PREDICTING CREDIT CARD DEFAULT

A PREPRINT

Group Name: GROUP 0
Department of Biomechanical Engineering
University College London
London, WC1E 6BT

January 18, 2022

1 Introduction

This report aims to present our approach and results towards credit card default prediction based on the Taiwan default payments dataset collected between April – September 2005, amidst the Taiwan credit crisis. The model classifiers used in this work are K-Nearest Neighbours, Random Forest, Linear Discriminant Analysis (LDA), Logistic Regression, Naïve-Bayes and Neural Network. In order to improve performance, two additional models have been created – one using the K-means Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance, respectively Principal Component Analysis (PCA) for dimensionality reduction. In addition, hyperparameter tuning has been implemented to further enhance the quality of the prediction models.

Model performance has been quantified in terms of classification accuracy, receiver operating characteristic (ROC) curve, confusion matrix, F1 score, precision and recall score.

2 Data Transformation and Exploration

First of all, the data has been preprocessed, transformed and explored, in order to prepare the implementation of the selected predictive models. The dimensionality, parameters, data types and data values of the CreditCard_training.csv dataset were examined, determining the necessary data cleaning and transformation procedures. Correlation plots and multiple formulas such as Pearson R test, t-test, Chi-Squared test and permutation test were then used to assess the relationships between variables and their importance to the credit card default outcome.

2.1 Data Processing

Preprocessing starts by examining the dimensions and data types of the dataset, confirming that there are not any unconformities, with all data types of integer type, respectively that there are not any null or missing values. In addition, there are not any duplicate records within the dataset. However, several nonconforming values are observed through examination and descriptive statistics, as follows:

- ‘PAY_0’ from the history of past payment headings is converted to PAY_1 for consistency.
- The unexplained codes 0, 5 and 6, respectively the explained code 4 from EDUCATION are combined within code 0 as ‘others’, since there are very few entries with these codes.
- The unexplained code 0 from MARRIAGE is combined within code 3 as ‘other’, since there are very few entries with value 0.
- ‘default payment next month’ heading is changed to a more consistent one – ‘DPNM’.
- The codes of the SEX category are changed from 1 and 2 to M and F respectively, using the category as a datatype.

- Since there are no ID duplicates, the ID is set as the dataframe index.

2.2 Data Exploration & Visualisation

Histograms, stacked bar and box plots are used to explore, visualise, and better understand the data and underlying trends, as well as suggestions regarding the model selection and results interpretation.

- **Default** - The bar plot of credit card default counts reveals that, as expected, there are far fewer defaulters (approx. 22%) than non-defaulters (approx. 78%), which leads to an unbalanced dataset. As such, it will be necessary to adapt the models through random sampling, oversampling and undersampling techniques, to improve their performance.
- **Credit Limit (‘LIMIT_BAL’)** -The histogram and box plot show only one significant outlier – similarly to the other outliers, since there is not a high magnitude difference, these are explained as being richer clients. The subsequent box plot shows that non-default status clients have a higher median credit limit than clients with default, as well as a higher dispersion across the credit limit range. Thus, a higher credit limit could indicate a lower likelihood of credit card default, yet there is weak negative correlation with default status.
- **Sex** – There are twice as many female clients compared to male ones (62.8 and 37.2% respectively). There is a slightly higher proportion of male clients with a default status (30%) than for female clients (26%).
- **Education** –Most clients (82%) have gone either to graduate school or university. Quite noticeably, only 6% of the ‘others’ clients have defaulted, however the category encompasses only 1.85% of total clients. About 20% of graduate school clients have defaulted, a slightly lower percentage than university (24%) and high-school (25%). Consequently, the education level and credit limit boxplot reveals, as expected, increasingly higher credit limits for clients with higher education levels. However, there is a very small positive correlation with default status.
- **Marriage** – Most clients are either single or married. A higher proportion (24%) of married clients have defaulted, compared with 21% of single clients and 23% of ‘others’. Nonetheless, it has a very low negative correlation (less than -0.1) with default status.
- **Age** – The box plot distributions of default status against age are similar, with a highly similar median. The defaulters category (1) has a slightly smaller age range for the lower quartile, with a higher age range of the higher quartile compared to non-defaulters. In contrast, non-defaulters reveal older outliers. Dividing clients in 6 age categories between 20-80, clients aged 70-80 and 30-40 have the lowest default percentages, at 18%, respectively 21%, whereas the highest default rates are found among clients aged 60-70 (29%), respectively 50-60 (24%).
- **Repayment Status** – Naturally, the number of delayed payments increases between Apr – Sep 2005. The repayment status highly correlates between consecutive months, with the correlation decreasing as the difference in time increases. The unexplained codes -2 and 0 have been preserved, since the analysis of bill and payment amounts suggest that i) -2 represents clients without credit cards (their bill amount is consistently 0, Figure 1) and ii) 0 represents clients whose payments are late by less than one month. There is low to moderate positive correlation with default status, with correlation increasing from April to September.
- **Bill Statement** –the minimum bill amount is a negative value, suggesting that some clients have. Quite interestingly, 30% of clients with a bill statement greater than their credit limit, respectively 38% of inactive clients (i.e., their bill statements are 0 in the last 6 months) have defaulted. There is relatively low to moderate positive correlation with credit limit, moderate to strong positive correlation between bill statement amounts, and almost zero correlation with default status.
- **Previous Payment** – There did not seem to be any inconsistencies or unusual values. Very low negative correlation with default status.

