Christine Gao
15, December 2022

Coding Assessment Write Up

Introduction

The purpose of this project is to conduct an analysis on credit card default risk by using Machine Learning models, given past customer behavior data.

Methodology and Data Exploration

The dataset contained 24001 entries (or customers) and 25 columns. These columns included demographic information such as sex, marriage status, age, and educational attainment. There was also information on the monetary amount of credit in dollars, history of past payments, history of payment delays, history of bill statements, and history of previous payments in dollars. Additionally, there was a binary outcome column indicating whether or not the customer had defaulted at the end of the 6 month period (October).

During the Data Exploration phase, the goal was to identify whether there were observable trends that impacted the likelihood of defaulting and whether any correlations existed between these variables. The main findings of note include:

1. **Educational attainment and limit balance:**
   Customers with high educational attainment tend to have a higher initial credit balance, and also have fewer defaults. However, there may exist some outliers in the 'others' group, which displays a very low number of defaulted customers (displayed in **Fig. 1**).

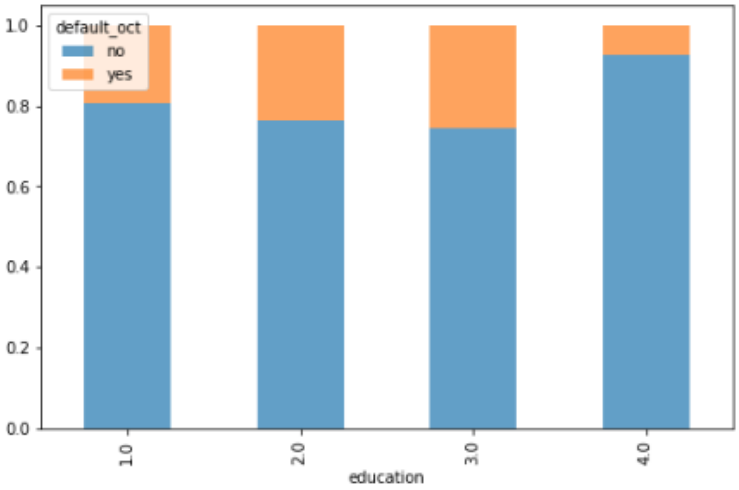

Figure 1. Education plotted against default frequencies

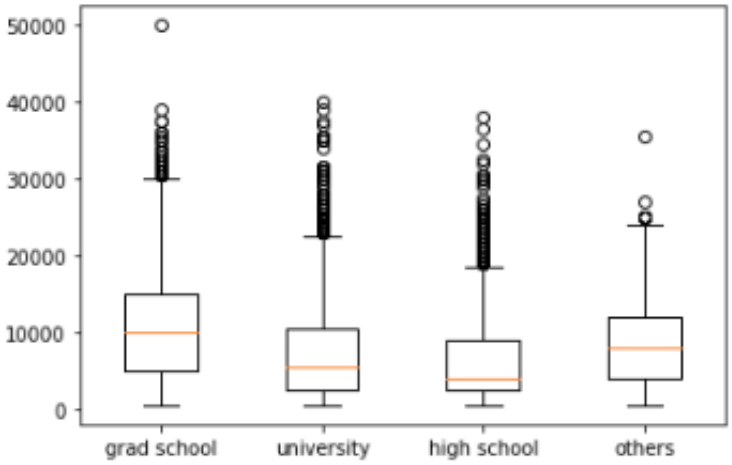

Figure 2. Education plotted limit balance

In **Fig. 2**, we take note that the median initial limit balance for customers with a graduate education is much higher than those with a university or high school education. However, the median limit balance in the "others" category is not much lower, and we observe that there exist a few outlier variables where the limit balance in the 'others' category as a little less than 40,000.

To interpret these findings, individuals that have lower than a high school education **may be less likely to receive the traditional credit card plan** (given they are already placed in a 'high risk' category). Thus a smaller fraction of individuals make up the 'others' category.

It is also likely that the **outliers in the 'other' category are customers who have attained beyond a graduate level education** (i.e. Doctorates) or **just have a very good credit score,** which is how they obtained such high limit balances on their credit card plans.

2. **Defaults and missed payments over time:**
   The second most interesting discovery made via the data was the trend over time on number of customers who missed payments leading up to the default month (October).

