CREDIT RISK ANALYSIS BASED ON LOGISTIC REGRESSION AND OTHER MACHINE LEARNING MODELS

Phuong Anh Tran

Abstract

Credit risk assessment is essential to maintaining the stability and sustainability of the financial system in banking and finance. In this study, using American Express data sets, we explored and compared the efficiency of several credit risk analysis techniques, including logistic regression and other machine learning models.

We conducted a comprehensive examination of the use of credit risk forecasting algorithms using data from American Express. Not only are we studying the process of logistic regression, but we are additionally investigating the potential of machine learning models such as Gradient Boosting Machines (GBM), Random Forest, and Decision Trees. Our goal is to determine which credit risk analysis model best predicts credit card defaults and which variables are important.

Our findings show that XGBoost is the most effective model, achieving high accuracy (0.97), precision (0.88), F1-score (0.91), recall (0.94) and AUC value of (0.94). Although logistic regression and other models also perform well, they are not up to par with XGBoost. Our findings indicate that the most important factors in predicting credit card debt risk are credit score, credit limit utilization rate, and number of days employed. Our research suggests practical applications for financial institutions in improving their credit risk analysis models. By using machine learning techniques like XGBoost, they can better identify and manage credit risk, minimizing losses due to bad debt.

1. Introduction

Credit risk, the risk of financial loss when borrowers fail to comply with financial obligations, has become an important issue for financial institutions. To deal with this risk, the use of machine learning models such as logistic regression, decision trees, etc. has become popular. These tools not only help predict credit risk but also help financial institutions better evaluate and manage their loans. Although there are many factors that contribute to credit risk, diligence in lending (credit assessment), continuous monitoring of customer payments, and other behavioral patterns can reduce the likelihood Non-productive asset accumulation (NPA) and fraud. And indeed, over the past few years, the number of unauthorized and fraudulent transactions has increased significantly, thus making it imperative that banks and financial institutions employ robust mechanisms to ensure performance. and safety in credit activities.

The past years have seen tremendous growth in the field of artificial intelligence, especially machine learning (ML) with improved access to the Internet, data and computers. While there are credit rating agencies and credit rating companies that provide their analysis of customers to banks for a fee, researchers continue to explore various machine learning techniques. to improve the accuracy of credit risk assessment.

In this study, we focus on credit risk analysis through applying logistic regression and other machine learning methods on a dataset from American Express. This dataset includes information about credit card holders, including demographic information, personal characteristics, and payment history. We aim to examine how different machine learning models perform in predicting credit risk and evaluate the factors that directly influence the assessment of credit card default risk.

Specifically, we will use logistic regression, decision trees, random forests, and gradient boosting models to build model predictions and evaluate their accuracy and interpretability. Our study indicates an important development in the use of machine learning to reduce credit risk in the banking industry. We found how these technologies can be applied by financial institutions to anticipate and prevent defaults. Our study contributes to the improvement of the risk management process by making it easier and more effective in addition to being theoretical in nature.

2. Theoretical background

2.1. Models

2.1.1. Logistic regression

Logistic regression is a statistical method used to predict the probability of a binary dependent variable based on one or more independent variables. In logistic regression, we use the logistic function to represent the relationship between independent variables and the probability of the dependent variable. Despite the name "regression", logistic regression is essentially a classification method.

The basic idea of logistic regression is to predict the probability of falling into a particular group or class based on observable characteristics. For binary problems, we use the logistic function (also known as the sigmoid function) to represent probability. The logistic function converts the input value to a value between 0 and 1, representing the probability of the event occurring.

In detail, given a set of independent variables $X_1, X_2, X_3, \ldots, X_n$ the logistic regression model estimates the probability of an event P of an event occurring. The logistic regression model assumes a linear relationship between the independent variables and the logarithm of the odds of the event:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

The applications of logistic regression are diverse, from predicting the likelihood of a customer purchasing a product to predicting the likelihood of a student passing an exam. In the financial sector, logistic regression is widely used in credit risk assessment, predicting loan defaults based on variables such as income, credit history, and many other factors. .