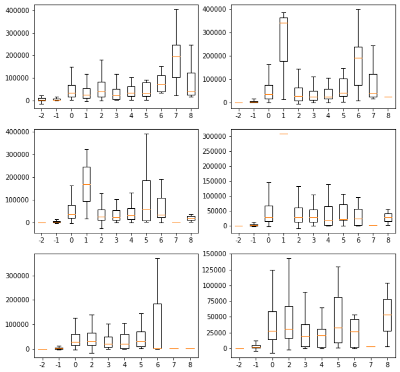

Figure 1: Bill Amount vs repayment status by months

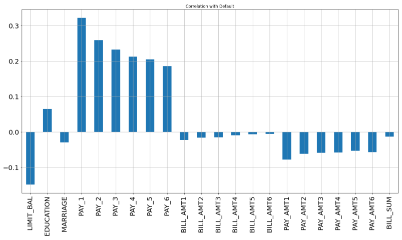

Figure 2: Feature correlation with default status

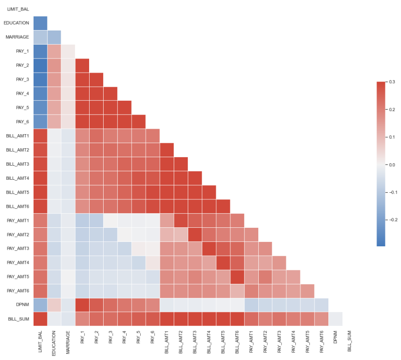

Figure 3: Correlation heatmap between features

3 Methodology Overview

3.1 Literature Review

Past literature on credit card default prediction has explored a variety of models, such as Artificial Neural Networks, K-nearest neighbours, Linear Discriminant Analysis, Logistic Regression, Naïve Bayes Classifiers and Random Forests (Sariannidis et al., 2020). The literature also suggested that the different classification methods had high preliminary accuracies, with the Naïve Bayes Classifier being the exception (Neema & Soibam, 2017). However, Naïve Bayes Classifiers also perform better than the other models at predicting defaults (Neema & Soibam, 2017). Alam et al. (2020) have showed a significant difference between the accuracy of the models on the unbalanced dataset, with 66.9% accuracy, and the dataset balanced through the K-means Synthetic Minority Oversampling Technique (SMOTE), with 89% accuracy respectively. From literature, a 5-fold cross validation has been commonly used to validate the training model (Sariannidis et al., 2020). As for the model evaluation, apart from the accuracy, the ROC, F1 Score and Recall have also been used to evaluate model performance (Yang & Zhang, 2018). A higher proportion of false negatives i.e., predictions of no default when there would, in fact, take place a default, is highly undesirable for the scope of credit card delinquency (Rabiul Islam et al., n.d.). As such, unlike the majority of literature focussing on accuracy, our most significant evaluation metrics will be the recall (i.e., a higher recall indicates a lower number of false negatives) and the confusion matrix (showing the number of false negatives). In addition, the F1-score, precision and the ROC curve will serve to complement the analysis and evaluation.

Therefore, based on past literature, this paper seeks to build upon the basic methodologies set out in the literature by exploring various resampling methods, focussing on the most appropriate performance metrics for the scope of credit card default prediction and, thus, risk analytics.

3.2 Model Selection

The models are focused on binary classification, which aims to predict if a client has defaulted on their credit card (1) or not (0). The models considered in this paper are Logistic Regression, Naïve Bayes Classifier (NBC), Linear Discriminant Analysis (LDA), K-Nearest Neighbours (KNN) and Neural Networks. Prior to training, all data is normalized via standard scaling. The training dataset, CreditCard_training.csv, is then split into training and validation datasets, where 70% of the data is set into training the model and the remaining 30% set aside for validation. All models are trained, validated, and tested via 4 different settings: i) PCA only (dimensionality reduction); ii) PCA and Random Undersampling; iii) PCA and SMOTE (oversampling); iv) PCA and SMOTE-ENN (oversampling and undersampling).

