CREDIT RISK EVALUATION USING LOGISTIC REGRESSION AND MACHINE LEARNING TECHNIQUES

Le Ngoc Anh

Abstract

One major problem that financial organizations, especially those in the banking industry, face is credit risk. One of the ways that can be done as an evaluation of the risk profile to handle this risk is building a quality scorecard model and default risk prediction. Logistic regression is the most common method due to its flexibility with many kinds of datasets. However, using this algorithm alone often does not give optimal results. In our research, a more efficient method is introduced machine learning techniques like Boosting Models and Tree-based model alongside logistic regression. Besides that, Weight of Evidence and Backward Elimination are also used in combination with Logistic regression for feature selection. Our proposed credit risk model is able to predict up to 97,1% AUC Score.

Keywords: credit risk, logistic regression, Adaboost, credit score

1. Introduction

After the 2007 and 2008 financial crisis, credit risk analysis had gain exposure and started to be put into practice by many banks. Financial institutions are required by Basel II to measure their exposure to credit risk using internal ratings, leading to banks improving their credit scoring systems, as small improvements could lead to significant profits.(Hand & Henley, 1997).

Logistic regression as well as traditional models (causal inference (Athey and Imbens, 2019), econometric analysis (Angelini et al., 2008)) are effective for financial problems. However, they might not be sufficient for huge datasets with complex features relationships. Therefore, machine learning algorithms are recommended as they are powerful dealing with such datasets (Varian, 2014).

Markov et al. (2022) conducted a systematic study wherein they examined 150 articles published between 2016 and 2021 to identify patterns in credit scoring techniques. The paper advances knowledge of the numerous statistical and machine learning methods used at different phases of credit rating. According to their research, sophisticated approaches like ensembles and neural networks are becoming more and more popular and has better performance than logistic regression.

In this research, we explore the use of both statistical and machine learning approaches to assess credit risk in the American Express dataset, which includes data on the demographics, and repayment histories of credit card customers. In addition to using logistic regression, we investigate ensemble techniques such as AdaBoost and XGBoost in order to determine the most effective model for loan default prediction. Through feature

selection procedures, the paper seeks to determine the most suitable model for credit scoring and risk analytics to give insightful information for the business.

2. Credit Risk Rating: Selected Algorithms

Logistic regression (Maddala, 1992), also known as Logit, is commonly used to estimate the performance of a data point that may belong to a certain class (for example, the loan whether it can be paid or not). In case the threshold of the model is set to 0.5, when the final prediction of a datapoint is > 50% (equals to the threshold), the model can confirm that the datapoint belongs to class 1 (label is 1) or vice versa (label is 0). This task is called binary classification (only 0 and 1). This model is widely applied in credit risk analysis, based on factors affecting the customer's credit capacity. The form of Logistic regression is:

$$\text{Log}\ \left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Random forests is a supervised machine learning proposed by Breiman in 2020, which constructs an ensemble of decision trees chosen at random to serve as a predictor. In Breiman's approach, Random Forest model uses a random sampling technique to select data points and features from a dataset. Each decision tree is constructed using n random records and m features, with the same sample selected multiple times. All trees are evaluated independently, and the final output is determined using Majority Voting or Averaging for classification and regression, with the most voted prediction result being the outcome.

AdaBoost algorithm (Freund and Schapire ,1995), takes as input a training set (x_1, y_1),…(x_m, y_m) where each x_i belongs to some domain or instance space X, y_i is in some label of set Y. AdaBoost repeatedly train base learning algorithm t times with all examples are set weights equally, after each round, weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set.

The Gradient Boosting method has the same idea as AdaBoost, which is to continuously train weak models. But we do not use the model error to calculate the weights for the training data but use the residuals. Starting from the current model, we try to build a decision tree that tries to fit the residuals from the previous model. The special feature of this model is that instead we try to match the target variable value then we will find a way to match the error value of the previous model. We will then add the training model to the prediction function to gradually update the residuals. Each decision tree in the model chain is very small in size with only a few decision nodes determined by the depth parameter.

Popular GBDT models like LightGBM, CatBoost (Category Boosting), and XGBoost (Extreme Gradient Boosting) each have special benefits. A popular model with a reputation for effectiveness and adaptability is XGBoost. It uses a strong optimization method to perform better across a range of jobs. LightGBM prioritizes efficiency and speed.

