Machine Learning: Credit-Scoring

Applying Naïve Bayes and KNN algorithms to classify credit-risk of cardholders

Islam Ibrahim & Saeed Ahmed
City, University London

I. Introduction

It is imperative that retail banks engage in robust credit-scoring to determine if a cardholder is likely to default. There is a great opportunity to apply Machine Learning (ML) techniques in this field to handle the analysis of vast amounts of credit data and minimise reliance on human input. This will allow retail banks and financial institutes to predict whether a customer is worthy of credit based on readily accessible data, such as their previous payments behaviour or their characteristics [6]. Credit scoring is the technique used by lenders to quantify the borrower's credit risk. In this case, credit risk is defined as a binary category as either:

- Credit-worthy – cardholder classified as would 'Not Default'
- Credit-unworthy – cardholder classified as would 'Default'

II. Objectives

- Apply two ML algorithms to predict credit-scoring of a cardholder by leveraging both their financial and personal attributes.
- Analyse which type of feature selection produces better results by comparing the accuracy of the ML model.

III. ML Techniques

The supervised learning algorithms of K-Nearest Neighbour (KNN) and Naïve Bayes (NB) are applied to conduct a binary classification prediction on the data.

Naïve Bayes

- Is based on Bayes theory.
- Assumes the effect of an attribute value on a given class is independent of the values of other attributes [4].
- **Advantage** provides a theoretical justification for other classifiers that do not explicitly use Bayes theorem
- **Limitation** results of Bayesian classification depends heavily on prior probabilities, which may not be available [2].

KNN

- Non-parametric classifier that learns by analogy.
- Using a training set a distance function is introduced between the observations to make predictions without building a model - the usual Euclidean distance function is used in this study [3].
- Given an unknown sample – KNN classifier searches the pattern space for KNN closest to unknown sample.
- **Advantage**: can achieve strong classification performance without prior assumptions of the training set distributions.
- **Limitations**: choice of k heavily influences the predictive accuracy of the model.

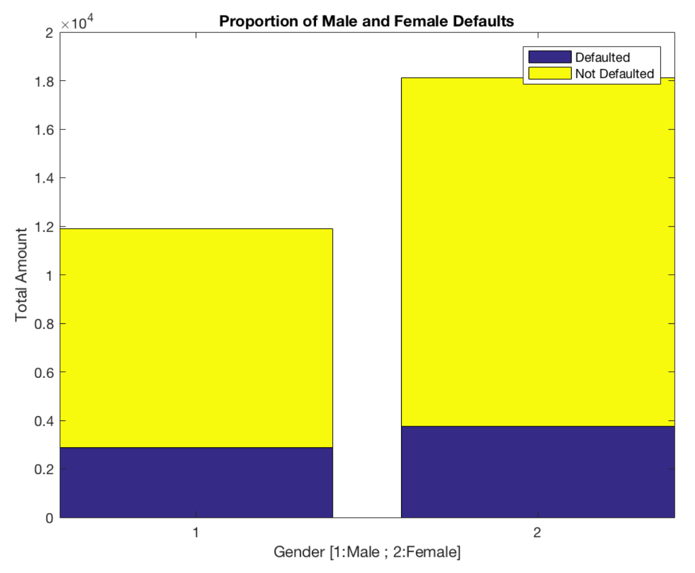
IV. Data

- This study used payment data of 30,000 credit card holders at a retail bank in Taiwan from the UCI Library[1], comprising both five-months of previous payment information and characteristics of the cardholder as the **predictor variables** which are a mix of continuous and discrete data. The target **response variable** is a binary classification that states whether the cardholder **'Defaulted'** or **'Not Defaulted'.** The techniques were applied to:

1. The cardholders' personal characteristics
2. The previous payment behaviour of the cardholder

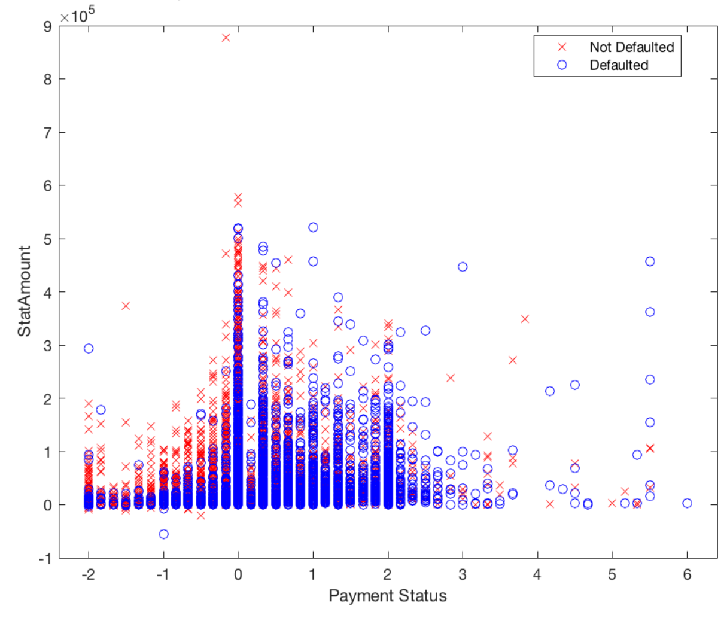
Personal Characteristics

- The first set of predictor variables were **discrete** comprising the cardholders age, gender, marital status and educational level.
- The figure below illustrates the proportion of cardholders in the data that defaulted and did not default based on their gender. Of those that default, 4% more were males than females.



Payment Information

- Second set of predictor variables were **continuous** comprising the prior payment behavior of the cardholder.
- Visualised below, where a 'payment status' of -1 means the cardholder paid duly and 1 indicating the cardholder was one month late and so on – where the majority of those who defaulted were skewed to the right, shown by the blue circles.



- The data was randomly divided, 60% of the data into a training set to train the algorithms and the rest into the test set to validate the model using the cross-validation partitioning function.

V. Hypothesis

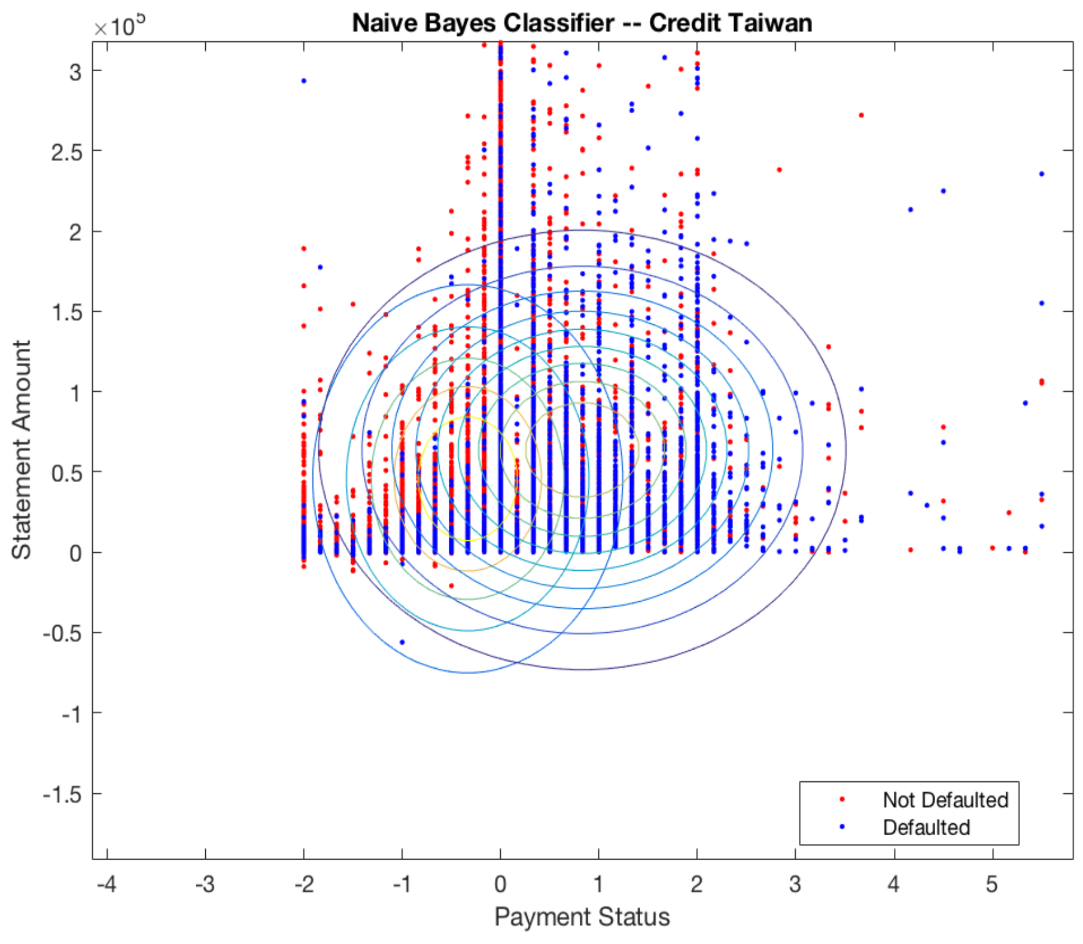
- Payment information will yield higher predictive accuracy than using personal characteristics.
- NB and KNN will yield similar predictive accuracies [2].