2.1.2. Decision trees

The decision tree algorithm is a fundamental machine learning technique renowned for its simplicity, interpretability, and effectiveness in modeling complex decision-making processes. Operating on a hierarchical tree-like structure, decision trees recursively partition the dataset into subsets based on the most informative features, ultimately leading to predictive outcomes or classifications. This algorithm has garnered significant attention in various domains, including finance, where it finds compelling applications in credit risk analysis.

2.1.3. Random forest

The random forest model is trained based on a combination of association rules (*ensembling*) and iterative sampling process (*boostrapping*). Specifically, this algorithm creates many decision trees, each decision tree is trained based on many different subsamples and the prediction result is voting from all decision trees. Thus, a forecast result is synthesized from many models so their results will not be biased. At the same time, combining forecast results from multiple models will have smaller variance than just one model. This helps the model overcome the overfitting phenomenon.

Random Forest's applications are diverse and popular in many different industries, including credit risk management, healthcare, transportation, and consumer and market. Compared to other machine learning models, Random Forest has its own advantages. Ability to handle large data effectively, can handle both numerical and categorical data flexibly and often produce accurate and stable classification results across different data sets.

However, Random Forest also has limitations such as model training time can be long for large data sets and requires care in tuning hyperparameters to avoid overfitting. Compared to other models such as Logistic Regression, Decision Trees, and Random Forest often has better performance on large and highly complex data sets, but requires large amounts of training data. and can be computationally expensive.

2.1.4. Gradient boosting machines (GBM)

Gradient Boosting Machines (GBM) is a powerful machine learning method built on the idea of combining multiple weak models to create a stronger model. In order to improve predictions for each component of the data that the current model predicts wrongly, GBM builds decision trees one after the other. This process is repeated until either an established perfect number of decision trees has been achieved or the prediction can cannot be improved.

Variants of GBM such as XGBoost, CatBoost and LightGBM have been developed to improve the performance and flexibility of the model. XGBoost focuses on loss function optimization and regularization to minimize overfitting and increase model accuracy.

CatBoost automatically processes category variables without pre-coding, reducing data preprocessing time. LightGBM optimizes computational speed and memory usage efficiently using the Leaf-wise optimization algorithm.

Applications of GBM and its variants are diverse, including credit risk prediction, fraudulent transaction classification, market trend prediction in finance, classification prediction in medicine, and many more. other applications in different fields. These algorithms often produce accurate and stable results on large and complex data sets, contributing to solving many real-world challenges.

2.2. Weight of Evidence (WOE) and Information value (IV)

2.2.1. Weight of Evidence (WOE)

The Weight of Evidence (WOE) is a statistical technique commonly used in credit risk modeling and other areas of predictive analytics. It is a measure of the strength of association between a predictor variable and a binary target variable.

The main purpose of WOE is to transform categorical or continuous predictor variables into a more meaningful form that is suitable for predictive modeling, especially logistic regression. WOE is calculated for each category or bin of a predictor variable and is defined as the natural logarithm of the ratio of the proportion of non-events (e.g., good customers) to the proportion of events (e.g., bad customers) within that category or bin.

Mathematically, the formula for calculating WOE is:

$$\mathrm{WOE} = \ln\left(\frac{\text{Proportion of non} - \text{events}}{\text{Proportion of events}}\right)$$

WOE offers several advantages, including interpretability, monotonicity, effective handling of missing values, and reduction of overfitting. However, it also comes with limitations such as loss of information due to discretization, the assumption of linearity, and dependence on the quality of binning.

In market risk analysis, WOE finds applications in assessing creditworthiness, predicting default probabilities, and managing credit portfolios. By transforming predictor variables into WOE values, financial institutions can gain insights into customer risk profiles and make informed decisions regarding lending, pricing, and risk management strategies.