Meanwhile CatBoost handles these features with a unique method that improves interpretability and accuracy. Additionally, it encourages balanced trees, which are less likely to overfit and can be computed more quickly.

2.1. Preprocessing Techniques

In binary classification problems, data imbalance is a prevalent concern, particularly in credit risk. An imbalanced dataset has an equal or greater than 60:40 disparity in data between the two groups, whereas a balanced dataset has a 50:50 data ratio.

When working with class-imbalanced data, preprocessing is crucial. Two effective methods for preprocessing feature selection and resampling. While feature selection is carried out in variable space by removing useless features, and continue to process the data with features that contribute to the model prediction only, clarifying the pattern of the class-imbalanced data.

2.1.1. Feature selection

In Anderson, 2007 and Witten and Frank, 2005 introduced Weight of Evidence (WOE) for transformation of nominal attributes into each bins with continuous values. WOE measures how good each grouped attributes can predict the desired value of the dependent variable. The formula of of Weight of Evidence is:

$$\text{WOE} = [\ln(\frac{\%Good}{\%\text{Bad}})]$$

WOE and IV (Information Value) have a close relationship. IV is a data exploration strategy that ascertains which variable in a data set influences or has predictive power over a certain dependent variable's value (0 or 1). The formula of Information value is:

$$\text{IV} = \sum(\%\text{good} - \%\text{bad}) * \text{WOE}$$

Information Value (IV)	Predictive Power
< 0.02	Useless
0.02 - 0.1	Weak predictors
0.1 - 0.3	Medium predictors
0.3 - 0.5	Strong predictors
> 0.5	Suspicious

Table 1: *Predictive power based on IV*

According to Siddiqi (2006), if IV < 0.02, the independent variable has no relationship with the dependent variable; From 0.02 to 0.1, they do not have too close a relationship with each other, and the independent variable is considered a weak predictors; only when IV ≥ 0.3, independent variable has a very close relationship with the dependent variable.

Backward Elimination recursively determines the optimum feature combination from a set of possible combinations based on p-value. After each loop, a feature will be removed if its p-value is greater than the significance level (0.05 in this research). Until only features with a p-value of less than 0.05 are kept, this iterative process must be repeated.

2.1.2 Resampling

To address class imbalance, random resampling of the training dataset can be used. Undersampling, which removes samples from the class that has more record (majority class), and oversampling, which duplicates examples from the minority class, are the two basic strategies.
SMOTE (Synthetic Minority Oversampling Technique) is an improved version of oversampling, which creates new instances by interpolating between positive examples while concentrating on the feature space. This approach, which was first presented in 2002 by Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, is an improved version of the oversampling method. This method enhances model performance and produces more accurate predictions as it reduces bias and captures significant characteristics of the minority class.

2.2 Metrics

The main evaluation metrics in this study include AUC value, F1 value, accuracy.
The confusion matrix is used to establish four variables:

(i) TP, which represents the number of positive samples that are correctly predicted by the classification model

(ii) FN, which represents the number of positive samples that are incorrectly predicted as a negative class by the model (also known as Type II error)

(iii) FP, which represents the number of negative samples that are incorrectly predicted as a positive class by the model (also known as Type I error)

(iv) TN, which represents the number of negative samples that are correctly predicted by the model.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

The "Area Under the Curve," or AUC, is a critical performance indicator for binary classifiers. Plotting the true positive rate versus the false positive rate at various threshold values, it shows the area under the Receiver Operating Characteristic (ROC) curve. A random classifier has an AUC of 0.5 and a perfect classifier of 1.

