Default Credit Predictive Analytics Model to Enhance Bank Decision-Making Using Big Data

Sherif Elsaied E. Gad
Ph.D. Candidate
Sadat Academy for Management Sciences
Doctorsherif2021@gmail.com

Prof. Nashaat Elkhameesy
Department of Computer Sciences
Faculty of Computers and Information
Sadat Academy for Management Sciences

Abstract

In the financial sector, figuring out a person›s creditworthiness may be exceedingly challenging. Individual credit scores or credit report data are a major component of traditional models that determine credit eligibility. If they can receive financial services at all, it will probably be at very high-interest rates and fees. This sort of strategy makes it exceedingly difficult for those with little to no credit to do so. Financial institutions have been looking for methods to include non-traditional data into the credit risk process to provide services to these people at more affordable prices while maintaining the risk at manageable levels. With an accuracy that is acceptable to financial institutions, Machine Learning (ML) could be utilized to estimate a person›s creditworthiness using atypical data. On the data supplied by Home Credit Group, we processed and engineered features. With an AUC of 0.7926 and 92% accuracy, the Light Gradient Boosting Machine (LGBM) model trained using this data can predict the probability of default using 5-fold cross-validation. The system produced promising results and outperformed all other state-of-the-art methods.

Keywords: Machine Learning; Credit Default Prediction; Light Gradient Boosting Machine; Big Data; Financial Sector

1. Introduction

Emerging and developing nations see several financial system advances during the decade, notably in the banking industry. These include digital financial services including agent banking, mobile banking, and online banking. However, a sizable portion of the population is not financially involved [1]. Approximately 63 million Americans are thought to be underbanked or unbanked in the United States [2]. Due mostly to deficient or nonexistent credit history, these people have had trouble establishing bank accounts and obtaining access to financial services like credit cards and loans. Most banks and financial organizations continue to utilize credit determination systems that are based on financial payback history, even though the majority of consumers do not even have credit scores. Therefore, communities with little access to banking services run the danger of being entirely shut out of the financial system, particularly in terms of credit [3].

One of the biggest financial issues facing banks and financial organizations is credit risk, often known as loan default risk. Credit risk is the possibility of a borrower›s inability to pay back a loan or complete the requirements of a contract, leading to a default on the loan or bankruptcy. The borrower›s credit profile, which includes the borrower›s solvency, credit type, maturity, loan amount, and other elements inherent in financial operations, determines the majority of the loan default risk [3].

Modifications to systems that carry out artificial intelligence (AI)-related activities are now referred to as machine learning [4]. Forecasting, robot control, planning, analysis, and recognition are a few examples of these. It explores data analysis and algorithm development for data forecasting. Machine learning is used to generate programs with their tuning parameters. They will thus perform better since they will be responding to early data. Machine learning is a rapidly developing technique that imitates the operation of the human mind. It accurately captures multi-level data and solves the selectivity-invariance issue [5]. Machine learning is used in many areas of life, but mainly in finance [6].

Big Data-based credit systems have begun to take center stage in banks and other financial organizations during the last several decades. By examining additional data about their prospects to improve a traditional score, these businesses aim to enlarge the typical credit criteria. To be more competitive, banks and other lending institutions may fully use big data and exploit huge data sources [7].

Numerous databases are said to contain a wealth of information about a bank›s customers and their financial activity, information that may be utilized to dramatically boost company performance. In reality, data may be one of the most important tools for any bank or financial institution–but only if such organizations can uncover the knowledge that lies inside it and draw conclusions from it. Data may be used to its maximum potential by using technical tools and scientific techniques. In the multidisciplinary discipline of data mining, information and insights are derived from both organized and unstructured data using scientific techniques, processes, algorithms, and systems [8]. In practice, using data science approaches may be crucial for resolving issues facing banks and other financial institutions. This may also be done by using data analysis to build models that keep the best clients at the least expensive prices [9].

Home Credit Group established a Kaggle dataset [10] to broaden the financial options for the segment of the population that lacks access to banking services. The dataset was created to test whether teams could come up with a reliable way to forecast a person›s creditworthiness using different types of consumer data. A financial firm›s success depends on being able to predict whether a consumer will pay back a loan. For this reason, a lot of studies have already been done to create the best credit risk assessment models that are suitable for every single firm. Big data and machine learning are being used by researchers to estimate a person›s creditworthiness with accuracy that is comparable to and sometimes better than previous approaches [11].

In this work, we suggested a model for predicting loan default. On the data supplied by Home Credit Group, we processed and engineered features. Data cleaning from null values and outlier analysis are conducted before data aggregation, which involves merging the input data and cleaning the output data of empty values. With an AUC of 0.7926 and 92% accuracy, the Light GBM machine learning model trained using this data can predict the probability of default using 5-fold cross-validation. The rest of the paper is organized as follows, related work explaining studies that used the same dataset is presented in the second section. The third section explains the proposed system methodology including, the dataset and the procedure of preprocessing analysis and visualization that has been done on the dataset. Also,the third section explains how the dataset is encoded and ML is used for prediction and presents the results achieved with the proposed system. Finally, we conclude the paper in the fourth section.

2. Related Work

To reduce their losses from uncollectible accounts, financial institutions must accurately analyze the credit risks of borrowers. To measure and assess credit risk objectively, they gather borrower data and create many statistical and machine learning methodologies. This topic has been the focus of several studies [11-15] because of its academic and practical relevance.

Gundogmus et al. [12] used customer data and loan outcomes for consumers who asked for loans to develop a model using the Adaboost algorithm. On 67 variables holding the customer›s financial information, we run different data cleaning, feature extraction, and feature selection studies. Based on the target variable, our scoring model was developed utilizing supervised learning and statistical machine learning. Customers› risk scores were assessed, and a variable cutoff value was established following the sample. With our risk ratings, they were classified as Fraud and Nonfraud. They had a 70.8% accuracy rate.