3.3 Dimensionality Reduction

Due to the large number of parameters (23 following data transformation) in the dataset and its multicollinearity, dimensionality reduction is recommended, reducing the risk of overfitting (Zhou et al., 2019). PCA is preferred compare to feature selection because it maximises the variance and projecting it into principal component without removing the whole feature as those uncorrelated features might be important for the target output to connect with other features. This can be done by computing the number of principal components needed for our dataset to have maximum variance.

3.4 Sampling Methods

Since there are large class imbalances within the dataset, these have been addressed through different sampling methods. The following methods were considered:

- **Random Undersampling - •** the simplest undersampling method, aiming to balance the data by eliminating samples randomly (Xu et al., 2020). Two other undersampling techniques, Tomek-Link and Edited Nearest Neighbours (ENN) also have been tested using the dataset but Random Sampler stands out the most. Hence, Random Sampler has been chosen as the technique for undersampling.
- **Synthetic Minority Oversampling Technique (SMOTE)** - it aims to "synthetically" create new minority observations, which could minimise the proportion of false negatives (Chawla et al., 2002). Another variant SMOTE, K-means SMOTE also have been tested against the dataset. It appears that the K-means SMOTE yields a better accuracy while SMOTE gives better Recall which should be prioritised in this case.
- **SMOTE-Edited Nearest Neighbours (ENN)**–it combines the SMOTE ability in "synthetically" create new minority observations and incorporate the ENN to remove samples in the majority class: in this case, the non-default class, by attempting to remove noise from the dataset (Wilson, 1972). One other combination were tested also, which is the SMOTE-Tomek Link and it show lower overall performance than the SMOTE-ENN.

3.5 Hyperparameter Tuning and Cross Validation

The "train_test_split" function is used to split the CreditCard_training.csv into the training data set and validation dataset. The validation of the dataset is done via a five-fold cross validation, which splits the whole dataset into 5 equal "folds". Among these five folds, four of it will be the training dataset, and the remaining one would be the validation dataset. The cross-validation scores of the dataset were then obtained over the five different combinations of "folds". Once the cross-validation scores were obtained, the hyperparameters of the models (if applicable) were obtained.

4 Model Training and Validation

4.1 K-Nearest Neighbours (KNN)

KNN classifier attempts to classify unlabelled observations with similarly labelled examples. This can be done by considering the Euclidean distances between a data point with its neighbours. The number of points considered is determined by the hyperparameter "k" which governs the number of "neighbours" chosen in considering the Euclidean distances. The hyperparameter is tuned by iterating through 1 to 100 neighbours and obtaining the error rate associated with it. As explored through previous research regarding unbalanced datasets, kNN undersampling generally outperforms other sampling methods and models (Beckmann et al., 2015), yet the tradeoff between precision and recall is rather sensitive (Zhang & Mani, 2003). To properly assess this tradeoff and the performance of kNN, performing SMOTE, random undersampling and SMOTE-ENN in conjunction is highly insightful.

