

Predictive Analytics in Credit Risk Management

By Nadia Isanga

EXECUTIVE SUMMARY

This report outlines the development of a predictive model to assess credit for a banking client, which is crucial for informed decision making on approving credit. The approach integrates industry standard methodologies with robust data processing to ensure reliability and accuracy in the predictions. Here are the key components and findings:

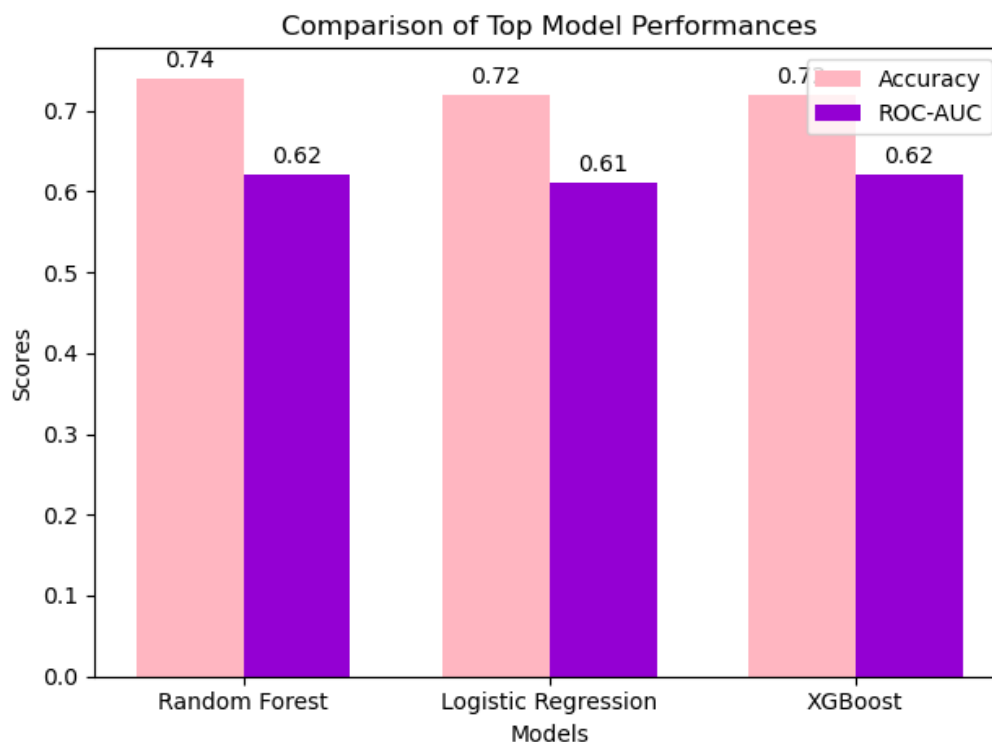


Figure 1: Accuracy and ROC-AUC of the top models considered.

- **Methodologies Employed:**
 - **Random Forest Classifier:** Selected for its robustness in handling large datasets and its capability to model complex relationships. It outperformed other models in terms of overall accuracy and provided valuable insights into feature importance.
 - **Logistic Regression:** Utilized for its transparency and speed, facilitating easy interpretation of the effects of various predictors on credit risk. Although less powerful, it provides essential benchmarks for probability-based risk assessments.
- **Data Pre-processing Techniques:**
 - **Handling Imbalances:** Employed Synthetic Minority Over-Sampling Technique (SMOTE) to address class imbalances, enhancing the predictive performance regarding minority class outcomes.

- **Feature Engineering:** Conducted thorough analyses to select relevant features, reducing dimensionality and focusing on the most informative predictors based on domain knowledge and statistical tests.
- **Key Findings:**
 - The Random Forest model, enhanced with hyperparameter tuning and class rebalancing, reached an accuracy of 74%.
 - Logistic Regression, though not as strong in performance metrics, provided clear insights into the likelihood of default, establishing itself as a valuable asset for initial risk assessments.
- **Limitations:**
 - The models do well, but they are not perfect and this maybe due to insufficient diverse data and class imbalances. Additional integrations and enriching the dataset with more varied inputs could enhance outcomes.
 - The models might oversimplify how different factors interact. Using more complex machine learning methods could help us understand these relationships better.
- **Further Analysis Possibilities:**
 - Incorporating macroeconomic indicators could help analyze how broad economic trends affect credit risks.
 - Employing deep learning techniques could also give a sharper edge in predicting outcomes more accurately.

