Credit Card Default Prediction

Merielyn Sher; Brown University DSI; Github: https://github.com/mgsher/data1030-project

I. Introduction

Given the nature of the credit card service, for the banks that provide this service, there has always been the risk of customers defaulting on their credits. With the increasing number of credit card holders, it is helpful to analyze existing data collected on them and build machine learning models to predict future defaults as a way to mitigate the risk for the banks. This research project builds on an existing dataset and aims at constructing a model to predict whether a credit card holder will default in the future.

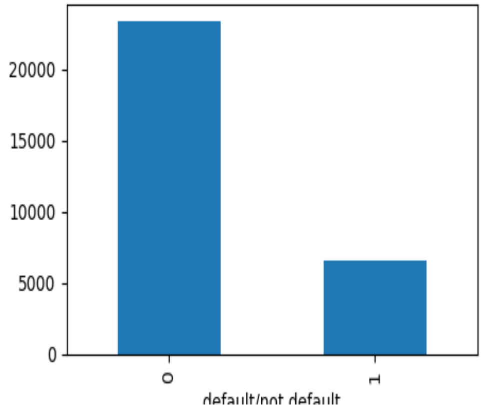
Specifically, the target variable for the machine learning model is whether a credit card holder will default in the following month; an output value of 1 indicates that the person will likely default on their credit while an output of 0 indicates otherwise. As a result, we are looking to construct a binary classification model.

The dataset can be found on both UCI and Kaggle. It was collected in October, 2005 from a bank (a cash and credit card issuer) in Taiwan, and the data was only on credit card holders of the bank. There are in total 23 features and 30,000 observations in the dataset. Here, we list the target variable as well as all 23 features and a detailed description for each of them.

- Whether a Holder Default (default): Indicates whether a credit card holder defaults; the data was collected in Taiwan for the month of October, 2005
- Amount of Given Credit (LIMIT_BAL; Taiwan dollar): the credit limit given to the credit holder each month
- Gender (SEX): 1 indicates male and 2 indicates female
- Education (EDUCATION): 1 indicates graduate school; 2 indicates university; 3 indicates high school; 4 indicates others
- Marital Status (MARRIAGE): 1 indicates married; 2 indicates single; 3 indicates others
- Age (AGE): provides the credit card holder's age in year
- History of Past Payment (PAY_1, . . . , PAY_6): provides past monthly payment records/statuses of the customers for six months leading up to October, 2005; This includes six features, with PAY_1 representing the repayment status in September, 2005 all the way to PAY_6 representing the repayment status in April, 2005. For each of the six features, a value of 0 indicates that the payment was made on time and values ranging from 1 to 9 indicates the number of months for which the payment was delayed
- Amount of Bill Statement (BILL_AMT_1, . . . , BILL_AMT_6): provides the net amount of statement in a given month measured in Taiwan dollar; Again, this includes six features representing the data for each of the six months right before October, 2005
- Amount of Previous Payment (PAY_AMT_1, . . . , PAY_AMT_6): The total amount paid back by the credit card holders in a given month measured in Taiwan dollar; It includes six features representing the data for each of the six months right before October, 2005

There has been several previous efforts at performing exploratary data analysis and preliminary modeling on the dataset.

II. Exploratory Data Analysis



The figure above shows the distribution of the target variable. There is a significantly larger portion of credit card holders who did not default, which is consistent with our expectation.

From the stacked bar plot shown above, there is a slight yet obvious trend where as a client's education level get higher, they tend to be less likely to default on their credit.

While we expected that people belonging to the younger age group might default more often than those who are older, the pattern is not significant as indicated by the violin plot here. The average values of age are roughly the same for the group that defaults versus the group that does not default.

The correlation heatmap in general shows relatively low correlation level among some of the features that we predicted might be interesting. However, the correlation between whether a person default and their past payment status appears to be higher than the other features.

III. Methods

A. Splitting and Data Preprocessing

I choose to apply a 80-20 train-test split on the dataset. A k-fold cross validation pipeline with four splits is set up on the