The most recent deep learning framework, DeepGBM, was used by Chen et al. [13]. Two components of the DeepGBM deep learning system, CatNN, and GBDT2NN are used to handle sparse category inputs and dense numerical features, respectively. They made use of the Home Credit Default Risk data collection from Kaggle. In this data collection, they have tested several experimental techniques. These investigations› end findings show that DeepGBM performs with a 0.75 Area Under the Receiver Operating Characteristics (ROC) Curve (AUC).

To improve the predictive performance of credit scoring, Rigo and Yamur [14] build an ensemble classification model based on machine learning methods. First, basic classifiers are chosen and fitted to the data sets, including Logistic Regression, Multivariate Adaptive Regression Splines, Support Vector Machines, Random Forest, and Gradient Boosting. Through these fundamental classifiers, a stacked generalization ensemble model is incorporated second. Four real-world credit scoring data sets are used to assess the model›s performance and efficacy in making predictions. The results showed that, in terms of several performance metrics, the stacking model performed marginally better than the classifiers using a single base model. It had a 0.7461 AUC and 91.94% accuracy on the dataset for home credit default risk.

A credit scoring multi-agent system named «CSMAS» was suggested by Tounsi et al. [15] for the prediction of issues in the credit scoring sector of data mining. This engine creates a data mining method based on the cooperation of intelligent agents utilizing a seven-layer multi-agent system architecture. Preprocessing and data forecasting are the foundation for CSMAS›s success. The initial layer is made to pull any data from different core banking systems, payment systems, credit bureaus, and other databases and data sources and store it in a big data platform. Three distinct subtasks–feature engineering, pre-processing data, and integrating various datasets–are the focus of the second layer. While handling missing values and handling outliers is the exclusive responsibility of the third layer. The amount of features in the initial collection of features is decreased in the fourth layer using dimensionality reduction methods. The fifth layer is used to create a model and generate predictions utilizing the latest iteration of gradient boosting algorithms (XGBoost, LightGBM, and CatBoost). The model›s assessment is planned for the sixth layer. The rating of new credit applicants is done by the seventh layer. A big dataset of Home Credit Default Risk from Kaggle Challenge (307511 records) is used to analyze the performance of CSMAS to estimate the risk of a loan application as a significant issue for banks. CSMAS obtained 92% accuracy and 0.7792 AUC using CatBoost.

Data from the Home Credit Group were processed by Beck [11] and feature engineering was done on the information. Then, using this data, a Light GBM machine learning model that was created by Microsoft [16] and makes use of gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) was trained. It had an AUC of 0.7759 when predicting the chance of default.

Using ensemble techniques based on trees, Egan [17] developed models for predicting credit default. It is shown that employing gradient boosting techniques instead of conventional credit default prediction models may increase model performance. The best XGBoost model is then selected and explained. They suggested a counterfactual extraction approach that is model-independent and explains the motivations behind a certain prediction. The extraction of the counterfactuals with the fewest contrasting attributes is the algorithm›s main goal. The suggested technique has an AUC of 0.7812 and a 71.4% accuracy rate.

For default prediction, Nabil [9] utilized the Home Credit Group dataset. The data was initially checked for missing and null values. The features are then preprocessed, encoded using encoding techniques, and certain characteristics are added using aggregation to produce a bigger dataset. Finally, some features are normalized and scaled. After preprocessing, the data is divided into training and testing sets to train a deep neural network architecture, and the system achieves 0.76 AUC on the testing set while obtaining 87.7% accuracy.

3. Methodology

In this section, the proposed system approach is explained, along with the dataset and the steps taken for preprocessing, analysis, and visualization of the dataset. Additionally, it describes how the dataset is encoded, how machine learning is utilized for prediction, and it presents the outcomes obtained with the proposed methodology.

3.1. Dataset

The data set [10] offered by Home Credit includes information from 7 sources with the following extensions:

- Application train and test tables: both constitute a single primary table that contains data about each loan application at Home Credit, separated into two files, one for the train (with TARGET) and the other for the test (without TARGET). Each loan has a unique id (SK ID CURR) and is arranged in a row. For the training, subset data has a goal that displays the number 0 as a symbol if the loan was repaid and the number 1 as a symbol if it was not.
- Bureau table: This table lists all of the prior loans that the customer has obtained from other lending institutions and that have been reported to the credit bureau (for clients who have a loan in the provided sample). Additionally, there are as many rows for each loan in the example given as there were credit lines held by the customer before the application date.
- Bureau balance table: The monthly balances of prior credits in the Credit Bureau are shown in this table. A single prior credit might contain numerous rows, one for each month of the credit term since each row represents a month of each previous credit reported to the credit bureau.
- POS CASH balance: A monthly overview of the customer›s prior POS (point of sale) and cash loans with Home Credit. Each month of a prior credit in the Home Credit (consumer credit and cash loans) associated with the loans in our sample is represented by a row in this table. By default, a single prior point of sale or cash loan might contain numerous rows.
- Credit card balance: The applicant›s past credit card balances with Home Credit are represented by the monthly information. One row in this table represents a credit card balance for each month, and several rows might be associated with a single credit card.
- Previous application: It comprises all prior Home Credit loan applications submitted by consumers

whose loans are included in the offered sample. The application data allows for many prior loans for each current loan. The feature SK ID PREV serves to identify each preceding application, which also contains one row.

- Installment payments: The application data contains the Repayment history for the prior Home Credit credits associated with the loans. Every payment that was paid has its row, and any missed payments have an additional row. It might be interpreted as meaning that one row corresponds to one installment payment or that one installment corresponds to one preceding credit payment.

