Analyzing Credit Risk Determinants: Machine Learning Approach Using Armenian Credit Registry Data *Spring 2024*

Author: Anahit Hakobyan
BS in Data Science
American University of Armenia

Supervisor: Gevorg Minasyan
The Central Bank of Armenia

***Abstract*—This paper studies the application of machine learning methods to determine the key parameters influencing default probabilities among bank customers in Armenia. This research project aims to develop a prediction model to minimize the risk of defaults, thereby enhancing risk management processes. By analyzing Armenian Credit Registry' data about borrowers' credit histories, demographic information, and credit scores, the study identifies critical determinants of borrowers' default. The analytical models evaluated in this research include logistic regression, random forest, and XGBoost. Model performance evaluation metrics AUC scores, ROC curves, F1, accuracy scores and confusion matrices are used to assess model efficency. The analysis of the machine learning models identify that XGBoost performs the best in terms of evaluation metrics and effectively identifies the key credit risk determinants. Future research can potentially improve the performance of these models by considering feature engineering and exploring more sophisticated modeling techniques.**

***Keywords*—credit history, consumer loans, risk class, precision, recall, accuracy, learning curves, confusion matrices, roc curve**

I. Introduction

In the recent years, the increased availability of datasets among the financial institutions gave an opportunity to the banks and credit organizations to improve their risk management strategies by analyzing credit risk determinants. The challenge of predicting customer defaults accurately continues to be an important study for those institutions who aim to optimize their credit allocations and minimize their losses. This paper explores the application of advanced machine learning techniques to model and predict default probabilities for consumer loans among bank customers in Armenia.Utilizing the dataset from the Armenian Credit Registry, this study systematically identifies and evaluates the most influential features leading to customer defaults. Given the complexity and variability of factors influencing credit risk, traditional statistical methods fail to capture the patterns among the data. Hence, this research implements three machine learning models logistic regression, random forest, and XGBoost which offer distinct advantages in handling large, diverse datasets with complex relationships among variables. Through a comparative analysis of these models, the study aims to determine which model most effectively predicts default among loans and understand the relative importance of various predictors in the credit scoring process. The effectiveness of these models are evaluated and compared using metrics like AUC scores, ROC curves, F1 scores, accuracy scores, and confusion matrices.

II. Data

A. Data Source - Armenian Credit Registry

For the purpose of analyzing credit risk determinants, the data from the Armenian Credit Registry was used. Credit Registry is a data system that contains all the information about borrowers and loans provided in all of the banks, credit organizations, and resident branch offices of foreign banks in Armenia. Being a member of the Armenian Credit Registry is mandatory. This collected information creates a credit history for each borrower in Armenia. Financial Group is required to submit information to the Credit Registry within 3 business days after the loan contract is signed. The Credit Registry data system contains all the types of loans that generate a monetary obligation. The owner and the user of the Registry's data is the Central Bank of Armenia [1].

B. Data Description

1) Borrower Information: This category includes demographic information about consumers like birthdate, gender, marital status, address, information about education, employment status, income, and also individual bank-specific features like FICO score and bank score.

2) Loan Features: This category includes information about the loan, such as the provided date, maturity date, loan interest rate, and amount.

3) Loan Details: This category includes information about the loan after some behavior, for example, the risk class of a loan, status, first and last classification dates, and actual maturity date.

III. Data Preprocessing

Since the data comes from banks, it includes both mandatory and optional columns. Additionally, besides mandatory

columns that are necessary to submit to the Armenian Credit Registry, banks also have their own set of compulsory columns that differ from bank to bank. However, for all the parameters that are not mandatory, employees populate these optional columns with values that can either carry no meaningful information, be unreal, or contain some random text. Consequently, the dataset contains many null values, missing values that aren't explicitly marked as null, and some values that are not possible to have. Thus, to prepare a dataset and make it suitable for analysis, extensive data processing and manipulations have been applied.

