk/download/pdf/4430264.pdf

5. D. J. Hand, W. E. Henley. Statistical Classification Methods in Consumer Credit Scoring: a Review, 2007. URL: https://doi.org/10.1111/j.1467- 985X.1997.00078.x

6. B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens & J, Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring, 2003. URL: https://doi.org/10.3390/info10120397.

7. Beibei Niu, Jinzheng Ren and Xiaotao Li. Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer- to-Peer Lending, 2019. URL: https://doi.org/10.3390/info10120397

8. Karim Amzile ,Mohamed Habachi. Assessment of Support Vector Machine performance for default prediction and credit rating, 2022. URL: https://halshs.archives-ouvertes.fr/halshs-03643738/

9. Zongyuan Zhaoa, Shuxiang Xu, Byeong Ho Kang, Mir Md Jahangir Kabira , Yunling Liu, Rainer Wasingera. Investigation and improvement of multi-layer perceptron neural networks for credit scoring, 2015. URL: https://doi.org/10.1016/j.eswa.2014.12.006

10. Paul Wanyanga. Credit Scoring using Random Forest with Cross Validation, 2021. URL: https://medium.com/analytics-vidhya/credit- scoring-using-random-forest-with-cross-validation-1a70c45c1f31

11. Lei Zhanf, Quankum Song. Multimodel Integrated Enterprise Credit Evaluation Method Based on Attention Mechanism, 2022. DOI: 10.1155/2022/8612759

12. Maisa Cardoso Aniceto, Flavio Barboza ,Herbert Kimura. Machine learning predictivity applied to consumer creditworthiness, 2020. URL: https://fbj.springeropen.com/articles/10.1186/s43093-020-00041-w

13. Prashant Gupta. Decision Trees in Machine Learning, 2017. URL: https://towardsdatascience.com/decision-trees-in-machine-learning- 641b9c4e8052

14. Leila Halawi, Amal Clarke, Kelly George. Decision Trees and Ensemble, 2022. DOI: 10.1007/978-3-030-89712-3_5

15. Tony Yiu. Understanding Random Forest, 2019. URL: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

16. Simon Bernard, Sébastien Adam, Laurent Heutte. Dynamic Random Forests, 2012. URL: https://hal.archives-ouvertes.fr/hal- 00710083/document

17. Rohith Gandhi. Support Vector Machine — Introduction to Machine Learning Algorithms, 2018. URL: https://towardsdatascience.com/support-vector-machine-introduction-to- machine-learning-algorithms-934a444fca47

18. Drew Wilimitis. The Kernel Trick in Support Vector Classification, 2018.

19. URL: https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

20. https://xgboost.readthedocs.io/en/stable/

21. https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

22. https://www.investopedia.com/terms/c/credit-worthiness.asp#:~:text=Creditworthiness%20is%20determined%20by%20several,determine%20the%20probability%20of%20default.

23. https://www.kaggle.com/code/rikdifos/credit-card-approval-prediction-using-ml/data

24. https://www.google.com/search?q=Feature+engineering&source=lmns&bih=794&biw=1535&hl=ru&sa=X&ved=2ahUKEwihpMan2ID9AhUh_SoKHbQYBWwQ_AUoAHoECAEQAA.

25. https://towardsdatascience.com/model-or-do-you-mean-weight-of-evidence-woe-and-information-value-iv-331499f6fc2.

26. https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

27.https://docs.tibco.com/pub/sfire-dsc/6.5.0/doc/html/TIB_sfire-dsc_user-guide/GUID-07A78308-525A-406F-8221-9281F4E9D7CF.html.