Figure 1 depicts the relationship between the data and the files:

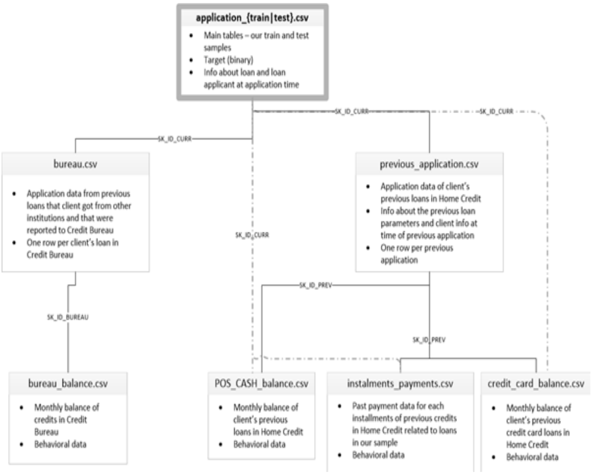
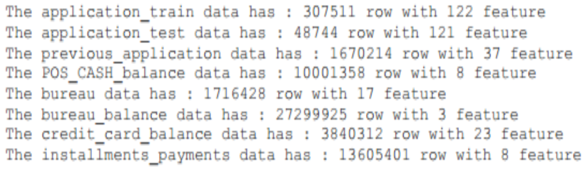

application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
SK_ID_CURR
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_PREV
SK_ID_BUREAU
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Figure 1: Home Credit Default Risk Dataset Organization [10]

The output of the data frame form code offers us a summary of the number of characteristics and records in each table of this data collection, as seen in the following example:



```
The application_train data has : 307511 row with 122 feature
The application_test data has : 48744 row with 121 feature
The previous_application data has : 1670214 row with 37 feature
The POS_CASH_balance data has : 10001358 row with 8 feature
The bureau data has : 1716428 row with 17 feature
The bureau_balance data has : 27299925 row with 3 feature
The credit_card_balance data has : 3840312 row with 23 feature
The installments_payments data has : 13605401 row with 8 feature
```

Figure 2: Dataset Size

3.2. Data Exploration and Visualization

Data exploration and visualization are then shown. Because understanding our data is the primary objective of this research, this portion is just exploratory; as a result, sophisticated analytics are not used in the search for new data or patterns. There are more than 212 variables total across all datasets. It would not add anything to the thesis to visualize and analyze each one, and it would also be difficult to read. As a result, only a limited number of variables will be chosen and covered in this thesis. The choice of variables, however, is not random; it is determined by the target variable›s relevance, the proportion of missing values, and what, in the judgment of the analyst and the subject matter expert, seems to be relevant to understanding the data and the company.

We started by studying the goal characteristic. The goal variable in the application training data set is the repayment status. We can see that the dataset is unbalanced from a straightforward analysis of this variable. 8.07% of the company›s clients had payment issues (encoded 1), which signifies that at least one of the loan›s initial payments was overdue for more than X (number of days). The small proportion demonstrates the company›s success in reducing loans with payment issues. 91.9% of instances fall under the category of «all other cases,» which includes persons who make regular payments as well as those who make no payments at all and maybe additional situations.

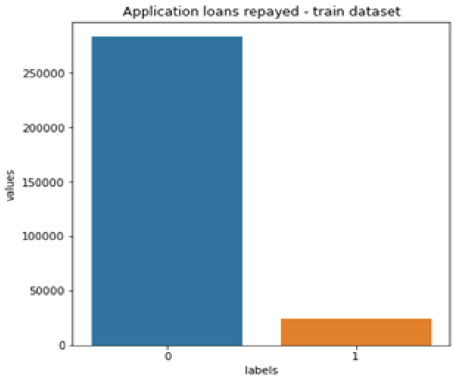


Figure 3: Target Distribution

Figure 4 displays the types of loans taken as well as the percentage of loans (by types of loans) with

TARGET values of 1. (Not returned loan). Revolving loans of the contract kind make up just 10% of all loans; nevertheless, when compared to how often they are repaid, a greater proportion of revolving loans are not.

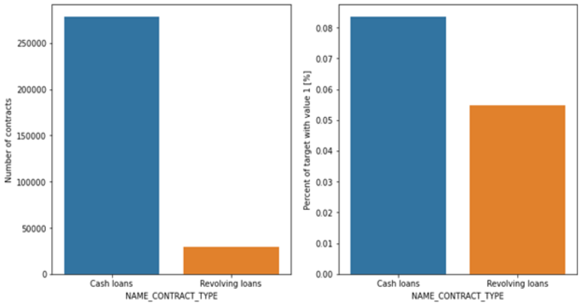


Figure 4: Loan Type

Figure 5 displays the gender of the customers as well as the percentage of loans (by client gender) with TARGET value 1 on a separate plot (not returned loan). The proportion of female customers is approximately two times that of male customers. When comparing the percentage of defaulted credits, men (10%) have a larger likelihood of not repaying their debts than women (7%).

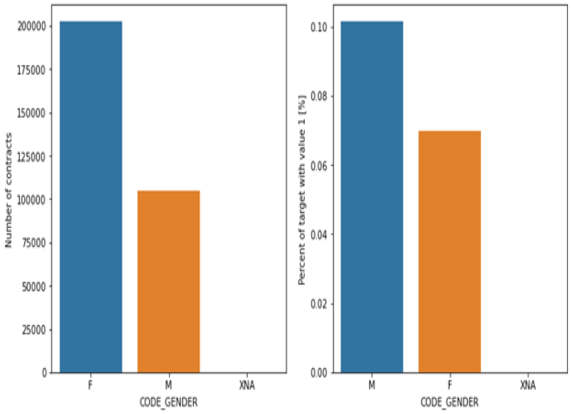


Figure 5: Client Gender

Figure 6 displays the flags that indicate whether a customer owns a vehicle or a piece of real estate as well as, on separate plots, the percentage of the loan›s worth that each of these flags represents (not returned loan). Nearly half of the clientele have a vehicle, compared to those who do not. Clients who own cars are less likely than those who don›t to default on payments for a vehicle. Not-repayment rates for both categories hover around 8%. More than twice as many customers own real estate as those who do not. The not-repayment rates for both groups (real estate owners and non-owners) are less than 8%.