A. Handling Outliers

Outliers are observations that deviate from the expected range and, as a result, produce extremely large residuals. These outliers affect the analysis results; therefore, addressing outliers is an important technique to ensure accurate results [2]. In this dataset, the following variables contained outliers: volume actual (this is the exact amount of the loan that was provided to the customer), age (calculated using birth date and the contract date, which indicates when the loan was issued to the customer), income, previous loans count (was calculated based on consumers' id and date), maturity and family members. To ensure that this dataset is suitable for the analysis, outliers were handled based on their domain and meaning.

Since the 'volume actual' field described the size of loans provided to the consumers in USD, EUR, and AMD, all loan values were standardized into AMD using the daily exchange rates from the Central Bank of Armenia for the period from 2002 to 2021. After standardizing loans to AMD, the histogram1 shows that the distribution of the loans is highly right-skewed. This suggests that while the majority of loans are between 100k and 500k, there are a few very large loans that push the mean higher. The presence of these large loans is unusual for consumer banking, indicating some commercial or special-purpose loans. Thus, these extreme loan amounts were excluded using the IQR method to better reflect the typical consumer loan banking process.

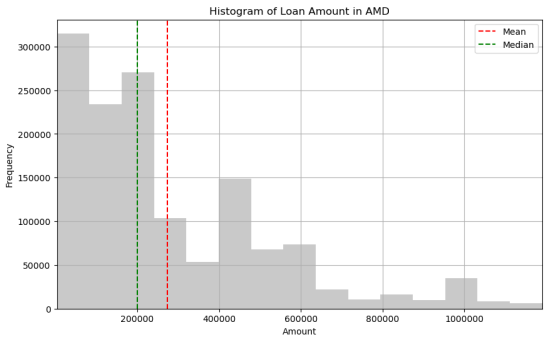

Fig. 1: Histogram of Loan Amount

The age column contains some anomalies and incorrect information. For example, there are records of loan borrowers who are younger than 16 years old but have a marital status of "married," which is inconsistent with the Family Code of the Republic of Armenia, Article 10 [3]. These entries were removed to maintain data integrity. Additionally, due to data entry errors, there are some observations containing age values that are significantly exceeding the normal lifespan. To address these, the IQR method was applied to identify and remove these extreme values, ensuring that the remaining data reflects a more accurate and realistic distribution of ages.

The number of family members column also exhibits some unrealistic values that can be due to data entry mistakes. This is because the information in this column isn't sourced automatically from databases like NORK or ACRA, but rather, it is manually inputted. This manual entry process increases the possibility of inaccuracies. Since these anomalous high values aren't representations of extremely large families but rather inaccuracy, these numbers have been replaced by the mean of family members in Armenia.

The last column for which outliers were removed is the maturity of the loan. The maturity of a loan refers to the length of time over which the loan was scheduled to be repaid. It is calculated from the final due date at which the borrower must pay back the total amount of the principal and any remaining interest. The maturity period of loans can differ based on the type of loan. In this dataset, which focuses on consumer loans, these are typically classified as short-term loans. Since there were very few extreme outliers, they have been removed from the dataset.

B. Creating New Parameters

Based on the existing parameters, several new variables were calculated to further improve the prediction of loan default probabilities. These parameters include age, the number of previously taken loans, and the creation of new columns to manage valuable missing data. These new columns indicate whether a customer has an income, has a FICO score, has had overdue payments, has undergone reclassification, or received a credit score from the bank.

FICO and bank scores are numeric representations of a person's ability to pay a loan based on his/her credit history, demographic information, and other factors. Financial organizations that are giving credit use credit scoring systems to predict the behavior of a consumer. Based on these scores, they determine whether to approve a loan, set the interest rates, and decide the repayment terms for an individual. Each bank can have its own scoring system, but the FICO score is the most widely used and universally accepted scoring system across various credit institutions (Avery & White, 2024). In this dataset, only 18% of the entries had a FICO score, making it impossible to use directly for analysis. However, to utilize the available information, a categorical variable was created to indicate whether a customer had a FICO score or not. This approach can influence the probability of default calculations and provides valuable information. Taking into account whether a loan was issued based on a FICO score can enhance predictive accuracy and reduce risk in lending decisions.

