University Of Dublin

School Of Mathematics and Statistics

COMPARISON AND EVALUATION OF DIFFERENT MACHINE LEARNING METHODS AT PREDICTING CREDIT CARD DEFAULT

By

Sidney Harshman-Earley
Denis O'Riordan

in partial fulfilment of ACM40960 -
Projects in Maths Modelling required
for the completion of the degree
MSc in Data and Computational Sciences

July 2022

Abstract

(the spacing is set to 1.5)

no more than 250 words for the abstract

- a description of the research question/knowledge gap what we know and what we don't know
- how your research has attempted to fill this gap
- a brief description of the methods
- brief results
- key conclusions that put the research into a larger context

Contents

1 Introduction

A credit card is a financial instrument issued by banks with a pre-set credit limit allowing customers to make cashless transactions. Each month or at other intervals, a statement is issued by the credit card provider with details of spending history, interest charged, balance and payment deadlines. They provide a very flexible line of credit, rewards and occasional cash-back to users and can be extremely cheap to use if managed correctly.

A default event occurs with regard to a credit card if a customer does not pay the amount they have due within 120 days of their latest statement date. The most significant risk to lenders is a large and unexpected number of customers failing to meet their credit repayment obligations. This is known as credit default risk.

According to the Federal Reserve, credit card usage declined significantly in the United States during the most recent pandemic.According to the G.19 Consumer Credit statistical release, revolving consumer credit fell more than $120 billion (11 percent) in 2020, the largest decline in both nominal and percentage terms in the history of the

Given that banking portfolios can be worth billions, even marginal improvements in preventing and reducing defaults can be considered significant.

Many banks and lenders are legally obligated to create models that predict credit defaults. These models advise banks capital requirements which ensure that the bank can absorb losses brought about by a significant increase in credit defaults.

The Big data paradigm has revolutionised the banking industry, changing the way financial institutions operate. The aftermath of the past financial crisis has been slowly rectifying and people are nowadays better off in terms of job opportunities and financial health.

Credit card default prediction is based on the historical data of credit card customers. The use of corresponding methods to predict and analyse credit card customer default behaviour is a typical classification problem.

There have been advances in Machine Learning (ML) techniques such that models now exist, that typically outperform Logistic Regression. These models have yet to be deployed because of their opaque (or Black-Box) nature.

The aim of this project is to apply different machine learning techniques to find

The following report describes the process undertaken to build, tune, evaluate and compare a number of models, beginning with an industrial size dataset, cleansing and formatting the data before using it to train, tune and evaluate the chosen models,

Credit card default prediction is based on the historical data of credit card customers.

Credit default prediction is central to managing risk in a consumer lending business. Credit default prediction allows lenders to optimize lending decisions, which leads to a better customer experience

and sound business economics. Current models exist to help manage risk. But it's possible to create better models that can outperform those currently in use.

2 Review of Literature

2.1 Taiwan Data

There have been numerous articles and papers within the scope of using ML methods to predict credit default, this includes the prediction for credit card default.

Studies examining credit card default have been concentrated, mainly using the data used originally used as part of Yeh & hui Lien (2009) . This data is currently freely available as the Default credit card clients Data Set on the UC Irvine Machine Learning Repository.[1] This data, collected in October 2005, is from a cash and credit card issuing bank in Taiwan, the targets were credit card holders of the bank. Among the total 30,000 observations, 22.12% are the cardholders who defaulted on payment. This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This data uses 23 variables as explanatory variables including a mix of personal information(age,gender,marital status and education level), amount of credit given, historical bill statements, and historical payment information.
This initial study, Yeh & hui Lien (2009), compared six classification algorithms - K-nearest neighbour, Logistic regression, Discriminant analysis Nave Bayesian classifier, Artificial Neural Networks, Classification Trees. In the classification accuracy, the results show that there are little differences in error amongst the six methods. The generated probability of default by the Artificial Neural Network most closely resembled the actual probability of default. The actual probability of default was estimated using a novel "Sorting Smoothing Method".

