University College 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

List of Figures

List of Tables

1 Introduction

A credit card is a financial instrument issued by banks and other financial institutions with a pre-set credit limit allowing customers to make cashless transactions. Each month or at different intervals, a statement is issued by the credit card provider with details of spending history, the interest charged, balance, and payment deadlines. As of May 2022, there were 1.5 million active credit cards in circulation in Ireland with daily spending exceeding €1 million.[1]. Further to that, according to research conducted by the Central Statistics Office in 2018, 12.7% of all households have credit card debt but is 21.0% for households with two adults and one to three children under 18.[2]

A default event concerning a credit card occurs if a customer does not pay the amount due within 120 days of their latest statement date. However, the deadline can depend on the issuing institution. 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. Credit default prediction is central to managing risk in a consumer lending business. Financial institutions are legally obligated to create models that predict credit defaults. Credit card default prediction is based on historical data of credit card customers. The use of models to predict and analyse credit card customer default behaviour is a typical classification problem. These models advise the banks capital requirements, which ensure that the bank can absorb losses brought about by a significant increase in credit defaults. Credit default prediction also 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 is it possible to create better models that outperform those currently in use? Given that banking portfolios can be worth billions, even marginal improvements in preventing and reducing defaults can be considered significant.

The big data paradigm has revolutionised the banking industry, changing the way financial institutions operate. Developments in technology has revolutionized the customer experience and has also granted access to a ever-growing volume of structured and unstructured information. Institutions can leverage this new information in many ways including predicting credit default. There have been advances in Machine Learning (ML) techniques such as representation learning methods e.g. Neural Networks, ensemble learning methods such as Random Forest and linear separating methods e.g. Support Vector Machines. These models have yet to be widely deployed because of their opaque nature i.e. Black-Box methods. From a regulation perspective, transparency and explainability are necessary when deploying credit default prediction models in practice. However, they may still provide value to financial institutions and may in the future be accepted as valid, clear and sound methods for predicting credit card default.

The aim of this project is to apply a range of advanced machine learning techniques to predict credit card default using the historical data of credit card customers. The following report describes the process undertaken to 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 based on the historical data of credit card customers.

[1] https://www.centralbank.ie/statistics/data-and-analysis/credit-and-debit-card-statistics/daily-credit-and-debit-card-statistics

[2] https://www.cso.ie/en/releasesandpublications/ep/p-hfcs/householdfinanceandconsumptionsurvey2018/debtandcredit/

2 Review of Literature

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.

2.1 Taiwan Data

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.[3] 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

[3] 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.

More advanced forms of Neural Networks have also been applied to the problem, Hsu et al. (2019) proposed using a Recurrent Neural Network (RNN) feature extractor with a Gated Recurrent Unit (GRU) on credit card payment history to leverage the time dependencies embedded in these dynamic features of historical credit card data. Recurrent Neural Networks (RNN) are specifically designed to use recursive architecture to extract patterns from input sequences (Goodfellow et al. 2016). They have been proven useful in applications that heavily rely on time-variant features. As a result, it is natural to consider RNN models as feature extractors for customer behaviour that often appear as sequences in financial data.

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.[4]

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[5]. 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.

[4]https://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/
[5]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, an online platform and community for data scientists.[6] One of Kaggles main features is it's competition platform.[7] 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 test and grow their data science skills while competing for prizes and gives hosts an outlet to tackle tough business problems by turning them into a competition and allowing Kaggles user base of 10 million to provide potential solutions.

Access to the data required registering with Kaggle and joining the American Express - Default Prediction competition.[8] American Express is a globally integrated payments company and the largest payment card issuer in the world.[9] 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.

3.2 Data Structure

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% defaulted.

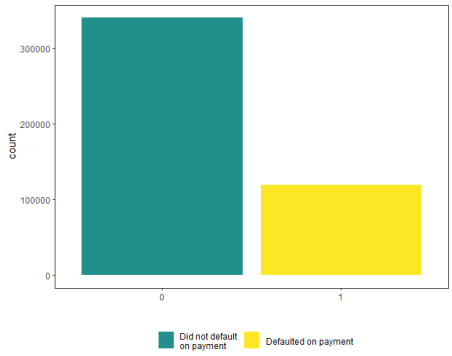

Figure 1: Breakdown of defaults in training data

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

train_data.csv

This is a tabular data set corresponding to the training data, consisting of 190 features. It contains multiple statement dates per customer_ID, each observation corresponds to a monthly statement for a customer. The data is ordered firstly by `customer_ID`, corresponding with the train_labels.csv file and secondly ordered chronologically by statement date. The dataset contains aggregated profile features for each customer at each statement date. There are 177 numeric features which are anonymised and normalized, and fall into the following general categories:

Table 1: Variable Categories

Variable Name	Category
D_*	Delinquency variables
S_*	Spend variables
P_*	Payment variables
B_*	Balance variables
R_*	Risk variables

There are 11 categorical variables - B_30, B_38, D_114, D_116, D_117, D_120, D_126, D_63, D_64, D_66 and D_68.

