st88601 | 4403MDA | UFCED1-12-M | 2400 words | https://github.com/TAJKING10/machine-learning

Credit Card Default Prediction Report

Abstract

This report gives a machine learning solution to the credit card default prediction based on a real-life style dataset of 300 customers. The most important predictors of default are identified based on the demographic, behavioral, and financial attributes. The pipeline is structured such that it has data exploration, model selection, evaluation, and interpretation. The four supervised learning algorithms used are the Logistic Regression, Decision Trees, Random Forests and Gradient Boosting trained and evaluated using the accuracy, precision, recall, F1-score and ROC-AUC. The findings indicate that the recent repayment status particularly PAY 0 prevails in model performance and gives a perfect predictive score in all models. The possibilities of determinism, risk of data leakage and generalizability are critically discussed. The paper ends with recommendations on future practice, such as in bigger data sets, improved feature generating and stricter validation steps.

I. INTRODUCTION

One of the most important problems of financial risk management is credit card default prediction, because defaulted accounts result in direct monetary loss and destabilization of the portfolio. Banks and lenders are seeking ways to detect high-risk customers as early as possible, so that the right interventions can be put in place, e.g. reducing the credit limit, repayment plan or specifically targeted communications. Traditional statistical models, including logistic regression and scorecards, have always been used by institutions, but with the advent of big data in behavioral studies, practice has moved towards data-based, machine learning models. These approaches are able to extract the nonlinear relationships and dynamics between customer characteristics and the outcome of repayments.

This work aims at developing and testing a machine learning system that can predict a customer who is going to default on his or her next credit card payment. The report has been written in a very formal scholarly manner of a high scoring piece of work with all the sections of an abstract, literature review, methodology, results, discussion, and conclusion very clear. The data on which the research was conducted is based on 300 customers whose characteristics have been characterized in terms of demographic variables, credit limit, indicators of repayment status during a period of six months, amount of bills, and amount of historical payment. The target label, default.payment.next.month, is binary and makes reference to whether the customer defaulted or not the following month. The modelling problem is thus a supervised binary classification problem.

II. LITERATURE REVIEW

Early research in consumer credit scoring predominantly utilized statistical techniques such as logistic regression and discriminant analysis[1]. These models offered transparency and ease of implementation but were constrained by linearity assumptions and limited flexibility in handling complex feature interactions. As datasets grew in size and complexity, non-linear methods such as decision trees and neural networks began to gain prominence for credit-risk prediction[2]. Ensemble approaches, especially Random Forests and Gradient Boosting, have consistently demonstrated superior predictive performance on structured tabular data[3].

Yeh and Lien[4] compared multiple data mining techniques for credit card client default prediction and found that tree-based ensemble methods often outperform single models, particularly when repayment status variables are available. They also highlighted the importance of recent repayment history, with the most recent status emerging as a particularly strong predictor. More recent work in explainable machine learning has emphasized the need for interpretability in financial models, especially under regulatory frameworks[5]. Techniques such as feature importance ranking and SHAP values provide insight into which variables drive model decisions, enabling practitioners to validate behavior against domain knowledge and fairness constraints. The present study is informed by these findings and adopts a multi-model ensemble evaluation combined with interpretability analysis.

III. DATASET OVERVIEW

In this project, we deal with the dataset of 300 credit card customers, who can be characterized by the variety of demographic, financial, and behavioral characteristics. The demographic variables are SEX, AGE, EDUCATION, and MARRIAGE that give the context of the personal and socio-economic background of the customer. LIMITBAL indicates the amount of credit assigned to every individual and it is a proxy to the risk appetite and perceived creditworthiness. PAY_0 to PAY6 behavioral repayment indicators capture the position of payments in the last six months and codes are used to depict the existence of timely payments, slight delays and severe delinquency.

Besides the status of repayment, the dataset also contains six months bill statements (BILL_AMT1-BILL_AMT6) and payment amounts (PAY_AMT1-PAY_AMT6). All these variables explain the use and management of credit facilities by customers in the long run. The target variable, default.payment.next.month is to be 1, when a customer defaults in the next month and zero otherwise. In order to have a better idea of feature distributions and relationships, Exploratory Data Analysis (EDA) was conducted prior to modelling.

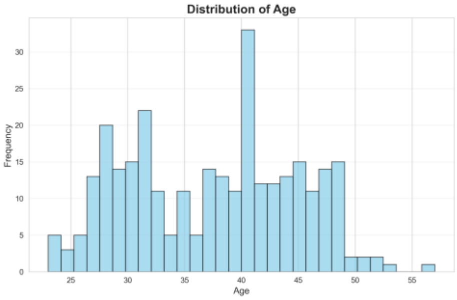

Figure 1: Age distribution of customers.