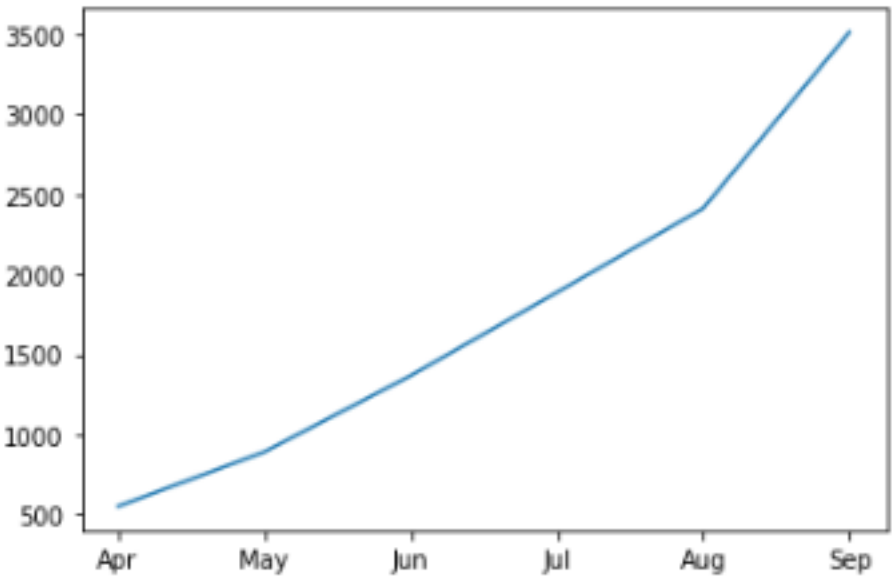

Figure 3. Total number of missed payments from defaulted accounts

Fig. 3 shows an increasing trend of delayed payments over time. It's unclear what year the data set is from, or whether or not there was a significant event that occurred in the marketplace that could have yielded such a noticeable increase in number of delayed payments. However, this could be important to note that the summer months are vacation months – increased travel, especially out of the country, could lead individuals to accidentally miss their payments, and thus we could consider the summer months 'high risk' months for credit card applications.

Model

Since this analysis seeks to predict the probability that a customer will default given their payment history and background, the best model to perform this task is a binary classification model that also yields the probability of each classified document. I used an XGBoost model from the built-in Scikit-Learn library.

To yield the best model possible, and also select the most important variables for the final classification, I followed the below model fitting steps:

- **Feature Selection** – Initially, my predictor variables included all the columns in the initial dataset except for 'customer_id' and 'default_oct' (as this is our target variable).
- **Train/Test Splitting –** For each model, I set a random_state and conducted an 80/20 split of train and test data. After splitting the data, the model is ready to be trained and assessed for fit.

Model Tuning and Outcome

To evaluate the fit and performance of the model, I assessed the accuracy, a confusion matrix, ROC curve, and three other metrics (precision, recall, F1). The initial pass that included all of the variable from the dataset yielded a relatively high accuracy of 81.63%.

Next, I wanted to review the feature importance sorted, which shows the corresponding F-score of each feature. The most interesting outcome of note was the F-scores of 'pay_1' (repayment status in September), and additionally the previous repayment status in earlier months. This indicates that in lieu of all the other variables, a customer's repayment history is a very important indicator of their likelihood of defaulting.