Other research on predicting credit card default is subsequent years utilised this data to train and evaluate model. The following is a sample of articles available, applying a wide variety of Machine Learning methods to this classification problem.
Another study, (Neema & Soibam 2017), took a similar approach in choice of methods but attempted to predict the best possible cost-effective outcome from the risk management perspective. Again, K-Nearest Neighbour Logistic Regression, Classification Trees and Discriminant Analysis were evaluated but also Naive Bayes and Random Forest classifiers were included. This was evaluated using a cost function which gave a higher cost to defaulters classified not correctly as they are the minority in the data. Defaulted payments can prove more costly to a bank rather than potential customers wrongly identified as a potential default case. It was concluded that original data with Random Forest algorithm is the best in terms of a good balance on cost versus accuracy.
Yang & Zhang (2018) introduces two new methods used to predict credit card default. Support Vector Machine (SVM) involves using a kernel function to map the predictor data into a high-dimensional feature space where the outcome classes are easily separable. XGBoost and LightGBM, forms of gradient boosted trees algorithm were used, as well as previously tried methods - Logistic Regression and Neural Networks. LightGBM and XGBoost were both deemed to have the

[1] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

best performance in the prediction of categorical response variables.
While other studies have used Neural Networks in predicting credit card defaults, the models used have been vague and little detail has been given on architecture or tuning of the model. Chou & Lo (2018), trialled a range of Networks, experimenting with two to five layers with number of processing units of 64, 32, 16 units. Neural Networks with three layers and 64 units recorded the highest accuracy of all configurations.

Due to a lack of credit card specific data pertaining to defaulting on payment, all available studies which predict credit card default utilise the Taiwan data which is both region specific, dated over 15 years and obtained at a time when credit card issuers in Taiwan faced a credit card debt crisis.[2]

2.2 Other Literature

In the scope of predicting defaults on other forms of credit such as mortgages, research has also been conducted using data collected by the Central Bank of Ireland, comprising four separate portfolios of over 300,000 owner-occupier mortgage loans of Irish lenders, Fitzpatrick & Mues (2016). It was found that boosted regression trees provided the best classification algorithms for mortgage default prediction.

A 2019 research thesis, Egan (2021), examined a number of high-performing methods in predicting credit default on a Home Credit dataset[3]. Home Credit is an international non-bank financial institution that specializes in lending to people with little or no credit history. This study examined methods that could be deemed explainable - - which consisted of a number of tree-based ensemble methods. The top performing model was deemed to be XGBoost, a form of gradient boosted trees algorithm.

[2]https://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/
[3]https://www.kaggle.com/c/home-credit-default-risk

3 Data

3.1 Underlying Data

Data used for this project was obtained from the Kaggle website which is a crowd-sourced platform and community to train and challenge data scientists to solve complex data science, machine learning and predictive analytics problems.[4] One of Kaggles main features is it's competition platform.[5] Kaggle allows users to organise and host competitions, these range from commercially-purposed prediction problems to more experimental research competitions. They give entrants the opportunity to grow their data science skills and compete for prizes and gives hosts a outlet to tackle a tough business problem by turning it into a competition and allowing Kaggles user base of 10 million to provide potential solutions to the problem.

Access to the data through this route requires registering with Kaggle and joining the American Express - Default Prediction competition.[6] American Express is a globally integrated payments company. The largest payment card issuer in the world, they provide customers with access to products, insights, and experiences that enrich lives and build business success.[7] This competition provides an industrial scale data set to build a machine learning model to predict credit card default using time-series behavioural data and anonymised customer profile information over the period of March 2017 to March 2018. The target binary variable i.e. default is calculated by observing 18 months performance window after the latest credit card statement, and if the customer does not pay due amount in 120 days after their latest statement date it is considered a default event.

Tabular data

[4] https://www.kaggle.com/getting-started/44916
[5] https://www.kaggle.com/docs/competitions
[6] https://www.kaggle.com/competitions/amex-default-prediction/data
[7] https://about.americanexpress.com/our-company/our-business/our-business/default.aspx

3.2 Data Structure

Four files are provided as part of the Kaggle competition

3.2.1 train_labels.csv

A list of unique customer identifiers `customer_ID` with the target label `target` indicating a default event with `target = 1` indicating a default, `target = 0` indicating no default. There are 458913 observations in this data meaning 458913 unique AMEX customers. 74.1% of customers in the data did not default on payment while 25.9% did