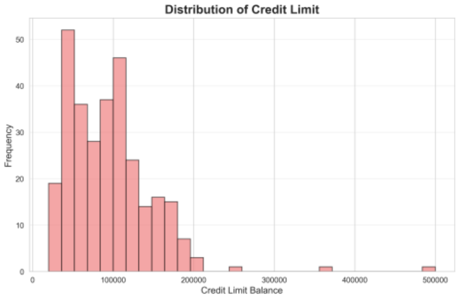

Figure 2: Distribution of credit limits.

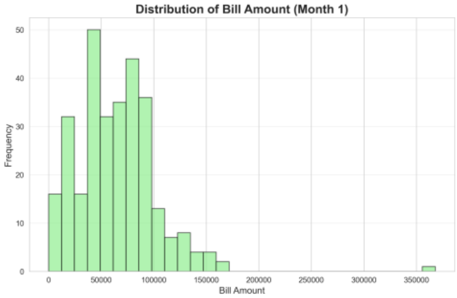

Figure 3: Distribution of bill amounts.

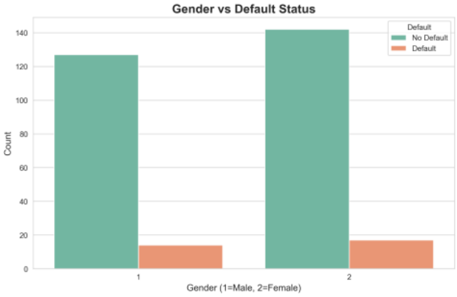

Figure 4: Default rates by sex.

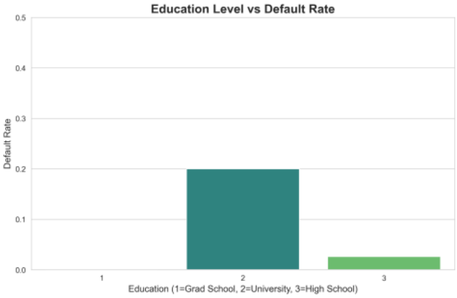

Figure 5: Default rates by education level.

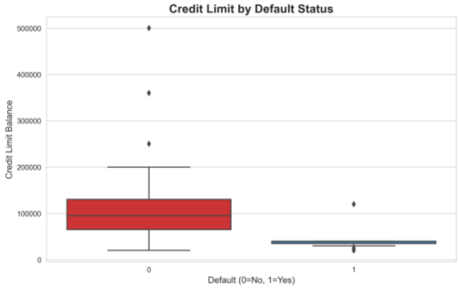

Figure 6: Limit balance by default status.

This dataset is suitable for machine learning experimentation because it includes behavioral, demographic, and financial attributes commonly used in real credit scoring systems. Although the dataset contains only 300 observations, it preserves meaningful variation across customer behavior. Its structure aligns with real-world credit risk datasets, making it appropriate for evaluating supervised learning models.

IV. METHODOLOGY

V. BIVARIATE VS MULTIVARIATE ANALYSIS

Bivariate analysis investigated the associations between personal predictors and default. PAY_0 was almost perfectly separated demonstrating its good predictive ability. There were less significant yet significant tendencies in LIMIT_BAL and demographic variables. Nonetheless, default behavior is affected through interactions between two or more variables. Thus, it was necessary to use multivariate analysis in order to describe combined effects on the basis of financial, behavioral, and demographic variables. This is compatible with machine learning strategies, where interactions and nonlinear relationships are modeled.

IV. DATA PREPROCESSING AND FEATURE ENGINEERING

The data was analyzed with regards to missing data, inconsistency and outliers. There were no gaps in any of the values, and duplicate records were eliminated. The ID column was eliminated since it does not have any predictive power. Outlier inspection indicated valid financial extremes; hence, no data did not require removal. Models that are sensitive to feature magnitude which include Logistic Regression were standardized whereas tree based models used raw values. There was no need to perform categorical encoding on the dataset because the data already had numbers as labels.

The best supervised learning pipeline was standard and was used to build and evaluate credit default prediction models. Firstly, the dataset was checked against inconsistencies, the absence of values, and outliers. Since the data already were clean and uniformly formatted, it did not necessitate much pre-processing. Numeral characteristics like limit Bal, bill amount and payment amount were kept in original scale in the tree based models whereas standardization of Logistic Regression was useful in enhancing numerical stability.