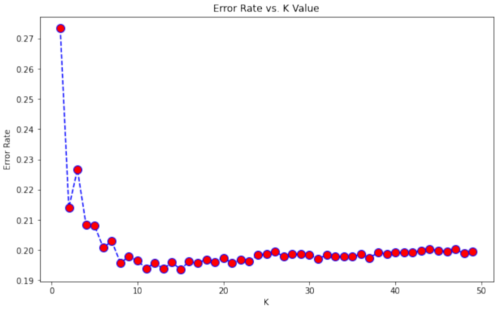

Figure 4: KNN+PCA k-value tuning

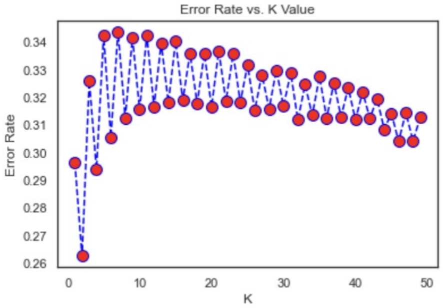

Figure 5: KNN+PCA+SMOTE k-value tuning

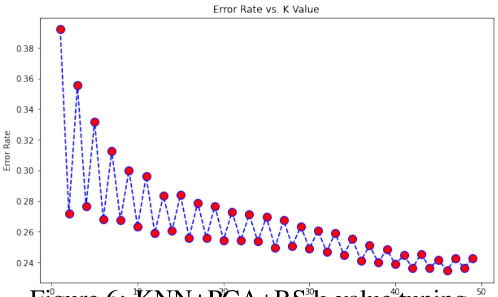

Figure 6: KNN+PCA+RS k-value tuning

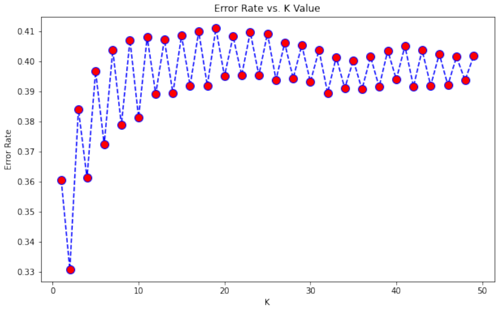

Figure 7: KNN+PCA+S-ENN k-value tuning

4.2 Neural Networks

The neural network (NN) has been implemented through the Multi-Layer Perceptron (MLP) classifier. Despite their lack of explanation capability regarding the classification process, which is highly important in the process of credit-risk evaluation (Baesens et al., 2003), their high predictive accuracy and generally lower proportion of false negatives (i.e., higher recall) (Lessmann et al., 2015) make NNs a relevant benchmark to the scope of retail credit scoring. Nonetheless, other research suggests that logistic regression is comparable in terms of percentage of both true positives and true negatives to NNs (Desai et al., 1996). Since the performance of generic NNs such as ours is inferior to that of customized models (Desai et al., 1996), future work might account for this effect and build an improved, customized NN.

4.3 Linear Discriminant Analysis

LDA has also been commonly used in risk prediction (Hand & Henley, 1997), as an alternative to logistic regression (Yeh & Lien, 2009). Even though approaches such as the hybrid integration between discriminant analysis and neural networks outperforms traditional LDA and logistic regression in terms of credit scoring accuracy (Lee et al., 2002), LDA alone has shown a relatively good performance for credit scoring (Yeh & Lien, 2009), outperforming other traditional individual classifiers, such as k-NN or Support Vector Machines (SVMs) (Lessmann et al., 2015). As such, our implementation uses an LDA with tuned hyperparameters and five-fold cross-validation.

4.4 Gaussian Naïve-Bayes

A Gaussian Naïve Bayes classifier has been used, assuming a normal distribution. Since it is based on Bayes theory and class conditional independence, computation is simplified, however the predictive accuracy will be strongly correlated with this assumption (Yeh & Lien, 2009). Thus, even though the training time will be several orders of magnitude lower than for the other models, the performance is also expected to be weaker. Nonetheless, this model serves as a guiding baseline.

4.5 Logistic Regression

Logistic regression aims to model the conditional probability of a given output by transforming the output into a continuous probability distribution via the "sigmoid" function which is bounded between 0 and 1 (Kirasich et al., 2018). For this dataset, after applying PCA and the appropriate sampling techniques, the output of the linear combinations of the principal components are then transformed into a continuous probability distribution. The trained model is then cross validated using five folds, obtaining the cross-validation score. The hyperparameter ("C" for the case of logistic regression) is then fine-tuned using the RandomisedSearchCV function from the sklearn package. The function is implemented to fine tune the hyperparameter "C" by iterating through each fold 200 times.

5 Results

5.1 Accuracy

Table 1 summarises the training accuracy, training time and testing accuracy of all the 20 models that have been implemented. These metrics have been obtained after implementing each relevant hyperparameter optimisation and cross validation, as well as the respective dimensionality reduction and sampling methods.

Table 1: Validation Accuracy, Training Time & Testing Accuracy of all models

No.	Model	Validation Accuracy	Training Time (s)	Testing Accuracy
1	KNN + PCA	0.805278	0.034865	0.820333
2	KNN+PCA+SMOTE	0.7375	0.064827	0.744833
3	KNN+PCA+RandomSampler	0.764306	0.012965	0.787667
4	KNN+PCA+SMOTE-ENN	0.678889	0.03299	0.699333
5	NN+PCA	0.810278	15.583414	0.825
6	NN+PCA+SMOTE	0.726389	68.586764	0.7335
7	NN+PCA+RandomSampler	0.705972	8.125329	0.717167
8	NN+PCA+SMOTE-ENN	0.655417	50.66965	0.675667
9	LDA+PCA	0.807917	0.041888	0.821167
10	LDA+PCA+SMOTE	0.670278	0.073803	0.704
11	LDA+PCA+RandomSampler	0.683333	0.022939	0.719333
12	LDA+PCA+SMOTE-ENN	0.545278	0.048869	0.5695
13	NB + PCA	0.782222	0.02294	0.800333
14	NB+PCA+SMOTE	0.326111	0.009974	0.329167
15	NB+PCA+RandomSampler	0.37375	0.005027	0.387
16	NB+PCA+SMOTE-ENN	0.314028	0.00895	0.314667
17	LR+PCA	0.804167	0.063877	0.817167
18	LR+PCA+SMOTE	0.665278	0.117694	0.702333
19	LR+PCA+RandomSampler	0.690556	0.04089	0.722333
20	LR+PCA+SMOTE-ENN	0.536806	0.057843	0.5655

Out of all the models, the neural networks and kNN offer the best, most consistent accuracy performance across all sampling techniques. Even though LDA and logistic regression offer lower accuracy, their performance is relatively good across all sampling techniques as well, apart from SMOTE-only and SMOTE-ENN.