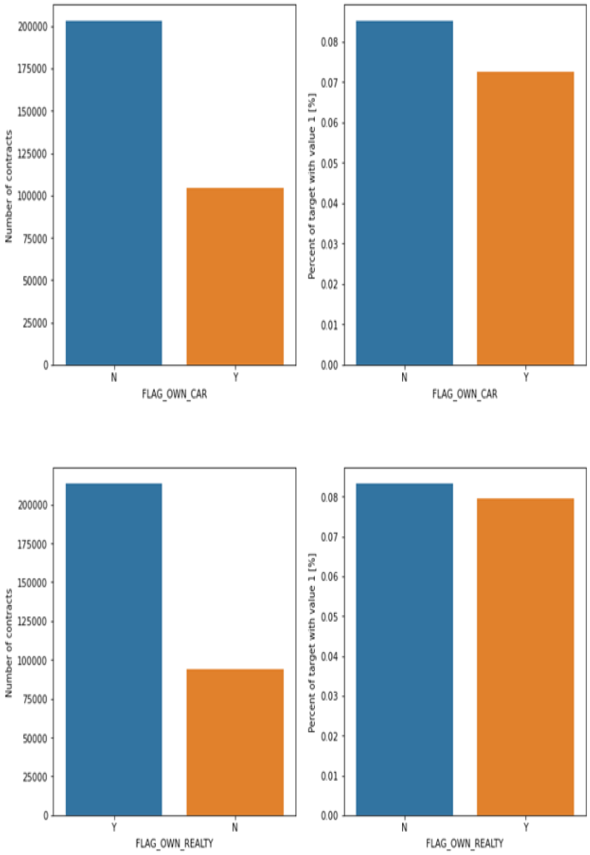


Figure 6: Client Flags

The number of children and customer status is shown in Figure 7. The majority of customers are married, then single or never married, then in civil unions. Civil marriage has the greatest rate of loan defaults (10%), while widows have the lowest percentages (exception being Unknown). Regarding the distribution of client children, the majority of customers applying for loans are childless. Customers with one kid have four times as many loans linked with them, consumers with two children have eight times as many debts associated with them, and clients with three, four, or more children are considerably more uncommon. Clients without children, those with 1, 2, 3, and 5 children all have payback percentages that are around the average (10%). The percentage of unpaid loans for customers with 4 and 6 children is higher than the national average (over 25% for households with 6 children). The percentage of loans that have not been returned for consumers with 9 or 11 children is 100%.

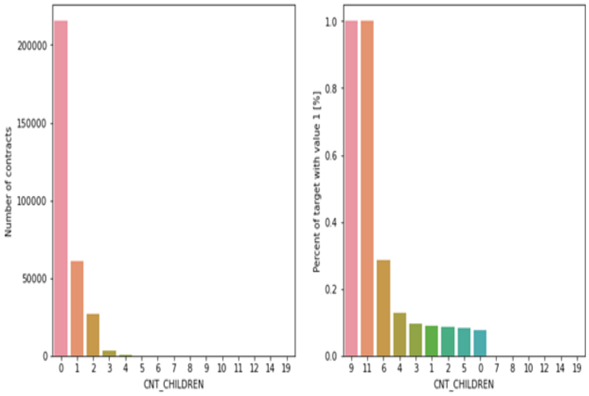


Figure 7: Client status and number of children

Figure 8 shows the number of family members of the client.

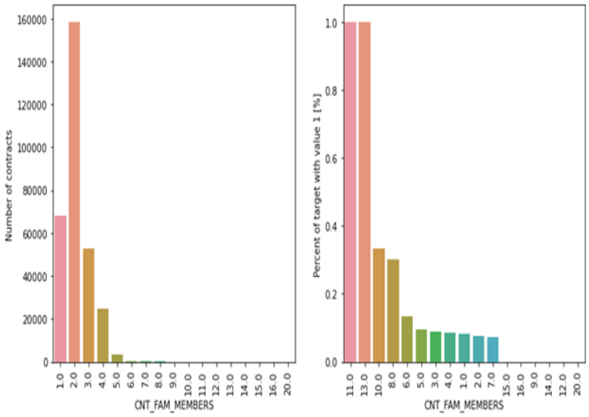


Figure 8: Client family members

The majority of clients had two family members, followed by one single individual, three families with one kid, and four households with four people. Families with 11 and 13 members have a 100% non-repayment rate for clients. Other households with 10 or 8 people have loan default rates that are higher than 30%. Repayment rates for families of six or fewer people are similar to the national average of 10%.Figure 9 investigate the numbers of clients with different income type.