This executive summary outlines the methods employed, the crucial steps in data processing, and consolidates the findings. The full report goes deeper into these points, offering detailed reasons for each method used and a thorough analysis of the results, along with strategic advice for the bank.

INTRODUCTION

Problem Statement

Credit risk management is crucial for banks to ensure stability and manage defaults effectively. This project focuses on developing a predictive model to assess the default risk of bank customers, which guides decisions on credit approvals. Accurate predictions are vital to maintain the bank's financial health and align with its risk management strategies.

The challenge lies in modelling the complex relationships between borrower attributes while addressing data imbalances—default instances are rarer than non-defaults. Two methodologies, Random Forest and Logistic Regression, were explored for their robustness and ability to handle diverse data. This report details the approach from data pre-processing to model evaluation, providing insights into the factors influencing credit risk to support informed decision-making.

METHODOLOGY

Data Pre-processing:

The initial step involved handling missing values and encoding categorical variables to prepare the dataset for modelling. Given the complexity of financial data, dimensionality was reduced by dropping less impactful features, determined through feature importance evaluation. This was crucial in mitigating overfitting and improving computational efficiency.

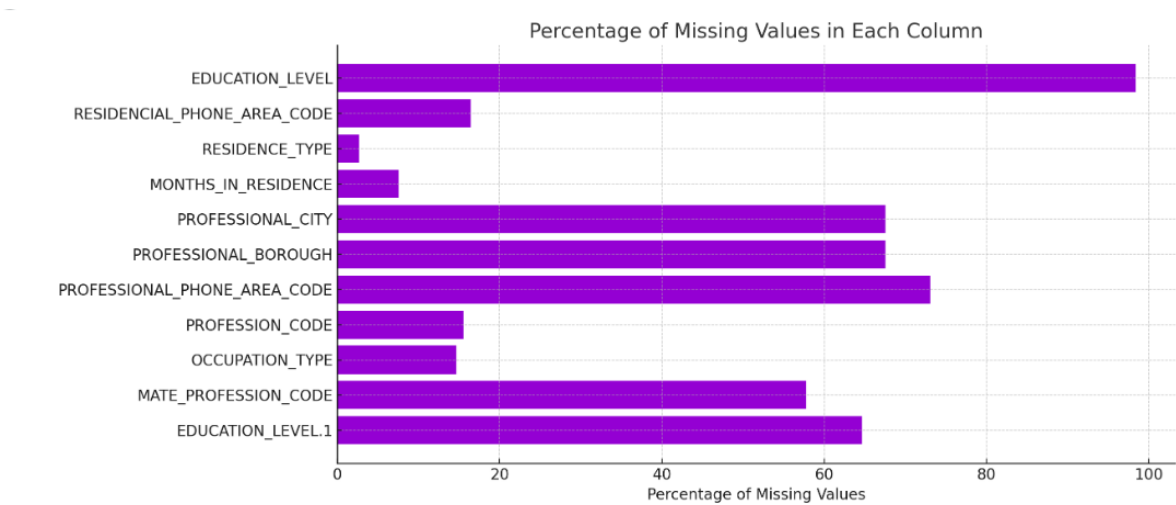
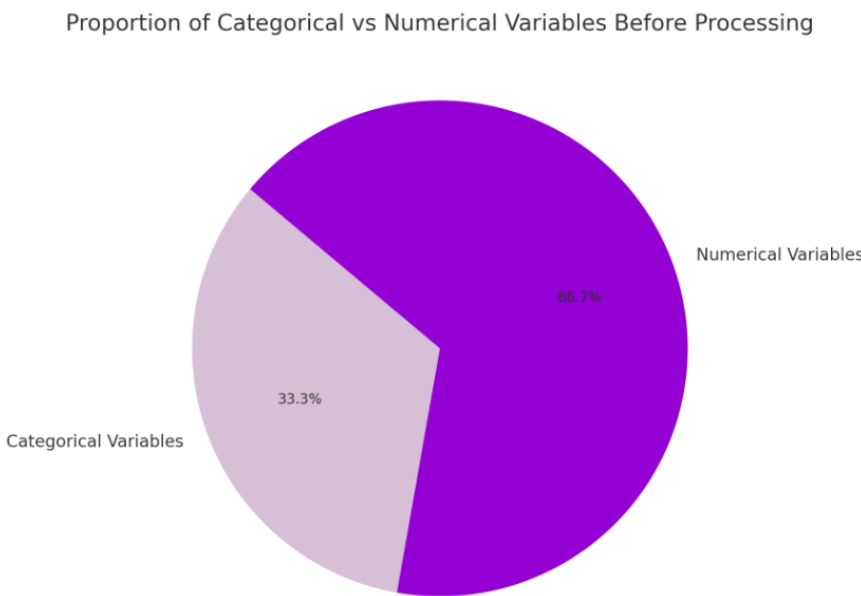


