Examining Predictions for Credit Card Defaults

DSI, Dongyan Sun

1. Introduction

Motivation

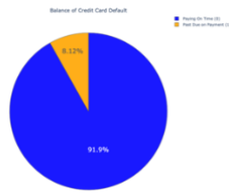
This analysis focuses on predicting whether an individual will experience overdue in credit card payments based on personal financial information such as revenue and payment history. Examining predictions for credit scoring models is meaningful, because it allows for improvement in lending decisions and promotes responsible risk management policies.

Dataset Description

American Express owns this dataset and shares it exclusively for educational use (Pradip Basak, 2021). [1]. The dataset contains 45,528 rows and 19 columns, and lacks both group structure and time structure. So that it renders the data points independent and identically distributed (I.I.D) data. Seven out of nineteen features contain null values, and, in overall, 4.43% of data points are missing.

Target Variable

Credit card default is selected as the target variable, which is coded with “1” representing past-due on payments and with “0” for paying on-time. Therefore, this categorical characteristic of target variable contributes to a classification problem within this task. Notably, 8.12% of clients fall into class 1 category, resulting in an imbalance.

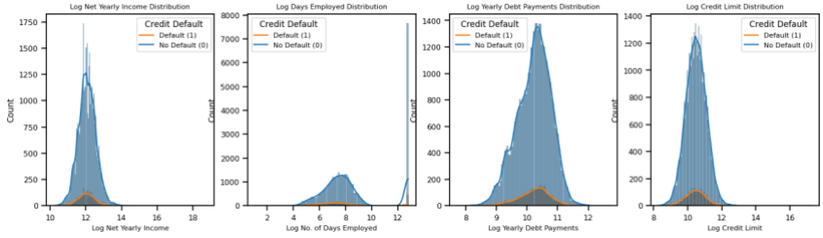


Previous Work

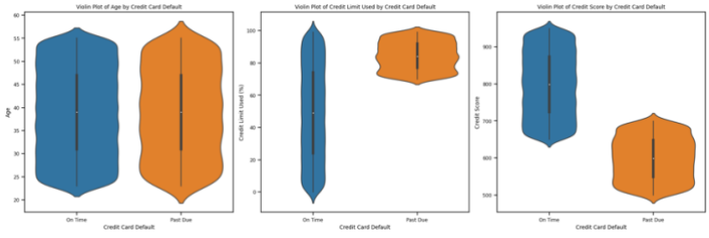
World Bank Group's “Credit Scoring Approaches Guidelines-Final-Web” document offered a comprehensive roadmap for credit evaluation. The author emphasized incorporating "supervised techniques such as regression, decision trees, random forests, gradient boosting and deep neural networks" (Knutson, M. L., 2020). [2]. Support vector machine plays an essential part of credit scoring as well.

2. Explanatory Data Analysis

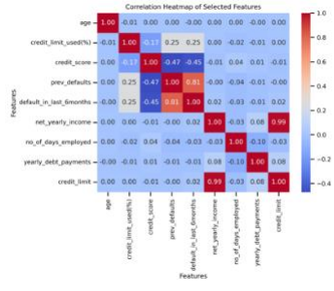
Four variables, number of days employed, yearly debt payments, net yearly income and credit limit, show extensive value ranges and skewed distributions. Those distributions can be visualized through histograms with kernel density estimates. These histograms illustrate how log transformation impacts value concentration across log transformed plots. In each plot, default cases remain consistently distributed as non-default cases do.



The violin plots demonstrate that the age distribution appears consistent between default and non-default cases. However, the credit limit used (%) varies more drastically among the group who pay overdue. And the credit score exhibits distinct patterns, suggesting potential correlation with credit card default.



The heatmap displays the correlation matrix for nine selected features, providing insight into potential relationships among variables. Strong correlations exist in certain pairs such as net yearly income versus credit limit, as well as default in last 6 months versus prior defaults. It is preferable to retain features instead of dropping them, because there could be nonlinear relationships or patterns. Dropping features would lead to missing any valuable observations or patterns that arise out of non-linearity alone.



3. Methods

ML Pipeline

A function is created to combine all the steps, including a pipeline containing preprocessor and machine learning algorithms, GridSearchCV for hyperparameter tuning and F-beta score as evaluation metric. This function facilitates exploring various algorithms such as logistic regression, decision tree and random forest. In loops through 10 random states, the model is fitted after preprocessing and splitting, and then the F-beta test score is calculated. At the end of iterations, the function returns a list of 10 test scores along with 10 best models for each algorithm.