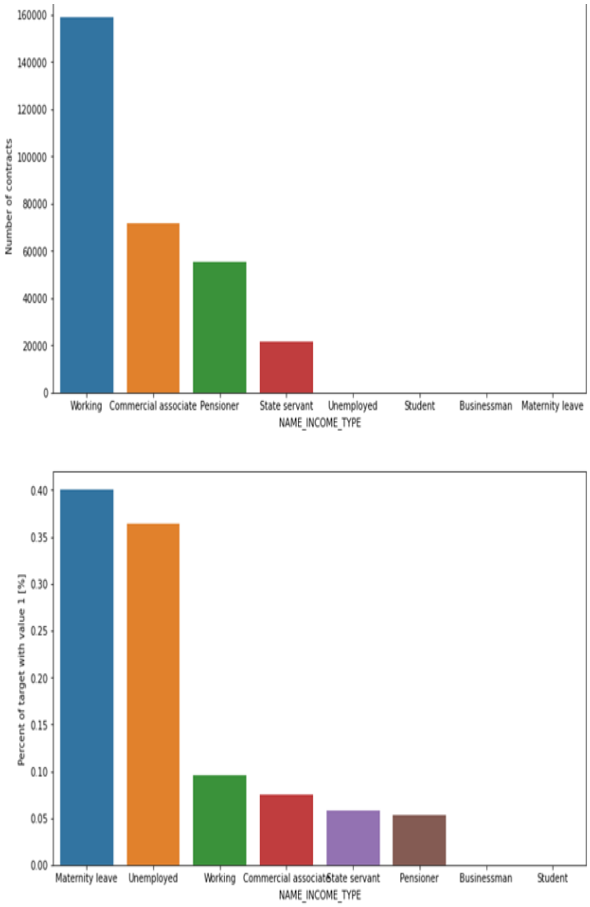


Figure 9: Client Income Type

The percentage of loans that were not repaid as a function of applicants› income types. The majority of loan applicants have employment-related income, followed by associates in business, retirees, and public employees. Maternity leave applicants had a nearly 40% rate of not repaying loans, followed by unemployed applicants (37%). The average loan default rate for the remaining income groups is 10%.Figure 10 investigates the occupation type.

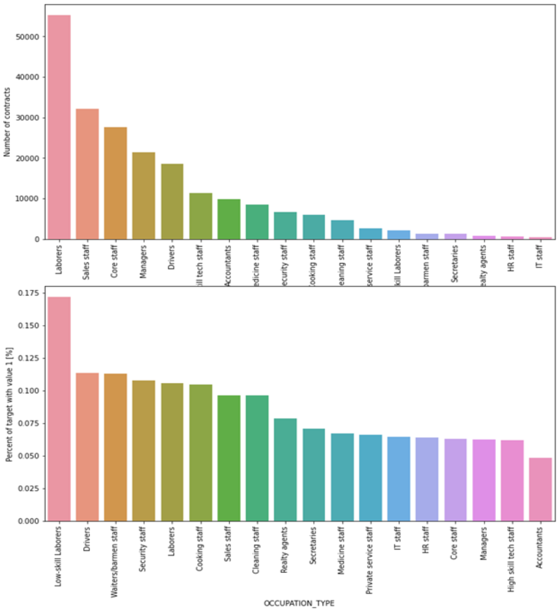


Figure 10: Client Occupation Type

Workers take out the most loans, followed by sales personnel. The least quantity of loans is taken by IT workers. Low-skilled laborers (over 17%) are the group with the largest percentage of unpaid loans, followed by drivers, waiters/bartenders, laborers, security personnel, and kitchen personnel.Figure 11 investigates the organization type.

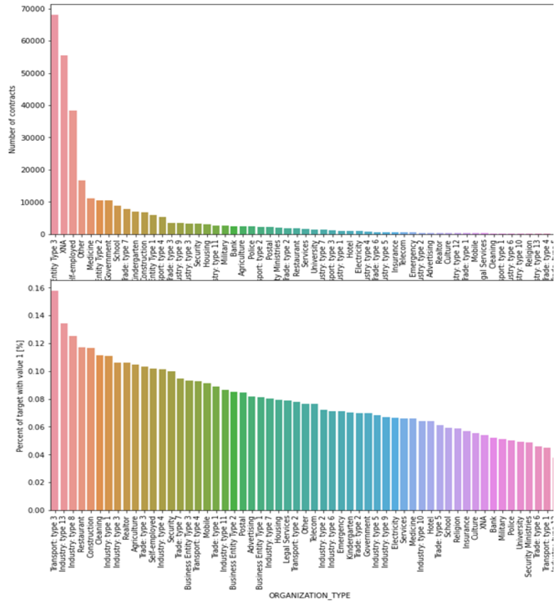


Figure 11: Client Organization Type

Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%), and Restaurant (less than 12%) are the companies with the greatest percentage of loans that have not been returned. Figure 12 looks at the client›s educational background. Secondary or secondary special education is the most common educational background of the clientele, followed by higher education. Possessing a college degree is rare. Although it›s uncommon, the Lower Secondary group has the highest percentage of loan defaults (11%). Less than 2% of those with academic degrees default on their loans.

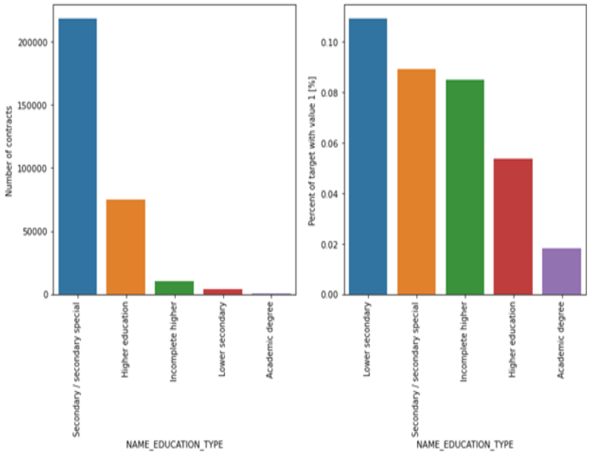


Figure 12: Client Education Type

Figure 13 investigates the housing type of the client.