VI. Results and Evaluation

In order to evaluate the results, first the ML models were applied to both sets of **predictor variables** described in Section II on the training set. The ML models were then evaluated using the test set and the accuracy of each model was determined. After evaluation of both ML algorithms the maximum accuracies are shown below:

	Predictive Accuracy	
	Payment History	Personal Characteristics
Naïve Bayes	77.7%	78.1%
KNN	80.4%	77.5%

From this table we can see that both algorithms produce relatively similar accuracies as stated in the hypotheses. Although KNN using payment behaviour achieves the greatest accuracy.

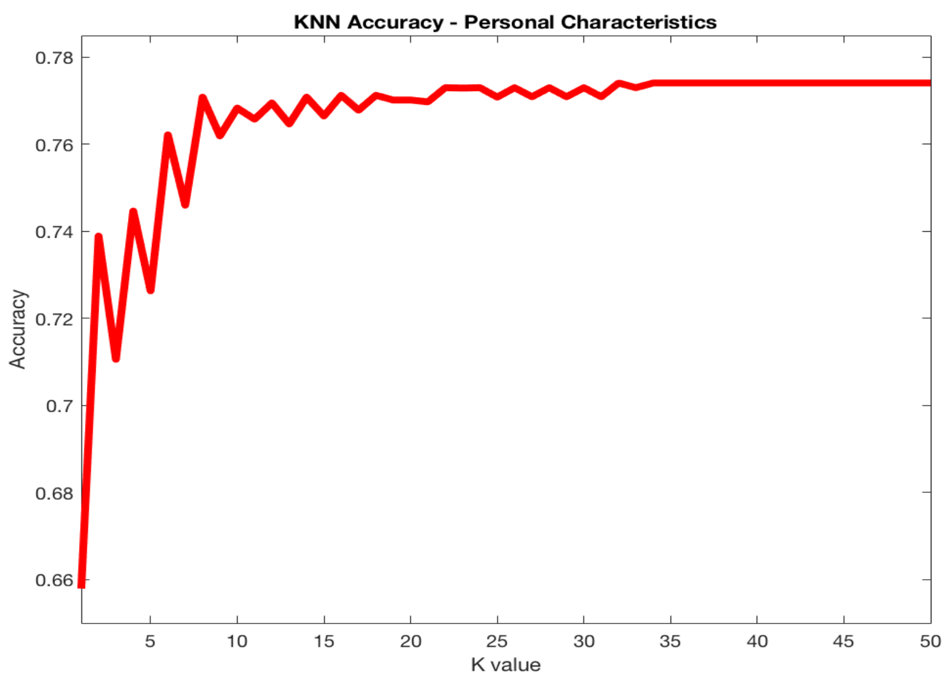


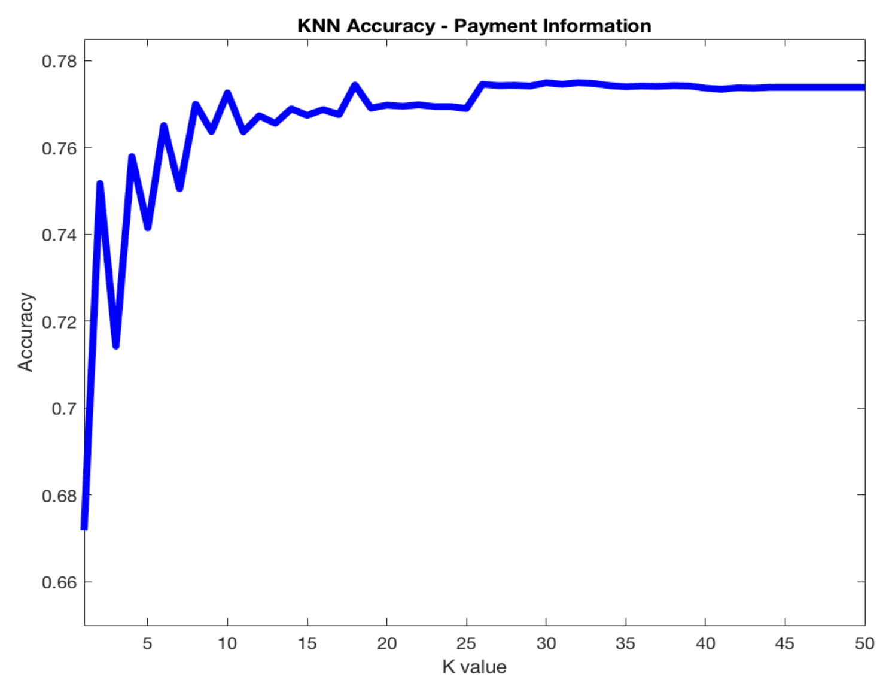
Naïve Bayes

- The Naive Bayes model gives an accuracy of approximately 78% for both payment histories as well as for general characteristics. However, payment histories will achieve a slightly higher level of accuracy, and the ability to visualise the clusters, as shown, makes it more beneficial for ML.
- The two clusters overlap significantly, although it is clear that cardholders who defaulted were skewed to the right of the 'Payment Status' axis. Which indicated a tendency to delay in payment.

KNN

- Based on the value of k, the k number with the smallest distance in the training set are assigned the same classification as the KNN in the test set. This is compared to the actual classification from the training set to give the model **accuracy** for that value of k.
- Predictive accuracy of the KNN algorithm was tested on k-values 1-50.





- The red-line shows the effect on the predictive accuracy of the KNN model at different k-values using personal characteristics of the cardholder as the predictor variables.
- The blue-line shows the same while using payment history of the cardholder as the predictor variables.
- Whereas the accuracy reaches a maximum and plateaus at k-value of circa 30 for both, an optimal k-value is deemed at k-value of 15 as increasing computational demand as a result does not produce a justifiable increase in predictive accuracy.

VII. Critical Review and Analysis

- The use of payment history as the feature selection for the ML models gives more accurate predictions for credit-scoring than that of personal characteristics.
- Higher bill amounts and a cardholders previous tendency to delay in payment indicates a higher probability of defaulting, thus allowing the ML algorithms to cluster the instances.
- Although KNN gives somewhat better accuracy, NB allows the ability to visualise the clusters and will permit greater interpretability.
