# Analyzing Credit Card Default

By

Garima Agarwal, Eduardo Herrera, and Ram Sridhar

## Abstract

The business of lending money has always carried inherent risks. Since the turn of the century we have experienced multiple financial crashes that have left the economy as whole including businesses and individuals in financial ruin. Finding methods to avoid such costly decisions is paramount to maintaining a successful credit lending business. others. The goal is to find the most efficient machine learning algorithm that leads us to discover the most relevant factors in our dataset to determine defaults. The results of comparing different models can lead us to understand what are the most influential features to determine whether one defaults on a loan.

## Introduction

[1]As we saw during the 2008 housing crisis, credit institutions lent money in the form of mortgages for homes with little or no regard as to whether those loans could be repaid. Since then, private lending institutions have had to evolve. Now companies must consider the negatives of high yield and high risk loans.

Companies take various factors into account when deciding whether to issue a loan. Credit scores, income, debt-to-income ratio, down payment, loan term, and collateral are all factors that companies consider when someone applies for a loan. Their goal is to use these factors to predict whether the potential borrower will be able to repay the loan and how much of a return will become of their investment.

In this project, we will analyze a dataset containing the defaults of credit card clients in Taiwan from the UCI Machine Learning Repository. Our goal is to create various machine learning models that can predict whether a given individual will default on their next credit card payment. We plan to create models using the following algorithms: KNN, Logistic Regression, Naïve Bayes, Decision Trees, SVM, and Random Forest.  Using this information companies can help inform their decisions as to which features contribute to loans that are more likely to result in default.

## Related Work

Credit card companies are a major contributor to the payments ecosystem. [2] Credit cards are the largest form of payment when looking at in-store payment type by volume. [2] In the US, credit cards form a total of 23.6% of total transactions in 2019. [2] This equals almost $4 trillion in volume in 2019

which is up from the previous year. Credit cards will continue to be a popular form of payment as companies expand reward systems to incentivize customers.

Our dataset contains credit card information for individuals in Taiwan. [3]Looking at the total number of circulating credit cards in Taiwan, we found that there are currently 49 million active credit cards in Taiwan. This is a trend that has seen the number of credit cards in Taiwan increasing steadily since 2009.

Predicting credit default is not a unique or uncommon task. There are various papers dedicated to building machine learning models to predict credit default. [4] We found one titled *Prediction of credit card defaults through data analysis and machine learning techniques*. The paper was written by several researchers from Manipal University in India. They ran five different models which were Random Forest, SVM, Logistic Regression and KNN. Their most accurate model ended up being Random Forest with an accuracy of 80% on test data. One of the main challenges they faced was dealing with such an imbalanced dataset. When predicting defaults, any data collected will tend to be highly unbalanced. The majority of people tend to pay their bills on time, and our target is to predict the rare case in which a default occurs.

## Models

In order to create models to predict the defaults of credit card owners, we built models around the following machine learning algorithms; Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbors.

Baseline

Our baseline model will be to always predict the most frequent class. As we will show in our Experimental Results section, this dataset is highly imbalanced. Of the 30,000 instances only 6,636 defaults occur. The remaining 23,364 out of 30,000 instances are on time payments. This translates to 22.12% of instances resulting in defaults while 77.88% of instances are normal on time payments. Thus, a model which predicts next month's payment will be on time every time would be correct 77.88% of the time. So any model we choose to implement must have an accuracy score higher than 77.88% to be useful.

Logistic Regression:

Logistic Regression is a classification algorithm that gives us a probability value between zero and one. It categorizes data into discrete classes by taking a linear relationship from a given set of labelled data and introduces a Sigmoid function which maps the prediction from a linear regression to a probability value of between zero and one. The Sigmoid function as well as the hypothesis function for a Logistic Regression is given by the following:

$$\textit{Sigmoid Function}: \quad g(z) = \frac{1}{1 + e^{(-z)}}$$

$$\textit{Hypothesis}: \quad h_\theta(x) = \frac{1}{1 + e^{(-\theta^T x)}}$$

Our final Logistic Regression model was tuned using the following parameters; c value of 1.3 x $10^6$ and using L2 regularization.

