Predicting Credit Card Risk

MSBA 305: Business Intelligence
& Decision Support

Term Paper

Respected Heinz Joerg Schwarz

Attempted by: Digvijay Bisla

1. Table of Contents

Abstract

Widely regarded as one of the key drivers of economic growth, credit plays a central role in the profitability of banking and finance institutions. This study aimed to predict the credit risk of credit card holders using hybrid random forest model as a decision support tool for credit risk assessment usage by a hypothetical bank of India. Two algorithms were evaluated: against the Random Forest, including Naive Bayes, and Decision Trees. The models were evaluated using accuracy and recall metrics. Based on the findings it was noted that the Random Forest algorithm had the highest accuracy of 97% and recall rate of 97%, indicating that the model has a 97% ability to reduce false negatives (i.e., predict clients classified as low-risk but actually high-risk).

2. Introduction

Credit risk comes about when the debtor fails or delays debt repayment, be it in its entirety or part of it against the agreement in the debt contract. According to (Nobanee, Shanti, Aldhanhani, Alblooshi, & Alali, 2022), credit risk is the likelihood that a legally binding contract would become worthless or decrease in value due to defaulting by a business or individual. Over the recent years, credit has been at the center of modern economics given the role of credit in the growth of the economy when used well (Butiuc, 2013). However, the misuse of credit can cause damages and disruption to the economy with immediate effects being felt by banks and other lending institutions whose total assets and ability to lend decreases (Opa & Tabe-Ebob, 2020). Given the effects of credit default, the prediction of credit risk is an important aspect of credit risk management. To this end, the question is, can accurate models be built to predict credit risk?

Modern finance is dependent on credit and trust. There have been extensive studies regarding the application of statistics and machine learning in predicting credit risk. (Shi, Tse, Luo, D'Addona, & Pau, 2022) argue that, accurately determining the likelihood of a borrower defaulting on their debts is a fundamental aspect in credit risk evaluation, which plays a significant role in assessing and predicting the risk involved. Ideally, advances in statistics and computing power have led to the adoption of new methods related to machine learning and deep learning (Nobanee, Shanti, Aldhanhani, Alblooshi, & Alali, 2022; Rozo, Crook, & Andreeva, 2023).

Objective

Various factors influence the credit default risk of debtors. The current study seeks to propose an optimal machine learning model for predicting the likelihood of credit defaulting for a hypothetical bank of India which has experienced an increment in credit card defaults, and it needs to reduce this trend by generating insights regarding the main factors that influence credit

default among credit card holders. In particular, the current study seeks to implement a hybrid random forest model as well as use descriptive analytics with a functional dashboard to generate an understanding of various attributes related to credit default among credit card holders.

3. Literature Review

Default Risk

Banks generate most of their profits through accepting deposits and giving out credit – the primary business operation of banks (Kwashie, Baidoo, & Ayesu, 2022). However, in the process of carrying out these crucial operations, financial institutions encounter various forms of risk. (Quang & Gan, 2018) argue that it is unlikely to completely eliminate risks, especially credit risks.

Credit risk is the probability of a credit holder failing to meet their obligations under the agreed terms (Shi, Tse, Luo, D'Addona, & Pau, 2022). Credit risks are the costliest risks for banks besides affecting the economic stability of economies since they erode the profitability of banks and marks the onset of a crisis (Naili & Lahrichi, 2022). By definition, it is therefore important to adopt efficient risk management strategies for banks or become part of their business models (Kwashie, Baidoo, & Ayesu, 2022) to ensure its profitability and soundness (Quang & Gan, 2018).

Risk Prediction

Risk prediction focuses on the various methods and models used for predicting the likelihood of default by credit borrowers (Oliinyk, Donchenko, Larionova, & Kapinos, 2019).

Statistical Models

Statistical models, such as logistic regression and discriminant analysis, have been widely used for credit risk prediction (Khemais, Nesrine, & Mohamed, 2016). These models are based on traditional credit scoring attributes, including but not limited to income, age, credit history, bank balance and outstanding debt. Often, statistical models are straightforward to implement and interpret, and have generally been found to provide accurate predictions in many cases.

