# Project 3 Summary

## A New Credit Card Offering? An Assessment via Machine Learning

(Predicting Credit Card Default by Classification)

Charlie Lew

**Abstract**

The purpose of this project is to leverage classification algorithms in order to predict whether or not a credit card holder will default. In order to present this as a business case, the story revolves around a fictitious financial institution called Obsidian Banking Corporation (OBC) who are in the midst of deciding if launching a new credit card to its current customers is a sound financial investment. Based on the current findings, the best classifier in predicting a default is Gradient Boosting with an Area Under the Curve (AUC) score of 0.78. This translates, while tuning the threshold parameter, to a maximum return of USD 86 million in the next six months with a fixed Annual Percentage Rate (APR) of 19.99%. Hence, with a positive return, the OBC has decided to move ahead with the new credit card program. The summary will also discuss next steps as how to improve the model further.

**Introduction**

The reported revenue on credit card interest to financial providers in the United States as of March 2019 is USD 113 billion with a near USD 1.03 Trillion in credit card balances. At Obsidian Banking Corporation (OBC), we are proud to part of this still growing sector. Our newly formed credit card services arm which issues the OBC Credit Card has been successfully operational for six months. Up until now, we have reported USD 653 million in revenue from interest with a net after defaults of USD 330 million. Hence, OBC is operating with a healthy balance sheet in the black. However, there has strong request from OBC's customer base to offer a new type of credit card that provides incentives such as travel points, cash back etc. In order to decide if OBC should move forward, machine learning classification algorithms will be utilized.

**Data, Tools & Methodology**

The OBC Accounts & Financials department were able to provide a dataset of 30,000 of current customers. The names of our customers have been anonymized to ID numbers to keep their identities confidential. Along with their IDs, other information such demographics, credit limit, education level, card balance and payment habits for the past six months are included. The last item that was also provided was whether or not a customer defaulted. This is important as this feature will designated as a target and machine learning will be used to predict whether or not a potential customer defaults on their payment. This will help us determine the potential return if we do embark on the new credit card campaign. Table 1 below shows the features and target in the dataset.

**Table 1: Feature and target of client data set**

| Feature | Type | Description |
|---|---|---|
| **Customer ID** | Integer | Anonymized customer ID |
| **LIMIT_BAL** | Integer | Amount of given credit |
| **SEX** | Integer | Gender, 1=male, 2=female |
| **EDUCATION** | Integer | Range from 1 to 6 |
| **MARRIAGE** | Integer | Marital status, range from 1-3 |
| **AGE** | Integer | Age in years |
| **PAY_1** | Integer | Repayment in September 2005 |
| **PAY_2** | Integer | Repayment in August 2005 |
| **PAY_3** | Integer | Repayment in July 2005 |
| **PAY_4** | Integer | Repayment in June 2005 |
| **PAY_5** | Integer | Repayment in May 2005 |
| **PAY_6** | Integer | Repayment in April 2005 |
| **BILL_AMT1** | Integer | Bill amount for September 2005 |
| **BILL_AMT2** | Integer | Bill amount for August 2005 |
| **BILL_AMT3** | Integer | Bill amount for July 2005 |
| **BILL_AMT4** | Integer | Bill amount for June 2005 |
| **BILL_AMT5** | Integer | Bill amount for May 2005 |
| **BILL_AMT6** | Integer | Bill amount for April 2005 |
| **PAY_AMT1** | Integer | Amount paid September 2005 |
| **PAY_AMT2** | Integer | Amount paid August 2005 |
| **PAY_AMT3** | Integer | Amount paid July 2005 |
| **PAY_AMT4** | Integer | Amount paid June 2005 |
| **PAY_AMT5** | Integer | Amount paid May 2005 |
| **PAY_AMT6** | Integer | Amount paid April 2005 |
| **Default** | Integer | 0=no, 1=yes, used as Target |

The analysis was exclusively performed in Python with Scikit-Learn, Pandas, Seaborn and Matplotlib. Several Machine Learning (ML) classifiers were used. The metric chosen to best represent how the classifiers are performing to capture a default or not is the Area Under the Curve or AUC. Table 2 below shows the classifiers used in this study and the corresponding AUC score. From Table 2 below, with the exception of Gaussian Naïve Bayes, the AUC scores are quite similar. In this case, Gradient Boost with an AUC of 0.783 was chosen as it scored the highest. It is highlighted in bold font in Table 2.