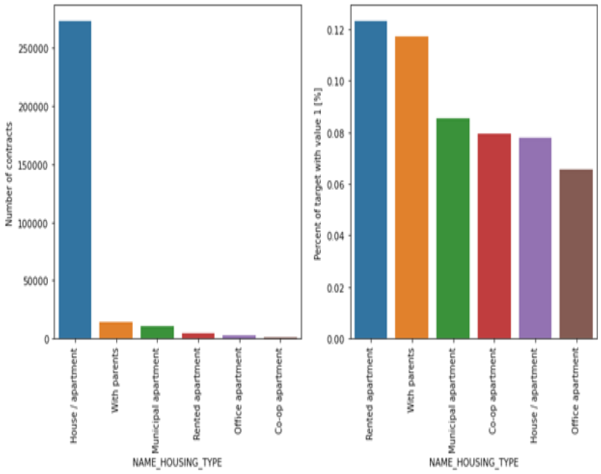


Figure 13: Client Housing Type

More than 250,000 credit applicants listed their place of residence as «House/Apartment.» The following categories have extremely few customers (With parents, Municipal apartments). Renting an apartment and living with parents have greater than 10% non-repayment rates among these groups.

The distribution of credit status is seen in Figure 14. The number of credits for each category is shown first (could be Closed, Active, Sold, and Bad debt). The majority of credits reported to the credit bureau (900K) have the status «Closed.» Active credits are in second place (a bit under 600K). Only a few include sold and bad debt. At the same time, customers with credit recorded to the credit bureau with bad debt have a 20% default on the current applications, with percent having TARGET = 1 from the total number per category. Customers that have credits that are Sold, Active, and Closed have a TARGET = 1 (default credit) percentage that is equal to or less than 10% (10% being the rate overall). Clients with credits listed at the credit bureau with closing credits have the lowest default credit rate. Since the percentage of applications defaulting with a history of Bad debt is twice as high as for Sold or Active and almost three times greater than for Closed, this indicates that the previously registered credit history (as registered at Credit Bureau) is a good predictor for the default credentials.

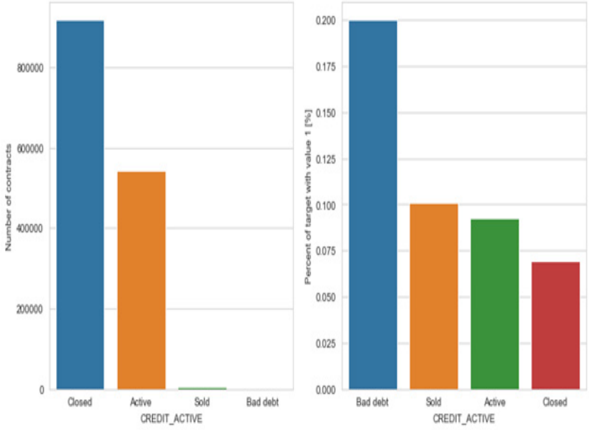


Figure 14: Credit Status Distribution

The total number of credits reported to the credit bureau in various currencies is shown in Figure 15. Additionally, the percentage of defaulting credits (for current applications) was broken down by the various currencies of previous credits for the same customer that was recorded at the Credit Bureau. Most credits are in currency 1. The percentage of defaulting customers varies greatly depending on the currency. The percentage of customers who default starts with currency 3, then moves on to currency 1, then currency 2, and so on. Nearly 0% of applications for customers with recorded credits with currency 4 default.

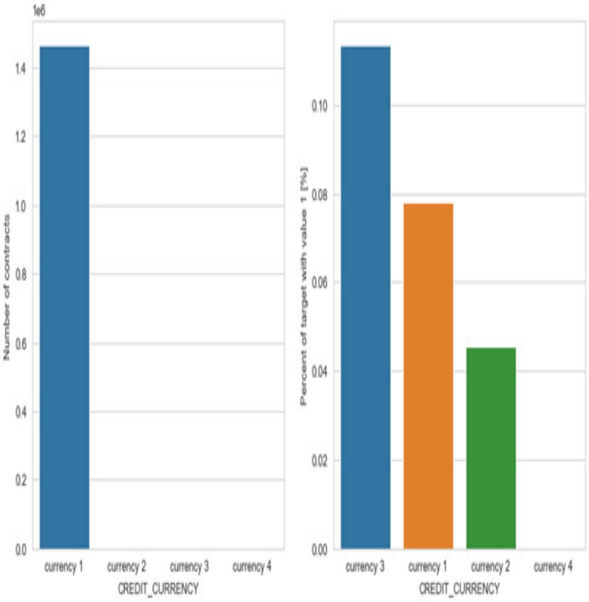


Figure 15: Credit Currency Type

The credit kinds for credits recorded at the Credit Bureau are shown in Figure 16.

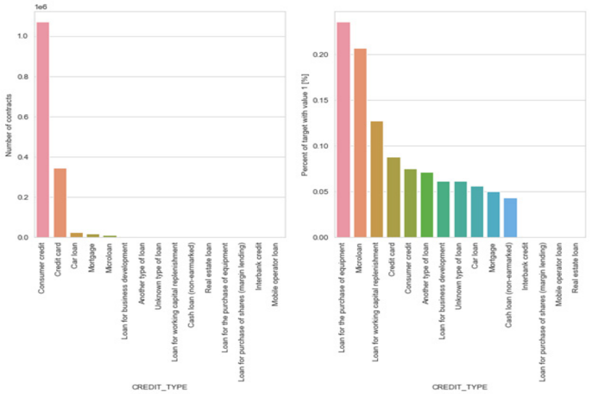


Figure 16: Credit Types

Consumer credit and credit cards account for the majority of historical credits reported to the credit bureau. Car loans, mortgages, and microloans are types of credit that are less common. There are just a few categories of historical credit types with a high percentage of recent credit failures, as shown by the following:

• Microloan - with over 20% current credit defaults;

• Loan for working capital replenishments - with over 12% current credit defaults;

• Loan for equipment acquisition - with over 20% current credit defaults.