The remaining two unaccounted for features are the variable S_2 which contains the date of the statement and `customer_ID` which can used to match to the target label.

5531451

Table 2: First five rows and columns from train_data.csv

customer_ID	S_2	P_2	D_39
0000099d6bd597052cdcda90ffabf56573fe9d7c79be5fbac11a8ed792feb62a	2017-03-09	0.9384687	0.001733339
0000099d6bd597052cdcda90ffabf56573fe9d7c79be5fbac11a8ed792feb62a	2017-04-07	0.9366646	0.005775443
0000099d6bd597052cdcda90ffabf56573fe9d7c79be5fbac11a8ed792feb62a	2017-05-28	0.9541803	0.091505397
0000099d6bd597052cdcda90ffabf56573fe9d7c79be5fbac11a8ed792feb62a	2017-06-13	0.9603836	0.002455224
0000099d6bd597052cdcda90ffabf56573fe9d7c79be5fbac11a8ed792feb62a	2017-07-16	0.9472484	0.002483014

Other Data

The competition data also includes two other files but are not within the scope of this project. test_data.csv is test data used to evaluate entries in the competition. However, labels for this data are not included an the evaluation for the models performance on this data is limited to the competitions own evaluation metric.[10] sample_submission.csv is a sample submission file for the competition in the correct format.

[10]https://www.kaggle.com/competitions/amex-default-prediction/overview/evaluation

3.3 Data Engineering

Brief bit on load, chunking and caching process?

3.4 Data Cleaning

- Remove NA columns,
- remove low variance columns
- remove highly correlated columns
- incremental PCA now with Imputation
  Anything else?

include the “why” - methods don’t work with missing values, very wide dataset - reduce dimensions to increase efficiency

merge label and data

4 Methods

4.1 Random Forest (RF)

Random Forests, Breiman (2001), are a very powerful ensemble classification method. Ensemble learning is an aggregation of predictions made by multiple classifiers with the goal of improving accuracy. The method uses Classification Trees and bootstrapping extensively. A random forest is, effectively, 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 groups of observations or 'buckets', by 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 $\sqrt{N}$ where N is the number of features 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)

Support Vector Machines, Boser et al. (1992) and Cortes & Vapnik (1995), is a classification method based on projecting the data into a higher dimensional space where distinction between the target variable groups is clearer. With regards to a binary classification problem, data is linearly separable if there exists a separating hyperplane in the input space of the training data that fully separates observations by the value of the target variable. This is known as a linear classifier. In reality data can be quite messy and such hyperplanes are not likely to exist for a specific problem, especially in higher dimension problems.

Support Vector Machines use a kernel function to map the training data of the input space into a high-dimensional feature space. Scenarios that are inseparable in the original space can be linearly classified in the high-dimensional space. A kernel function $K()$ is a generalization of the inner product of the form: $K(x_i, x_h) = \phi(x_i)^T \phi(x_h)$ where $\phi()$ is some mapping function of the data and x_i,x_i are generic input observations. To avoid the explicit mapping to the higher-dimensional space Q

which could be computationally inefficient, the kernel computes the inner products via the kernel in the original input feature space $\mathbb{R}^V$. The kernel function returns the inner product between two points in the enlarged feature space with little computational cost even in very high-dimensional spaces. Kernelizing methods can help to account for non-linear patterns and separate groups and classes for clustering or classification.

The predictive performance of SVM can be very sensitive to the choice of the kernel and the cost, which is a measure of tolerance a function has to observations violating hyperplane and its margins.A classifier with a large cost will strive to getting all the (training) data points classified correctly, conversely, a classifier with a small cost will be tolerant to a certain degree of misclassified observations.

Types of kernels + chosen

4.3 Deep Neural Network (DNN)

Neural networks are machine learning methods that simulate the biological learning mechanism of the brain (Goodfellow et al. (2016), Zhang et al. (2021)). Neural networks allow to learn representation features encompassing complex relations between inputs and output. Representation learning is the use of Machine Learning to not only learn the mapping from representation features to output, but also to learn the representation itself. Deep Learning methods introduce representations that are expressed in terms of other simpler representations, thus enabling to derive complex concepts out of simpler concepts.