Figure 2: Columns with missing values before data pre-processing.



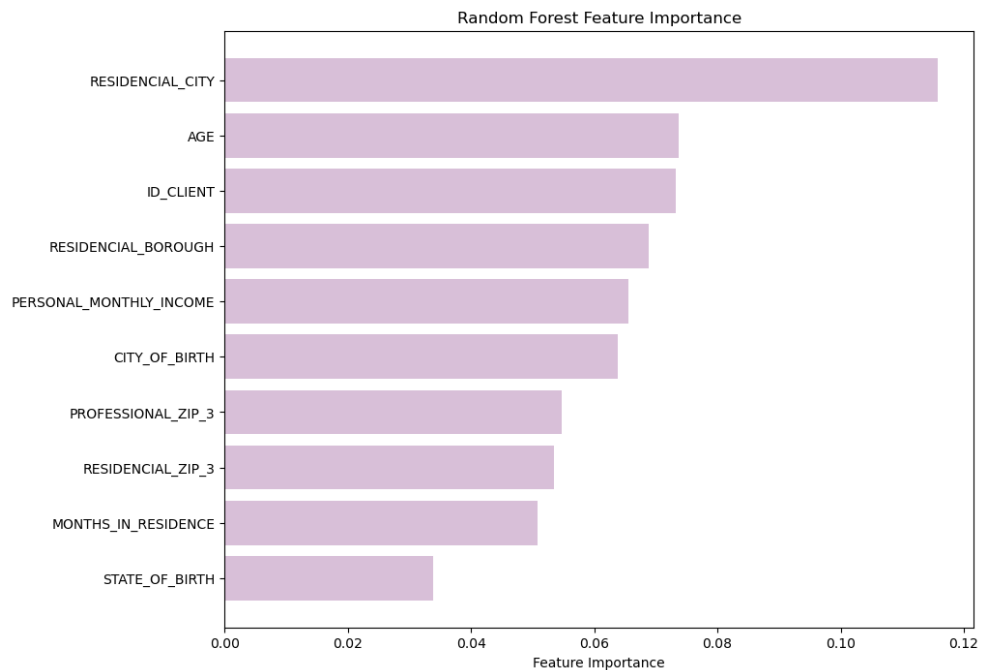


Figure 3: Top 10 predictors for the model.

Model Choice and Techniques:

- **Random Forest:** This method, effective in handling unbalanced datasets, constructs multiple decision trees and aggregates their predictions. It ranks the importance of various attributes, making it suitable for identifying key predictors of credit risk.
- **Logistic Regression:** Used for its simplicity and efficiency in estimating probabilities, making it effective for binary classification tasks like credit default prediction.

MODEL	PARAMETERS
Random Forest	n_estimators = 100, random_state= 42
Logistic Regression	Max_iter = 1000, random_state = 42
KEY	n_estimators -# Number of trees in the forest. random_state -# seed for the random number generator max_iter-# Maximum number of iterations for the solver to converge.

