***Predicting Credit Card Defaults Using Machine Learning***

17038086 | UFCFMJ-15-M | WORD COUNT

# Introduction

It is business critical for financial institutions to be able to accurately predict whether a credit card customer will default on their next payment. Defaults have a direct impact upon both profitability and compliance, while influencing customer relations and how lenders manage risk. Alongside this, regulators require information around credit decision-making, ultimately meaning that predictive capability and interpretability are crucial. Model selection is constrained due to regulators, it is critical that models are explainable for audit, further supporting the need for interpretability (Guidotti et al, 2018).

Traditionally, credit scores have relied on statistical models such as logistic regression. This approach has limitations, it assumes linearity and treats varying factors independently, this results in not being able to capture complex patterns. For example, a younger individual could be seen as to be of a slightly higher risk, however when you combine a younger individual with a poor repayment history this clearly presents a far riskier proposition. Factors in isolation do not give the same level of depth in terms of predictability.

Machine Learning (ML) methods, therefore, are better equipped to handle such interactions between factors. Techniques such as Decision Trees, Random Forests and Support Vector Machines can learn non linear boundaries in an automatic fashion. However, results must be carefully evaluated, they can vary in interpretability, can overfit and can also have a large computational cost associated with them.

This study, investigates multiple machine learning algorithms to predict credit card defaults, utilising the UCI Credit Card Default dataset. The specific objectives are as follows:

1. To analyse the dataset and explore ther relationships between demographic, financial and behavioural attributes.
2. To evaluate a range of supervise learning algorithms in terms of predictive performance.
3. To identify the most suitable model, and to critically analyse the least suitable model.
4. To reflect on limitations and suggest future improvements.

# Dataset Overview

The dataset used in this study is the UCI Credit Card Default dataset (Yeh and Lien, 2009) accessed via OpenML website. An initial problem is caused by generalisability of the models and their results on this dataset, as the data is specifically recorded in Taiwan, this therefore is highlighted as a limitation and models should be tested on more global data in future iterations.

The dataset contains 30,000 observations with 24 attributes. Due to attributes containing demographic, financial and behavioural features it can be utilised well in terms of credit risk modelling in this study. Attributes include the following:

* + Demographic: Age, Sex, Education & Marriage status.
  + Financial: Credit Limits
  + Payment History: Pay\_0 to Pay\_6 (Records of payment status)
  + Bill Statement: Bill\_amt1 to Bill\_amt6 (Amount billed each month)
  + Payment Amount: Pay\_amt1 to Pay\_amt6 (Amount paid each month)

The target variable, default, is binary with 1 indicating default on the next payment and 0 indicating otherwise. The class distribution is rather imbalanced, 22.1% defaults and 77.9% non-defaults, this ultimately means that evaluation metrics in recall, F1 and precision become more valuable that purely accuracy alone.

Preprocessing: A number of preprocessing techniques were used prior to modelling:

* + The ‘ID’ attribute was dropped due to it only serving as a record identifier.
  + Scaling: Continuous attributes were standardized to zero mean and unit variance, utilizing Z-Score scaling. This meant models sensitive to the varying feature magnitude (SVM and KNN) could perform.
  + Encoding: Categorical variables such as age, marriage status and education were already encoded within the dataset and retained in this format.

Overall, the dataset contains a large collection of credit card clients with a diverse set of attributes. Both its size and heterogeneity make it suitable for ML methods for prediction of credit card defaults.

# Exploratory data analysis

To better understand the dataset, exploratory data analysis (EDA) was carried out across the demographic, financial and behavioural attributes. This analysis supports identification of important patterns and correct modelling strategies.

## Univariate Analysis

*Figure 1: Age Distribution*

The age distribution (Figure 1) is right skewed, with most of the clients clearly being aged between 40 and 25 years of age. Defaults are higher among younger clients, this suggests that age contributes to prediction. The lack of samples for the older generation may limit the models generalisability. This highlights potential limitations of purely linear models, such as logistic regression, in capturing interactions between age and repayment behavior. Nevertheless, logistic regression remains a key baseline in credit scoring, valued for its interpretability and regulatory acceptance.