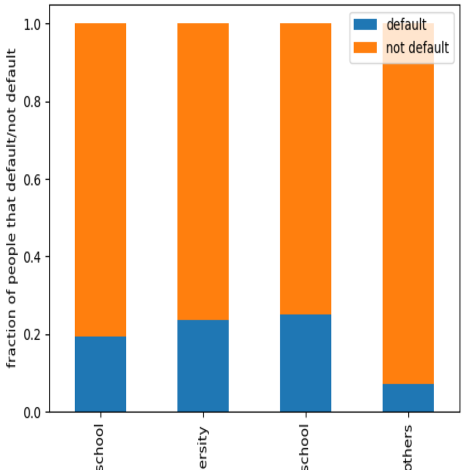


Fig. 1. Stacked bar plot of default versus education level

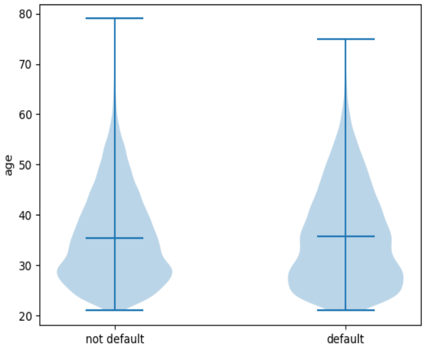


Fig. 2. Violin plot of age versus default or not

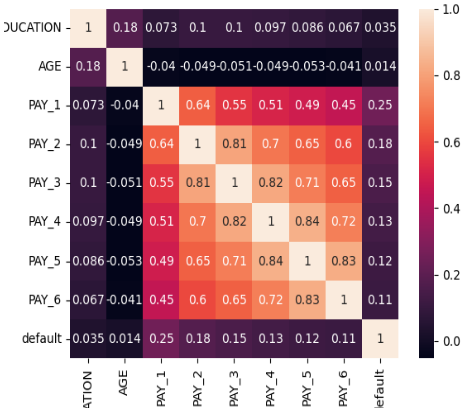


Fig. 3. Correlation heatmap

training data. Specifically, a stratified 4-fold cross validation is used because of the imbalanced nature of the target variable.

The first step in the overall architecture is data preprocessing. Here, ordinal encoder and one-hot encoder are respectively applied to the ordinal feature EDUCATION and the categorical features GENDER, MARRIAGE, PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, and PAY_6. MinMax scaler is applied on AGE, which has lower and upper boundaries, and standard scaler was applied on LIMIT_BAL, BILL_AMT1 through BILL_AMT6, and PAY_AMT1 through PAY_AMT6. The preprocessing pipeline resulted in 27 features.

B. *Models and Hyperparameter Tuning*

Overall, we used k-nearest neighbors, logistic regression, support vector machine, and XGBoost to respectively train the model. Their performances are compared against one another, and the detailed results will be shown in the next section. We also performed grid search on several possible values of the hyperparameters for each of the models. The specific hyperparameters that we tuned for each of the model are shown in the table below.

Model	Tuned Hyperparams
KNN	n_neighbors: [3,6,9,15]
Logistic Regression	C: np.logspace(-3,3,6)
SVC	C: [1, 10] gamma: [0.1, 1]
XGBoost	Learning_rate: [0.01, 0.05, 0.1] max_depth: [1,3,6, 10] reg_lambda: [1e-1, 2, 10]

We used F1-score as the evaluation metric during the training process. This decision arose natually from the imbalancedness of the target variable. Since our main goal is to predict whether a credit card holder is going to default on their credits in the near future, it is important that our metric emphasizes on the model's ability to capture true positives. Accuracy, on the other hand, provides balanced information on the model's performance of capturing both true positives and true negatives, and thus it might get pulled towards a higher value since the target value is imbalanced.

IV. Results

As shown in the bar plot above with the performance of all four ML models, the standard deviations of the models across the five random states are all relatively small. The mean f1-scores of the models, therefore, are all significantly above the baseline value. Out of the four models, the XGBoost model appears to have the best performance with a mean f1-score of 0.4996 and a standard deviation of 0.0073. Therefore, the baseline f1-score is around $\frac{0.4996-0.3622}{0.0073} = 18.82$ standard deviations below the performance of the XGBoost model.

The perturbation feature importance graph above shows two features, 'ord_PAY_1' and 'std_BILL_AMT1', that appear to exert significant more influence on the model than the other features. This is intuitive to think about since if a credit card holder defaulted on their credit in the past month, there is a significant higher risk that they will also default in the following month.