Figure 17 examines how the number of credit days is distributed (registered at the Credit bureau). The credit length (in days) is fluctuating between less than 3000 days (with a local high of around 2000 days) and lesser numbers of days with increasing regularity, peaking at about 300 days (or less than one year).

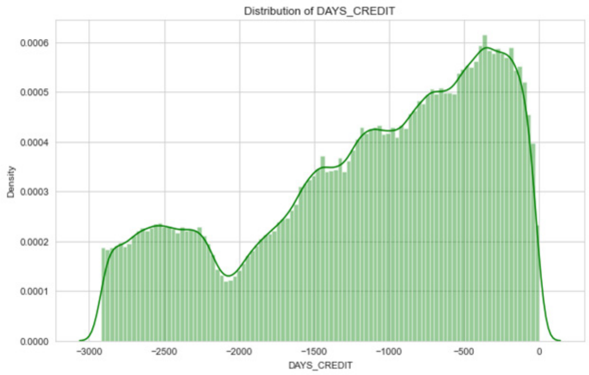


Figure 17: Credit Duration

We are now looking into the applicants› clientele. In the instance of cash loans, Figure 18 explores the purpose of the loan. The majority of contracts are for repairs, other urgent requirements, purchasing a used automobile, building a home or an addition, and not identified/not available categories. Clients with a history of prior applications had the highest percentages of defaults when those prior applications were for cash loans to refuse to specify the aim, which is 23% (which makes perfect sense), hobbies (20%), and car repairs (which is 18%).

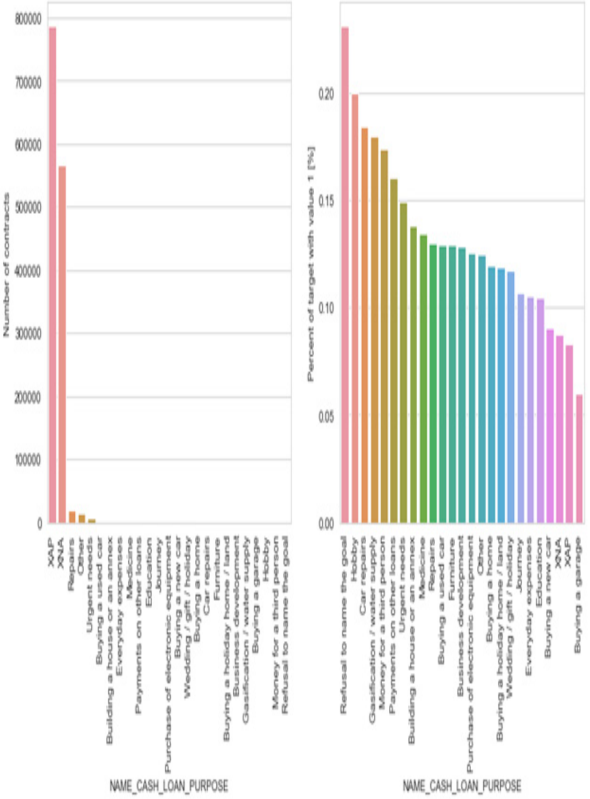


Figure 18: Cash Loan Purpose

Figure 19 shows the status of the previous application.

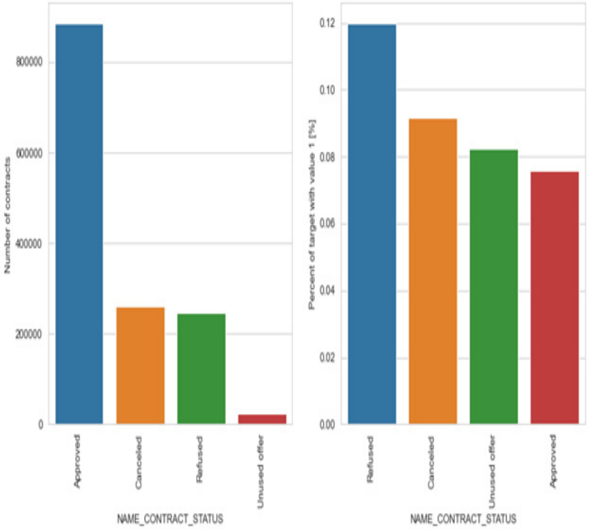


Figure 19: Previous Application Status

The most common contract statuses for earlier applications are Approved (850K), Canceled, and Refused (240K). There are just 20K offers with the status Unused. Clients who have a history of prior applications have the highest percentages of defaults when their contract statuses in the past have been Refused (12%), Canceled (9%), Unused offer (8%), and Approved (lowest percentage of defaults in current applications, with less than 8%).

3.3. Prediction

Now that we have got some more understanding of the data, we go furthermore and do some data preparation. First, we removed rows with values not present in the test set which are ‹CODE_GENDER› = ‹XNA›, ‹NAME_INCOME_TYPE›= ‹Maternity leave› and ‹NAME_FAMILY_STATUS› =›Unknown›. Next data cleaning is done by removing empty features. Then outlier analysis is done to remove some outliers. We must convert category variables into numerical values since ML models need numerical input. Categorical features with binary values are labeled encoded to 0 or 1. These are ‹CODE_GENDER›, ‹FLAG_OWN_CAR›, and 'FLAG_OWN_REALTY›. The rest of the categorical features are one hot encoded [18].

The next step is data aggregation [19]. Combining two or more attributes (or objects) into a single attribute is referred to as aggregation (or object). Aggregation›s goal is to minimize the number of objects or characteristics. Since there are several loans associated with each applicant›s ID, aggregating our data was required for each table. For instance, in the Bureau and Bureau balance data we have, each row represents an old loan that is connected to the current loan by an ID (for example, SK ID BUREAU). To assess the average, total, maximum, and lowest values for the loans of each unique customer ID, numerical variables were aggregated. These aggregations will result in the creation of new columns.