Firstly, all models consistently increase their accuracy during testing significantly, apart from the Gaussian Naïve-Bayes model. Separate analysis of the resulting accuracies rendered that the models employing PCA only, without any other sampling techniques, have performed the best in terms of both validation and testing accuracy – over 80% for all models, apart from NB. PCA and Random Sampling, respectively PCA and SMOTE, have offered the next most accurate results – above 65% and, again, apart from NB. Indeed, the Naïve-Bayes model represents the outlier, as it severely underperforms across all sampling techniques.

5.2 Precision/Recall, F1-score and ROC curves

Table 2: Precision/Recall, F1-score and ROC curves of all models

Model	Precision (V)	Precision (T)	Recall (V)	Recall (T)	F1(V)	F1(T)	ROC-AUC (V)	ROC-AUC (T)
KNN	0.634492	0.67803	0.306021	0.28278	0.412898	0.399108	0.627604	0.623435
KNN+S	0.409474	0.392013	0.391682	0.379937	0.400381	0.38588	0.614431	0.611177
KNN+RS	0.475843	0.496965	0.52576	0.517378	0.499558	0.506966	0.679413	0.688663
KNN+S-ENN	0.367084	0.370549	0.600869	0.608215	0.455744	0.460526	0.651123	0.665958
NN	0.640643	0.671429	0.346369	0.334123	0.449637	0.446203	0.645183	0.645198
NN+S	0.414889	0.402346	0.543141	0.541864	0.47043	0.461797	0.661175	0.663306
NN+RS	0.406018	0.399721	0.678461	0.678515	0.508018	0.503075	0.696181	0.703009
NN+S-ENN	0.354709	0.35786	0.659218	0.676145	0.461238	0.468015	0.656769	0.675842
LDA	0.697917	0.741855	0.249534	0.233807	0.367627	0.355556	0.609201	0.606025
LDA+S	0.365527	0.37834	0.6437	0.626382	0.466277	0.471743	0.660819	0.67557
LDA+RS	0.376431	0.395709	0.632526	0.626382	0.471978	0.485015	0.665252	0.685287
LDA+S-ENN	0.299687	0.294668	0.772191	0.746445	0.431795	0.422535	0.626031	0.634313
NB	0.51296	0.528476	0.528243	0.49842	0.520489	0.513008	0.691837	0.689747
NB+S	0.239428	0.229138	0.924271	0.921801	0.380332	0.367039	0.538983	0.546241
NB+RS	0.250172	0.239974	0.900683	0.879147	0.39158	0.377033	0.561274	0.567267
NB+S-ENN	0.237374	0.226767	0.933582	0.932859	0.378508	0.364844	0.534513	0.541102
LR	0.698225	0.743516	0.219739	0.203791	0.334278	0.319901	0.596182	0.592496
LR+S	0.363325	0.377358	0.659218	0.631912	0.468461	0.472534	0.663121	0.676539
LR+RS	0.383541	0.398167	0.630664	0.617694	0.476995	0.484211	0.669242	0.684005
LR+S-ENN	0.293977	0.292479	0.763501	0.746445	0.424504	0.42028	0.617481	0.631778

All models were applied with PCA, and S in this table means SMOTE, where V means validation set and T means test set

Since the dataset provided is imbalanced in nature, a metric which is insensitive to changes in class distribution is needed to evaluate the model in an unbiased manner. ROC curves in particular are not dependent changes to class distributions, as it only depends on either the True Positive Rate (TPR) which can be calculated from the positive predicted column for the and the False Positive Rate which can be calculated from the negative predicted column (Fawcett, 2003). Therefore, if there is a skew in class distribution, or the proportion of positive to negative instances, ROC will not be affected. Another metric which are insensitive to changes in class imbalances is the Recall score, which is effectively the True Positive Rate. Therefore, ROC and Recall are given precedence when selecting the models.

Quite surprisingly, the **NB model** performs the best in terms of recall for both the validation and testing data sets, with the following NB models registering the three highest recalls: SMOTE-ENN (0.93049 in testing), SMOTE and Random Sampling. Nonetheless, paired with the lowest AUC scores, a lower-than-average F1-score, respectively a very low accuracy (under 0.27 for all the top three ranking NB models) and precision during validation and testing alike, it does not manage to balance the requirements of credit-risk prediction. As such, the most consistent NB model would be the PCA-only one, as it balances the recall, accuracy and F1-score appropriately, achieving a high AUC score. The **Logistic Regression** models perform quite well and consistently in terms of recall, precision, F1-score and AUC score, with the exception of the PCA-only technique. In addition, their classification accuracy is rather unsatisfactory. The **LDA models** seems to perform similarly to Logistic Regression, offering slightly smaller values of recall, precision, F1-score and AUC score, respectively a slightly higher classification accuracy. The **KNN model**s underperform LDA and Logistic Regression in regard to AUC score, recall and F1-score, in spite of their superior classification accuracy. Finally, the **NN models** seem to outperform all other models, offering the highest and most consistent AUC, F-1 scores and recall values, in combination with high classification accuracies.