**Figure 1:**

A graph of age by default status

AI-generated content may be incorrect.

*Figure 2: Credit Limit Distribution*

Credit limits are also positively skewed. Most customers hold low limits, with a few outliers in terms of high limits. Defaults are more common in lower limit groups, which is in line with expectations, as banks generally allocate higher limits to those customers deemed as low risk.

Figure 2:

A graph of credit limit

AI-generated content may be incorrect.

*Figure 3: Target Variable*

As shown in Figure 3, the dataset is imbalanced, this means accuracy alone will not be the evaluation method, recall, F1-Score and precision should be analysed when evaluating models.

**Figure 3:**

A graph with a red and blue rectangle

AI-generated content may be incorrect.

## Bivariate Analysis

Figure 4 shows the two distributions of age between defaults and non-defaulters. Each of their medians lie at a similar level (around 34 years of age), and each share similar interquartile ranges. These boxplots suggest that age as a standalone is not a strong driver in terms of default behaviour, predictive capacity may only appear when it interacts with another variable,

**Figure 4:**

A diagram of a number of blue rectangular objects

AI-generated content may be incorrect.

## Multivariate Analysis

The 3D scatterplot (Figure 5) combines age, credit limit and repayment status, coloured by default status. While there are larger clusters in certain areas such as lower age and higher repayment delays, there is clearly a lot of overlap between the defaulters and non-defaulters. The suggestion derived from this is that no combination of three attributes will distinguish classes, which further supports the use of ML models to try to capture the complex patterns.

**Figure 5:**

A graph with red and blue dots

AI-generated content may be incorrect.

The skew in continuous features, and clear class imbalance supports the use of non-linear models as well as models beyond accuracy (recall, precision, F1). The use of precision-recall is supported via He & Garcia (2009) as a more informative metric in class imbalance.

While bivariate analysis does reveal some patterns, defaulting behaviour is primary explained via interactions across multiple attributes or variables, making multivariate analysis more appropriate within this context.

# Model development & evaluation

## Modelling Approach

The aim at this stage is to evaluate several machine learning algorithms to predict credit card defaults. A range of supervised learning classifiers were implemented in Python using Scikit-learn. The chosen models represent both linear and non-linear classifiers, allowing a comparison between baselines and more flexible approaches that can handle more complex relationships. The models chosen were as follows:

* Logistic Regression
* Support Vector Machine (SVM) with RBF kernel
* k-Nearest Neighbours (KNN)
* Naïve Bayes (GaussianNB)
* Decision Tree
* Random Forest

## Algorithm Selection

The ability to handle class imbalance and non-linear relationships is crucial in this scenario with the dataset being utilized. Since 22% of the dataset is defaulters, models that balance recall and precision are most appropriate.

Logistic Regression: Logistic regression remains the traditional baseline in credit scoring, widely adopted by financial institutions due to its simplicity, transparency, and regulatory acceptance (Markov et al., 2022). However, it assumes a linear relationship between predictors and the log-odds of default, which restricts its ability to capture complex interactions between features. Logistic Regression is included as a benchmark, despite its linearity, because regulators require explainability in financial models.

Naïve Bayes: This model is based on Bayes theory, with the ‘naïve’ assumption that all features are independent of one another (Raschka,2014). While rarely the top performer due to it’s simplistic approach, this method is computationally efficient and useful as a baseline model.

Support Vector Machine: SVM’s are margin based classifiers introduced by Vapnik, which work by maximizing the separating hyper parameters between different classes (Cortes & Vapnik, 1995). This model can handle complex relationships between features, specifically of use in terms of credit risk whereby demographic, behavioural and financial factors interact. SVMs have been applied consistently in credit risk assessment, and often show competitive performance (Goh, 2019).