The new dataset after aggregation need also to be cleaned. We did cleaning steps by removing empty features. Also, feature reduction is done and the final list of features after these operations is 1762 features. Training data has 307500 instances while the testing set has 48744 instances. But the dataset that is originally split into training and testing is unbalanced so we needed to consider that while training so we used 5-fold cross-validation [20] to solve the imbalanced class problem. AnLGBM [16] algorithm is used for training and testing and the parameters are then fine-tunedto achieve better results. The proposed system achieved 92% accuracy and 0.792675 AUC. Table 1and Figure 20 shows a comparison between the proposed system and other state-of-the-art systems that used the Home Credit Default dataset. The system outperformed all the state-of-the-art techniques and achieved promising results.

Table 1: Proposed Model Performance Comparison

Method	AUC	Accuracy (%)
Rigo and Yağmur [14]	74.61	**91.94**
Tounsi et al. [15]	77.92	92
Egan [17]	78.12	71.4
Nabil [9]	76	87.7
Proposed System	**79.26**	**92**

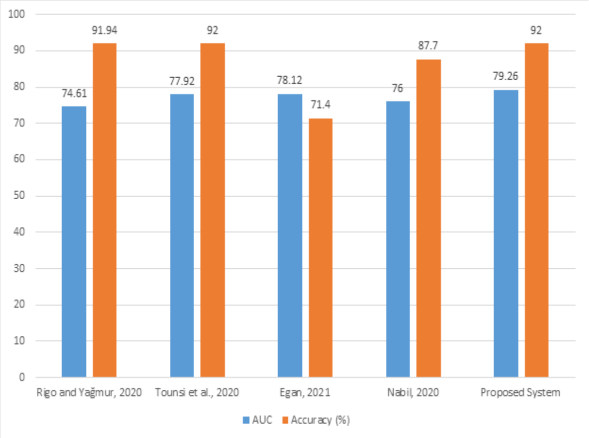


Figure 20: Proposed Model Comparison

4. Conclusion

Machine learning technology has recently advanced quickly in the field of credit rating. In this work, we suggested a model for predicting loan default. On the data supplied by Home Credit Group, we processed and engineered features. Data cleaning from null values and outlier analysis are conducted before data aggregation, which involves merging the input data and cleaning the output data of empty values. With an AUC of 0.7926 and 92% accuracy, the LGBM machine learning model trained using this data can predict the chance of default using 5-fold cross-validation. The system produced promising results and outperformed all cutting-edge methods. To increase prediction performance in credit scoring, deep learning algorithms will be integrated with basic classifiers and ensemble models.

References

[1] M. Qamruzzaman and W. Jianguo, "Financial innovation and economic growth in Bangladesh," Financ. Innov., 2017, doi: 10.1186/s408540-0070-017-.

[2] F. D. I. Corporation, "2017 FDIC national survey of unbanked and underbanked households." Federal Deposit Insurance Corporation Washington, DC, 2018.

[3] M. Alam, "Risk prediction of loan default using knowledge graph," 2022.

[4] J. P. Simon, "Artificial intelligence: scope, players, markets and geography," Digit. Policy, Regul. Gov. , 2019, doi: 10.1108/DPRG-080039-2018-.

[5] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," Electron. Mark., 2021, doi: 10.1007/s125252-00475-021-.

[6] U. Kose, "Using artificial intelligence techniques for economic time series prediction," in Contemporary Studies in Economic and Financial Analysis, 2019. doi: 10.1108/S1569375920190000101002-.

[7] K. Mungai and A. Bayat, "The impact of big data on the South African banking industry," 2018.

[8] V. Dhar, "Data science and prediction," Commun. ACM, 2013, doi: 10.11452500499/.

[9] A. Nabil, "Data Science in FinTech: credit risk prediction using Deep Learning," ETSI_Informatica, 2020.

[10] Home Credit Group, "Home Credit Default Risk DataSet," Kaggle, 2018.

[11] P. Beck, "Predicting Loan Default Likelihood Using Machine Learning," 2021.

[12] Y. E. Gundogmus, M. Nuhuz, and M. Tez, "Risk-based Fraud Analysis for Bank Loans with Autonomous Machine Learning," in y-BIS 2019 Conference Book: Recent Advances n Data Sc ence and Bus ness Analyt cs, 2019, p. 143.

[13] X. Chen, X. Liu, Z. Liu, P. Song, and M. Zhong, "A deep learning approach using DeepGBM for credit assessment," 2019. doi: 10.11453366194.3366333/.

[14] Dall›asta Rigo, Elif Ya□mur. Evaluation of stacking for predicting credit risk scores. Diss. Applied Data Science, 2020.

[15] Y. Tounsi, H. Anoun, and L. Hassouni, "CSMAS: Improving Multi-Agent Credit Scoring System by Integrating Big Data and the new generation of Gradient Boosting Algorithms," 2020. doi: 10.11453386723.3387851/.

[16] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," 2017.

[17] C. Egan, "Improving Credit Default Prediction Using Explainable AI," Dublin, National College of Ireland, 2021.

[18] Rodríguez P, Bautista MA, Gonzalez J, Escalera S. Beyond one-hot encoding: Lower dimensional target embedding. Image and Vision Computing. 2018 Jul 1;75:2131-.

[19] Boubiche S, Boubiche DE, Bilami A, Toral-Cruz H. Big data challenges and data aggregation strategies in wireless sensor networks. IEEE access. 2018 May 3;6:2055871-.

[20] Fushiki T. Estimation of prediction error by using K-fold cross-validation. Statistics and Computing. 2011 Apr;21(2):13746-.