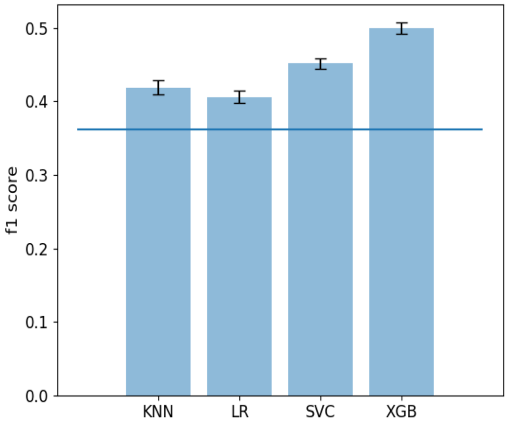


Fig. 4. Performance of the ML Models

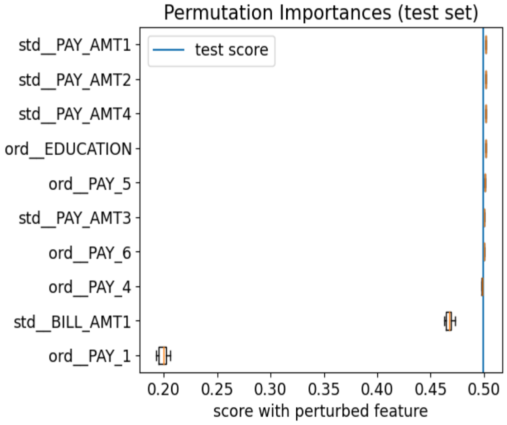


Fig. 5. Perturbation Feature Importance

The amount of their bill statement in the previous month is indicative of their current income and expenses, which are both influential factors that affect their risks of defaulting in the following month.

Both XGBoost cover metric and SHAP summary plot show relatively consistent result with what is displayed in the perturbation feature importance graph. This boosts my confidence that the PAY_AMT_1 and BILL_AMT_1 are indeed two of the most influential features on the predictive power of our model.

The SHAP local feature importance plot shows that the 6th datapoint has a predicted value of -1.13. An ord_PAY_1 of 0.0, indicating that the person did not default in the previous month, significanly pushed them towards the unlikely-to-default end of the spectrum. A negative bill statement value of the previous month, on the other hand, slightly pushes back in the opposite direction. This is consistent with what we have seen in the global feature importance plots.

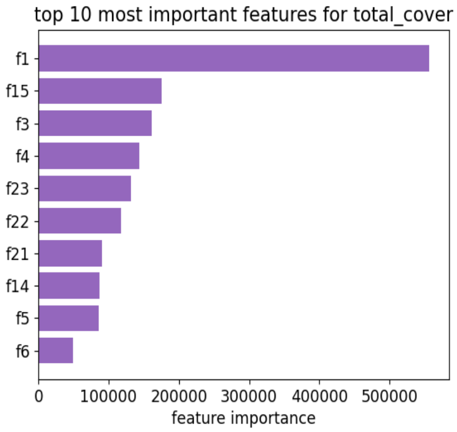


Fig. 6. XGBoost Cover Feature Importance

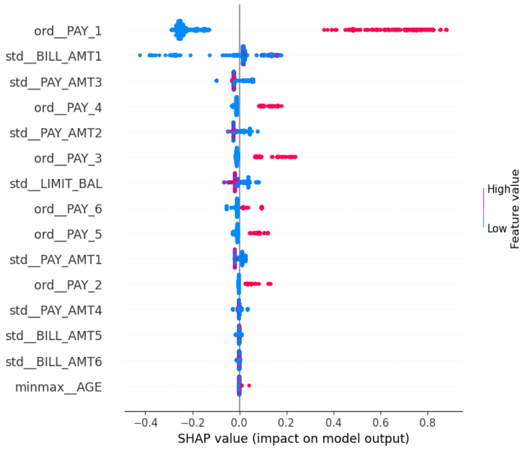


Fig. 7. SHAP Summary Plot

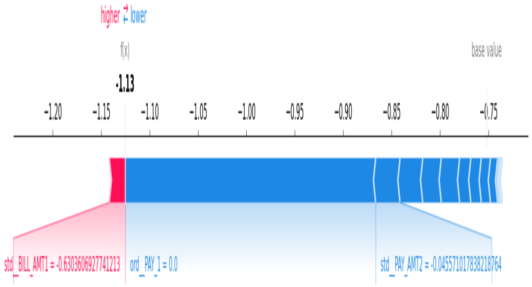


Fig. 8. SHAP Local Feature Importance Plot of Index 6

V. Outlook

One of my next steps would be to explore other similar datasets with more potentially interesting features, like more demographic features, which may assist us to make predictions more into the future.

I also hope to explore more model architectures and potentially make use of deep learning models. One possibility would be to adopt 1D convolutional neural network to take the time locality of the past bank records into account.

With more time and computing power, I also believe it would be helpful to search a larger hyperparameter grid.

VI. Github Repository

https://github.com/mgsher/data1030-project

References

[1] I-C. Yeh, C-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Systems with Applications, https://doi.org/10.1016/j.eswa.2007.12.020, March 2009. Accessed 13 October 2022.