Predicting credit card defaults

using machine learning techniques

# Introduction

Credit card are issued by banks or financial companies which allow clients to borrow money to pay for goods or services. Credit card companies maintain vast databases on cardholders and have the condition that cardholders pay back the company on time. Since the companies make the payment on the client’s behalf it is in their best interest that they get their money back from the cardholders.

There are several ways the issuers make profit, but it mainly comes from annual fees to cardholders, transaction fees on purchases or transfers, processor fees but the largest profit portion comes from interest fees. Interest fees are charged when balance on the card is not paid on time. To read more on how these fees are calculated you can go on the [link.](https://www.fool.com/the-ascent/credit-cards/how-do-credit-card-companies-make-money/)

The card fees are calculated based on credit risk which is determined by credit history of individual. However, with the case of this dataset the reality is quite different. What happens when we don’t have any background checks and only have limited credit history? How will we know if a client can pay back the bill at all?

## background

To encourage the flow of capital, banks in Taiwan began extensively encouraging the population to apply for credit cards. They lowered the requirements for approvals mainly targeting young people still studying, who had no job or financial knowledge.

By February 2006 Taiwan had a credit card debt of $268 billion USD! This led to a range of nationwide social issues such as homelessness, illegal drug sales, violent debt collection especially towards lower income families and even increased suicide because they were not able to pay off the debt. ½ million people were coined by the term ‘Credit card slaves’ which referred to someone who could only pay the minimum credit card debit back each month stuck in a never-ending cycle.

The problem was only resolved when the government stepped in to sort the problems and implemented stricter rules that monitored who could be approved for credit cards as well as how much limit they can get.

**This is how we get our main problem statement and business motivation:**

**How can we accurately predict the probability of a customer defaulting especially when we have very limited to no credit history information?**

So, in this classification problem from the perspective of risk management our stakeholders are credit card companies whose main motivation is not the accuracy of the prediction model but rather the recall rate. The recall measures the proportion of actual positives that were identified correctly. We are more interested in the number of false negatives being as low as possible and want the number of true positives to be high as possible. Basically, a customer who has been incorrectly predicted as likely to not default when in reality they are likely to default, poses the biggest risk to the bank.

## data source and schema

Normally credit card data is hard to find as it contains private personal details such as marital status, employment status, income amount etc. This dataset was donated to UCI Machine learning repository from Department of Information Management, Chung Hua University, Taiwan however it is also available on Kaggle.com. They are both the same but the original on UCI website has a fuller data dictionary of the two. You can view the original data from([here](https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)).

This dataset contains information on default payments, demographic factors such as age or sex, credit data such as history of repayment, bill statements of credit card clients in Taiwan from April 2005 to September 2005. There are originally 30,000 rows and 25 columns. The target column is a categorical variable that records default payments next month column. The credit data is all numerical continuous data, and the demographic data barring age is categorical data that is originally converted into a numbering format.

## Pre-processing and eda

On initial glance there did not appear to be much cleaning to do as there were no null values or duplicates. However, upon closer inspection there was a range of issues that needed to be sorted before going forward. This included incorrectly labelled columns, various unknown values not mentioned in data dictionary as well as the biggest problem. The data was originally categorical, but it had been converted into numerical values in ascending order. This was a problem as it could cause the model to recognise them as ordinal values with a ranking system when in fact, they were all equal status to each other. So, I needed to one hot encode the columns into binary values. In this case I used pandas function get-dummies to do the conversion and then dropped any insignificant values such as education-others as it represented a minute portion of the whole data frame. Before doing this, I grouped together unknown values into 1 value e.g., education there was only 4 values mentioned but there were 3 extra unknown ones, so I replaced them all to be 1 value.

For EDA I first did a correlation matrix to get a general overview of which features would be most important in creating models. I found that with the default column PAY\_1 -3 had the highest correlation coefficient followed by limit balance. All the bill amount, pay and pay amount columns were highly correlated with the column parallel to itself. This is expected as if you can’t pay your bill in August then its highly likely you can’t pay it the Chart, bar chart

Description automatically generatednext month either. Then I did a bivariate analysis on every feature against the default column. This helped to get a better idea of both the distribution of the data but also what characteristics did customers who defaulted on payments shared.

For example, in the graph shown you can see that women form a much larger portion of people who pay their balance on time compared to men. Some columns such as required a log transformation as they had power law distribution where the data was skewed towards 0, before they gave any useful insights.

## Findings and conclusions

Chart, treemap chart

Description automatically generatedFrom my EDA I found that single women who had completed higher education were the group least likely to default. Whereas from both EDA and feature importance I found features such as Pay\_1, bill\_amt1 and limit balance were the best indicators for default payment. If someone is already started to struggle to pay their credit card last month, they are more likely to be unable to pay next month.

My models struggled to get very good recall rates for the test set. From practical standpoint this was what was most important to get right as we wanted the portion of correctly identified positives to be high as possible and false negatives to be low as possible. Reducing dimensionality reduced the accuracy of some models but optimizing hyperparameters gave the best results. The highest rate I manged to get was for XGBoost at 78% recall rate (TPR), with about 12% being recorded as false negative as shown on the left.

## Final summary

To conclude this project is essentially what most banks employ nowadays to screen who gets approved for credit cards and ensure customers taking out a credit card has some sort of backing to be able to pay off the borrowed money. The ethics of this can also be argued as it could be discriminatory to use only demographics to determine approval.

Next steps could include trying to increase true positive rates and lower false negatives. This would be by trying a neural network machine learning model. It is predicted to have the best accuracy and recall rate out of rest of the ml techniques.