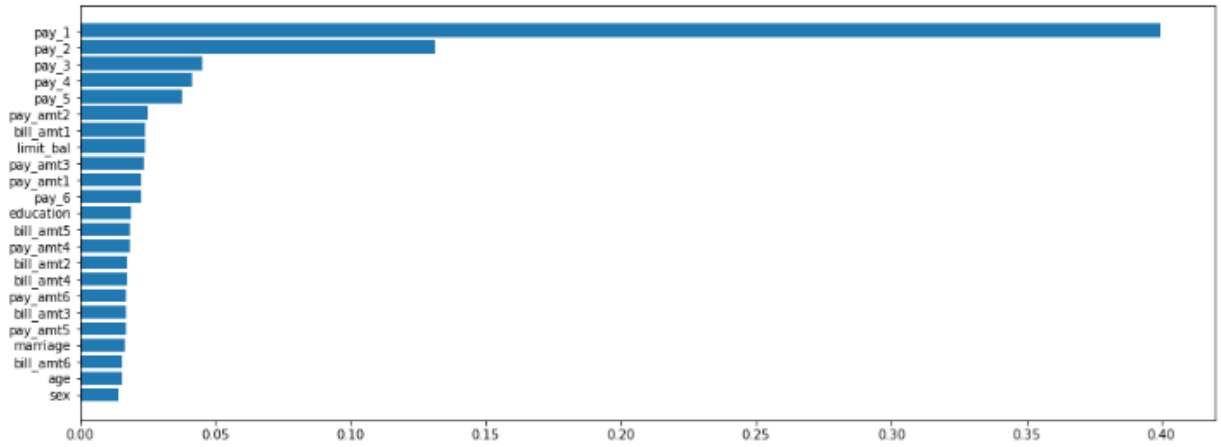

Figure 4. Feature importance sorted in descending level of F-score

```
Thresh=0.014, n=23, Accuracy: 81.02%
Thresh=0.015, n=22, Accuracy: 80.65%
Thresh=0.016, n=21, Accuracy: 81.07%
Thresh=0.016, n=20, Accuracy: 80.98%
Thresh=0.017, n=19, Accuracy: 80.90%
Thresh=0.017, n=18, Accuracy: 80.80%
Thresh=0.017, n=17, Accuracy: 80.88%
Thresh=0.017, n=16, Accuracy: 80.80%
Thresh=0.017, n=15, Accuracy: 80.98%
Thresh=0.018, n=14, Accuracy: 80.67%
Thresh=0.019, n=13, Accuracy: 81.42%
Thresh=0.019, n=12, Accuracy: 80.88%
Thresh=0.023, n=11, Accuracy: 80.86%
Thresh=0.023, n=10, Accuracy: 80.98%
Thresh=0.023, n=9, Accuracy: 81.42%
Thresh=0.024, n=8, Accuracy: 81.15%
Thresh=0.024, n=7, Accuracy: 81.21%
Thresh=0.025, n=6, Accuracy: 81.50%
Thresh=0.038, n=5, Accuracy: 81.86%
Thresh=0.041, n=4, Accuracy: 81.59%
Thresh=0.045, n=3, Accuracy: 81.63%
Thresh=0.131, n=2, Accuracy: 81.63%
Thresh=0.400, n=1, Accuracy: 81.46%
```

Figure 5. Accuracy of the model varied by number of features selected.

From **Fig. 5,** I also tested whether the accuracy of the model could be improved using a simpler model (to adjust for a potentially poor fit as a result of too many variables). Surprisingly, the accuracy of the model did not vary wildly as the number of features was reduced. Ideally, since our goal is to create as least-complex of a model as possible, the trade off from 81.63% to 81.86% is marginal in terms of our model. Thus, I reduced my final model to take in only the top

5 most important features in terms of F-score, which are the previous 5 payment months leading up to October ('pay_1' through 'pay_5').

By selecting the top 5 features (which also happens to be the history of past payment), we make the assumption **that the best indicators of a customer default will be their past payment history.** This might hold **implications in the future for when deciding at what month a customer may run the 'risk' of defaulting, given their previous payment history.**

Finally, I reassessed the XGBoost model using the selected features, and yielded an accuracy of 81.88% (again, not too different from our initial features selected).

Conclusion

From the exploratory data analysis, we find that between individuals, demographic characteristics such as sex, marriage status, and educational attainment actually do not have a noticeable impact on risk of defaulting**. From the model, we realize that payment history and likelihood of being able to make a payment on time are the most important indicators of a customer's likelihood of defaulting.**

Some limitations of this model to note: given the time constraint**, the model could have benefited from hyperparameter tuning**, which would have improved the best model architecture. Additionally, there are some variables that were not explored too much in depth, such as **age**, and **balance limit**. Age is generally a very high factor in terms of credit score evaluation; balance limit could indicate whether or not an individual is already a high-risk customer and could be at risk for defaulting later on.

In the future, there are many other classification models that can be applied to this project to better predict possibility of defaulting, and more insights to be learned from demographic information of each customer. The credit card industry is a volatile one, but there are many surprising insights and trends we can learn from such valuable information.