**Table 2: Regression classifiers and corresponding AUC scores**

| Classifiers | AUC |
|---|---|
| k-Nearest Neighbours (kNN) | 0.749 |
| Logistic Regression | 0.727 |
| Naïve Bayes (Bernoulli) | 0.737 |
| Naïve Bayes (Gaussian) | 0.629 |
| Support Vector Machine (SVM with RBF) | 0.714 |
| Random Forest (RF) | 0.764 |
| Extra Random Trees | 0.752 |
| AdaBoost Classifier | 0.749 |
| CatBoost Classifier | 0.782 |
| ***Gradient Boost (GB)*** | ***0.783*** |
| XGBoost (XGB) | 0.780 |

The AUC is acceptable but at this point does not provide any meaningful insight on the possible return on investment. Figure 1 below shows a confusion matrix laying out the cost-benefit of each prediction set based on the GB classifier.

| **Actual** \ **Predicted** | No Default | Default |
|---|---|---|
| No Default | **True Negative (TN)**<br>*Predict*: No Default<br>*Actual*: No Default<br>*Action*: Offer to customer<br><br>*Revenue*: $28,000 from interest | **False Positive (FP)**<br>*Predict*: Default<br>*Actual*: No Default<br>*Action*: No offer to customer<br><br>Loss: $0 |
| Default | **False Negative (FN)**<br>*Predict*: No Default<br>*Actual*: Default<br>*Action*: Offer to customer<br><br>Loss: -$48,500 from defaults | **True Positive (TP)**<br>*Predict*: Default<br>*Actual*: Default<br>*Action*: No offer to customer<br><br>*Cost*: $0 |

**Figure 1:** Confusion matrix for cost benefit analysis

From Figure 1, the return is as follows depending on how the model is able to capture the defaults or not. First, a true positive (TP) implies that the model was able to capture defaults and therefore in our case, we would not offer a credit card loan to the customer. Hence, no loss or benefit here. On the diametric end, a true negative (TN) implies that we predicted no default and the actual customer defaulted. Therefore, we want to offer this customer the credit card loan and assuming this person accepts, OBC makes revenue on interest. A false positive (FP) implies an incorrect prediction on the default and thus still no offer to the customer. Finally, a false negative (FN) means an incorrect no default prediction and thus the bank has to cover the customer's remaining balance. Figure 2 below shows the same confusion matrix but with the number of predicted true positives etc. in each cell. The numbers are determined by tuning the threshold to maximize profit.

| | |
|---|---|
| **TN**<br>**$28,000**<br>**4,476** | **FP**<br>**$0**<br>**211** |
| **FN**<br>**-$48,500**<br>**861** | **TP**<br>**$0**<br>**452** |

**Figure 2**: Predicted numbers in each confusion matrix cell

If we multiply the dollar amount and the number in each cell and sum all of them together, the net revenue is comes up to USD 87 million assuming an APR of 19.99%. Hence, via this prediction, it is strongly recommended that OBC moves forward with the new credit card campaign.

## Closing Remarks & Future Work

This study required the use of classification algorithms to predict a credit card default in order to determine if our bank OBC will move forward on a new credit card offering. The best classifier so far is Gradient Boost (GB) with area under the curve (AUC) score of 0.783. Examining the confusion matrix, the maximum net return is USD 87 million. Hence, OBC should move forward with the new credit card program. In the future, the model can be improved further by examining stacking or ensembling and by investigating the use of neural networks. Finally, an internal web portal can be setup where the Finance department can use this application in order to determine if further new programs should move forward.