Advanced Techniques

- **Polynomial Features:** Used to identify how different features interact and affect credit risk, enhancing the model's input to possibly reveal complex patterns.
- **Class Imbalance Handling:** Techniques like SMOTE were used to help the model generalize better by ensuring it trained on a balanced set of data.
- **Threshold Adjustment:** Different decision thresholds were experimented with to refine the trade-off between precision and recall, which is essential in situations where both false positives and false negatives have substantial consequences.
- **Ensembling and Stacking:** Combining predictions from both models through ensembling methods provided a robust prediction by leveraging the strengths of each

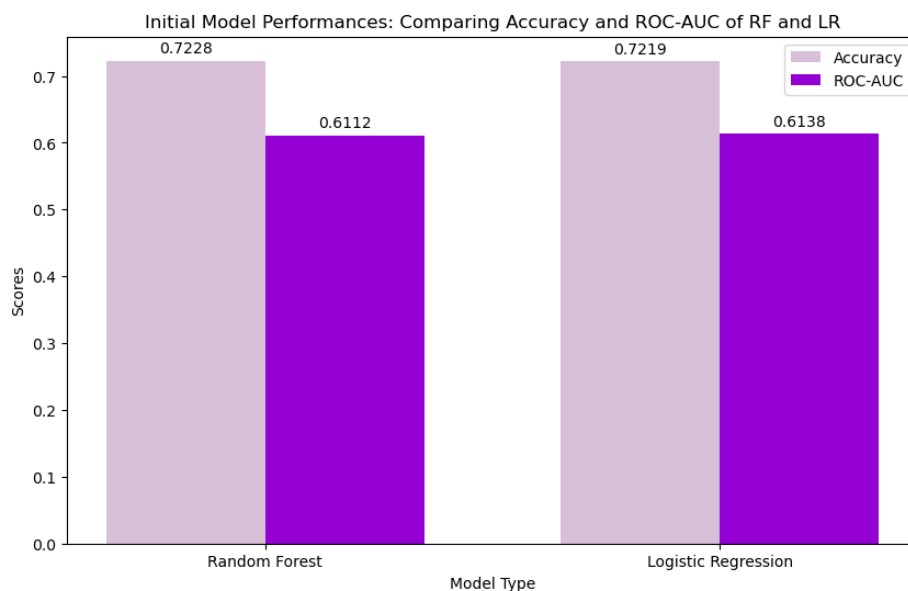
model. Stacking combined the predictions in a way that took into account the different biases and variances of each model.

These methodologies were integrated thoughtfully to build a predictive framework that aligns with the banking sector's requirements and adheres to practical constraints of model performance and interpretability. The approaches were tested to identify which provided the best accuracy and AUC, guiding the final selection of techniques.

RESULTS AND ANALYSIS

The evaluation of Random Forest (RF) and Logistic Regression (LR) models, alongside various enhancement techniques, highlighted the challenges and potential of each approach in predicting credit default.

The chart below gives us a visual summary of the initial performance metrics of both models.



Random Forest:

- **Initial Performance:** Achieved an accuracy of 0.7228 and ROC-AUC of 0.6112. It showed strong precision but weaker recall for the minority class (defaulters).
- **With Class Weight Adjustment:** Minor improvement in recall for defaulters but overall precision and recall remained low for this group.
- **Using SMOTE:** Improved balance between precision and recall, reflecting a better generalization but still presented limited effectiveness in identifying defaulters.
- **Threshold Adjustment:** Attempting to optimize the F-score resulted in no improvement, indicating challenges in balancing sensitivity and specificity.
- **GridSearchCV Optimization:** Improved model's recall significantly (0.6312) but reduced overall accuracy, showing a trade-off between general accuracy and class-specific identification.
- **With Polynomial Features:** There was no significant improvement, showing that adding complexity didn't reveal any new useful patterns for the model.

Logistic Regression:

- **Initial Performance:** Slightly lower accuracy (0.7219) and similar ROC-AUC (0.6138) compared to RF, struggling similarly with low recall for defaulters.
- **Blended Model Using Advanced Techniques:** Showed a slight increase in accuracy and ROC-AUC, yet the improvement in identifying defaulters was minimal.
- **Ensemble Techniques:** The ensemble approach does not substantially improve the overall performance concerning the minority class. It mostly retains the characteristics of the Random Forest model, likely due to the averaging method used.