Income also possesses valuable information about borrowers' behavior because individuals with a stable income are generally more likely to meet their loan repayment obligations. In the dataset, only about 18% of the entries have income information. However, since the requirement of customer income varies between banks and for different types of loans, the dataset reflects this inconsistency. Some entries have income data, while others do not for the same person. To address these gaps, missing income information was filled out based on the customer ID. However, there is still inconsistency since it's not possible to identify whether the person has no income or whether the bank did not require that information from the customer. Therefore, another column, named has_income, was created to identify whether there was information about the customer's income when taking up the loan.

Following the same idea two other parameters were created to further clarify creditworthiness of a customer. The first one indicates whether the customer has had any overdue days on the loan, and the second one indicates whether customer's risk class was changed.

C. Categorical Variables

The dataset contains a parameter regarding marital status, which originally included four categories: married, divorced, never married, and widowed (lost husband/wife). However, since over 94% of the entries were classified as either married or not, these categories were consolidated into two—married and single. This simplification helps to smooth the analysis and better reflect the distinctions in the model, enhancing the predictive accuracy and interpretability of the results. Following a similar approach, a newly created parameter that indicates the number of previously taken loans was categorized into three groups to facilitate analysis and interpretation: no previous loans, up to five loans, and more than five loans.

D. Target Variable

To predict whether the loan will go default or not, it is necessary to identify which loan is considered to be defaulted for the bank. For the target variable two possible options were considered which found to be highly correlated with each other. Based on CBA's law on the classification of loans and receivables of banks, loans having the last risk class (5-th risk class) are considered to be fully depreciated whose accounting in the balance sheet as assets is no longer appropriate [4]. Consequently, a default parameter was created where a loan is marked as 1 if it is in the last risk class, indicating a default, and 0 otherwise.

E. Train-test split, standardization

Before running the model it is important to prepare the data for the analysis and make sure that it is suitable for the machine learning algorithms. The data has several categorical columns that have either two categories or more than two. All the categorical columns that had two possible types including marital status (married/single), ispe (yes/no) and gender (female/male), where mapped to 0 and 1. The other columns including bank id, loan type, currency, education, previous loans, year and quarter were converted into dummy variables. As a result for each category of a feature binary columns were created providing a numeric input for the modeling part. Then the dataset was divided into training, validations and testing set using stratified split to address the class imbalance. Afterwards, the numeric columns were standardized using to have zero mean and unit variance. This preprocessing was applied to the training data, and the same transformation was later applied to the validation and test data to maintain consistency.

IV. Exploratory Data Analysis (EDA)

It is important to conduct Exploratory Data Analysis (EDA) before developing the model to get insights and understand the underlying patterns of the dataset. Firstly, the visual of the dependent parameter is depicted to understand the distribution of the default parameter.As mentioned previously, there are two options for defining this parameter: one based on the reclassification date of loans and another based on the 5-th risk class. Figure 2 shows the distribution of y for the second option which is used in the modeling part. It is that the majority of cases (approximately 88%) are classified as non-default (0), and only 11% account for defaulted loans. This shows a clear imbalance in the y parameter. However, this distribution was expected in this dataset since banks aim to minimize risk by primarily lending loans to individuals who less likely to default.

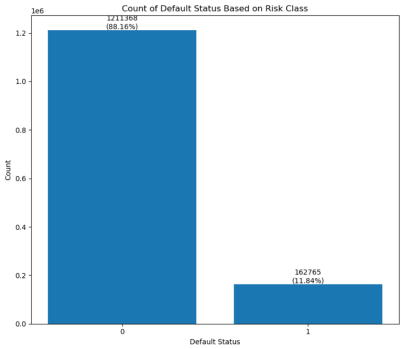

Fig. 2: Distribution of Default Parameter

Additionally, FICO score distribution was considered for the analysis. Overall, FICO score has normal distribution with mean and median close to each other. Figure 3 shows the distribution of FICO score for defaulted and non-defaulted loans.