All the other models present their most consistent and balanced performance for the Random Sampling technique, with the SMOTE technique ranking appearing to be second. Hence, it can be observed that the PCA-only technique renders the best classification accuracies, whereas Random Sampling consistently offers the most advantageous combination of AUC score, F1-score and recall, whilst rendering a high classification accuracy. The SMOTE technique results in rather average values and classification accuracy, whilst SMOTE-ENN seems to generally underperform across all evaluation metrics.

5.3 ROC curves

Since the dataset provided is imbalanced in nature, a metric which is insensitive to changes in class distribution is needed to evaluate the model in an unbiased manner. ROC curves in particular are not dependent changes to class distributions, as they only depend on either the True Positive Rate (TPR) and the False Positive Rate (FPR). Therefore, if there is a skew in class distribution, or the proportion of positive to negative instances, ROC will not be affected (Fawcett, 2003). As the AUC score was presented in the previous section, the ROC curves of each technique and model will be outlined below for the validation dataset.

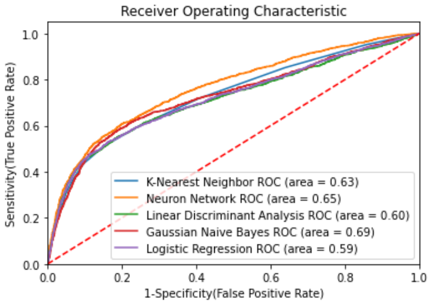

Figure 8: ROC curve PCA only models

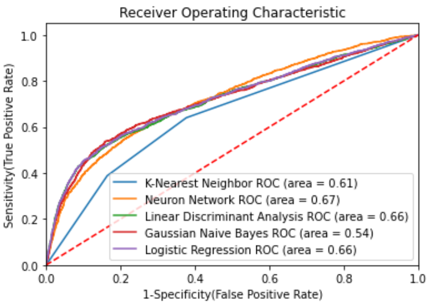

Figure 9: ROC curve PCA+SMOTE only models

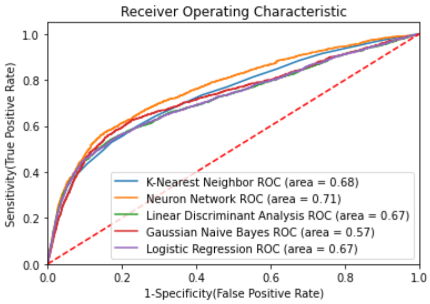

Figure 10: ROC curve PCA+RS only models

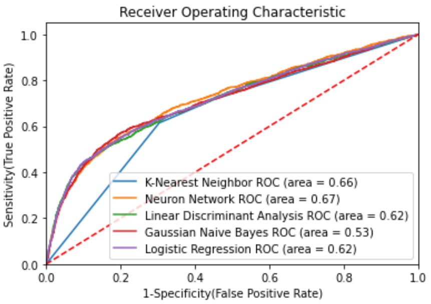

Figure 11: ROC curve PCA+SMOTE-ENN only models

Through the visual interpretation of the ROC curves, it can be confirmed that, as expected, Random Sampling consistently offers the best performance and trade-off between sensitivity and specificity. PCA-only also offers a satisfactory performance, whereas the SMOTE and SMOTE-ENN techniques render relatively weaker, more specific performances.

5.4 Final Ranking

Following the analysis of these evaluation metrics, based on their balance, as well as the higher significance of AUC score, recall and F1-score, the models outlined in have been chosen as offering the best performance for each type of classifier. It should be noted again that the Random Sampling technique renders the best performance across all classifiers, apart from Naïve-Bayes.

Table 3: Validation Accuracy, Training Time & Testing Accuracy of all models

Model	Accuracy (T)	Recall (T)	F1 Score (T)	ROC (T)
NN + PCA + RandomSampler	0.705972	0.678515	0.503075	0.703009
NB + PCA	0.782222	0.49842	0.513008	0.689747
KNN+PCA+RandomSampler	0.764306	0.517378	0.506966	0.687304
LDA+PCA+RandomSampler	0.706833	0.630332	0.483783	0.688663
LR+PCA+RandomSampler	0.690556	0.617694	0.484211	0.684005

5.5 Confusion Matrices

The confusion matrices for the validation data sets of the highest performing models and techniques are outlined below.