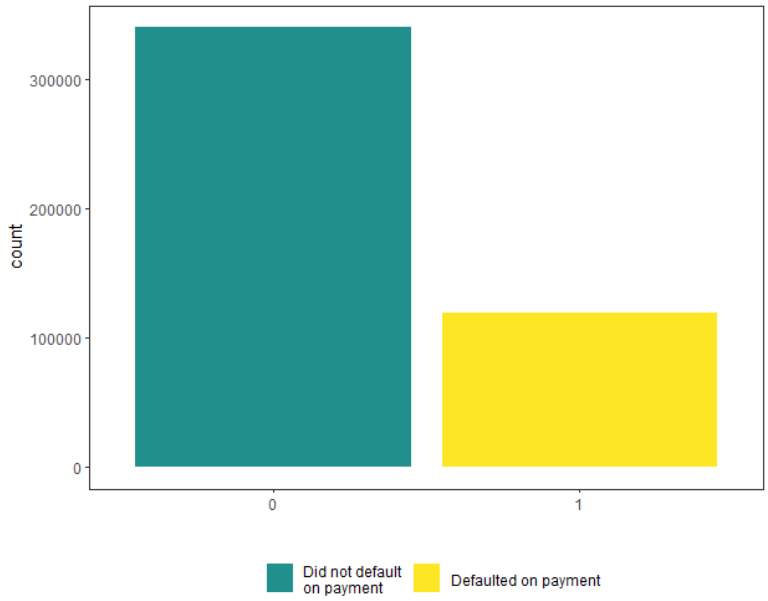

Figure 1: Breakdown of defaults in training data

3.2.2 train_data.csv

3.2.3 test_data.csv

3.2.4 sample_submission.csv

3.3 Data Engineering

3.4 Data Cleaning

4 Methods

4.1 Random Forest (RF)

Random Forests (RF) are a very powerful ensemble classification method. The method uses classification trees and bootstrapping extensively. Ensemble learning is an aggregation of predictions made by multiple classifiers with the goal of improving accuracy. A random forest is a random collection of classification trees estimated on random subsets of the data.

A Classification Tree is an iterative process of splitting the data into partitions based on values of the observations, and then splitting it up further on each of the branches.
The classifier is trained in order to produce "pure" buckets, minimising the entropy or spread of the target variable in each bucket. The majority value of the target variable in a bucket is then used for predictions. The main disadvantage is that Classification Trees suffer from over-fitting and bias and using a single Classification Tree would present an over simplistic model.

In a Random Forest, each Classification Tree is grown independently, and each individual Classification Tree has an equal vote as to what the outcome is. Random forests provide an improvement by means of a small tweak that reduces the dependence among the trees. The main idea is to use only a random subset of the predictor variables at each split of the classification tree fitting step, this is generally taken as $\sqrt{N}$ but can be specified by the user. If we build classifiers on subsets of the variables, then they will behave more independently than if we build them on all of the data. This increases diversity and averaging results across independent classifiers will be more stable than averaging results on dependent ones.

The main parameter of the Random Forests to be tuned is the number of Classification Trees used and is tuned to prevent overfitting and improve efficiency.

4.2 Support Vector Machines (SVM)

linear classifiers, where a linear combination of input features and parameters is employed to define the plane of maximum separation between classes

using a kernel function to map the data X of the input space into a high-dimensional feature space

A support vector classifier seeks to maximize the distance between the classes

inseparable in the original space can be linearly classified in the high-dimensional

space.

4.3 Deep Neural Network (DNN)

5 Results

6 Discussion

introduce dynamic element? Recurrent Neural Network (Hsu et al. 2019)

7 Conclusion

8 Bibliography

Chou, T. & Lo, M. (2018), 'Predicting credit card defaults with deep learning and other machine learning models', *International Journal of Computer Theory and Engineering* **10**, 105–110.

Egan, C. (2021), Improving credit default prediction using explainable ai, Master's thesis, Dublin, National College of Ireland.
URL: *http://norma.ncirl.ie/5146/*

Fitzpatrick, T. & Mues, C. (2016), 'An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market', *European Journal of Operational Research* **249**(2), 427–439.
URL: *https://www.sciencedirect.com/science/article/pii/S0377221715008383*

Hsu, T.-C., Liou, S.-T., Wang, Y.-P., Huang, Y.-S. & Che-Lin (2019), Enhanced recurrent neural network for combining static and dynamic features for credit card default prediction, *in* 'ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 1572–1576.

Neema, S. & Soibam, B. (2017), The comparison of machine learning methods to achieve most cost-effective prediction for credit card default.

Yang, S. & Zhang, H. (2018), 'Comparison of several data mining methods in credit card default prediction', *Intell. Inf. Manag.* **10**(05), 115–122.

Yeh, I.-C. & hui Lien, C. (2009), 'The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients', *Expert Systems with Applications* **36**(2, Part 1), 2473–2480.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417407006719*

9 Abbreviations

Table 1: List of Abbrevivations

Abbreviation	Meaning
AMEX	American Express
DNN	Deep Neural Network
NN	Neural Network
RF	Random Forest
SVM	Support Vector Machines

List of Figures

List of Tables