From the figure it is evident that the median for non-defaulted loans is slightly higher than 600, suggesting that customers who do not default tend to have good FICO scores. The median FICO score for the defaulted customers is visibly lower. Also, the interquantile range for non-defaulted loans is wider.

Figure 4 shows the distribution of loan volume for the two groups. This boxplot helps to compare the size of a

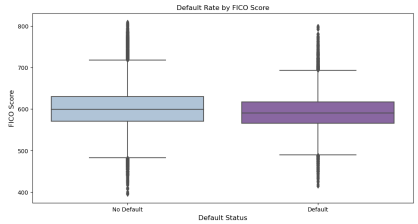

Fig. 3: Distribution of FICO Scores

loan amount taken by customers who defaulted versus those customers who did not. As opposed to usual expectations, the median loan volume for the default group is slightly higher than for non-default group.

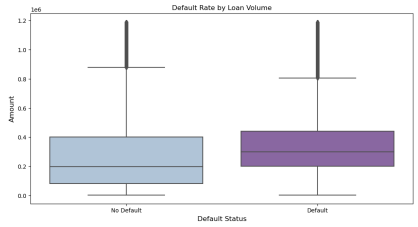

Fig. 4: Distribution of Loan Volume

The visualization suggests that while loan amount does play a role in explaining the default parameter, the distribution is not dramatically different between the two groups.

Lastly, the correlation matrix 5 is depicted to show the relationship between different features in the dataset. It is evident that some features are highly correlated with each other which can lead to the issue of multicollinearity, therefore it is important to eliminate one of the highly correlated variables.

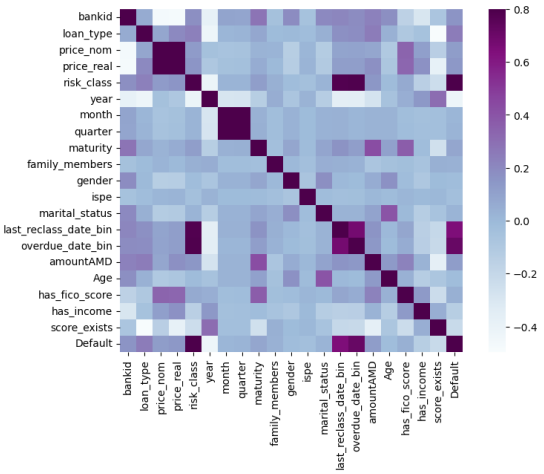

Fig. 5: Correlation matrix

From the matrix we can see that nominal (price nom) and real interest rates (price real) are highly correlated with each other. To address the issue, we only keep the feature of real interest rate. Additionally, from the two parameters of month and quarter that are highly correlated, only the indicator of quarter was kept for the further analysis. Lastly, there are three variables highly correlated with each other: last_reclass_date_bin, overdue_date_bin and the risk_class. Since the default parameter was creating based on risk class, we remove the risk class. And from the two variables in case of which the first one indicates whether the loan has ever changed its risk class, and the second indicates whether there were overdue days related to the loan or not, the last one is kept for the modeling.