For instance, a logistic regression model is used to calculate the odds ratio of a defaulter using the formula:

$$\text{logit} = \log\left(\frac{p}{1-p}\right)$$

Where p is the probability of success. The logit can be modeled as a linear function of the predictor variables, which is known as the logistic regression equation (Memic, 2015). The logistic regression equation takes the form:

$$\text{logit} = \beta_0 + \beta_{1x1} + \beta_{2x2} + \ldots + \beta_{kxk}$$

where $\beta 0$, β_1, β_2, ..., β_k are the regression coefficients that are estimated from the data, and x_1, x_2, ..., x_k are the predictor variables.

However, statistical models have been known to fail to capture the complexity of the underlying relationships between the predictor variables and the likelihood of default.

Machine learning models

Machine learning (ML) models, such as decision trees (Wu, 2022), random forests (Madaan, Kumar, Keshri, & Jain, 2021), support vector machine (Moula, Guotai, & Abedin, 2017) and neural networks (Gao, Xiong, Xiong, & Xiong, 2021), have become increasingly popular for credit risk prediction in recent years. In practice, ML models are designed to learn complex relationships between the predictor variables and the likelihood of default from the data. (Gambacorta, Huang, Qiu, & Wang, 2019) notes that ML models tend to provide improved predictions compared to statistical models, especially when the number of predictors involved is large. However, in cases where the training data is imbalanced, they may also suffer from overfitting or lack of interpretability, and they may require more computational resources (Tran & Dang, 2021; Singh, Ranjan, & Tiwari, 2021; Khatir & Bee, 2022).

Hybrid models

In credit default prediction, hybrid models combine the predictive ability of both statistical and machine learning models (Liu, Zhang, & Fan, 2022). For instance, a statistical model can be used to pre-process the data, such as imputing missing values or reducing the dimensionality of the data, after which a machine learning model can be used for the final prediction. Studies like (Chi, et al., 2020; Li, Stasinakis, & Yeo, 2022) demonstrate that hybrid models generate relatively improved predictions compared to either statistical or machine learning models alone.

Alternative data and data augmentation

Studies have been conducted focusing on alternative data sources such as social media, mobile phone data, and health data, have been used to augment traditional credit scoring variables for improved credit risk prediction (Wei, Yildirim, Bulte, & Dellarocas, 2015). These alternative data sources can provide valuable information on a borrower's behavior and lifestyle that may not be captured by traditional credit scoring variables. However, the application of alternative data also raises several challenges, ranging from data quality, data privacy, to data interpretability (Tanant, 2022).

Model Evaluation

Model evaluation plays a central role in the creation of an optimal prediction model. (Bastos, 2020) places more emphasis on a model's recall which is the proportion of true positive predictions among all actual positive observations in the data computed as:

$$\text{Recall} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

where True Positives (TP) are the number of instances that are correctly classified as positive by the model while False Negatives (FN) are the number of instances that are actually positive but are incorrectly classified as negative by the model. Ideally, *recall* is important when the goal is to reduce the number of individuals who are predicted as non-defaulters (Bastos, 2020).

4. Methodology

Data

Data proposed for this study includes information on default payments, demographic information, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The dataset contains 30,000 observations and 25 variables. It is acknowledged that the original data set belongs to (Lichman, 2013). Table 1 below provides an overview of the characteristics of the data.

Table 1: Data characteristics

Data characteristic	**Observation**
Demographic Information	Age, education, and marital status of the credit card clients.
Credit information	Amount of credit limit, the balance on the credit card, and the number of months since the last default
Payment history	Payment status for the previous six months (indicating whether the payment was on time, delayed for one month, etcetera.).
Bill statements	amount of bill statement for the previous six months, and the amount of previous payment.
Target variable	"Default payment next month", which indicates whether the credit card client will default on their payment in the following month (1 = default, 0 = not default).

The target variable as shown in figure 1 below is imbalanced with approximately 22% of the observations being labeled as default.