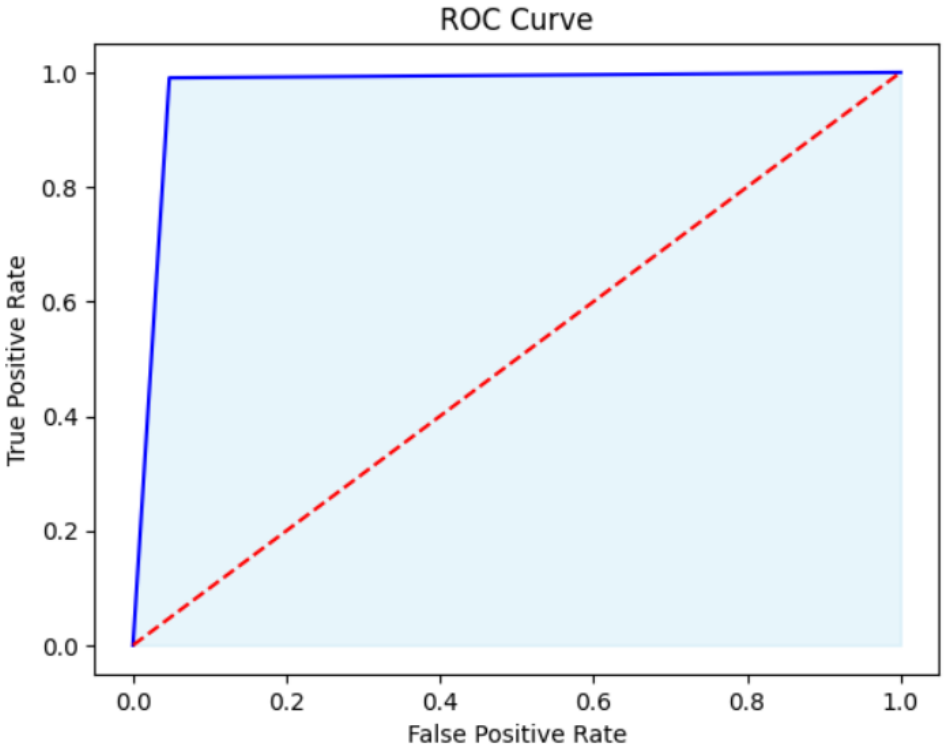


Figure 1. The higher the AUC value, or the closer the ROC curve is to the upper left corner, the better the classifier's

Through harmonic mean, the F1-score provides a synthetic measure of precision and sensitivity. This measure outperforms accuracy in the situation of unbalanced datasets.

The F1-score computation method is shown in the following equation:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3. Data*sets*

The data for this study originated from the "AmExpert 2021 CODELAB - Machine Learning Hackathon," a competition hosted by American Express on HackerEarth, a platform for coding challenges. This dataset reflects real-world data used by American Express, a financial services company offering various payment products and services. While the original dataset contained 45,528 observations across 19 features, this study utilized a subset of 30,000 rows while maintaining all 19 features for analysis.