2.2.2. Information value (IV)

When evaluating credit risk, information value (IV) is commonly used. This helps financial organizations evaluate and manage risk before making loan approvals. It evaluates the ability of independent variables to distinguish between various outcomes, such as good and bad credit risks, in order to measure their predictive potential. IV offers insightful information about how closely the variables relate to the target variable of interest.

In this study, the most important independent variables for credit risk prediction are determined using IV. Higher IV variables are usually considered to have a significant impact on the prediction outcomes and therefore to be prioritized when building a model. IV helps remove variables with low predictive value and concentrate on employing more significant variables by providing information on each variable's predictive power. As a result, the prediction model performs better and requires less computing power.

The formula to calculate IV is as follows:

$$\mathrm{IV} = \sum (\%\ \text{of positive} - \%\ \text{of negative}) \times \mathrm{WOE}$$

Table 01 provides conventional interpretation of IV values and assessing the predictive strength of independent variables in predictive modeling applications.

Table 01. Conventional Interpretation of IV

Information Value (iv)	**Meaning**
< 0.02	Unless for prediction
0.02 – 0.1	Weak predictor
0.1 – 0.3	Medium predictor
0.3 – 0.5	Strong predictor
> 0.5	Suspicious or too good predictor

2.3. Metrics

A number of metrics, including accuracy, precision rate, recall rate, F1 value, and AUC value, can be used to assess a model's performance. This study's primary evaluation criteria are the accuracy, precision, recall, F1-score, and AUC value.

One important tool for evaluating machine learning models' performance in classification tasks is the confusion matrix. It groups the dataset's actual and predicted classes to provide an in-depth overview of the model's predictions. Important metrics like accuracy, precision, recall, and F1-score can be measured by analyzing the confusion matrix. In credit risk management, these metrics are crucial for leading accurate decision-making processes.

Table 02. Confusion Matrix

		Predict	
		0	1
True	0	TN	FP
	1	FN	TP

In order to define these metrics, we first establish four variables: (i) TN (T*rue Negative*), which represents the number of observations correctly predicted as non-default; (ii) TP (*True Positive*), which represents the number of observations correctly predicted as default; (iii) FP (*False Positive*), which represents the number of observations incorrectly predicted as default when they are actually non-default; (iv) FN (*False Negative*), which represents the number of observations incorrectly predicted as non-default when they are actually default.

From the Confusion Matrix, accuracy is commonly used and it represents the proportion of correctly classified instances to the total number of instances. The formula to calculate accuracy is as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Additionally, precision and recall are important metrics that provide more nuanced insights into the performance of a model. Precision measures the proportion of true positive predictions among all positive predictions made by the model, and it is calculated using the formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall, on the other hand, measures the proportion of true positive predictions among all actual positive instances in the dataset, and it is calculated using the formula:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Furthermore, the F1-score is a harmonic mean of precision and recall, providing a single metric that balances both precision and recall. It is calculated using the formula:

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

Moreover, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a critical metric for evaluating the discriminative power of a binary classification model. It measures the area under the ROC curve, which plots the true positive rate against the false positive rate across various probability thresholds. A higher AUC-ROC score indicates superior discriminatory performance, with a perfect classifier achieving a score of 1 and a random classifier having a score of 0.5.

In conclusion, these metrics collectively provide a comprehensive evaluation of the performance of binary classification models, particularly in the context of credit risk assessment. Accuracy offers a broad overview of model effectiveness, while precision and recall provide insights into specific aspects of model performance. The F1-score balances precision and recall, offering a single metric to assess overall model efficacy. Finally, the AUC-ROC metric evaluates the model's ability to rank instances correctly, essential for making informed decisions in credit risk management. Together, these metrics enable financial institutions to develop robust predictive models that accurately identify default risks, ultimately contributing to the maintenance of a healthy credit portfolio and the mitigation of potential losses.