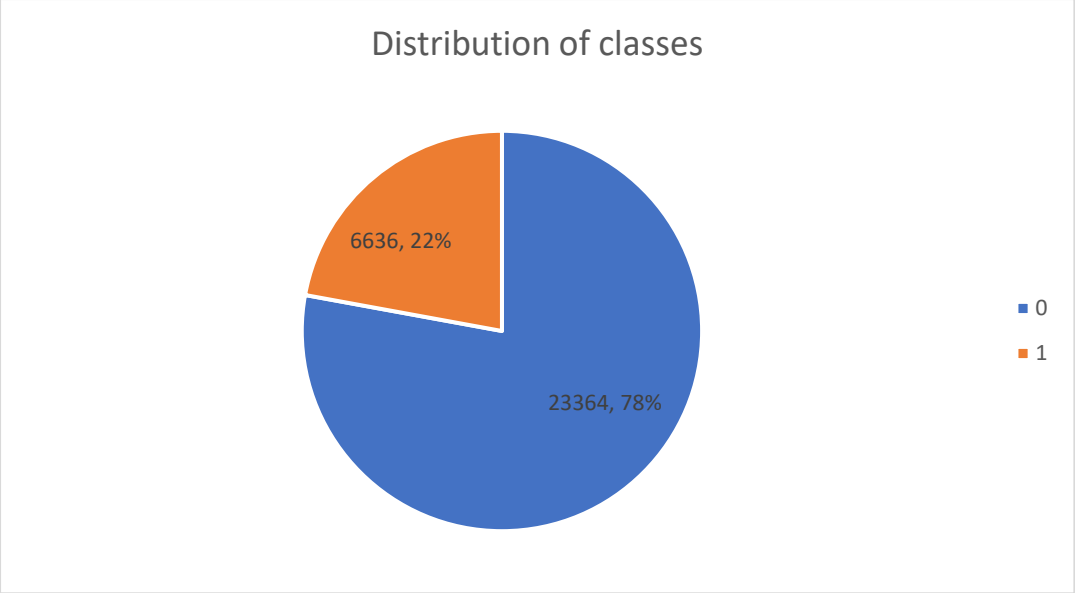

Figure 1: Distribution of classes

As a result, the hybrid prediction model will involve class balancing before generating predictions. Both oversampling and under-sampling methods will be implemented and the respective performances evaluated.

Data Analysis

Machine Learning Model

The random forest (RF) prediction model was proposed as the second stage of the hybrid model. In practice, an RF model is an ensemble method that uses multiple decision trees to make predictions. Every decision tree in the RF model is trained on a bootstrapped sample of the data, and the predictions of the individual trees are combined to generate the final predictions (Schonlau & Zou, 2020). Since the RF model is robust to outliers and can generate a non-linear relationship between the attributes, it generally performs well for credit default prediction.

In this study, an RF model with optimized hyperparameters will be used to predict the probability of defaulting.

Descriptive Analysis

Various visualizations will be created to provide an overview of the various customer attributes in relation to credit risk. In particular, this study will create a demographical functioning dashboard to enable the examination of the demographic attributes of defaulters and non-defaulters.

5. Analysis Results

Dashboarding

The questions aimed to be answered by the dashboard are related to the proportion of defaulters by gender, education level and marital status to assess the credit risk level of various demographics.

Distribution of defaulters by age

As shown in figure 2 below, the average age for defaulters is approximately 36 years while non-defaulters are aged approximately 35 years.