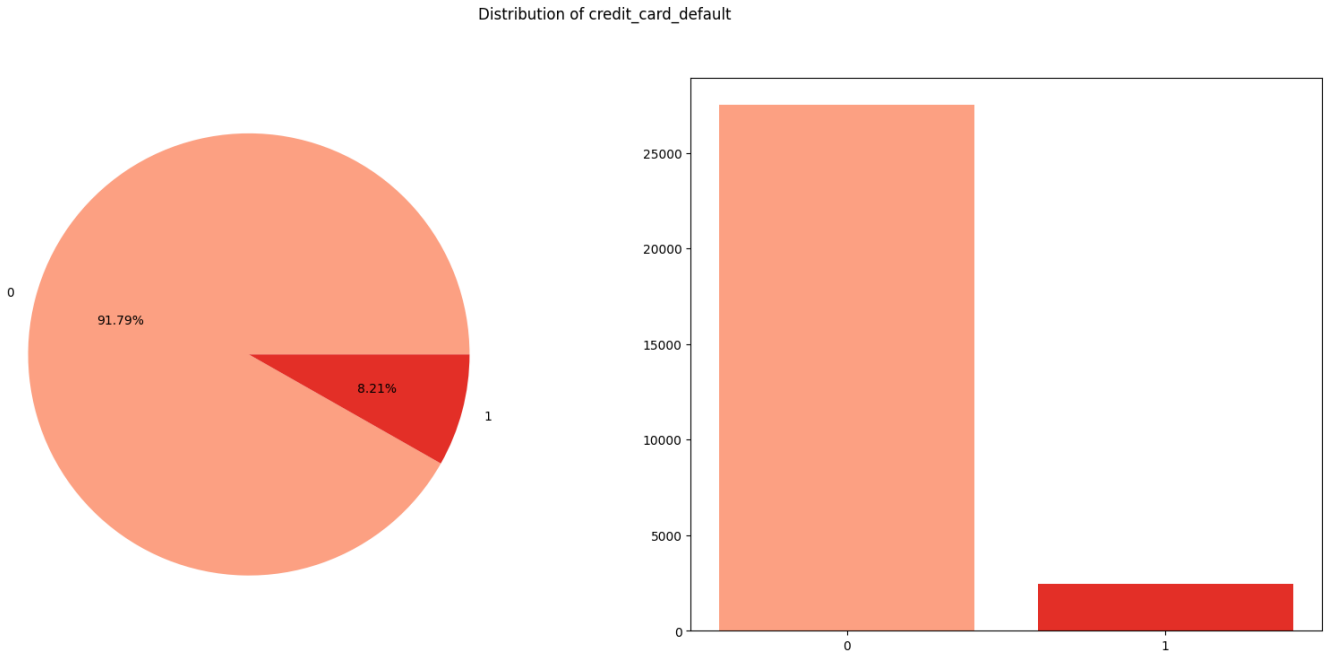
Table 1: AmExpert 2021 dataset

Variable	Meaning	Type
customer_id	Customer ID	object
name	Name of customer	object
age	Age of customer	int
gender	Gender of customer: Female (F), Male (M)	object
owns_car	Whether own car or not? (T/F)	object
owns_house	Whether own house or not? (T/F)	object
no_of_children	Number of children	float
net_yearly_income	Net yearly income	float
no_of_days_employed	Number of days employed	float
occupation_type	Occupation type	object
total_family_members	Total family members	float
migrant_worker	Is migrant worker or not? (0/1)	float
yearly_debt_payments	Yearly debt payments	float
credit_limit	Credit limit	float
credit_limit_used(%)	Percentage of credit limit used	int
credit_score	Credit score	float
prev_defaults	Number of previous defaults	int
default_in_last_6months	Whether default in last 6 months or not? (0/1)	int
credit_card_default	Target variable: Credit card default or not? (0/1)	int

4. Exploratory Data Analysis (EDA)

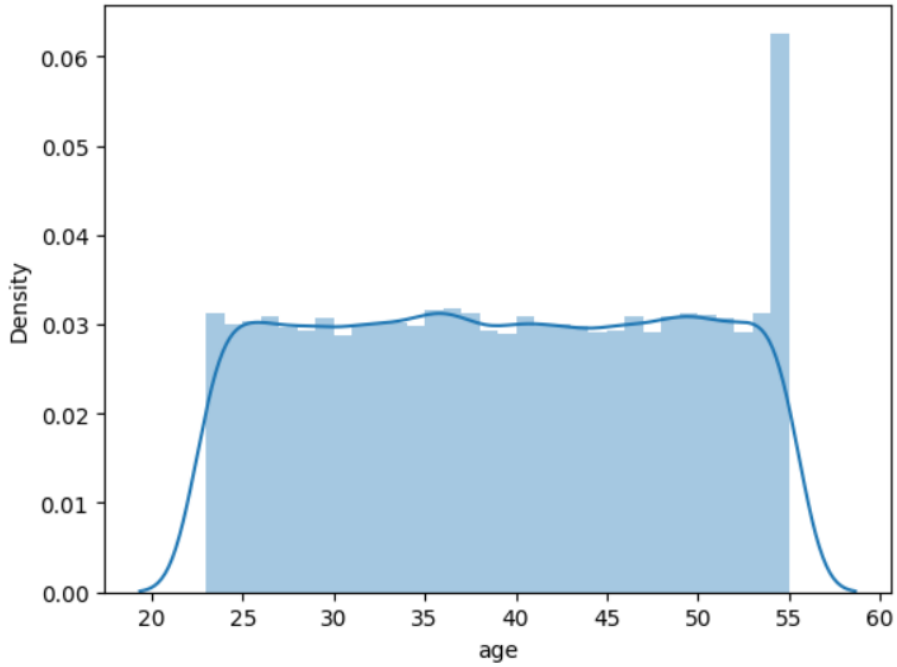
4.1 Univariate analysis

Figure 2. Distribution of target



When examining the "credit_card_default" target variable, we observe that this dataset is highly imbalanced with the ratio 92:8 (~92% card is default) raising the need to use resampling techniques like SMOTE.

