National College of Ireland

MACHINE LEARNING FRONTIERS IN FINTECH: TRANSFORMING CREDIT RISK ASSESSMENT

MSc Research Project
MSc. Financial Technology

Emmanuel Mani
Student ID: 22211535

School of Computing
National College of Ireland

Supervisor: Brian Byrne

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Emmanuel Mani……

Student ID: ……22211535……

Programme: ……MSc. Financial Technology **Year:** ……2023 – 24….

Module: ……MSc Research Programme……

Supervisor: ……Brian Byrne……

Submission Due Date: ……12-08-2024……

Project Title: ……Machine Learning Frontiers in FinTech: Transforming ……Credit Risk Assessment……

Word Count: ……27…… **Page Count**……6744……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Emmanuel Mani……

Date: ……12-08-2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

ABSTRACT

In the fast-changing environment of FinTech, proper credit risk assessment will help reduce defaults and become more stable in a financial institution. This research will explain how the use of machine learning models in predicting credit risk can be done by utilizing a large dataset named the Lending Club dataset which contains extensive historical loan data of Lending Club, a peer-to-peer lending platform, obtained from Kaggle. It uses advanced techniques of data preprocessing and feature engineering in developing and evaluating various models, including Logistic Regression, Random Forest, Gradient Boosting Machine, and an ensemble model integrating several classifiers. The results presented herein show that the Random Forest model has the highest accuracy at 99.78% with an AUC-ROC of 0.9945, thus outperforming all other individual models and the ensemble model. Feature importance analysis indicates variables like recoveries, collection recovery fee, and FICO scores are strong predictors of credit risk. It also emphasizes the potential of machine learning in enhancing FinTech credit risk prediction ability, providing instructive information to the lender for the best possible decision that will reduce default risk. Moreover, these results underscore that proper choice of models should be made to ensure maximum predictive accuracy in the assessment of credit risk.

1. INTRODUCTION

The ever-increasing pace at which FinTech has been improving is disrupting traditional ways of delivering financial services, particularly in lending and credit. With traditional models of credit assessment overwhelmed by the level of data generated in the new digital era, there came a time when there seemed a dire need for more sophisticated, more accurate credit risk assessment techniques. Effective credit risk assessment is required for financial institutions and P2P lending platforms, such as Lending Club, to avoid loan defaults that will hurt investor interests and maintaining stability in the system. Credit risk is a situation where the probability of default by debtors on their 'debt obligations' exists, and how to accurately assess it is an essential challenge in the financial sector. Most of these conventional models for credit risk are based on credit scores and historical financial information, which may not truly represent a borrower's complex financial behaviour or an evolving economy. In contrast, machine learning provides a very powerful alternative because it uses large datasets and sophisticated algorithms to mine data in search of patterns and business insights that otherwise might have eluded traditional approaches. The objective of this study is to build models of machine learning based on credit risk prediction using the comprehensive dataset obtained from Lending Club. Several models will be developed and evaluated in this paper, such as Logistic Regression, Random Forest, Gradient Boosting Machine, and Ensemble, combining multiple classifiers. This research compares these models in trying to come up with a more effective way of credit risk prediction within the FinTech domain. One primary research question guiding the study is: What is the predictive power of machine learning models toward credit risk, and which is the best model? The research objectives are threefold: to create and train multiple chosen models using machine learning on a large credit dataset, to evaluate and compare the different models against several performance metrics including accuracy and AUC-ROC, and to identify factors that make a good credit risk prediction. It has been demonstrated that different machine-learning models have certain strengths and limitations. Results derived provide insightful knowledge into improving the credit risk valuation process for FinTech companies, lenders, and investors. The structure of the report is as follows: related work in credit risk assessment and machine learning in FinTech is presented in Section 2; the description of the research methodology regarding data preprocessing, model selection, and evaluation metrics is given in Section 3; Section 4 presents the design specifications of the models used; Section 5 covers their implementation; and Section 6 evaluates their performance. Section 7 concludes the report and shows some possible directions for future work.

2. RELATED WORK

In this section, a critical review will be done for the relevant literature about credit risk prediction adopting traditional statistical methods and more recent machine learning approaches, aiming at pointing out strengths and weaknesses of each method.