- Optimum k-value of 15 was inferred in terms of predictive accuracy for the KNN model - where additional computational demand of higher k-values is not conducive of sufficient increases in predictive accuracy.
- This ties in with literature where larger k values reduce effect of noisy observations in the training set, thus achieving higher accuracy until an increase in k has negligible effect on accuracy [7].

Limitations

- Generalising according to personal characteristics can raise ethical issues, thus cementing the choice of payment information as the predictive feature.
- The models depend on cardholder data that may not be readily available to a bank.
- Calculating the probability of default as a continuous probability, like that devised in [5], is of more value to a financial institution than a binary classification conducted in this study.

VIII. Further Work

- There is scope to combine both the personal characteristics and payment information as the predictive features, which could potentially yield greater predictive accuracy.
- A similar study concluded that Artificial Neural Networks was the optimal ML technique that accurately estimated the probability of default [2]. Therefore this algorithm could be applied to this dataset to possibly give a more precise estimation of credit-worthiness.
- A hybrid KNN-NB model was implemented on a similar data-set for credit-scoring and could be applied to gain better accuracy for future study [1].

References

[1] Li, F.C., 2009, August. The hybrid credit scoring strategies based on knn classifier. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on* (Vol. 1, pp. 330-334). IEEE.

[2] Yeh, I.C. and Lien, C.H., 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), pp.2473-2480.

[3] Jiawei, H. and Kamber, M., 2001. Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, *5*.

[4] Peterson, M.R., Doom, T.E. and Raymer, M.L., 2005, June. GA-facilitated classifier optimization with varying similarity measures. In *Proceedings of the 7th annual conference on Genetic and evolutionary computation* (pp. 1549-1550). ACM.

[5] Keramati, A. and Yousefi, N., 2011, January. A proposed classification of data mining techniques in credit scoring. In *the Proceeding of 2011 International Conference of Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia, Jurnal* (pp. 22-4).

[6] Chen, M.C. and Huang, S.H., 2003. Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, *24*(4), pp.433-441.

[7] Islam, M.J., Wu, Q.J., Ahmadi, M. and Sid-Ahmed, M.A., 2007, November. Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. In *Convergence Information Technology, 2007. International Conference on* (pp. 1541-1546). IEEE.

URL:

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