Naïve Bayes

The Naïve Bayes classifier is a machine learning model that applies Bayes Rule to form a probabilistic classifier which assigns a class label $\widehat{y} = C_k$ for some *k* as follows:

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

Each instance is estimated as a probability of belonging to a class, and the maximum calculated class probability will be used to assign a class label.

Support Vector Machine

The Support Vector Machine classifier maps our features into a higher dimensional space where we can easily separate our classes with a hyperplane. The soft margin equation for the support vector machine classifier is shown below:

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i - b)\right)\right]$$

Our tuned SVM value used a C value of -1 and a Gamma value of 0.1 for it's tuned parameters.

Decision Tree

Decision Tree models provide high interpretability in addition to high accuracy. The models determine the best feature in a given dataset to split the data on. This creates a node with a subset of data, and the model continues creating new nodes by splitting each subset of data until we reach the point where the predictive accuracy is maximized while minimizing the number of splits and nodes. Our final Decision Tree models had a max depth of 3, 4 as the number of features to consider when looking for the bes split, 5 for the minimum number of samples at each leaf node, 3 for the minimum number of samples to split an internal node, and gini for the splitting criterion.

K-Nearest Neighbor

The KNN algorithm predicts whether a given instance belongs to a given class based on its k-nearest neighbor. The value of k determines the number of examples closest to each instance are used for predicting its class. In order to determine which points are closest to a given point, we use a distance formula to calculate the nearest neighbor. The formula for Euclidean distance is as follows:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

After parameter tuning, we found 19 as the best k value for our model.

## Experimental Results

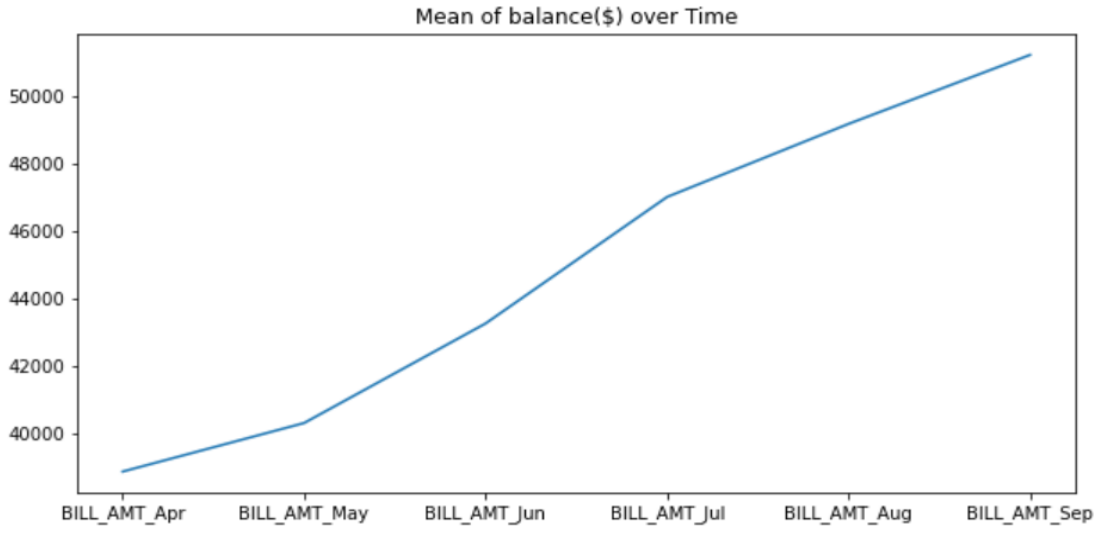
Dataset contains the credit of the customer, their gender, education, marital status, age, the repayment status of the customer over the previous 6 month period (April to September), the amount of bill statements(April to September) and the amount of previous payment over the same period.
The dataset comes from the UCI Machine Learning Repository. It consists of 30,000 rows as well as 24 features including our target. The target, default payment, is binary. It's coded as Yes = 1 and No = 0. Below is a description of the remaining 23 features.

- LIMIT_BAL : Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- SEX: Gender of each individual (1 = male; 2 = female).