3. Data

3.1. Data Analysis

3.1.1. Data description

The data used in this study is from the "AmExpert 2021 CODELAB – Machine Learning Hackathon" competition, organized by American Express and HackerEarth (the dataset can be accessed **here**). The data set consists of the following files:

train.csv	This file contains information about a company's customers, with the size of 45,528 lines and 19 columns. Each line represents a customer, and each column contains specific attributes or characteristics related to the customer's profile.
test.csv	This file contains the same information as the training data set but with fewer columns. It consists of 11,383 lines and 18 columns. Participants were asked to use their machine learning models to predict the risk of loss of credit for customers in this data set.
sample_submission.csv	This file provides a template for contest participants to submit their predictions. It has a size of five lines and two columns, with one column representing the customer's ID and another column for predicted credit loss status.

For this study, a subset of 30,000 rows and 19 columns have been used from the original dataset. For every column in the dataset, Table 03 provides an extensive description.

Table 03. Columns Description

NO.	COLUMN NAME	DESCRIPTION
1	customer_id	Customer ID
2	name	Name of customer
3	age	Age of customer
4	gender	Gender of customer: Female (F) or Male (M)
5	owns_car	Whether a customer owns a car: Y/N
6	owns_house	Whether a customer owns a house: Y/N
7	no_of_children	Number of children of a customer
8	net_yearly_income	Net yearly income of a customer (in USD)
9	no_of_days_employed	Number of days employed
10	occupation_type	Occupation type
11	total_family_members	Number of family members
12	migrant_worker	Is migrant worker or not? – 0/1
13	yearly_debt_payments	Yearly debt payments (in USD)
14	credit_limit	Credit limit (in USD)
15	credit_limit_used(%)	Percentage of credit limit
16	credit_score	Credit score of a customer
17	prev_defaults	Number of previous defaults
18	default_in_last_6months	Whether default in last 6 months or not? – 0/1
19	credit_card_default	Target variable: Default or not? – 0/1

Dataset has 6 categorical features and 13 numeric features.

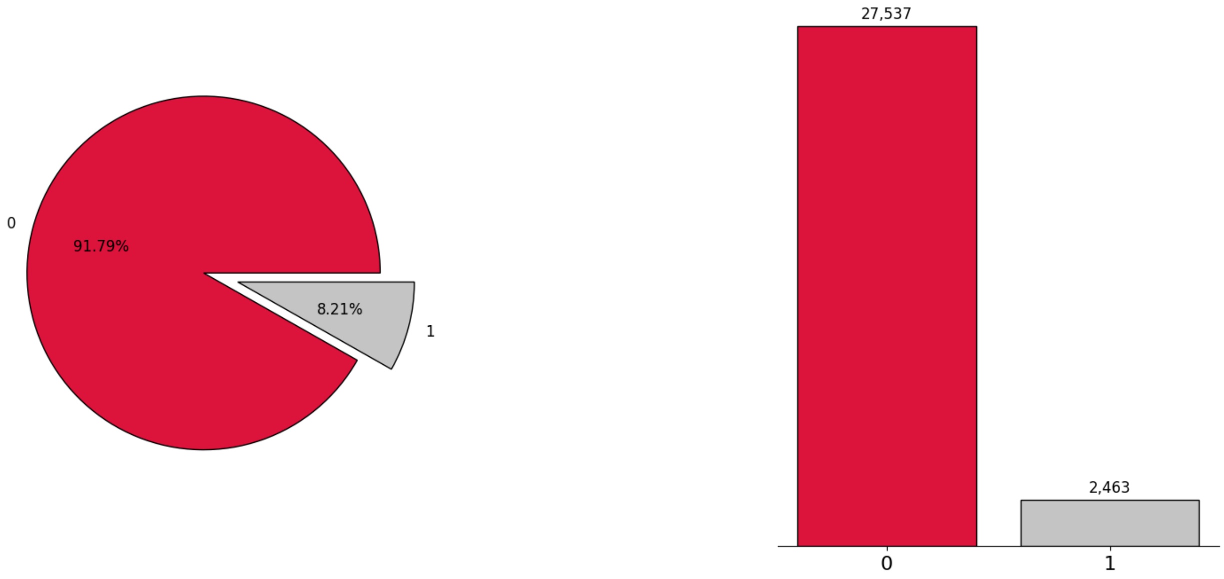
The "credit_card_default" variable serves as the target variable in the dataset, indicating whether a credit card holder has defaulted on their payments or not. It is a binary variable with two possible values:

- 0: Indicates that the credit card holder has not defaulted, meaning they have made timely payments on their credit card.
- 1: Indicates that the credit card holder has defaulted, meaning they have failed to make payments on their credit card as required.