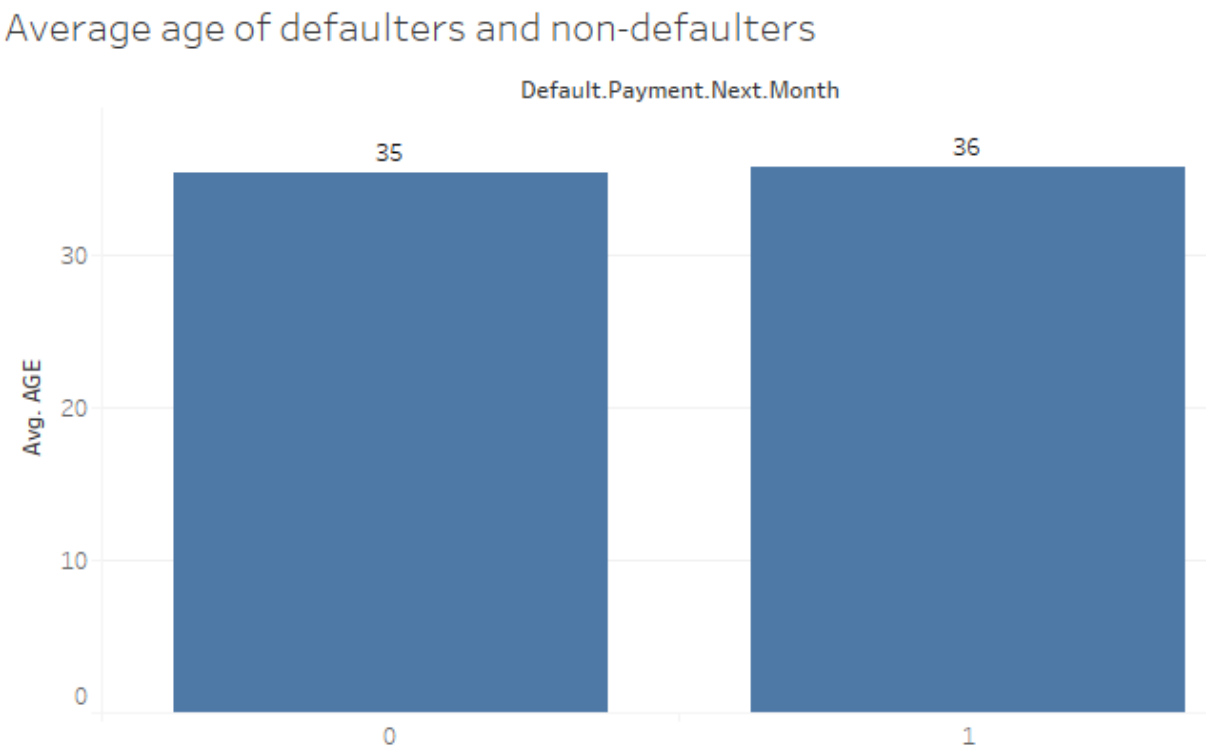

Figure 2: Average age of defaulters and non-defaulters

Proportion of defaulters by gender

In figure 3 below, 1 = male; 2 = female. Overall, 24.17% of male customers are defaulters while 20.78% of the female card holders are defaulters indicating that on average, there are more male defaulters among male clients than female defaulters as a proportion of the female clientele.

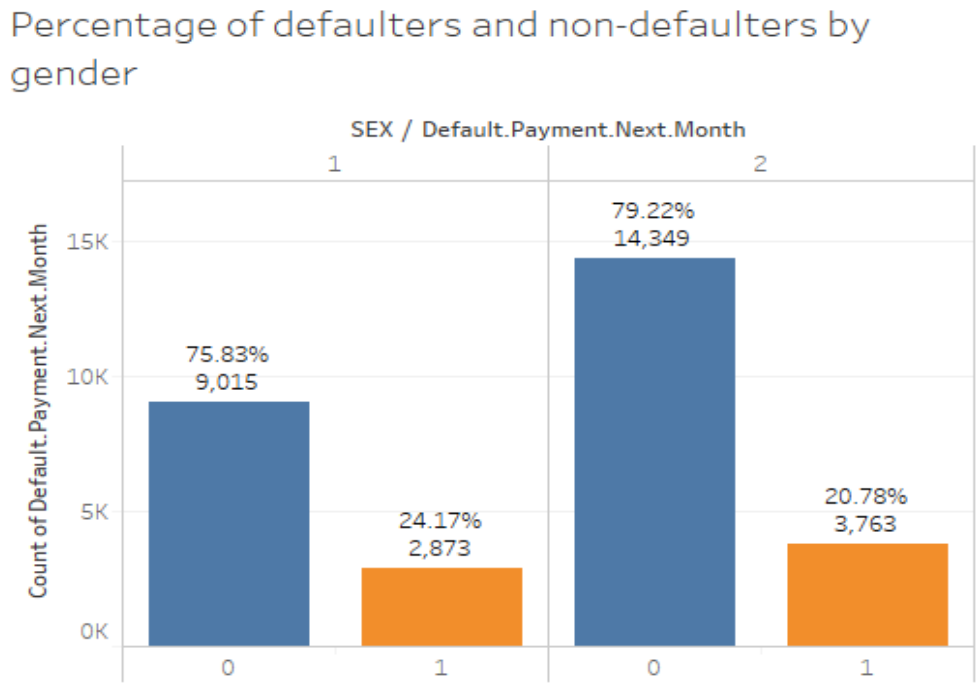

Figure 3: Percentage of defaulters and non-defaulters

Marital status and defaulting

11.14% of the card holders who are single defaulted while 10.69% are married and only 0.28% fall into the *other* category (*see figure 4*).