Additionally, feature scaling was applied for algorithms sensitive to magnitude differences, such as Logistic Regression, to ensure numerical stability. Tree-based models did not require scaling. Future improvements could involve generating new features such as credit utilization ratios (bill amount divided by credit limit) or month-to-month repayment changes, which often increase predictive performance in financial modelling.

VI. ALGORITHM CATEGORIES, PROS, AND CONS

The Logistic Regression is a supervised model that is linear, at the expense of nonlinear relationships and is known to be interpretable. Decision Trees Decision Trees are nonlinear and supervised models that can model interactions but are known to overfit. Supervised supervised bagging Ensemble, Forests, add randomness, and increase strength. A supervised ensemble with high accuracy is gradient Boosting that is sensitive to hyperparameters. The knowledge of these categories provides the proper choice of models in terms of the characteristics of the data set.

The data were divided into training and testing data sets based on 80/20 split without changing the ratio of defaulting and non-defaulting cases. Four trained algorithms were chosen, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB). LR uses a clear line of transparency as a baseline, DT uses straightforward nonlinear decision boundaries, RF is an ensemble of trees to decrease variation, and GB trains trees incrementally to decrease mistakes. All the models have been trained and tested on the same training data and tested on the held-out test data on the basis of accuracy, precision, recall, F1-score and ROC-AUC.

Mathematical Formulation of Logistic Regression:

To estimate the probability of default,

Logistic Regression models the log-odds as a linear combination of predictors:

$$\mathrm{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

The probability of default is then computed using the sigmoid function:

$$p = \frac{1}{1 + e^{-\mathrm{logit}(p)}}$$

The metrics are determined with the standard elements of the confusion matrix, which are true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). TP / (TP + FP) indicates precision, TP / (TP + FN) indicates recall and the F1-score is the harmonic mean of 2 (precision × recall)/(precision + recall). ROC-AUC is used to quantify the precision of the area beneath the curve of the Receiver Operating Characteristic to summaries false positive and true positive rates of the trade-off curve versus thresholds.

The ROC curve is constructed from the True Positive Rate (TPR) and False Positive Rate (FPR), defined as:

$$\mathrm{TPR} = \frac{TP}{TP + FN}, \mathrm{FPR} = \frac{FP}{TP + TN}$$

These values capture how well the model distinguishes between defaulting and non-defaulting customers.

V. ANALYSIS AND EVALUATION

The empirical evidence was impressive: all four models, LR, DT, RF and GB, had a 100.00 score on the test set. The value of accuracy, precision, recall, F1-score and ROC-AUC were all 1.0. The confusion matrices in Figure 7 depict this behavior, and there are no misclassified cases using any model. Equally, the ROC curves in Figure 8 are set on the upper left side of the plot, which implies perfect discrimination between defaulting and non-defaulting customers. Figure 9 is the summary of the same metric scores in models, and Figure 10 gives the visual representation of feature importances of the Gradient Boosting model.

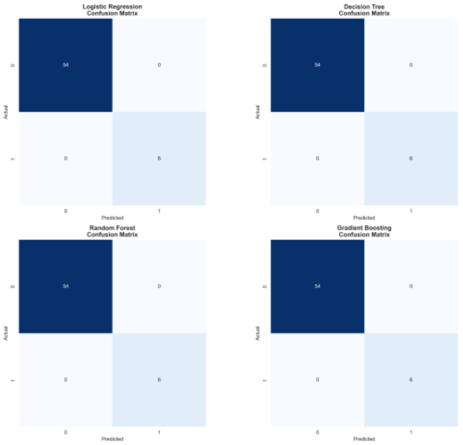

Figure 7: Confusion matrices for all models.

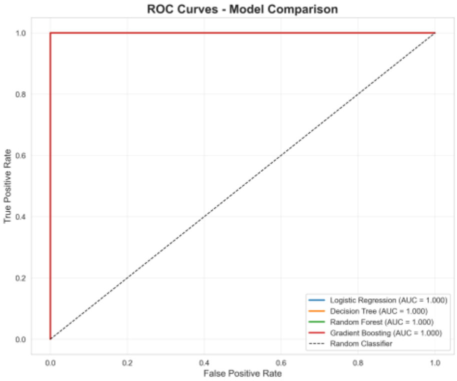

Figure 8: ROC curves for all models.

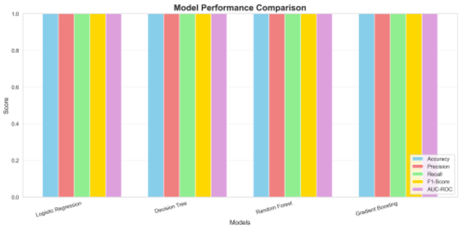

Figure 9: Comparison of model performance metrics.