V. Modeling

A. Logistic Regression

The first model considered for the analysis is logistic regression. Logistic regression is a supervised machine learning algorithm designed for classification [6]. Firstl, logistic regression model with its default parameters was utilized, leading to 94.52% accuracy on the train set and 95.53% accuracy on the validation set. Then the confusion matrix was defined to get the numbers of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions which showed that model correctly predicted TN and TP, respectively defaults and non-defaults with some small cases of errors. Finally, the classification report was considered to get information about the precision, recall and f1-score of the model. Precision which is the representation of ratio between predicted positive observations to the total predicted positives is 0.751. Recall which shows the ratio of correctly predicted positive observations to all actual positives is 0.79. Afterwards, hyperparameter tuning using Grid Search was done to find the optimal settings for the logistic regression model. With the newly found model AUC (Area Under the Curve) value is 0.97, which is quite close to 1indicating that the model is able to distinguish between the two classes effectively. Figure 6shows the ROC curve of the model where x stands for the False Positive Rate (FPR) and y stands for the True Positive Rate (TPR).

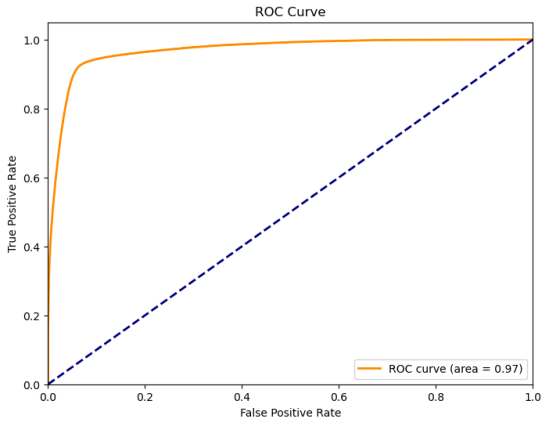

Fig. 6: ROC Curve Logistic Regression

Figure 7 shows the feature importance analysis found with the model. We can see that the among the most impactful in

determining the model's predictions are the age of a customer, whether the bank checked customer's bank score and fico before lending the loan. Also, some external factors had their impact on the prediction process among which are the year, bank, loan type and the type of currency. These suggest that the timing plays a significant role in determining whether the customer is going to pay the loan or not. Also, for some banks it is evident that their customers usually are more likely to pay their loans than customers from other banks. Thus, banks' targeted customer groups differ in their tendency of paying their loans.

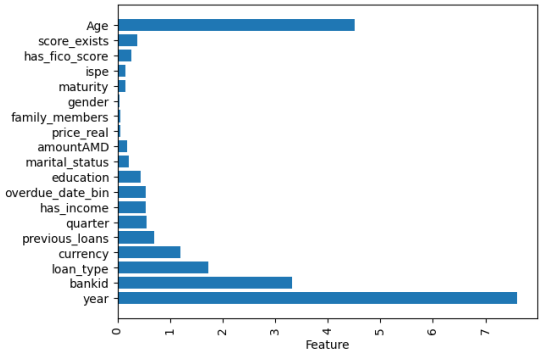

Fig. 7: Feature Importance Analysis Logistic Regression

B. Random Forest

Random forests are ensemble methods for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data [7].

Firstly, Random Forest model was trained on the train data using its default parameter which lead to high level of accuracy (95.76%) on validation with precision of 0.83 and recall of 0.80 on test data. Then using GridSearchCV, model systematically explored some hyperparameters to find the optimal parameters for the Random Forest model. After tuning the parameters, there was a slight improvement i the accuracy, precision and recall scores for the defaulted class.

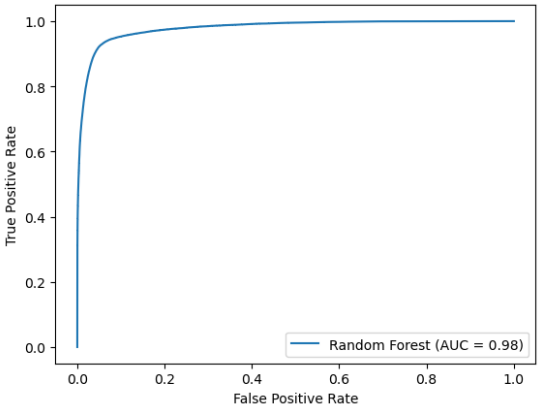

Fig. 8: ROC Curve Random Forest

Figure 8 shows the ROC Curve and indicates that AUC has a value of 0.98 which is higher than in case of logistic regression suggesting that model discriminates between the defaulted and non-defaulted classes. Thus, training, tuning, and validating the Random Forest model led to evaluation metrics' results that show that this model can be considered as a high-performing classifier.