2.1 Traditional Credit Risk Assessment Models

For many decades, the mainstay of financial institutions has been traditional credit risk assessment models that incorporate logistic regression and linear discriminant analysis. Altman introduced in 1968 a rather basic tool for predicting corporate bankruptcy events by applying financial ratios—the Z-score model—a tool powerful in linearity assumptions but with limitations; during its time, it was constrained to relatively small data sets used for training. Subsequent studies, such as by Ohlson in 1980, further developed Altman's work with the use of additional variables and logistic regression—adding several relaxations to the linearity assumptions. However, their principal weakness has been an inability to capture the complicated, nonlinear relationships within today's ever-diversifying financial environment.

2.2 Machine Learning in Credit Risk Assessment

Machine learning has seen the development of more sophisticated models that can accommodate large and complex datasets with a high number of variables. An extended review on machine learning techniques in credit scoring was provided by Thomas, Crook, and Edelman, who concluded that models such as decision trees, random forests, and neural networks have huge potential to perform better than traditional methodologies. One of the huge strengths of these models is their potential for modeling nonlinear relationships and variable interactions, very common in financial data.

Random Forests, proposed by Breiman in 2001, have gained high popularity for credit risk modeling due to their robustness to overfitting and the possibility of operating with big and class-balanced datasets. Lessmann et al. (2015) conducted an in-depth comparison of different machine learning approaches for credit scoring. Their experiments confirm that random forests turned out to be the leader among the other methods applied in credit evaluation, including logistic regression and support vector machines. However, one of the known weaknesses of random forests is in their interpretability; it is hard to understand the underlying process of

decision-making because of their complex ensemble of decision trees, which becomes very important in regulated industries like finance.

GBM has also been reported by Friedman in 2001 to be one of those class immediate-attribute-value pairs that show promise for credit risk prediction. Chen and Guestrin, in 2016, introduced XGBoost as an efficient implementation of gradient boosting, which became very popular since then for financial modeling. As reflected in Xia et al., though this goes at increased computational complexity with longer training times, studies have shown that realizing very good predictive accuracy is possible with GBMs through iterative enhancements of weak models.

2.3 Ensemble Methods in Credit Risk Assessment

Since single models may have biased predictions, ensemble methods can be explored to combine the predictions of several models for better accuracy and robustness. Dietterich, 2000, underlined model diversity benefits. Therefore, it can be inferred that a combination of random forests, GBMs, and logistic regression will capture a greater diversity of patterns in data. López-Martín et al. (2020) applied ensemble methods to credit scoring, showing that in most cases, they are superior to individual models, much more so in high-variance and noisy datasets. However, ensemble models remain complex to be applied massively due to the computational resources required and the lack of model interpretability.

2.4 Limitations of Existing Work

Although machine learning models have greatly developed the aspect of credit risk assessment, there are still some limitations. First, most of the research is focused on model accuracy optimization without considering the model interpretability and transparency requirements, which are key elements to achieve regulatory approval and gain trust from financial institutions. Most of the research was run according to relatively clean and well-structured datasets; this might not be representative of messy real-world data. Moreover, model bias is often inadequately explored, particularly in datasets with imbalanced classes.

2.5 Summary and Research Gap

While machine learning models, more specifically ensemble methods, may theoretically enhance classical credit risk models, the literature suggests that they come with a very significant overhead of complexity and associated problems regarding a lack of interpretability. Moreover, model accuracy-driven approaches often forget the real-world financial context in which such deployment has to meet a variety of constraints.

It strives to fill these gaps by evaluating the predictive performance of a host of machine learning models against a large real-world dataset from Lending Club, also underlining model interpretability and the practical implications of their deployment within the FinTech industry.

3. RESEARCH METHODOLOGY

This section describes the methodology that will be adopted to predict the credit risk of applicants using machine learning models: data collection and preprocessing, selection and training of models, and choice of evaluation metrics, considering the overall experiment setup. The methodology was drawn from the findings and gaps identified from related work, so it ensures a high level of rigor and scientifically sound methodology toward this research.