3.1.2. Exploratory Data Analysis (EDA)

We see an imbalance between defaulters and non-defaulters when looking at the "credit_card_default" target variable. The proportion of those who have not defaulted is 91.79%, whereas the number of people who have is just 8.21%.

Figure 01. The distribution of target variable



3.2. Data Preprocessing

3.2.1. Missing values handling

After EDA, we find that most missing values (less than 2%), thus we replace all missing values using statistical methods. That is meaning that we often fill in the null by using the mode for categorical variables and the median for numerical variables.

3.2.2. Features selection with Information value (IV)

IV (Information Value) estimation plays a pivotal role in credit risk analysis, serving as a cornerstone technique for understanding the predictive power of individual characteristics in credit scoring models. In the realm of credit risk assessment, where the accurate evaluation of default risk is paramount, IV serves as a robust tool for discerning the relationship between various borrower attributes and the likelihood of default.

By quantifying the strength and direction of the association between each characteristic and the outcome variable (commonly default or non-default), IV provides insights into which factors are most influential in determining creditworthiness. This information is indispensable for financial institutions, allowing them to prioritize and refine their risk assessment strategies.

In practice, IV estimation involves partitioning continuous variables into bins or categories, followed by the computation of IV scores for each bin. These scores encapsulate the discriminatory power of the variable in predicting default, with higher IV values indicating stronger predictive capability.

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

Table 04. IV values of VAR_NAME

	VAR_NAME	**IV**
1.	age	0.000002
2.	owns_house	0.000070
3.	yearly_debt_payments	0.000080
4.	credit_limit	0.000470
5.	no_of_children	0.000976
6.	total_family_members	0.002461

7.	net_yearly_income	0.003464
8.	owns_car	0.004524
9.	prev_defaults	0.006260
10.	migrant_worker	0.012561
11.	gender	0.048503
12.	occupation_type	0.096320
13.	no_of_days_employed	0.120459
14.	credit_score	0.422381
15.	default_in_last_6months	0.600895
16.	credit_limit_used(%)	1.148924

Specifically, each row in the table displays the name of a variable (VAR_NAME) along with its corresponding IV value. A higher IV value indicates that the variable is more important in predicting the target of the model, while a lower IV value may suggest that the variable has less influence.

With a threshold of IV < 0.02, the variables "age", "owns_house", "yearly_debt_payments", "credit_limit", "no_of_children", "total_family_members", "net_yearly_income", "owns_car", "prev_defaults", and "migrant_worker" will be removed from the model, as they do not meet the desired level of importance. The remaining variables, with IV values greater than or equal to 0.02, will be retained for use in model building.

3.2.3. Data imbalance handling

Fraud analysis often divides data into two groups: fraud and non-fraud, set as binary variables (value 0 - if the observation is not fraudulent and value 1 if the observation is fraudulent). However, fraud data differs from conventional data sets, with low fraud rates making the data imbalanced, making it difficult to implement binary data analysis methods.

Therefore, in this study, the data preprocessing technique - Synthetic Minority Oversampling Technique (SMOTE) was used to solve the problem of imbalanced data.