The chart below shows the performance metrics of the ensemble model alongside other models and shows that there is no substantial improvement.

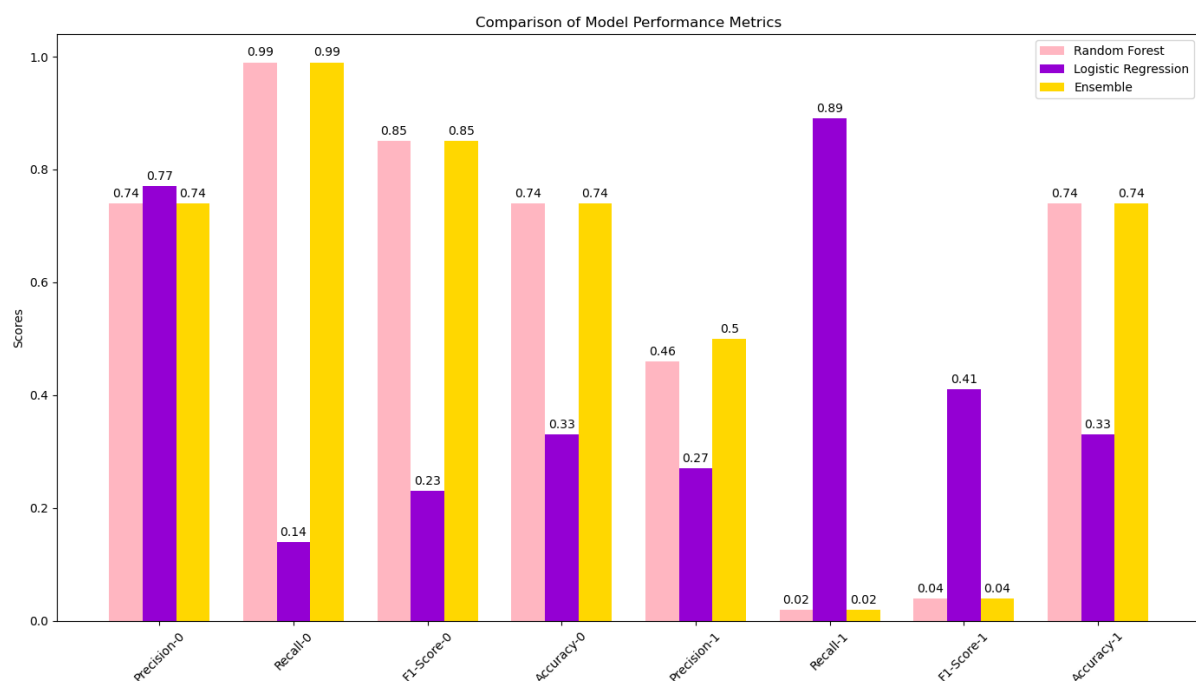


Figure 4: Comparison of Model Metrics with class weights adjusted and ensemble technique applied.