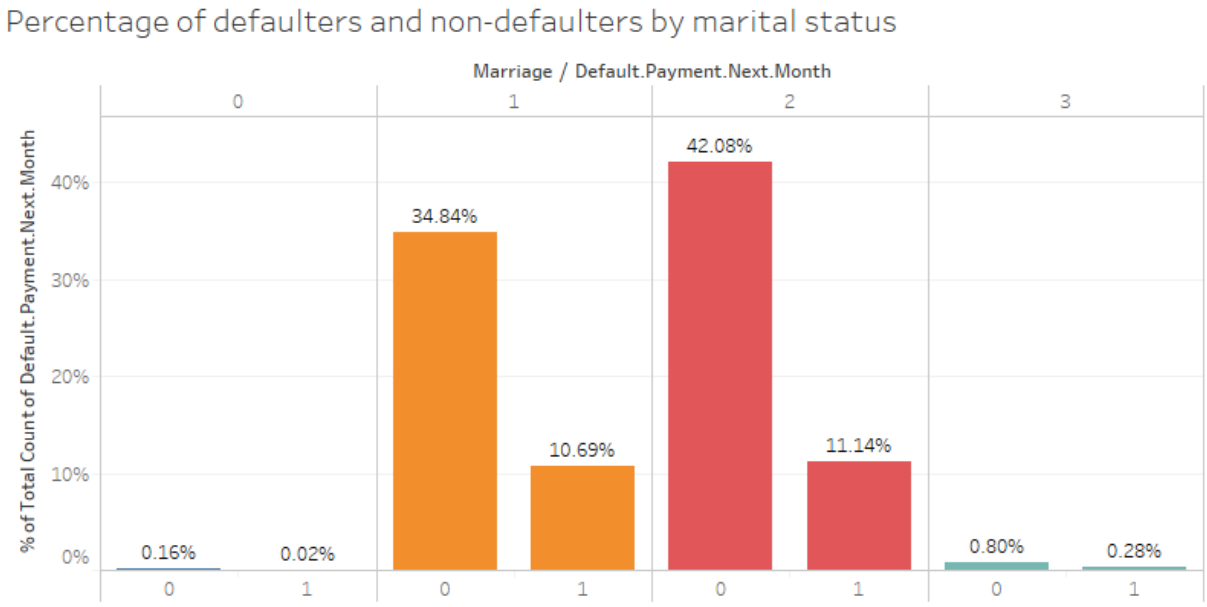

Figure 4: Percentage of defaulters and non-defaulters

Education level

Figure 5 below provides an overview of the percentage of defaulters and non-defaulters by education level. As observed from figure 5 below, the top three highest proportion of the defaulters attained university education (11.10%), 6.79% attained graduate school education while 4.12% attended high school respectively.