Lastly, feature importance analysis (figure 9) was done for the random forest classifier which highlights credit risk determinants for the default parameter. The consideration of these parameters will lead to minimization of the risks, since the model shows good performance on identifying the probability of the default. Among the features, we can see that the age, the identification of the legal person, interest rate of the loan and the number of family members have high predictive value for the default status.

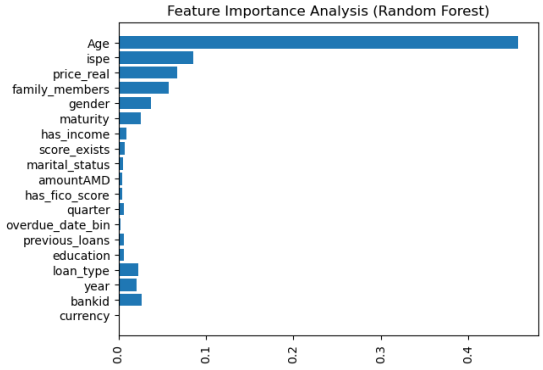

Fig. 9: Feature Importance Analysis Random Forest

C. Extreme Gradient Boosting

Extreme Gradient Boosting is a scalable, distributed gradient-boosted decision tree that provides a parallel tree boosting, and it is the leading machine learning library for regression, classification, and ranking problems [8]. The final model that was implemented is XGBoost. The initial results on the validation set produced 95.76% accuracy, 0.90 average precision and 0.89 average recall. These indicates robust performance on the default classes. Afterwards, hyperparameter tuning was implemented to find the optimal parameters of learning rate, maximum depth and number of estimators for the model. With the tuned parameters, model achieved a validation accuracy of 96.12%. Confusion matrix was also improved compared to the confusion matrices of logistic regression and random forest. Also, from the figure 10 we can see that the model got AUC score of 0.98. This indicates that the model has an excellent predictive capability.

Figure 11 shows the learning curves which are important for understanding model behavior over increasing dataset sizes. These curves indicate that while the training score decreases, the cross-validation score increases with more training data, with a score of approximately 0.96.

Figure 12 bar chart visualizes the relative importance of features determined by the Extreme Gradient Boosting algorithm

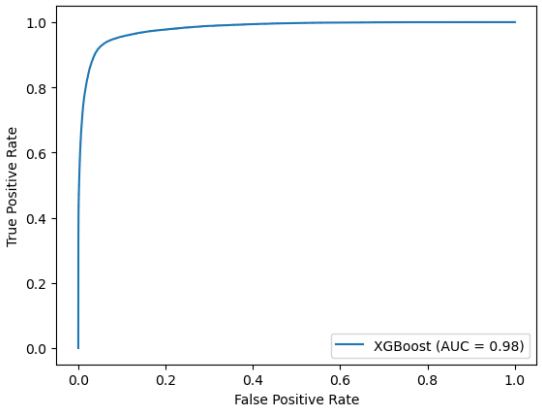

Fig. 10: ROC Curve XGBoost

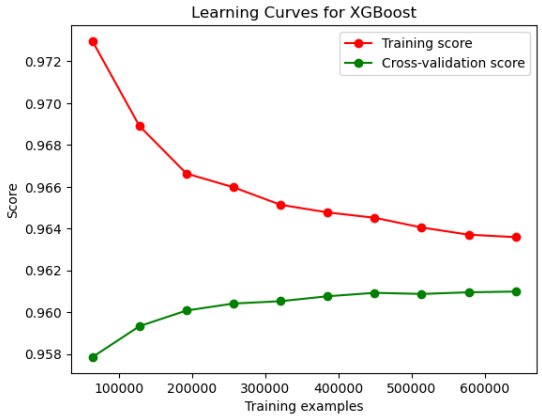

Fig. 11: Learning Curves XGBoost

in predicting the default of a loan. Among the highest importance are age, loan type, bank id and whether the customer is a legal entity. And the features of marital status, indication of having overdue days, number of family members, indication of income, real interest rate and some other parameters show minimal to no importance for this model.