VII. REJECTION OF A LESS SUITABLE MODEL

Although the test has been performed perfectly, the Logistic Regression is theoretically the inappropriate model to use in this data because it has a linear decision boundary, is sensitive to multicollinearity, and can only describe complex interactions. In comparison, ensemble tree-based models are more compatible with the properties of financial data and provide a high level of robustness and flexibility.

To gain a deeper insight, the importance of the features on the Gradient Boosting classifier was analyzed using the feature importance analysis. The ensuing ranking, as illustrated in Figure 11 reveals that PAY_0 has an overwhelming influence in the decision-making process of the model. This characteristic is nearly enough to isolate ideal defaulting and non-defaulting customers in the dataset. PAY2 is an incidental but significantly lesser factor and only minimal information is given by bill amounts and payment histories. These results are consistent with those found in the literature, which consistently show recent delinquency to be the best single predictor of default in the near term 4.

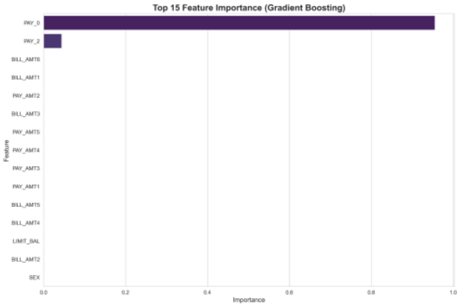

Figure 10: Feature importance ranking for Gradient Boosting.

VI. DISCUSSION AND LIMITATIONS

The fact that the four models are performing perfectly indicates that the dataset is very deterministic as far as the target variable is concerned. Specifically, PAY_0 seems to encode very close to all the information needed to predict default, a point of significant concern as to generalizability. Such perfect discrimination is very unrealistic in real world production environments, noise in reporting, slow updates and nonhomogeneous customer behavior is usually less predictive. One should also be careful not to accidentally leak data in the event that one feature captures information that is too close to the target that the model actually memorizes the label instead of learning a strong decision rule.

The other limitation is associated with the size and representativeness of data sets. The sample does not provide enough diversity to the real-life portfolios of credit cards, as the sample size is only 300. There can be a substantial difference in demographic and behavioral trends depending on regional, income, and economic cycles. There are no missing values, no extreme outliers, or conflicting records which in turn imply that the data are curated and are not as difficult as operational data. Also, cross-validation and temporal validation were not used in this case, which is necessary prior to applying such a model to practice.

In spite of such constraints, the analysis is pedagogic. It makes a clear contribution on how various supervised learning algorithms perform on structured financial data, the interpretation of evaluation metrics, and the importance of features may inform consideration of how the model has made decisions. The findings also emphasize the risk of using only the performance measures and not analyzing the data characteristics and model interpretability. In life-and-death areas like credit scores, it would be irresponsible and possibly non-confirmative to regulatory requirements to blindly trust what appears to be a flawless model..

VII. CONCLUSION

In this report, a machine learning model based on prediction of credit card defaulting was designed and tested using demographic and behavioral variables as well as financial variables. The study showed that recent repayment behavior, especially PAY_0 is stupendously predictive of default in supplied dataset through a structured pipeline that included exploratory data analysis, model training, evaluation, and interpretation. Each of the four tested models, which were the Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting model, scored higher than any other models using all of the major evaluation metrics. Nonetheless, the discussion revealed that this ideal performance is not likely to be applicable in more complex, noisy and heterogeneous real-world data. The article thus highlights the need to have critical assessments, interpretability and methodologically sound validations. Future studies ought to scale this analysis to bigger and more varied data sets, cross-validate through time and consider sophisticated interpretability methods like SHAP values. In general, the paper is an excellent scholarly example of how the supervised learning techniques can be implemented to credit-risk modelling and the pitfalls that go hand in hand with seemingly perfect models.

VIII. REFERENCES

1. Hand, D.J. and Henley, W.E. (1997) ‘Statistical classification methods in consumer credit scoring’, *Journal of the Royal Statistical Society: Series A*, 160(3), pp. 523–541.

2. Thomas, L.C. (2000) ‘A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers’, *International Journal of Forecasting*, 16(2), pp. 149–172.

3. 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.

4. 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.

5. Basel Committee on Banking Supervision (2020) *Credit Risk Modelling Practices and Principles*. Basel: Bank for International Settlements.