Decision Trees: Are highly interpretable, a useful positive aspect to this type of model as within financial risk assessment decision transparency is paramount. However they are prone to overfitting which reduces their generalisability with new data (Geron, 2009).

Random Forests: Random Forests were introduced by Breiman (2001) that aggregate multiple trees to improve robustness. This approach directly improves the overfitting tendancy with Decision Trees through averaging across a number of models. While Decision Trees offer interpretability, Random Forests are included as they overcome overfitting through ensembling, making them stronger candidates for predictive performance.

k-Nearest Neighbours: KNN is a non-parametric method that classifies based on the nearest data points. Although kNN is simple and provides a useful benchmark, its sensitivity to class imbalance and the curse of dimensionality make it unsuitable as a primary credit risk model. Comparative benchmarking in credit scoring confirms that kNN typically underperforms more advanced methods, though it remains a useful baseline for model comparison (Lessmann et al., 2015).

## Cross Validation Approach

As there was a class imbalance in the dataset (22.1% defaults, 77.9% non-defaults), evaluation was carried out using **5-fold stratified cross-validation**, ensuring that each subset of the data maintained the same proportion of defaulters and non-defaulters. This prevents biased estimates of model performance and ensures robustness across different splits of the data.

At each iteration, the model was trained on 80% of the data and tested on the remaining 20%. An average was then taken across the 5 runs to give an accurate estimate on performance

* **Accuracy**
* **Precision**
* **Recall**
* **F1-score**

Accuracy alone is not sufficient due to the imbalance already mentioned. Therefore, particular emphasis was placed on recall and F1-score, which are more informative in this context. As Saito and Rehmsmeier (2015) note, precision and recall analysis is more informative than ROC analysis when dealing with class imbalance, hence the additional importance stressed upon F1 results.

## Model Evaluation & Results

The average performance across 5-fold stratified cross-validation is displayed below (Figure 6). The Support Vector Machine model achieved the best performance, with the highest F1 Score (0.54) and the strongest balance between precision (0.49) and recall (0.58). This shows it was most effective at identifying defaulters, which is the critical objective in credit risk prediction. In terms of credit risk, this is a valuable weighting, as false negatives (failure to identify a defaulter) are more harmful and costly than false positives. SVMs perform well because they are robust to high-dimensional feature spaces and capture complex interactions, unlike Logistic Regression which is linear.

Naïve Bayes produced the highest Recall (0.65), therefore it captured the most true defaulters over the other models. However, the lower precision (0.40) meant that this model is recognizing many non-defaults as at risk of default, too many false positives in practice is to damaging.

Logistic Regression performed at a similar capacity to Naïve Bayes, with high recall (0.64) but low precision (0.38). There was over-prediction here, again reducing it’s real world application in recognition of defaulters.

Random Forest performed best in accuracy (0.81), however produced extremely low recall (0.34), ultimately making it inaccurate in highlighting defaulters which is of no use in a credit risk finding use case.

K-Nearest Neighbours, showed moderate prediction (0.55) yet low recall (0.34). This could be due to the nature of the imbalanced classification, again though this model is not tuned to recognize enough defaulters.

Decision Tree, performed the weakest overall with the lowest F1 Score (0.38) and poor performance in Precision and Recall (both 0.38).

**Figure 6:**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **F1 Std** |
| --- | --- | --- | --- | --- | --- |
| SVM (RBF Kernel) | 0.776933 | 0.496473 | 0.581071 | 0.535393 | 0.004494 |
| Naive Bayes | 0.701500 | 0.402359 | 0.657182 | 0.495574 | 0.020871 |
| Logistic Regression | 0.689267 | 0.380087 | 0.641199 | 0.477244 | 0.001875 |
| Random Forest | 0.816200 | 0.665563 | 0.340116 | 0.450129 | 0.010178 |
| K–Nearest Neighbours | 0.793533 | 0.553796 | 0.345689 | 0.425589 | 0.006843 |
| Decision Tree | 0.729367 | 0.388468 | 0.388636 | 0.388424 | 0.012933 |

A number of numbers on a white background

AI-generated content may be incorrect.