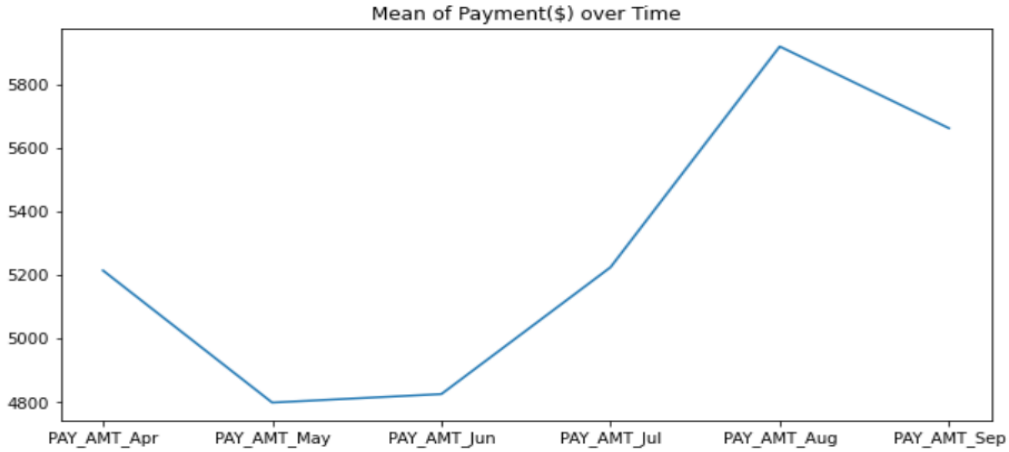
- EDUCATION: Education level for each individual(1 = graduate school; 2 = university; 3 = high school; 4 = others).
- MARRIAGE: Marital status (1 = married; 2 = single; 3 = others).
- AGE: Age (year).
- PAY_0 - PAY_6: Payment records from April to September, 2005. PAY_0 = the repayment status in September, 2005; PAY_2 = the repayment status in August, 2005; . . .;PAY_6 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- BILL_AMT1 - BILL_AMT6: Amount of bill statement (NT dollar). BILL_AMT1 = amount of bill statement in September, 2005; BILL_AMT2 = amount of bill statement in August, 2005; . . .; BILL_AMT6 = amount of bill statement in April, 2005.
- PAY_AMT1 - PAY_AMT6: Amount of previous payment (NT dollar). PAY_AMT1 = amount paid in September, 2005; PAY_AMT2 = amount paid in August, 2005; . . .;PAY_AMT6 = amount paid in April, 2005.

Taking a quick look at the data we see a mix of categorical and continuous variables. One unexpected and welcome surprise was the lack of missing values among all the variables.

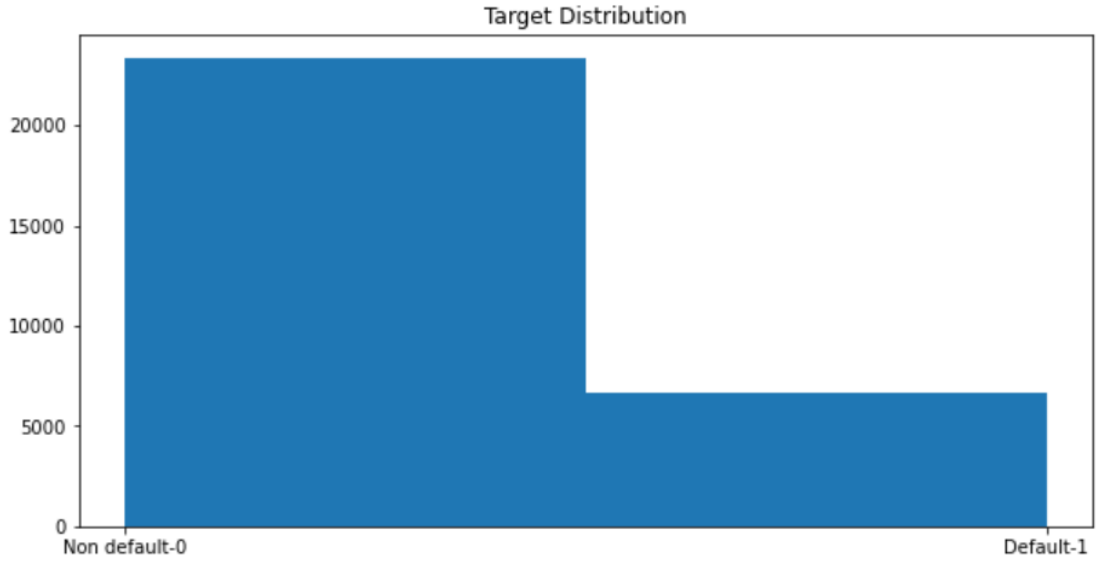
Average Balance over Time($):- . The graph below shows the average amount of balance over the previous 6 months (April to September). The amount of bill statements are observed to increase as the months change from April to September.