The SMOTE method is implemented by increasing the fraud rate in imbalanced data from the KNN (K Nearest Neighbor) algorithm introduced by Batista et al. (2004). This method generates synthetic samples for the minority class by creating new data points based on existing ones in the minority class. More specifically, the additionally generated observations with value 1 have metric properties close to the original fraud observations. The KNN algorithm uses observations that are close to each other to create a group and find the central observation of the group. Based on the distance between observations, the method finds K neighboring observations that meet the requirements such that the distance from that observation to the center observation is not greater than the set maximum distance. The result creates additional new observations in between the original fraud observations.

The target variable "credit_card_default" is distributed equally between the values of 0 and 1 after the application of SMOTE.

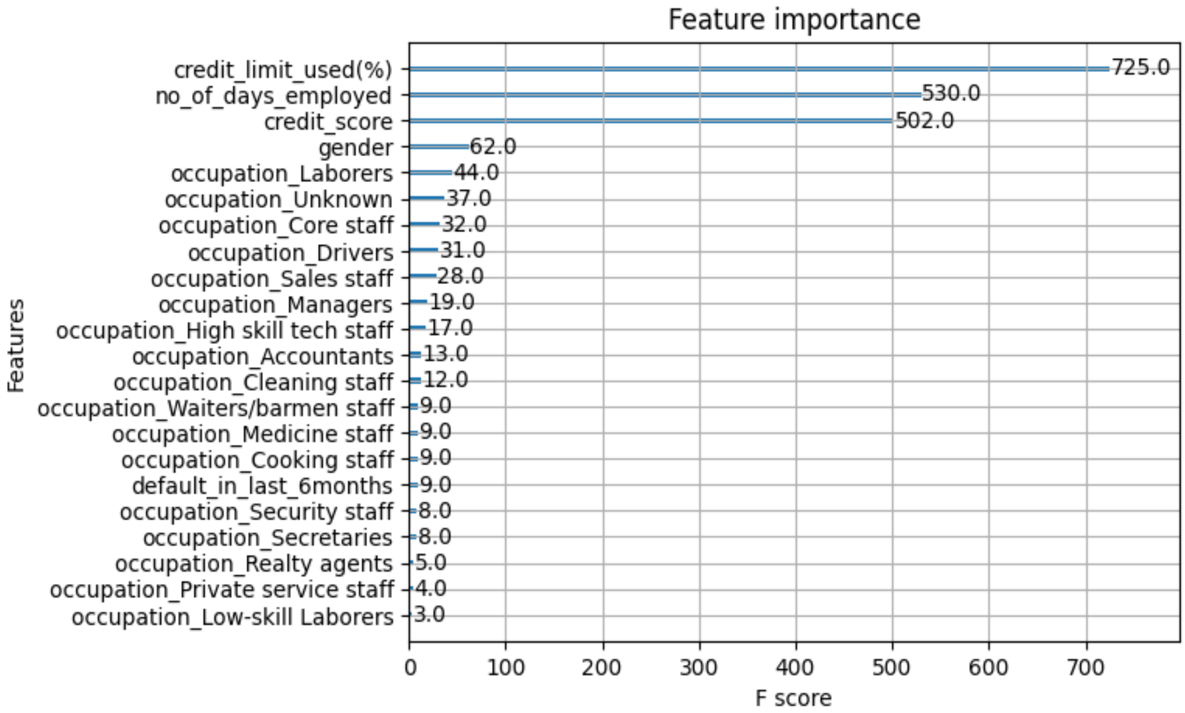
3.3. Modeling

After preprocessing step, the dataset was divided into a training set and a test set in a 70/30 ratio. Test set has 9000 rows, training set 21,000 rows.

As mentioned before, we train the model using 6 different methods on this dataset: Decision Tree, Random Forest, CatBoost, XGBoost, and LightGBM. We also fit the WOE binning dataset in Logistic Regression to see the results.

After the training timing, we analysed and evaluated the customer attributes according to how much they affected the model's prediction performance using feature_importances from the sklearn package. As seen in Figure 02, the credit score, the number of days employed and the percentage of credit limit used are some of the most important features.

Figure 02. Feature Importance



4. Results

The table below provides details about each model's name and its corresponding performance metrics.