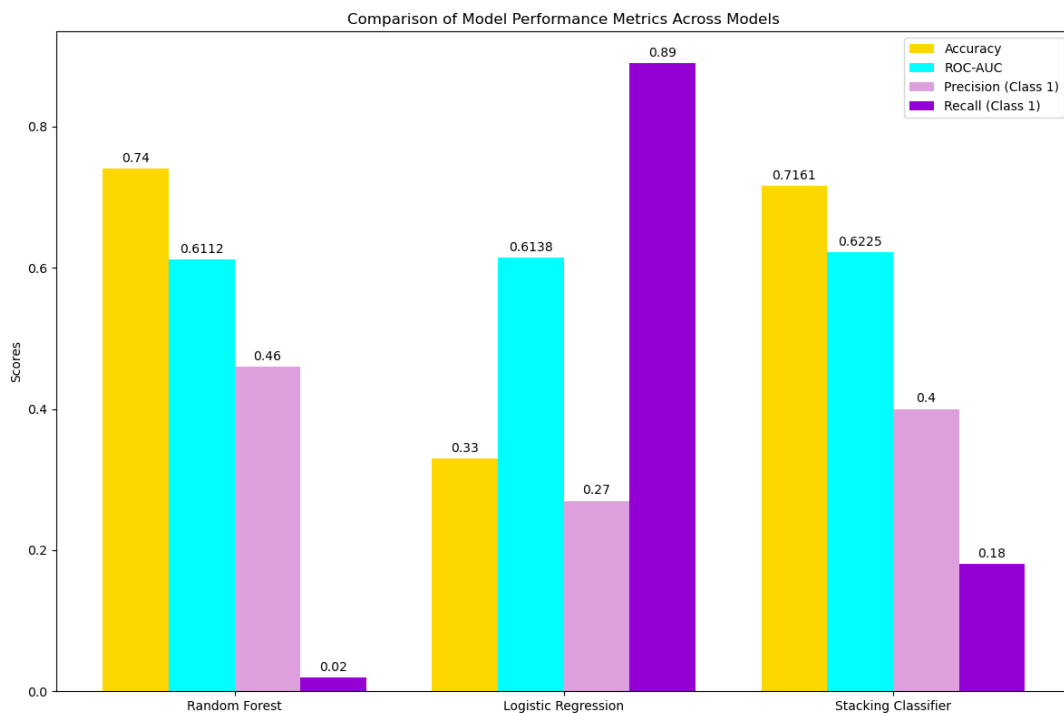
Key Observations:

- **Class Imbalance:** Continuously posed a significant challenge, as techniques like SMOTE, class weights, and threshold adjustments showed limited success.
- **Model Sensitivity:** Both models struggled to identify the minority class effectively. While RF showed a higher capability in handling mixed data types and capturing nonlinear relationships, LR provided faster computation and easier interpretation but with lower sensitivity to the minority class.
- **Technique Limitations:** Using advanced techniques like polynomial feature creation and adjusting thresholds didn't really change the results, suggesting that the features and model complexity were already fully utilized with the existing settings.
- After integrating both class balancing and feature engineering, the model exhibited a decrease in overall accuracy to 0.6673 and ROC-AUC to 0.5970. This suggests that

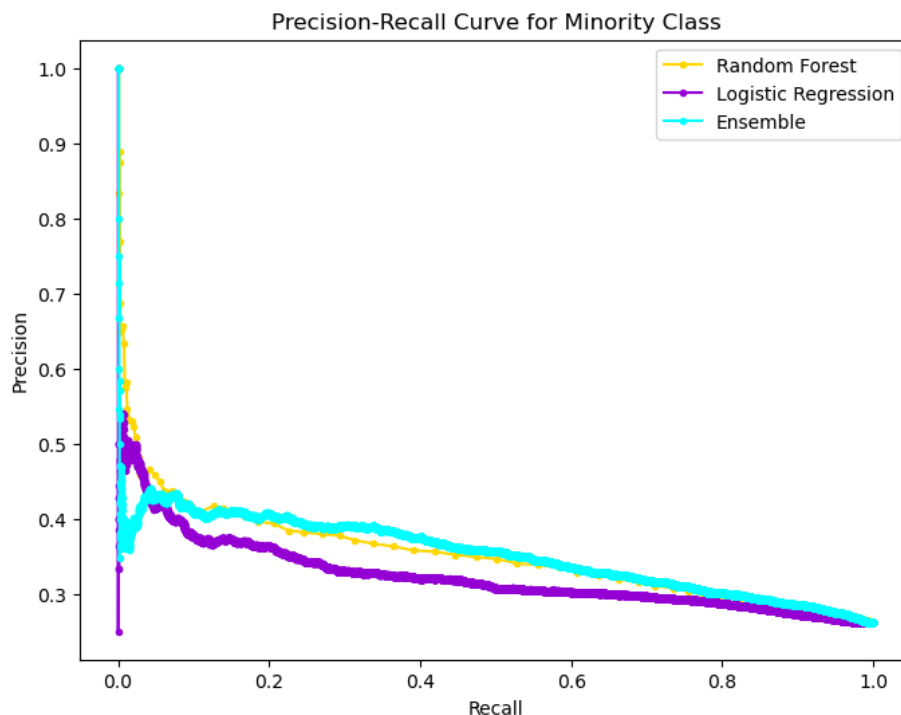
while addressing imbalance is crucial, it is equally important to refine the quality and interaction of features to maintain predictive performance.