Missing Value and Imputation

Three out of nine continuous features contain missing values. Imputation with mean is recommended for logistic regression model, as the data is missing completely at random. On the contrary, decision tree, random forest and XGBoost are entrusted with managing missing values internally, which simplifies the process of training with this dataset.

Splitting

Given small size, imbalance, and I.I.D nature of data set, it is suggested to implement StratifiedKFold in XBoost model to split data into 60% for training set, 20% for validation set and 20% for test set. Taking this strategy allows to preserve class distribution and mitigate risks associated with high bias or variance. Regarding logistic regression, decision tree and random forest, data is divided between other and test sets using an 80 to 20 stratified split. And then StratifiedKFold with 4 folds is employed for further examination on other set.

Preprocessing

It is encouraged to encode categorical data with OneHotEncoder, scale continuous features with MinMaxScaler and standardize four features with tailing distribution by using StandardScaler. 16 features have increased to 60 features due to the transformation process.

Algorithms and Parameters

Based on reference from publications and practical experience, four machine learning algorithms are selected, with three of them exhibiting nonlinear properties. Reasonable parameters and rationale behind their suggested settings, as well as their used and proposed values, are summarized in Table 1 as below.

Algorithm	Rationale for the Selection	Parameters	Rationale for the Selection	Best Parameters
Logistic Regression	effective in handling binary classification	C [0.01, 0.1, 1, 10]	allow fine-tuning of the regularization strength	C =0.1
Decision Tree	mitigate overfitting and capture non-linear patterns	max_depth [5, 10, 15, None]	prevent the decision tree from becoming overly complex	max_depth =5
Random Forest	robust to noise and overfitting	n_estimators [50, 100, 150, 200]	controls the trade-off between low training error and low testing error	n_estimators =100
XGBoost	mitigate overfitting and capture intricate patterns	learning_rate [0.01, 0.1, 0.3, 0.5]	influence the contribution of each tree allow control over the step size	learning_rate =0.01

Evaluation Metric

Using.predict_proba is recommended as it offers greater control of decision threshold in an unbalanced dataset. Most importantly, employing F-beta score with beta of 0.5 as evaluation metric is beneficiary since it aligns perfectly with preference to prioritize maximizing true positives. It is also preferable to prioritize reducing false positives over reducing false negatives, because prioritizing precision over recall contributes to effective allocation of valuable financial resources.

Measurement of Uncertainty

Decision tree, random forest and XGBoost models possess non-deterministic aspects that add additional unknowns during training sessions. Their algorithms inherently incorporate randomness for model outcomes across 10 random states' runs.

4. Results

Baseline

Based on the assumption that the baseline model predicts all points as class 1, this formula computes the baseline at 9.754%:

$$\text{Baseline F-beta Score} = 1.25 * \text{Precision} * \text{Recall} / (0.25 * \text{Precision} + \text{Recall})$$

Table 2 demonstrates the number of standard deviations above the baseline, through which decision tree and random forest signify their superiority to surpass the baseline model, along with higher numbers of 279.63 and 294.09, respectively.

Model	Mean of Test Score	Standard Deviation of Test Score	Coefficient of Variation	Number of Std Above Baseline
Logistic Regression	0.94033	0.003162	0.003363	266.48
Decision Tree	0.94008	0.003012	0.003205	279.63
Random Forest	0.94045	0.003711	0.003946	227.12
XGBoost	0.93896	0.002861	0.003047	294.09

Test Scores

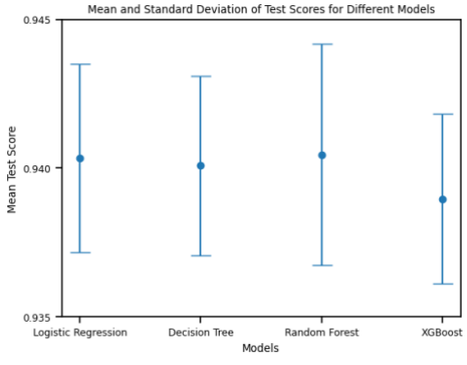
Table 2 illustrates that logistic regression, decision tree, random forest and XGBoost all achieve mean of test scores between 93.89% and 94.03%. The standard deviation of test scores ranges from 0.0028 to 0.0037, along with coefficient of variation between 0.0030 and 0.0039. These findings show all models perform consistently well, exceeding the baseline of 9.754%.