In practice, a neural network is a series of layers, made up of nodes with subsequent layers separated by activation functions.

4.4 Evaluation

5 Results

6 Discussion

Context of Results

Compare to previous studies?

One of the main obstacles for contextualising the results is the lack of information of the variables included as part of the data. The only detail of the features was the category to which it belonged, see Variable Categories table. Due to this, little inference can be made on the variables and the main outcome of this study is the performance of the models. Particularly it would be difficult to comment on the Delinquency and Risk variables as there is no further information to what they constitute and is impossible to infer any further information from them.

The choice of models employed for this project was relatively arbitrary. A selection of three well known high-performing models which were seen to perform reasonably well in previous studies using the Taiwan data were chosen. One ensemble learning, one deep learning model and one separation model were comparatively evaluated. This unfortunately does not allow us to benchmark the performance against internal models used by AMEX or other banks. This benchmark would allow analysis on whether any of the models employed in this study could provide value and risk mitigation. Alternatively a relatively simple and transparent model such as logistic regression could provide a benchmark for the models.

A small sample of potential models were used in this project. Models used in studies on other data that may improve performance independently or in tandem to other models. include K-Nearest Neighbour,Linear Discriminant Analysis,Logistic Regression, XGBoost and LightGBM.

Other methods of evaluating and comparing the models could be employed in future. Neema & Soibam (2017) used a cost function when comparing methods to predict customers credit card default. A higher penalty or cost was given to defaulters classified incorrectly. A range of cost factors were used to identify a model with the best accuracy in predicting a defaulter in a cost- effective manner. _____ compared models using a Receiver Operating Characteristic (ROC) curve and calculating the Area Under the Curve (AUC). The ROC curve plots true positive rate (Sensitivity) versus false positive rate (1 - Specificity) and it illustrates the diagnostic ability of a binary classifier. It works on binary classifiers that produce a probability of an outcome. The ROC determines how often will a randomly chosen 1 outcome have a higher probability of being predicted to be a 1 outcome than a randomly chosen true 0. The larger the area under the curve - AUC, the better is the discrimination. A perfect classifier would have AUC = 1. A classifier not better than random guessing would have AUC = 0.5 and the corresponding ROC curve would resemble a plot of the identity function ($y = x$).

For **XYZ** models an observation was, and hence a prediction was made for, each monthly statement detail for a customer. There were up to 13 separate observations in the training data for each customer with the possibility of contradicting predictions made for a customer. For example in one month of a customers data may have been predicted to default on payment while another months did not predict default. Only one guess should be made per customer, multiple contradicting predictions

are impractical. Hence, only one prediction should be made per customer. A majority vote system for each customer could be employed to give a singular prediction as to whether default will occur. Alternatively a weighted vote, with increased weighting for default predictions could create a more risk-adverse system with potential defaults more likely to be caught.

7 Conclusion

This extends model choice to include black-box models that are better able to identify the non-linear relationships between the target variable and the predictor variables. This means that banks and lenders have the potential to use models which could help them reduce expenses brought about by loan defaults and/or identify and avoid lending opportunities that are too risky.

8 Bibliography

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992), A training algorithm for optimal margin classifiers, *in* 'Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory', ACM Press, pp. 144–152.

Breiman, L. (2001), 'Random forests', *Mach. Learn.* **45**(1), 5–32.
URL: *https://doi.org/10.1023/A:1010933404324*

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.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Mach. Learn.* **20**(3), 273–297.
URL: *https://doi.org/10.1023/A:1022627411411*

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*

Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep Learning*, MIT Press.
http://www.deeplearningbook.org.

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*

Zhang, A., Lipton, Z. C., Li, M. & Smola, A. J. (2021), 'Dive into deep learning', *arXiv preprint arXiv:2106.11342* .

9 Appendices

9.1 Appendix A: additional tables

Insert content for additional tables here.

9.2 Appendix B: additional figures

Insert content for additional figures here.

9.3 Appendix C: code

Insert code (if any) used during your dissertation work here.

10 Abbreviations

Table 3: List of Abbrevivations

Abbreviation	Meaning
AI	Artificial Intelligence
AMEX	American Express
AUC	Area Under the Curve
DNN	Deep Neural Network
GRU	Gated Recurrent Unit
ML	Machine Learning
NN	Neural Network
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SVM	Support Vector Machines