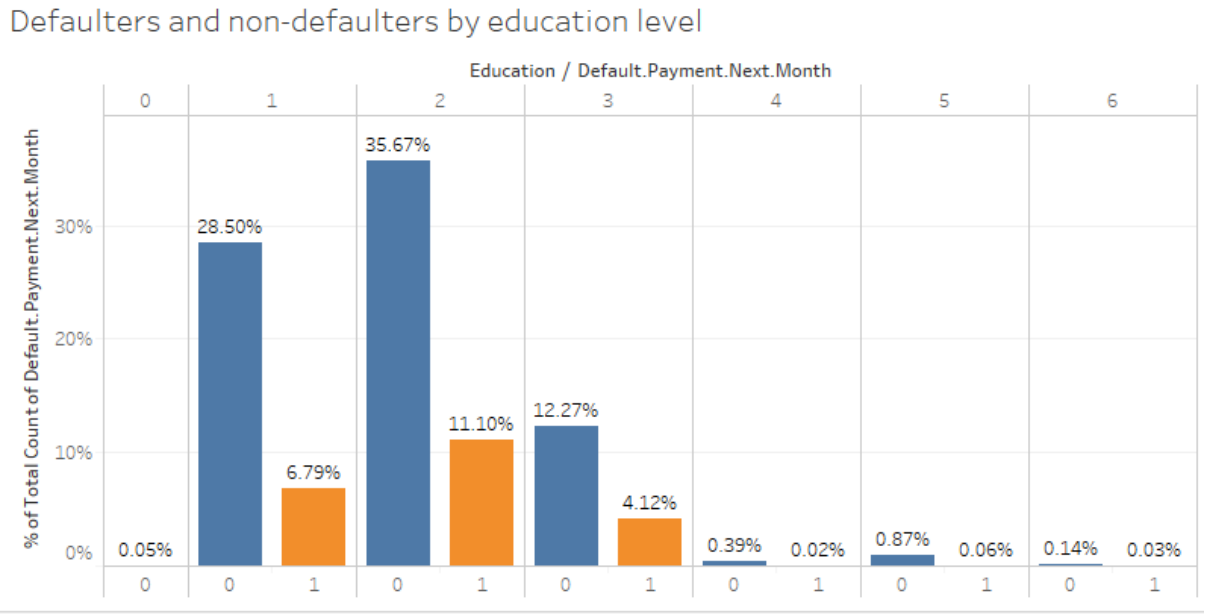

Figure 5: Defaulters and non-defaulters by education level

Dashboard

A functional dashboard was developed for the above visualizations which allow filtering by default status enabling the user to drill down the dataset for better generation of insights. The dashboard is used to generate the following insights to the high-risk group of card holders based on a user's interaction. For instance figure 6 below shows the resulting dashboard for users who defaulted.

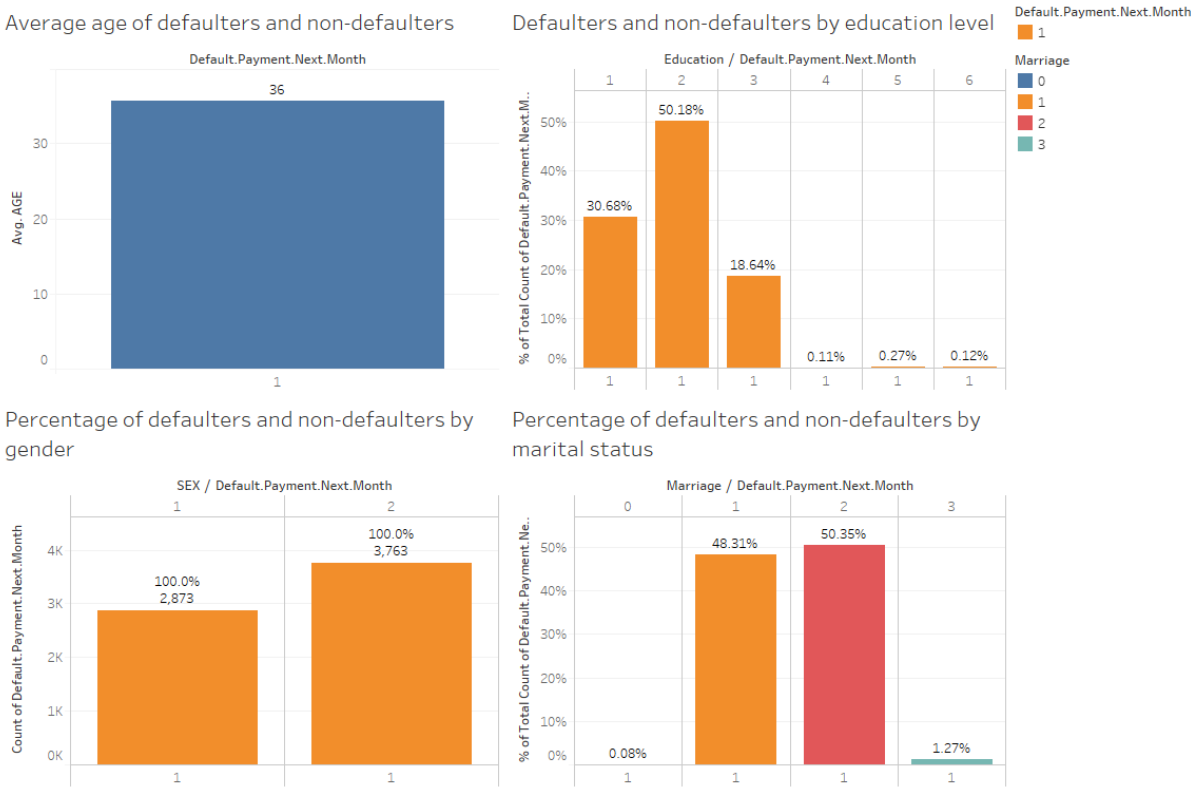

Figure 6: Dashboard Overview

Machine Learning

Figures 7 and 8 below show the distribution of the observations after resampling to improve class imbalance through oversampling and under sampling.