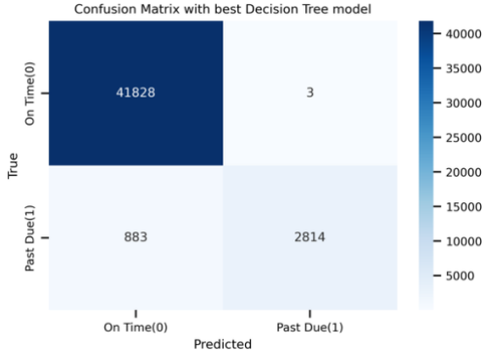
Best Model

Decision tree stands out as the best in performance. Although its mean test score of 94.01% may have been slightly less impressive, its high number of standard deviations above baseline (279.63) stands as testament to its superiority, suggesting not only accurate predictions but also consistency and robustness.

Result of Best Model Interpretation

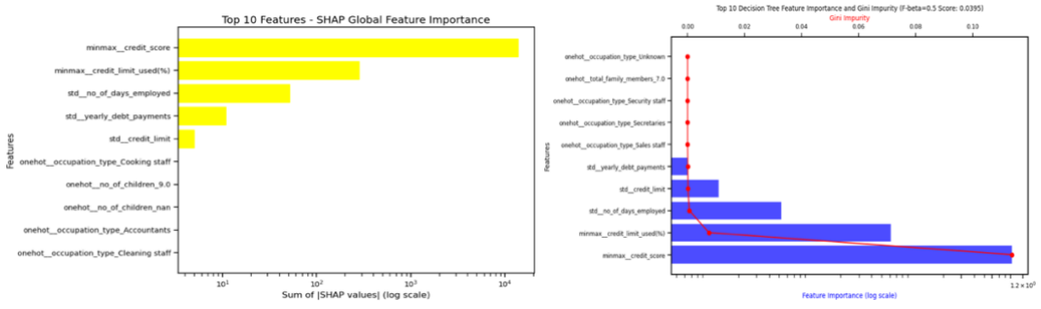
In confusion metric, decision tree model boasts impressive performance in prediction with an F-beta score of 0.9402, showing its success at balancing precision and recall. A precision rating of 0.9989 illustrates this success in accurately predicting positive instances with only 3 false positives. A recall rating of 0.7612 may have been slightly lower, but the F-beta score indicates an appropriate trade-off between precision and recall in overall.





Global Feature Importance

4 metrics, including Permutation Importance, SHAP value, Gini Impurity and .feature_importances_, are employed to calculate global feature importance. The yellow bars in left-side plot as below stand for SHAP value while the blue bars and red circles in right-side plot represent metric of .feature_importances and Gini Impurity, respectively. Both plots show that credit score, has the most significant feature importance in the overall model prediction. Following behind are number of days employed, credit limit used (%), credit limit and yearly debt payments, with their impacts which are much lower.



Local Feature Importance

SHAP force plots often show the top five features for various data points. For example, as shown with No.9000 data point, credit score is highlighted with blue arrow to demonstrate its most negative and lower-than-average contribution to output prediction. While credit limit used (%) and yearly debt payment on red arrow show positive and higher-than-average impact on prediction.

Most and Least Important Features

Assessing the overall importance of features in a model, both global and local feature importance converge upon credit score as the most influential variable. In contrast, gender and age appear as least influential factors that shape predictions.

5. Outlook

It is imperative to assess the potential of the model to be overfitting. Exploring different hyperparameters and values becomes crucial in finding a more balanced spot that minimizes overfitting or underfitting. Notably, although aligning with domain knowledge in lending decisions, the variable, credit score, stands out as extremely influential variables. Therefore, it is reasonable to use feature scaling or weighted models to help control its weight and impact when training the dataset again.

Heatmap analysis uncovers strong correlations among specific pairs, prompting a need to experiment with dropping certain features. Furthermore, it is prudent to consider handling missing values with alternative approaches. Examining various beta values for evaluation metric and evaluating their impacts can also provide insights for fine-tuning the model.

6. Reference

[1] Pradip Basak. (2021). AmExpert CodeLab 2021. [online] Kaggle. Available at: https://www.kaggle.com/datasets/pradip11/amexpert-codelab-2021/code [Accessed 2021].
[2] Knutson, M. L. (2020). Credit Scoring Approaches Guidelines - Final Web. [online] World Bank Group. Available at: https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/CREDIT-SCORING-APPROACHES-GUIDELINES-FINAL-WEB [Accessed 2 April. 2020].

7. Github Repository

https://github.com/EmmaSun19902023/Examining_Predictions_for_Credit_Card_Defaults_Project/tree/main