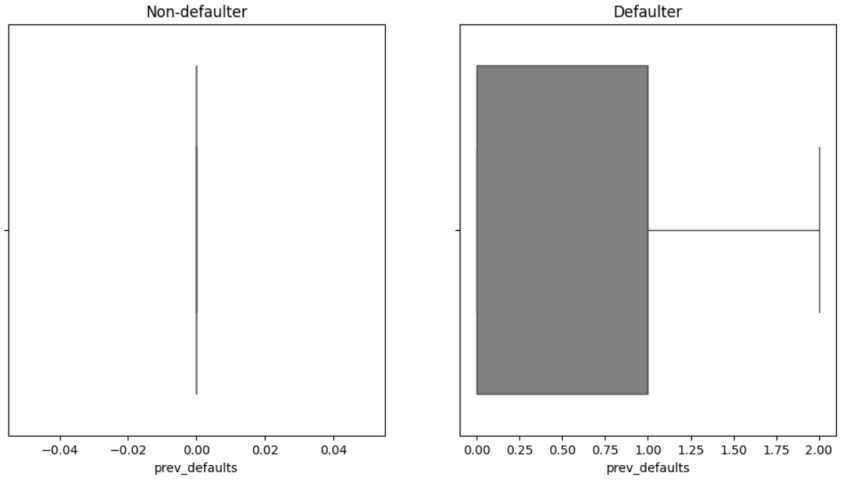
Figure 3. Histplot of age



Customers in the sample range in age from 23 to 55. This indicates that each of them is of working age and most likely held a job at the time of the credit loan.

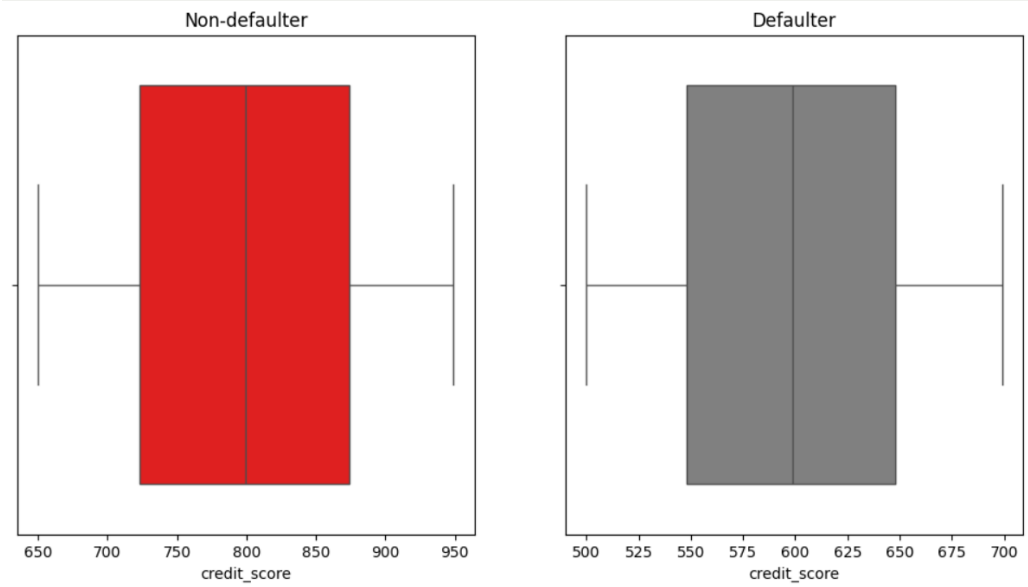
4.2 Bivariate analysis

Figure 4. Boxplot of prev_defaults on credit card default



Individuals with no prior defaults have a history of fulfilling their financial obligations, suggesting a higher chance of paying in full again. Therefore, individuals with more previous defaults are more likely to default again compared to those with no prior defaults.

Figure 5. Boxplot of credit_score on target



Individuals with high credit score, ranging from 725 to 875 are likely keep their credit utilization ratio (amount owed divided by credit limit) low and have a history of making payments on time. This suggests a lower risk of default compared to those with no defaults (higher credit score).

5. Data Preprocessing

Following EDA, we see that most of the missing values are less than 2%. Therefore, we impute all missing values by filling in the null in the numerical features by median and the categorical variable by mode.