Similarly to the evaluation metrics, the matrices reveal that the NN model with PCA and Random Sampling has the most balanced performance, as it optimises the prediction outcomes, i.e., false negatives (with the second lowest number of false negatives), false positives (third lowest number), true positives (again, second best performing) and true negatives (third highest). As such, our approach prefers a more robust, balanced performance across the entire classification. Even though the LR model with PCA and Random Sampling has the lowest number of false negatives, it is more imbalanced, as it renders the lowest number of true negatives and the highest number of false positives.

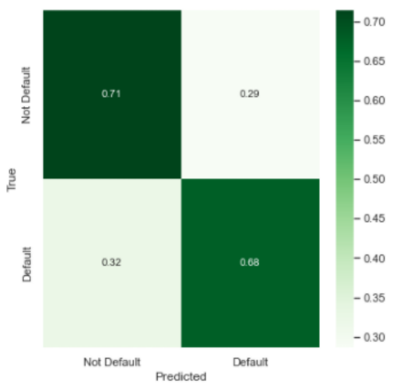

Figure 12: NN+PCA+RS

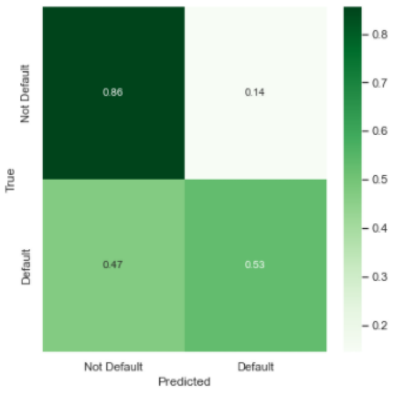

Figure 13: NB+PCA

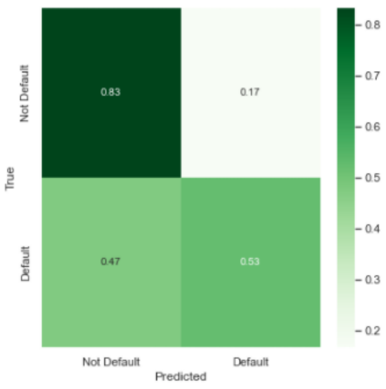

Figure 14: KNN+PCA+RS

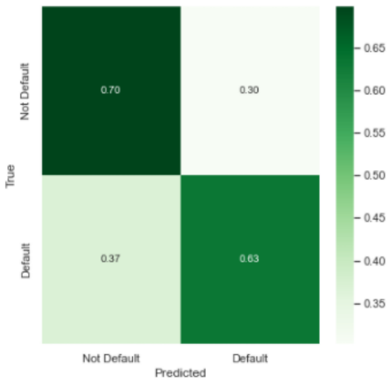

Figure 15: LDA+PCA+RS

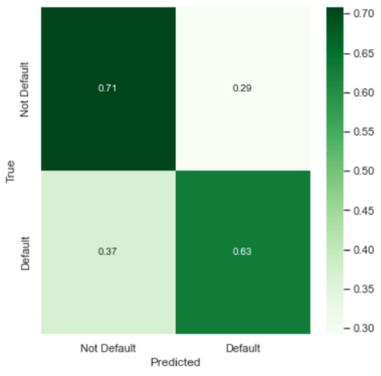

Figure 16: LR+PCA+RS

6 Final Prediction on Test Set

The final predictions on the test set are obtained through the NN model which uses PCA and Random Sampling, as outlined in its confusion matrix below. The optimal hyperparameters are further derived through 100 iterations of a randomised search, using RandomizedSearchCV. It achieves a reasonable proportion of false negatives (32%), respectively true positives (68%). It also minimises false positives (only 27%), i.e., the proportion of clients who are falsely predicted to default. As such, besides minimising false negatives, it is also very relevant to minimise false positives, preventing false alarms and further nuisance for the clients. Conversely, true negatives, i.e., clients who are correctly predicted not to default, are maximised at 73%. As outlined in Table 3, the testing accuracy of this model is high, at around 77.46%, with an AUC score of 0.7139, a recall of 0.609 and an F1-score of 0.5328.