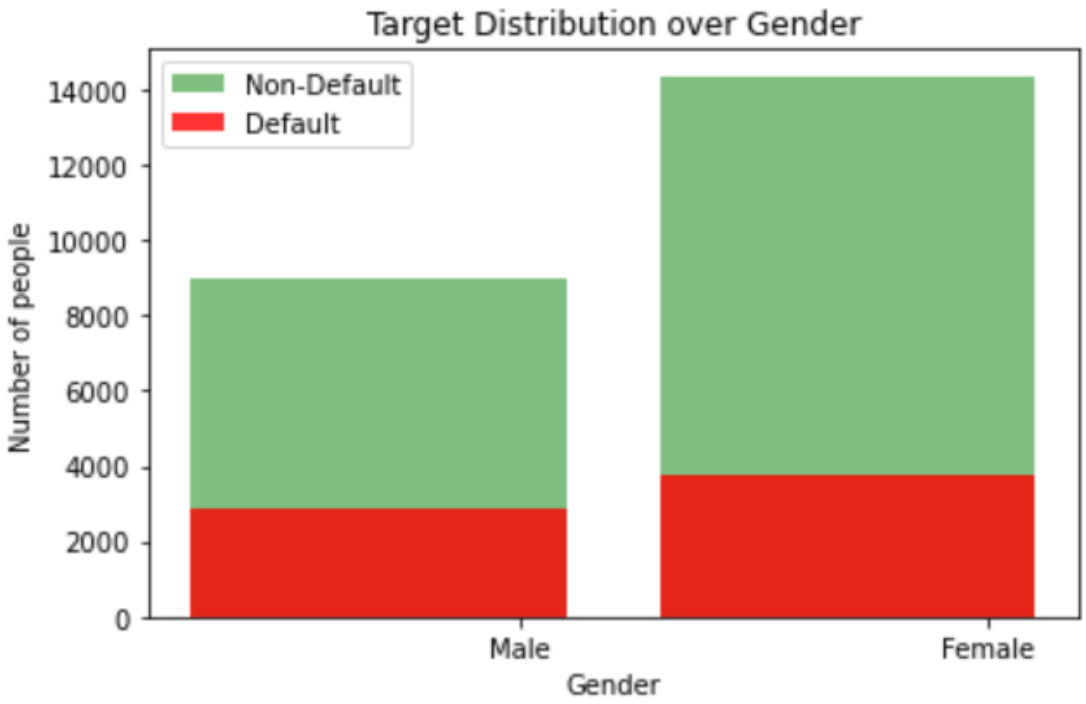
Average Payment over Time($):- . The graph below shows the average amount of payment over the previous 6 months (April to September). The amount of bill statements are observed to fluctuate as the months change from April to September. The mean payment drops from 5200 from April to almost below 5000 in May and June. It then rises above 5200 in July, reaching a high of above 5800 in August. There is no clear trend.