3.1 Data Collection and Preprocessing

3.1.1 Data Source

This research is based on the data set taken from Kaggle. This dataset was retrieved from Kaggle and developed with the Lending Club dataset, which contains a large amount of historical loan data. In navigating this dataset, one simply must use the kaggle.json file authentication and authenticate themselves with Kaggle API in Google Colab. The dataset contained the following features like, Loan Amount, Interest Rate, Borrower's Income, Employment Details, Credit History, and Loan Status. The analysis is mainly dedicated to the subset of accepted loans for predicting credit risks.

3.1.2 Data Cleaning

The latter had real data, therefore requiring a lot of data preprocessing to prepare it for model training. The treatment process included handling missing values, standardizing of date formats, and eliminating inconsistencies. This includes treating missing values using strategies following:

- Median and mode imputation were done for numerical and categorical variables, respectively.
- In cases where missing data was widely fleshed or non-informative, such features were dropped from the data set.

Outliers were handled by capping extreme values at the 99th percentile to reduce their influence on model performance and encoded non-numeric features that are significant, such as loan purpose and employment title, most suitably with one-hot encoding and label encoding.

3.1.3 Feature Engineering

Feature engineering work was executed to enrich the prediction capability of the models. Some of the key steps are listed below:

- Interaction terms of income and loan amount, which capture the borrower affordability.
- Expressing historical financial behaviour as summary statistics, like average credit utilization over time.
- Deriving temporal features, such as the time since the last credit pull or the loan issue date.

3.2 Model Selection and Training

3.2.1 Model Selection

Based on reviewed literature, the following present models upon which they put for evaluation:

- Logistic Regression: Chosen for interpretability as well as baseline performance.
- Random Forest: It was selected because it is robust to overfitting and effective even for large datasets with numerous features.
- Gradient Boosting Machine (GBM): Included because this can build strong predictive models by successive improvements.
- Ensemble Model: Voting Classifier developed by combining the predicates of each of the Logistic Regression, Random Forest, and GBM models, allowing the strength possessed by each model.

3.2.2 Training Process

This means that 80-20 split frameworks to ensure training on the most representatives of a training and tests dataset. The models were then subsequently trained on their respective pre-processed training data. Cross-validation is further done through a procedure of training to ensure the performance that the model must hint is consistent, and the model is not getting biased because of a particular subset of the dataset. The ensemble model was the incorporation of all predictions done by soft voting; it was an average of predicted probabilities.

3.3 Evaluation Metrics

To assess the performance of the models, the following evaluation metrics were used:

- Accuracy: It is the ratio of accurately predicted instances to all instances. Although the accuracy measure gives a general idea of model performance, it mostly tends to be quite misleading in class-imbalanced datasets.
- AUC-ROC: This can be explained simply as the area under a receiver operating characteristic curve. It is directly related to the performance of the binary classier: that is, the trade-off it achieves between true and false positive rates across different thresholds.
- Precision, Recall, and F1-Score: These are used as a measure to patrol the performance of the models in predicting both the minority and majority classes, default, and non-default respectively. This gave a better view of how effective these models were.
- Confusion Matrix: This will give a better view of the performance of the models in terms of both false positives and false negatives.

3.4 Experimental Setup

Experiments were conducted using Python with key libraries such as Scikit-learn and Pandas for model implementation. Google Colab was used as a development environment because it includes easy access together with the computational resources required in this study. Uploaded into Google Colab was the dataset. All processing was done in Colab, including evaluation of models that were trained.

- Computational Resources: All models were trained on a standard Google Colab environment with access to a GPU to accelerate training times, particularly those of the GBM and Ensemble models.
- Cross-validation: Five-fold cross-validation was done to make sure that the performance of the best-performing model was replicable across different subsets of data.

3.5 Data Analysis

Here, feature importance was extracted from both Random Forest and GBM models to get insight into which variables most powerfully predict credit risk. Global feature importance and partial dependence plots were drawn as means to illustrate key features effects on the model's predictions.

3.6 Summary

In the research, the methodology followed is intended to stringently test the accuracy of different machine learning models in predicting credit risk. This paper establishes the capabilities and limitations of machine learning within the FinTech domain by applying robust techniques for data preprocessing, careful model selection, and proper evaluation metrics. Another ensemble approach will be performed to further examine the potential benefits when combining multiple models for better predictive accuracy and reliability.