In column “gender”, there is only one outlier with values “XNA”. Since the name of customer in this observation is “Ernard”, which is exotic and can be hard to distinguish between Male and Female so we decided to drop this row.

i. *Features selection with IV & Backward Feature Elimination*

The IV computations for continuous variables were performed after grouping them into bins, and the results are presented in the table below:

Table 2: Information value and its rank compare with threshold 0.02

column	IV	rank
credit_score	1.821389	suspicious
credit_limit_used(%)	0.865080	suspicious
prev_defaults	0.702775	suspicious
default_in_last_6months	0.600895	suspicious
no_of_days_employed	0.133247	Medium
occupation_type	0.096320	Weak
gender	0.048503	Weak
yearly_debt_payments	0.028216	Weak
age	0.015057	Useless
net_yearly_income	0.013038	Useless
migrant_worker	0.012561	Useless
no_of_children	0.008318	Useless
credit_limit	0.007721	Useless
total_family_members	0.006816	Useless
owns_car	0.004524	Useless
owns_house	0.000070	Useless

Since we need to validate Logistic Regression on the original dataset first before fitting any Machine Learning models for better comparison, variables ranked as “Useless” will be eliminated before creating WOE features

ii. *Data transformation*

We also use binary encoding for the “gender” variable and using get_dummies method from Pandas to make dummies variables for “occupation_type”.

Our dataset has columns in different forms with different range, for example credit_limit_used(%) ranging from 0 to 100, credit score ranging from around 600+. Maruma et al. (2022) state that MinMaxScaler is frequently applied in credit risk analysis to guarantee that variables in model are equally scaled. Thus, we utilize Sklearn's MinMaxScaler on our dataset.

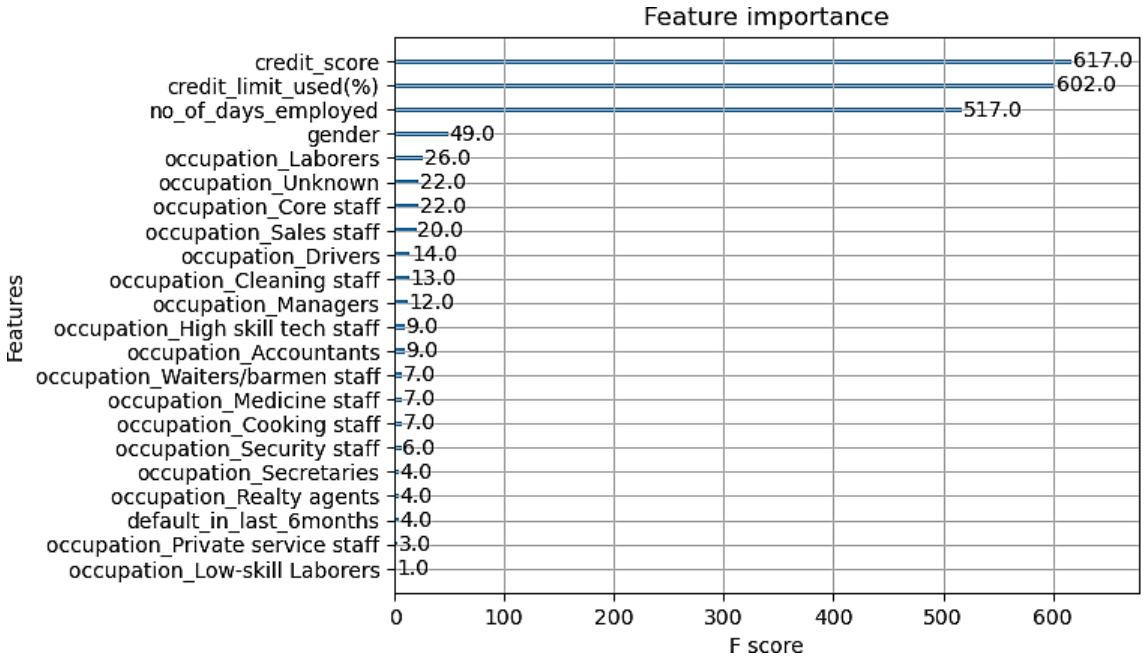
By oversampling using SMOTE, “credit_card_default” now has 38552 rows with both labels 1 and label 0

6. Modeling

After preprocessing step, we split the dataset into two parts with ratio 70/30: training set and test set. Training set has 21000 rows, test set has 9000 rows. The training set is a large dataset used to train a machine learning model. This is the dataset from which the model will learn and extract important features to remember. The test set is a dataset used to validate if the results achieved by the model after training are truly effective or not.