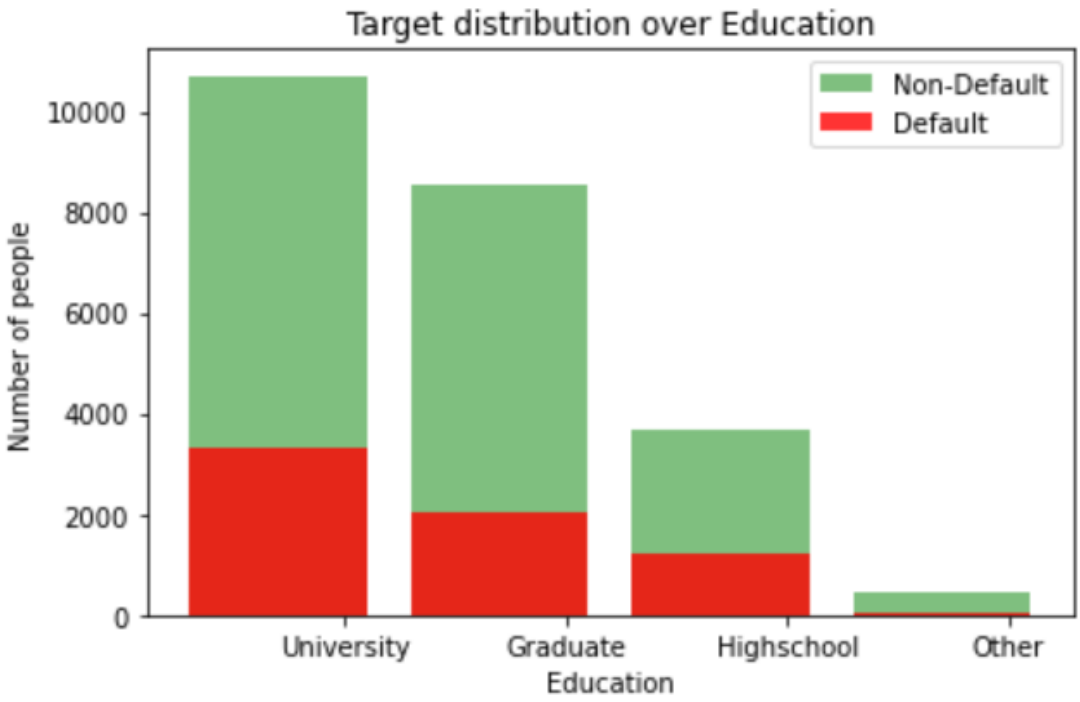

Target Distribution:- The target distribution is highly skewed. The number of non-defaults (23,364) are far greater than the number of defaults which is 6,636. This is not surprising since most people pay their credit card bills on time. In total defaults make up about 22.12% of all instances in the dataset.



Target Distribution over Gender:- The bar graph compares the number of defaults based on the gender. Red represents the Defaults and green, the Non-defaults.




Target Distribution over education:- The bar graph compares the number of defaults based on their education. Red represents the Defaults and green, the Non-defaults. Most defaults are from people having education level of University closely followed by Graduate. Non-defaults are almost similar for these 2 categories.