- A notable trial involved re-configuring the stacking classifier with logistic regression as the meta-classifier, which led to a slight improvement in model sensitivity (recall) for the minority class, reflecting a more balanced prediction capability across classes.

The chart below shows the performance metrics of the stacking classifier alongside other models and helps us to visually assess the improvements and trade-offs mentioned above.



The curves below show the trade- off between precision and recall across all the models.



From the curves above:

- **Recall:** If identifying as many defaulters as possible is crucial (high recall), even at the cost of more false positives, Logistic Regression seems to perform better. The ensemble model also offers a balanced approach, which may be preferable depending on the costs associated with false positives.
- **Precision:** If it's more important to be correct when predicting defaulters (high precision), especially at lower recall levels, the Random Forest starts stronger but doesn't maintain its performance as recall increases.
- **Balance:** No model is the best across all thresholds. The choice for the bank will depend on the acceptable balance between missing some defaulters (lower recall) and incorrectly labelling non-defaulters as defaulters (lower precision).

The results underscore the complexity of credit risk modelling, especially in handling imbalanced datasets. Both models showed strengths in specific areas but also pointed to critical limitations in sensitivity to predicting defaults, which is pivotal in the credit risk context.

FURTHER ANALYSIS POSSIBILITIES

Exploring beyond the current scope of the study could provide deeper insights and potentially enhance the predictive power of the models used for credit risk assessment. Below are some avenues for future analysis.

- **Advanced Feature Engineering - Temporal Features:**

Incorporating time-based features allows models to capture the dynamic nature of credit risk, to understand how risk profiles change over time. This can provide crucial insights into the seasonal patterns in financial behaviours, offering a more nuanced understanding of risk.

- **Alternative Ensemble Techniques - Meta-modelling:**

Meta-modelling involves using the outputs of multiple predictive models as inputs into a secondary, overarching model. This technique is particularly effective in capturing the complex interactions between predictions from various models, enhancing the overall predictive power and reliability when assessing credit risk.

- **Advanced Machine Learning Techniques - Reinforcement Learning:**

Reinforcement learning represents a significant advancement in predictive modelling, as it enables models to learn and adapt from new data continually. This approach is especially advantageous in the fast-evolving domain of credit risk, where models can adjust their strategies based on real-time customer behaviour and external economic changes, leading to more accurate predictions.

These further analyses not only extend the predictive capabilities of the models but also open new dimensions of understanding that could be crucial for tailoring financial products and managing risks more effectively.

CONCLUSIONS AND RECOMMENDATIONS

This analysis applied Random Forest and Logistic Regression to predict customer default risk. While neither model achieved perfect performance, Random Forest demonstrated a stronger capacity for handling the imbalanced dataset and Logistic Regression provided valuable insights due to its interpretability, which is essential for decision-making processes in banking.

Key Recommendations:

- **Prioritize Random Forest** for its relatively better performance and capability to rank feature importance, aiding in more informed decision-making.
- **Employ Logistic Regression** where model transparency is paramount, facilitating straightforward explanations for credit decisions to stakeholders.
- **Continuous Model Refinement:** Given the fluidity of financial behaviour, it is imperative to keep models updated with the latest data trends.
- **Future Exploration:** Additional models, especially those tailored for imbalanced data, should be explored to potentially improve upon the results presented.

Limitations: It is crucial to acknowledge that both models encountered challenges with the imbalanced dataset, and the results may carry inherent biases. The metrics achieved indicate room for improvement and do not suggest that one model is categorically superior to the

other. Consequently, a strategy that involves periodic reassessment of the models with fresh data and continued refinement is recommended.