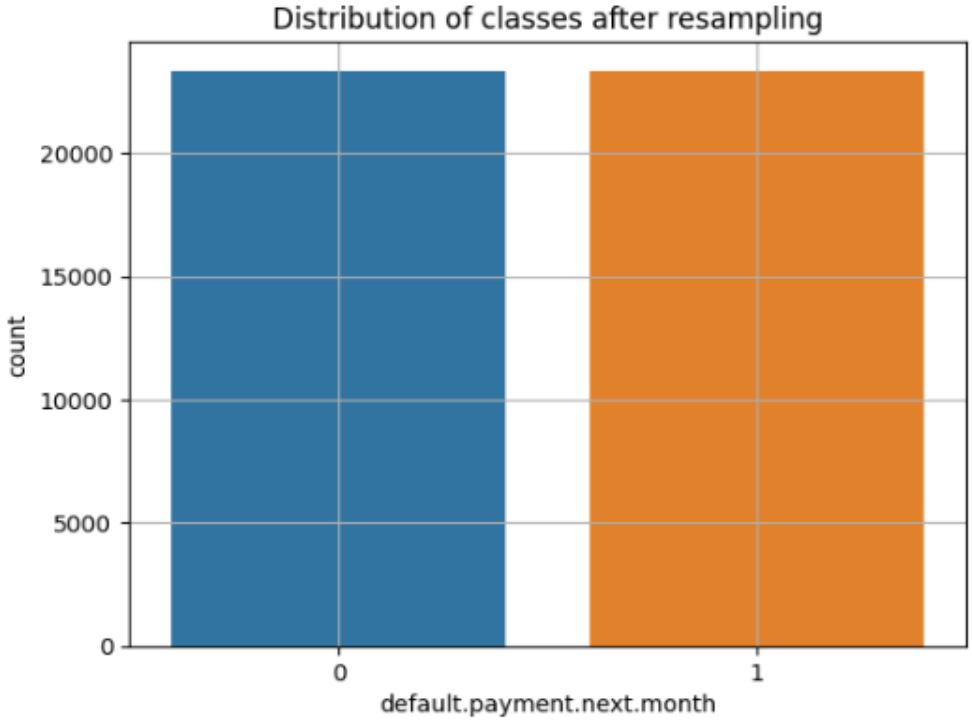

Figure 7: Class distribution after oversampling

Figure 8: Class distribution after under-sampling

Table 2 below shows the performance of the hybrid prediction model across the three data preparation methods.

Table 2: Model performance

Hybrid method	**Recall**	**Classification accuracy**
Oversampling	0.92	0.92
Under-sampling	0.97	0.97
No sampling	0.92	0.95

As noted from table 2 above, the random forest (RF) attains the optimal performance when trained with the under sampled data with a classification accuracy of approximately 97% with a recall score of 97% indicating that the model makes 97% correct identification of individuals who defaulted. Interestingly, non-balanced data had the same recall rate of data balanced suing over-sampling.

Other models including a standard decision tree and nave bayes were used for performance comparison. Table 3 shows the model performance of the RF model relative to the decision tree and naïve bayes.

Classification model	Recall	Classification accuracy
Random Forest	0.97	0.97
Naïve Bayes	0.59	0.54
Decision Tree	0.61	0.61

From table 3 above, it is observed that the random forest model attains the highest recall which improves the ability of the model to remove negatives (97%) compared to the Naïve Bayes (59%) and Decision Tree (61%).

6. Conclusion

The current study sought to aimed to predict the credit risk of an Indian bank given a dataset related to credit card clients using machine learning algorithms. To this end, several approaches were adopted including 2-step hybrid machine learning models. The steps included *step 1: resample the data, step 2: fit and evaluate the model on the resampled data*. The algorithms evaluated included Random Forest, Naive Bayes, and Decision Trees, and compared their performance using both the accuracy and recall metrics.

Following the analysis and comparison process, it was observed that the Random Forest algorithm had the highest recall (97%) and an accuracy of approximately 97%. The high recall rate attained by the model indicates that the RF model was successful in reducing the number of false negatives, i.e., the clients that were classified as low-risk but were actually high-risk.

Based on these results, we therefore conclude that the Random Forest algorithm can be used as a reliable tool for credit risk assessment by the hypothetical bank of India which has been experiencing an increment in credit card defaults.

7. References

Bastos, R. (2020, October 14). *Credit Risk Analysis with Machine Learning*. From Towards Data Science: https://towardsdatascience.com/credit-risk-analysis-with-machine-learning-736e87e95996

Butiuc, L. M. (2013). The Impact of Credit on Economic Growth in the Global Crisis Context. *Procedia Economics and Finance, 6*(2013), 25-30.