For the first phase: Train and test for original data, we apply Logistic Regression after every preprocessing steps except IV Score and binning features, to test out the effectiveness of this model in both scoring and risk analysis. After that, we compare it with Logistic regression applying WOE and IV. Finally, we use 5 more tree-based models and boosting models: Random Forest, CatBoost, XGBoost, LightGBM and AdaBoost in credit risk predicting only.

Figure 5. Features importance



2.2 Results

Table 5. Models and prediction metrics result

Model	Accuracy	F1-score	AUC
Logistic Regression	0.94	0.856	0.960557
Logistic Regression using WOE	0.97	0.897	0.87
Random Forest	0.967	0.90	0.9293
LightGBM	0.967	0.90	0.929901
CatBoost	0.968	0.90	0.9351
XGBoost	0.965	0.90	0.938910
AdaBoost	0.95	0.88	0.9714

AdaBoost has the largest AUC value. This superiority is due to their better handling of complex datasets, capturing intricate data patterns. In contrast, Random Forest cannot adapt as well to this data, leading to lower AUC Score of 92% AUC score.

7. Conclusions and discussions

This paper predicts an organization's creditworthiness by analyzing performance of six classification algorithms using American Express Dataset. We validate our model using accuracy, precision, recall, F1, and AUC metrics. The results suggest that after applying WOE and IV, the accuracy of scorecard increases significantly from 94% to 97%. But with risk analysis, models not using WOE have more stable performance with AUC score.

In future work, we aim to explore the impact of hyperparameter tuning on model performance. We will investigate techniques such as grid search and randomized search to identify optimal configurations for both deep learning and machine learning models. Additionally, we will experiment with forward model selection approaches to determine the most relevant features for the task. Furthermore, we will compare the performance of deep learning methods against traditional machine learning algorithms using established statistical tests. This comprehensive analysis will provide valuable insights into the relative strengths and weaknesses of these approaches for our specific problem.

References

Abellán, J., & Castellano, F.J. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. Expert Syst. Appl., 73, 1-10.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609.

Altman, E.I., and Sabato, G. (2007). Modelling Credit Risk for SMEs: Evidence from The US Market. ABACUS, 43 (3), 332-357.

Breiman, Leo. 2000. Some Infinity Theory for Predictors Ensembles. Technical Report; Berkeley: UC Berkeley, vol. 577.

Butaru, Florentin, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W. Lo, and Akhtar Siddique. 2016. Risk and risk management in the credit card industry. Journal of Banking and Finance 72: 218–39.

Chen, T.; Guestrin, C., (2016), "XGBoost: A Scalable Tree Boosting System". 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16.

Gahlaut, A., Tushar, & Singh, P.K. (2017). Prediction analysis of risky credit using Data mining classification models. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-7.

Galindo, Jorge, and Pablo Tamayo. 2000. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. Computational Economics 15: 107–43.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017).

Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. Consumer credit-risk models via machine-learning algorithms. Journal of Banking and Finance 34: 2767–87

Maruma, C., Tu, C., Nawej, C. (2022). Banking Credit Risk Analysis using Artificial Neural Network. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Seventh International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, vol 447. Springer, Singapore. https://doi.org/10.1007/978-981-19-1607-6_76

Pandey, T. N., Jagadev, A. K., Mohapatra, S. K., & Dehuri, S. (2017, August). Credit risk analysis using machine learning classifiers. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1850-1854). IEEE.

Satchidananda, S. S., & Simha, J. B. (2006). Comparing decision trees with logistic regression for credit risk analysis. International Institute of Information Technology, Bangalore, India.

Siddiqi, N. (2006). Credit Risk Scorecards, Developing and Implementing Intelligent Credit Scoring. Hoboken, NJ: John Wiley & Sons, Inc.

Yazdanfar, D., and Nilsson, M. (2008). The Bankruptcy Determinants of Swedish SMEs. Institute for Small Business and Entrepreneurship, Belfast, Ireland.

Zdravevski, E., Lameski, P., Kulakov, A., & Gjorgjevikj, D. (2014). Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets. 2014 Federated Conference on Computer Science and Information Systems, 387-394.

Appendix

Link to code