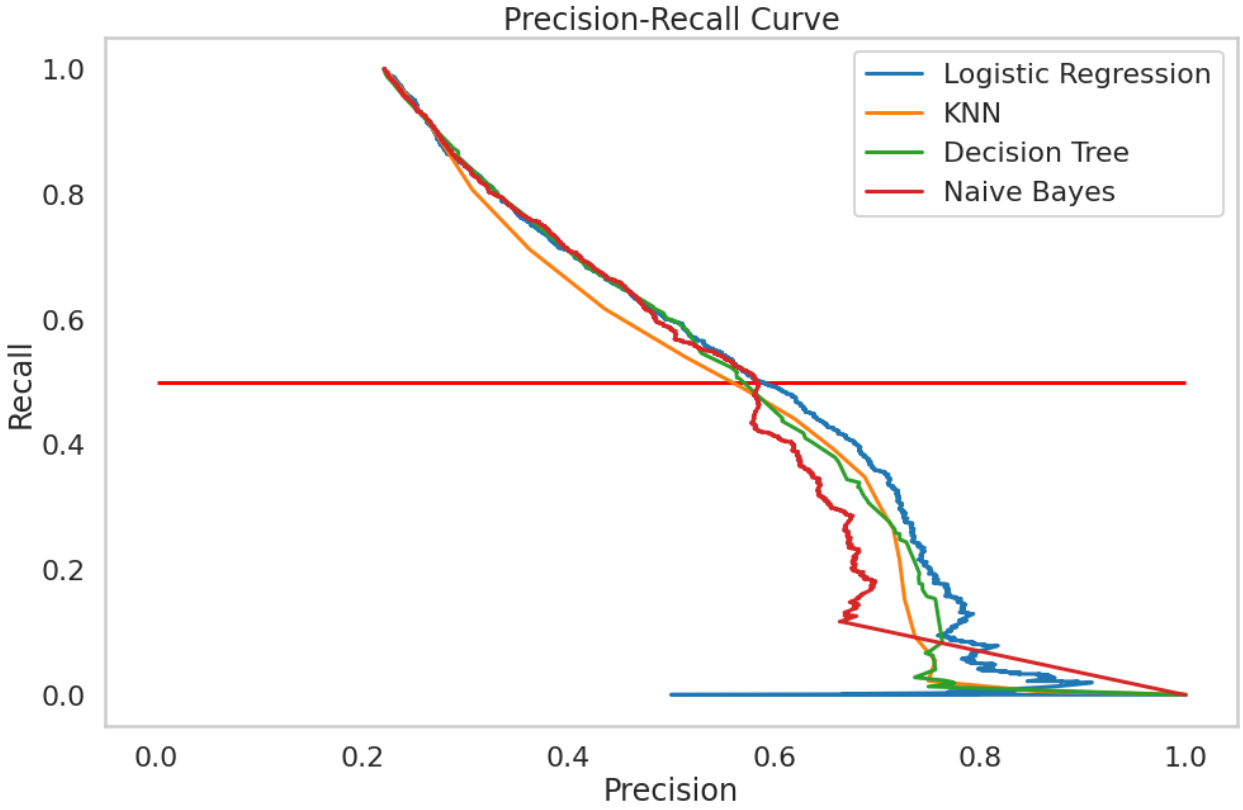



Model Results:

The table below summarizes the accuracy, precision, recall, and AUC of each of our models. Each of the models we ran outperformed the baseline across every metric listed. Our logistic regression and SVM are top performing models. SVM generated better accuracy and precision scores while logistic regression had a better recall score. Given the imbalance in the dataset, Accuracy is not an appropriate metric for model evaluation. The more imbalanced a dataset is, the higher an accuracy score can be achieved by just predicting the dominant class. Our objective is to predict credit defaults which are the minority class in this dataset. Thus, we need to ignore accuracy when evaluating our models.

| | Baseline | Logistic Regression | KNN | SVM | Decision Tree | Naive Bayes |
|---|---|---|---|---|---|---|
| Accuracy | 0.779 | 0.824 | 0.821 | 0.826 | 0.819 | 0.806 |
| Precision | 0.000 | 0.694 | 0.688 | 0.704 | 0.683 | 0.683 |
| Recall | 0.000 | 0.368 | 0.348 | 0.367 | 0.339 | 0.228 |
| AUC | 0.500 | 0.771 | 0.754 | N/A | 0.767 | 0.765 |

Below is a graph showing the Precision-Recall curve for our models. In it we see logistic regression outperforming the rest of our models.




## Conclusion and Discussion

Once again, we are predicting credit defaults. The cost of a false negative is far higher than the cost of a false positive. That is to say the cost of misclassifying a default as paid off is higher than misclassifying a paid off loan as a default. Our best performing model should produce the lowest number of false negatives. We want to prioritize maximizing recall at the cost of precision. Thus, recall is the metric we use for model evaluation. Based on this, our logistic regression model was our best performing model.

There are several improvements we could have made in our analysis. Oversampling is a technique commonly used when dealing with unbalanced datasets. However, we decided not to implement this. We also could have done more feature engineering such as creating new interaction features that could improve the predictive power of our model.

## SOURCES

[1] Three Challenges Faced By Today's Private Money Lenders And How To Weather Them (forbes.com)
[2] Credit Card Industry Overview: Analysis & Trends in 2021 (businessinsider.com)
[3] • Taiwan: credit card number 2020 | Statista
[4]Prediction of credit card defaults through data analysis and machine learning techniques.