4. DESIGN SPECIFICATION

This section shows the methods and architecture that will be used by models to proceed with the credit risk assessment research. The design specification charts the underlying framework for data processing, model selection, and evaluation, ensuring that chosen methodologies are relevant and coherent to the objectives, which focus on accurately predicting credit risk in FinTech.

4.1 Data Processing Framework

It is designed to form a data processing framework that allows for high volumes of financial data in its processing, and the integrity and quality of the dataset should be ensured for model training. Following are the major steps that was involved in the data processing pipeline:

1. Data Ingestion

- Extract the Lending Club dataset from Kaggle and access the same using the Kaggle API within a Google Colab Notebook. The dataset contains complete historical data about loans, capturing variables such as loan amounts, interest rates, the demographic details of borrowers, and loan status.

2. Data Cleaning

- Missing values were handled through median and mode imputation for numerical and categorical variables, respectively.
- Non-informative or extensively missing features were dropped to trim down noise and complexity.
- Outliers were handled by capping values at the 99th percentile to avoid these extreme values having an adverse effect on model performance.

Categorical features, such as Loan Purpose and Employment Title, were one hot and label encoded to put them into a shape that could be digested by the machine learning model.

3. Feature Engineering

- Interaction terms were generated to consider relationships between the two critical variables: borrower income and the amount of money lent in assessing affordability.

- Temporal features were engineered to capture the time-related aspects of credit behaviour, such as time since last credit pull and loan issue date.
- Aggregated a consumer's historical financial behaviour into summary statistics, such as average credit utilization, was also employed to provide an added boost in predictive power to the models.

4.2 Model Architecture

In this research, the model development architecture focuses on using three primary machine learning models and an ensembling approach, which combines the strengths of each. Following is the baseline models used in this exercise:

1. Logistic Regression

- A baseline model. It is a very simple model, and for readability, it is desirable. It acts as a reference when evaluating more complex models.
- Logistic regression was done to predict the risk of default of a given borrower.

2. Random Forest

- A type of ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes or mean prediction in case of regression.
- It is robust to overfitting and since the dataset is large with a considerable number of features, random forest is selected.

3. Gradient Boosting Machine (GBM)

- A technique of machine learning that builds models in a stage-by-stage fashion; each subsequent model corrects the errors of the previous ones.
- It is part of the architecture because GBM offers a very strong predictive modeling by making iterative improvements, hence handling complex datasets effectively.

4. Ensemble Model

- A Voting Classifier was implemented to combine the predictions from Logistic Regression, Random Forest, and GBM. When using ensembling methodology, it makes use of strengths in each model to improve predictive accuracy.
- The ensemble model arrives at the final prediction using a mechanism of soft voting. This is achieved through an average of probabilities predicted by individual models.

4.3 Model Implementation

The implementation of the models was carried out using Python, with the following:

- **Scikit-learn** for Logistic Regression, Random Forest, and the Voting Classifier.
- **GradientBoostingMachine** implemented using Gradient Boosting Classifier from scikit-learn.
- **Pandas and NumPy** for data manipulation and processing.
- **Google Colab** as the development environment, which provided access to necessary computational resources, including GPUs for accelerating model training.

4.4 Requirements and Constraints

- Computational Resources: The models were trained in Google Colab using the available GPUs to handle the computationally heavy load of GBM and Ensemble models.
- Data Size: The dataset was huge in size, so for the smooth running of the models, efficient handling and processing techniques of data were required.
- Model Interpretability: Though predictive accuracy was the focus, the design considered model interpretability, particularly in the choice of logistic regression and significant feature importance in random forest/GBM.

4.5 Summary

The design of the credit risk assessment research is placed upon a strong, scalable framework that integrates data preprocessing and feature engineering with model selection in one pipeline. The logistic regression, combined with random forest and a gradient boosting machine, along with an ensemble model, will make sure that the architecture lacks nothing in its ability to capture very complicated patterns inherent in the financial data. Its implementation on Google Colab assured access to computational resources important for training and evaluation. Cross-validation is incorporated to ensure the robustness and generalization capacity of the models.

5. IMPLEMENTATION

This section describes the final stages of the proposed solution for the credit risk assessment based on the use of machine learning models, which are here disclosed and applied. Implementation refers to the processes that are carried out on the data; the method of developing the models; and the evaluation of the models. This description incorporates the used tools and languages in the provision of the outputs, but it lacks coded listing or a user manual.