## Most & Least Suitable Models

The F1 score ( the mean of precision and recall) is best in a class imbalance scenario due to rewarding models which catch defaulters (recall) and avoid incorrect flagging of safe customers (precision).

Black text on a white background

AI-generated content may be incorrect.

SVM achieved the highest F1 score (0.54) with good balance between precision (0.49) and recall (0.58). Clearly, this model will be most suited in recognizing defaulters without over-prediction on non-defaulters, this is critical in a credit risk scenario. Figure 7 clearly outlines this SVM’s outperformance in F1-Score in comparison to it’s peers.

**Figure 7:**

A graph showing a number of different numbers

AI-generated content may be incorrect.

By contrast, Decision Trees underperformed, which is in alignment with their known and stated tendency to overfit. In a financial setting, regulators highly value transparency, however this models poor performance in terms of predictive power and overall instability make it unsuitable.

# Conclusions

This study demonstrates that machine learning can certainly provide improvements upon traditional methods. Among the models that were tested, Support Vector Machines provided the best balance between capturing defaulters and avoiding false alarms.

The key conclusion here, is that not one specific model will be consistently deemed appropriate to an institution, guidance on model selection should be based upon a businesses priority in whether they deem missed defaulters or avoidance of over-flagging customers as most important.

# Limitations And Future Work

While the results are promising, there are still limitations:

* + Limited hyperparameter tuning: Settings utilised were mostly default, future work could include or apply grid search to change model capacity.
  + Class Imbalance: The data was skewed at 22% default. Working on a separate dataset to improve model accuracy with less imbalance could reap better results.
  + SVM lacks in terms of transparancy, future models could utilise explainable AI tools (SHAP) for better explanation.
  + Dataset Scope: Data was collected from one institution in Taiwan (Yeh & Liean 2009). Broader, more recent, or better spread (age) datasets would be a useful test of generalisability.

**References:**

* Breiman, L. (2001) ‘Random forests’, Machine Learning, 45(1), pp. 5–32.
* Cortes, C. and Vapnik, V. (1995) ‘Support-vector networks’, Machine Learning, 20(3), pp. 273–297.
* Géron, A. (2019) Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. 2nd edn. Sebastopol, CA: O’Reilly.
* Goh, J. (2019) ‘Applications of support vector machines in credit risk modeling’, International Journal of Data Science and Analytics, 8(4), pp. 321–332.
* Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018) ‘A survey of methods for explaining black box models’, ACM Computing Surveys, 51(5), Article 93. doi: 10.1145/3236009.
* He, H. and Garcia, E.A. (2009) ‘Learning from imbalanced data’, IEEE Transactions on Knowledge and Data Engineering, 21(9), pp. 1263–1284. doi: 10.1109/TKDE.2008.239.
* Lessmann, S., Baesens, B., Seow, H.V. and Thomas, L.C. (2015) ‘Benchmarking state-of-the-art classification algorithms for credit scoring’, European Journal of Operational Research, 247(1), pp. 124–136.
* Markov, A., Seleznyova, Z. and Lapshin, V. (2022) ‘Credit scoring methods: Latest trends and points to consider’, Expert Systems with Applications, 198, 116804.
* OpenML (2013) ‘Credit Card Default Dataset (ID 42477)’. Available at: <https://www.openml.org/d/42477> (Accessed: 20 August 2025).
* Raschka, S. (2014) ‘Naive Bayes and text classification I – introduction and theory’, arXiv preprint arXiv:1410.5329.
* Saito, T. and Rehmsmeier, M. (2015) ‘The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets’, PLoS ONE, 10(3), p. e0118432. doi: 10.1371/journal.pone.0118432.
* 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.