Chi, G., Uddin, M. S., Habib, T., Zhou, Y., Islam, M. R., & Chowdhury, M. A. (2020). A hybrid model for credit risk assessment: empirical validation by real-world credit data. *Journal of Risk Model Validation, 14*(4).

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2019). How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm. *BIS Working Papers, 834*(2019), 24.

Gao, X., Xiong, Y., Xiong, Z., & Xiong, H. (2021). *Credit Default Risk Prediction Based On Deep Learning.* Research Square.

Khatir, A. A., & Bee, M. (2022). Machine Learning Models and Data-Balancing Techniques for Credit Scoring: What Is the Best Combination? *Risks, 10*(169).

Khemais, Z., Nesrine, D., & Mohamed, M. (2016). Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression. *International Journal of Economics and Finance, 8*(4).

Kwashie, A. A., Baidoo, S. T., & Ayesu, E. K. (2022). Investigating the impact of credit risk on financial performance of commercial banks in Ghana. *Cogent Economics & Finance , 10*(2022).

Li, Y., Stasinakis, C., & Yeo, W. M. (2022). A Hybrid XGBoost-MLP Model for Credit Risk Assessment on Digital Supply Chain Finance. *Forecasting, 4*(1), 184-207.

Lichman, M. (2013). *UCI machine learning repository*. From UCI: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Liu, J., Zhang, S., & Fan, H. (2022). A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. *Expert Systems with Applications, 195*(2022), 116624.

Madaan, M., Kumar, A., Keshri, C., & Jain, R. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series Materials Science and Engineering*. Reseach Gate.

Memic, D. (2015). Assessing Credit Default using Logistic Regression and Multiple Discriminant Analysis:Empirical Evidence from Bosnia and Herzegovina. *Interdisciplinary Description of Complex Systems , 13*(13), 128-135.

Moula, F., Guotai, C., & Abedin, M. (2017). Credit default prediction modeling: an application of support vector machine. *Risk Manag, 19*(2017), 158-187.

Naili, M., & Lahrichi, Y. (2022). Banks' credit risk, systematic determinants and specific factors: recent evidence from emerging markets. *Heliyon, 8*(2).

Nobanee, H., Shanti, H., Aldhanhani, H., Alblooshi, A., & Alali, E. (2022). Big data and credit risk assessment: a bibliometric review, current streams, and directions for future research. *FINANCIAL ECONOMICS, 10*(2132638), 1-18.

Oliinyk, A., Donchenko, T., Larionova, K., & Kapinos, H. (2019). Modeling Credit Risk in Banking. *Proceedings of the 6th International Conference on Strategies, Models and Technologies of Economic Systems Management (SMTESM 2019).* Atlantis Press.

Opa, V. O., & Tabe-Ebob, W. T. (2020). *The Effects of Loan Default on Commercial Banks Profitability: Case Study BICEC Limbe.* Research Gate.

Quang, N. T., & Gan, C. (2018). Bank Risk Management: A Regulatory Perspective. In A. G. Hessami (Ed.), *Perspectives on Risk, Assessment and Management Paradigms.* InTechOpen.

Rozo, B. J., Crook, J., & Andreeva, G. (2023). The role of web browsing in credit risk prediction. *Decision Support Systems, 164*(2023), 113879.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting communications on statistics and Stata, 4*(450).

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications, 34*(2022), 14327–14339.

Singh, A., Ranjan, R. K., & Tiwari, A. (2021). Credit Card Fraud Detection under Extreme Imbalanced Data: A Comparative Study of Data-level Algorithms. *Journal of Experimental & Theoretical Artificial Intelligence, 34*(2013), 1-28.

Tanant, F. (2022, September 29). *Social Media Credit Scoring: Pros, Cons, and How to Do It.* From SEON: https://seon.io/resources/social-media-credit-scoring/

Tran, T. C., & Dang, T. K. (2021). Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection. *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM).* IEEE.

Wei, Y., Yildirim, P., Bulte, C. V., & Dellarocas, C. (2015). Credit Scoring with Social Network Data. *Marketing Science, 35*(2), 234-258.

Wu, W. (2022). Machine Learning Approaches to Predict Loan Default. *Intelligent Information Management, 14*(5).