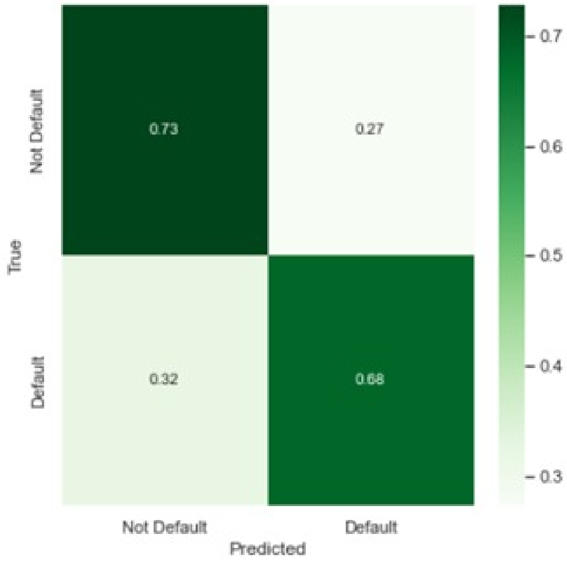

Figure 17: Confusion Matrix of the NN + PCA + Random Sampling model on the testing dataset

7 Conclusions

From the confusion matrix, the default rate predicted by the best performing model is equivalent to approximately 14.3%, compared to the test set's true default rate of 21%. Hence, even though the true positive rate of 68% and the false negative rate of 32% offer a satisfactory result, this represents a starting point for future work and improvements. Even though better prediction performances in terms of true positives and false negatives have been obtained by the Naïve-Bayes models, their classification has been the lowest among all analysed models and, as such, unsuitable for credit scoring prediction. Nonetheless, the chosen Neural Network using PCA for dimensionality reduction and the Random Sampling method is comparable in performance to other research, such as Yang et al. (2018), which uses 10-fold cross-validation – as such, our model renders a 0.7171 accuracy, compared to their 0.8176, respectively an AUC score of 0.7030, compared to 0.7735.

Our suggested improvements would firstly focus on the sampling technique, as others, such as Cluster Centroid undersampling and Borderline-SMOTE (oversampling), have been shown to offer an accuracy, recall and AUC score all above 80% (Alam et al., 2020).

Secondly, future work might explore different models which have proven effective, such as the tree-based Random Forest, Gradient Boosted Decision Trees (Alam et al., 2020) or LightGBM (Yang et al., 2018), respectively other hybrid classifiers, such as the Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis(Akkoç & Soner, 2012).

Finally, our consensus is that future work should focus on balancing the performance in a manner suitable for credit risk scoring and beneficial towards the lending institution itself, prioritising the reduction of false negatives (thus, the potential incurred loss), whilst optimising false positives (and the potential nuisance created for clients).

References

[1] Akkoç, & Soner. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. European Journal of Operational Research, 222(1), 168–178. https://doi.org/10.1016/J.EJOR.2012.04.009

[2] Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. IEEE Access, 8, 201173–201198. https://doi.org/10.1109/ACCESS.2020.3033784

[3] Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation.

[4] Beckmann, M., Ebecken, N. F. F., & Pires de Lima, B. S. L. (2015). A KNN Undersampling Approach for Data Balancing. Journal of Intelligent Learning Systems and Applications, 07(04), 104–116. https://doi.org/10.4236/JILSA.2015.74010

[5] Chawla, N. v., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/JAIR.953

[6] Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research, 95(1), 24–37. https://doi.org/10.1016/0377-2217(95)00246-4

[7] Fawcett, T. (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. HP Invent, 27. https://doi.org/10.1.1.10.9777

[8] Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. Data Science Review, 1(3), 9.

[9] Lee, T. S., Chiu, C. C., Lu, C. J.,& Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. Expert Systems with Applications, 23(3), 245–254. https://doi.org/10.1016/S0957-4174(02)00044-1

[10] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/J.EJOR.2015.05.030

[11] Neema, S.,& Soibam, B. (2017). The comparison of machine learning methods to achieve most cost-effective prediction for credit card default. Journal of Management Science and Business Intelligence, 9264, 36–41. https://doi.org/10.5281/zenodo.851527

[12] Rabiul Islam, S., Eberle, W., & Khaled Ghafoor, S. (n.d.). Credit Default Mining Using Combined Machine Learning and Heuristic Approach.

[13] Sariannidis, N., Papadakis, S., Garefalakis, A., Lemonakis, C., & Kyriaki-Argyro, T. (2020). Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ML) techniques. Annals of Operations Research, 294(1–2), 715–739. https://doi.org/10.1007/s10479-019-03188-0

[14] Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man and Cybernetics, 2(3), 408–421. https://doi.org/10.1109/TSMC.1972.4309137

[15] Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. Journal of Biomedical Informatics, 107(May), 103465. https://doi.org/10.1016/j.jbi.2020.103465

[16] Yang, S., & Zhang, H. (2018). Comparison of Several Data Mining Methods in Credit Card Default Prediction. Intelligent Information Management, 10(05), 115–122. https://doi.org/10.4236/iim.2018.105010

[17] Yeh, I.-C., & 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, 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

[18] Zhang, J. P., & Mani, I. (2003). KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. Proceeding of International Conference on Machine Learning (ICML 2003), Workshop on Learning from Imbalanced Data Sets. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1603053

[19] Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A: Statistical Mechanics and Its Applications, 534, 122370. https://doi.org/10.1016/J.PHYSA.2019.122370