REFERENCES

1. Kuhn, M. and Johnson, K., *Applied Predictive Modeling*, 1st ed. [ebook] Publisher Name, year of publication, accessed 1 April - 21 April 2024.
2. Kuhn, M. and Johnson, K., 2019. Feature Engineering and Selection: A Practical Approach for Predictive Models. 1st ed. [ebook] Publisher Name, Available at: <http://www.feat.engineering/>, accessed 1 April – 21 April 2024
3. OpenAI, 2024. *ChatGPT*. Available at: <https://openai.com/chatgpt> [Accessed 27 March - 22 April 2024].
4. GitHub, 2024. *GitHub: Where the world builds software*. Available at: <https://github.com> [Accessed 27 March - 22 April 2024]
5. James, G., Witten, D., Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning*. [Accessed 27 March - 22 April 2024]

APPENDICES

Acknowledgments:

In the development of this project aimed to create a predictive model to assess credit approval for a bank, I utilized OpenAI's ChatGPT as a generative AI tool to assist in various stages of the project, including generating initial ideas, developing problem-solving strategies, and overcoming coding challenges. While ChatGPT offered valuable insights and suggestions, the final decisions and actual content of this report were my original and individual work.

LET'S TRY SVM AND ANN TAUGHT IN CLASS, LET'S GO BACK TO USING THE REDUCED DATA SET, OR SHOULD WE USE THE BALANCED DATA SET?



Data Analytics

For trying out SVM (Support Vector Machine) and ANN (Artificial Neural Network), it's generally a good practice to consider the nature of the data and the specific challenges you've faced so far, such as class imbalance and model performance with different features.

Data Set Choice:

- **Reduced Dataset:** This dataset has fewer features, potentially reducing complexity and computation time, especially beneficial for SVM, which can become quite resource-intensive with high-dimensional data. However, reduced feature space might miss out on capturing complex patterns, which could be crucial for model performance.
- **Balanced Dataset:** Using a dataset balanced through methods like SMOTE can help address class imbalance, improving performance metrics for the minority class. This is particularly useful for models like ANN, which can learn complex patterns but might also learn the bias towards the majority class present in imbalanced data.

Recommendation:

- **SVM:** Start with the reduced dataset to see how well SVM can classify with less computational burden. SVM is sensitive to feature scales, so make sure to scale your features before training.

Figure 5: Example of using generative AI, to help in decision making.



You

I have done random forest and XG boost but not logistic, should I initialize it first before we proceed. Remember I have a new dataset that was reduced



Data Analytics

Yes, before setting up the ensemble, you should ensure that the logistic regression model is appropriately initialized and trained with your current dataset. Given that logistic regression can be sensitive to the scale of input features, especially when features vary widely in magnitude, it's often beneficial to scale the data.

Here's how you can initialize and train the logistic regression model with your reduced dataset:

1. **Scale the Features:** Logistic regression benefits from feature scaling because it uses gradient descent as an optimization algorithm.
2. **Initialize the Model:** Set up the logistic regression model.
3. **Train the Model:** Fit the model to your training data.



Figure 6: Use of generative AI to develop strategies.



You

```
-----
ValueError                                Traceback (most recent call last)
Cell In[49], line 5
      2 import numpy as np
      4 # Predict probabilities
----> 5 probabilities = rf_classifier.predict_proba(X_test_reduced)[:, 1]
      7 # Calculate precision, recall, and thresholds
      8 precision, recall, thresholds = precision_recall_curve(y_test, probabilities)

File ~\anaconda3\Lib\site-packages\sklearn\ensemble\_forest.py:862, in
ForestClassifier.predict_proba(self, X)
    860 check_is_fitted(self)
    861 # Check data
--> 862 X = self._validate_X_predict(X)
    864 # Assign chunk of trees to jobs
    865 n_jobs, _, _ = _partition_estimators(self.n_estimators, self.n_jobs)

File ~\anaconda3\Lib\site-packages\sklearn\ensemble\_forest.py:602, in
BaseForest._validate_X_predict(self, X)
    599 """
    600 Validate X whenever one tries to predict, apply, predict_proba."""
    601 check_is_fitted(self)
--> 602 X = self._validate_data(X, dtype=DTYPE, accept_sparse="csr", reset=False)
    603 if issparse(X) and (X.indices.dtype != np.intc or X.indptr.dtype != np.intc):
```

Figure 7: Use of generative AI to troubleshoot errors.