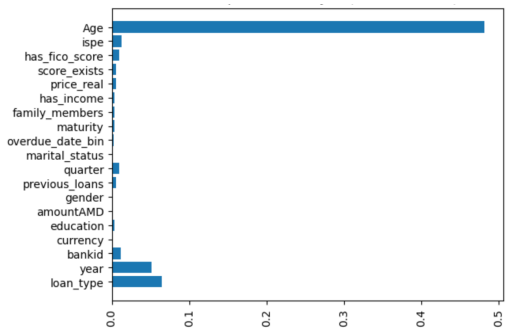

Fig. 12: Feature Importance Analysis XGBoost

VI. Conclusion

Three machine learning models were implemented to find a model with the best predictive ability for the default parameter. Accuracy scores on the test data show that for the logistic regression accuracy was 0.945, for the random forest 0.95 and for the extreme gradient boosting 0.96. Precision scores were 0.751 for logistic regression, 0.83 for random forest and 0.84 for extreme gradient boosting. Also, recall was 0.79 for logistic regression, 0.82 for the random forest and 0.83 for the extreme gradient boosting. AUC score also follows the same pattern having a slight increase from 0.97 to 0.98 for XGBoost. These evaluation metrics show that among the models evaluated, XGBoost emerged as the most effective. Additionally, this model identified the key credit risk determinants determining loan's default. Age is one of the key determinants since it shows financial stability. Younger borrowers often have lower levels of financial stability and income. As they enter middle age, they reach the peak of their earning potential which reduces their default risks. Also, loan types are among the highest determinants since different types of loans carry different levels of risk based on their structure, purpose, and terms. Finally, year also shows significant importance, which shows that the time period when the loan was issued affects the likelihood of default since economic conditions across years influence the probability of default.

References

[1] The Central Bank of Armenia. Procedure "Creation of Credit Registry and Participation of Banks, Credit Organizations, Resident Branch Offices of Foreign Banks in Credit Registry" https://www.cba.am/Storage/EN/regulations/CREATION%20OF%20CREDIT%20REGISTRY.pdf

[2] Sarkar. S. , Midi. H. and Sohel R. (2011). Detection of Outliers and Influential Observations in Binary Logistic Regression: An Empirical Study. Journal of Applied Sciences, 11: 26-35. https://scialert.net/abstract/?doi=jas.2011.26.35

[3] ARLIS. The Republic of Armenia (2013). ABOUT AMENDING THE FAMILY CODE OF THE REPUBLIC OF ARMENIA https://www.arlis.am/documentview.aspx?docid=83374

[4] Avery. D. & White. A. (2024). What is a FICO score and why is it important? https://www.cnbc.com/select/what-is-fico-score/

[5] ARLIS. The Republic of Armenia (2013). ON THE APPROVAL OF THE PROCEDURE "ON CHANGES AND ADDITIONS TO THE CLASSIFICATION OF LOANS AND RECEIVABLES OF BANKS OPERATING IN THE TERRITORY OF THE REPUBLIC OF ARMENIA AND THE FORMATION OF POSSIBLE LOSSES" https://www.arlis.am/DocumentView.aspx?docid=19798

[6] Jurafsky. D. & James.H. M.(2024) Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/5.pdf

[7] Bianu. Analysis of a Random Forests Model. https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

[8] NVIDIA. XGBoost. https://www.nvidia.com/en-us/glossary/xgboost/