Table 04. The prediction results of the models

Model	Accuracy	Recall	Precision	F1-Score	AUC
Logistic Regression	0.9464	0.96	0.80	0.86	0.9585
Decision Tree	0.9667	0.93	0.87	0.90	0.9258
Random Forest	0.9650	0.94	0.86	0.90	0.9378
CatBoost	0.9684	0.93	0.88	0.90	0.9347
XGBoost	0.9693	0.94	0.88	0.91	0.9420
LightGBM	0.9667	0.94	0.87	0.90	0.9437

The prediction results of the models on the dataset are summarized in Table 04. XGBoost achieved the highest accuracy rate at 96.93%, with other metrics including Recall at 94%, Precision at 88%, F1-Score at 91%, and AUC at 94.20%. These findings demonstrate that XGBoost outperforms other models in credit risk classification on this dataset.

5. Conclusions and discussion

In this study, we performed an in-depth analysis of credit risk using various models based on machine learning such as XGBoost, CatBoost, Random Forest,... and Logistic Regression. The American Express dataset, a popular source of data in this domain, was used for this investigation. We examined performance using a variety of metrics, such as recall, f1-score, accuracy, precision, and AUC, to identify the best performing model. Our findings highlight the strengths of the XGBoost model over the other models and highlight the importance of gradient boosting methods for credit risk analysis.

The practical impact of this study on finance, banking, and credit institutions, banks, is crucial. In fact, the use of machine learning models in credit risk analysis is becoming widespread. Specifically, machine learning models can be applied to assess risk and make decisions about giving credit to customers. By analyzing variables such as credit history, income, and personal information, these models can predict customer solvency and make credit decisions accurately and effectively. In addition, they can also be used to detect fraud in financial transactions by analyzing samples and other risk factors, helping financial institutions to detect and prevent fraud effectively. Using machine learning models in credit risk analysis can also optimize the lending process by predicting a customer's credit sustainability. By automatically assessing risk and identifying potential applicants, this model reduces processing time and enhances customer experience during the loan process.

In the future, this research could expand and develop in many different directions. The research could continue to compare the performance of different machine learning models such as Support Vector Machines, Neural Networks, and Deep Learning to determine which model works best in different situations and can improve accuracy and reliability in credit risk analysis. The refinement and improvement of existing credit risk analysis models by adding new variables and adjusting the parameters of the model can also help improve the performance and accuracy of the models in credit risk prediction. Furthermore, the discovery and development of new credit risk analysis models using advanced methodologies and techniques such as Transfer Learning, Ensemble Learning, and AI could also be a potential direction, bringing new breakthroughs in this field and improving the predictability of models.

In conclusion, this study is a clear demonstration of the power of machine learning in credit risk analysis. By adopting the best model, XGBoost, financial institutions can improve the quality of their credit risk analysis model. Furthermore, there is the possibility of conducting further research in this area to adapt existing models and develop new models that integrate new data sources and characteristics, thereby enhancing the accuracy and precision of credit risk analysis.

Preferences

Bhatore, S., Mohan, S., & Reddy, Y. R. (2020, April 1). *Machine learning techniques for credit risk evaluation: a systematic literature review*. Journal of Banking and Financial Technology. https://doi.org/10.1007/s42786-020-00020-3

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020, January 1). *Foundations of data imbalance and solutions for a data democracy*. Elsevier eBooks. https://doi.org/10.1016/b978-0-12-818366-3.00005-8

N. (2021, December 17). *EDA, SMOTE, Feature Selection, Hypothesis*. Kaggle. https://www.kaggle.com/code/nandha13/eda-smote-feature-selection-hypothesis

M. (2024, March 6). *Credit card default classification*. Kaggle. https://www.kaggle.com/code/mikolajhojda/credit-card-default-classification

V. (2021, December 29). *AMEX Credit Card Default Prediction (98% Accuracy)*. Kaggle. https://www.kaggle.com/code/vishnu0399/amex-credit-card-default-prediction-98-accuracy