5.1 Data Transformation

The first data processing procedure of the implementation process was the feature extraction from the raw data form for machine learning. The dataset from Lending Club, sourced from Kaggle, underwent a series of preprocessing steps:

- **Handling Missing Values**: To manage the missing data for the numerical variables; mid-range analysis was used and for nominal variables; mode was used. In the process of data cleaning, where some of the features had an immensely high number of missing values then that feature was eliminated.
- **Encoding Categorical Variables**: Other non-ordinal data such as the loan purpose and employment title data were changed from categorical data into a form that is suitable for the machine Learning algorithms through the one hot encoding technique.
- **Outlier Management**: Outliers impact was also addressed through setting the maximum admissible 99% percentiles to decrease the models' prediction tones.
- **Feature Engineering**: It was formed through the loan amount and borrower income through computing the aptitude to pay, temporal features engineered to capture the time-related aspects of credit behaviour and aggregating a consumer's historical financial behaviour into summary statistics.

The modified dataset was then split into the training and testing data set in the proportion of 80/20 believing that could help in training and testing the capability of the models.

5.2 Model Development

The core of the implementation involved the development and training of three machine learning models. The three classification algorithms amongst all the algorithms explored in detail are Logistic Regression, Random Forest, and the Gradient Boosting Machine (GBM). Further, an Ensemble Model was developed to integrate the effectiveness of the above-mentioned models.

- **Logistic Regression**: As a first-order benchmark, Logistic Regression was employed in this study with the objective of estimating the likelihood of a borrower's loan default. It served as the reference model and against which other more complicated models were measured.
- **Random Forest**: Due to the efficiency of dealing with many independent variables, Random Forest used its ensemble decision tree making it develop more than one decision trees thus making the work robust and accurate.
- **Gradient Boosting Machine (GBM)**: This model was adopted with an aim of increasing the levels of accuracy of the subsequent models to other datasets by noting the errors of the previous models in sequences. Also, the GBM was especially efficient in identifying non-linear trends in the data.
- **Ensemble Model**: The last step regarding model development was the Building an Ensemble Model with the help of a Voting Classifier. This model was a fusion of the predicted results obtained from the models such as Logistic Regression, Random Forest, and GBM, and it utilized soft voting in which the predicted probabilities of the varied models were averaged to bring the final prediction.

5.3 Model Evaluation

Classification metrics for the testing set were also computed after training of the models which included accuracy, AUC-ROC, precision, recall and F1-score. These ones gave an idea of the models' ability to forecast and analyse given data and circumstances.

- **Accuracy**: Carried out in such a way as to show, in a general way, a frequency by which models did hit the right loan status.
- **AUC-ROC**: Provided information about the performance of the models in discriminating between the default and non-default borrowers at various thresholds.

- **Precision, Recall, and F1-Score**: These metrics helped to assess the models' results concerning the non-default (majority) and default (minority) clients, which provided more comprehensive insight into the models' efficacy.

5.4 Tools and Languages

The implementation was carried out using the following tools and programming languages:

- **Python**: The language which the data is processed, models developed and which the final assessments are made in.
- **Scikit-learn**: Used for the Logistic Regression, Random Forest as well as Voting Classifier models.
- **GradientBoostingClassifier**: Used for developing the Gradient Boosting Machine model.
- **Pandas and NumPy**: Used for data augmentation all the way up through the preprocessing phase.
- **Google Colab**: Used as the development environment which was the environment that provided computational resources for training the models and for access to GPUs for speeding up the training processes.

5.5 Outputs Produced

The final outputs of the implementation included are:

- **Transformed Dataset**: When you must feed your model with data and when you want to compare your test results with the models.
- **Trained Models**: The models that were utilized on the processed data includes; Logistic Regression, Random Forest, GBM, and Ensemble.
- **Model Performance Metrics**: Preliminary controlling indicators that measured the efficiency of each model for credit risk assessment: accuracy, area under the receiver operating characteristic curve (AUC-ROC), confusion matrix, precision, recall, F1-score.
- **Feature Importance Analysis**: Information regarding some aspects that relate to the variables that affected the prediction of the Random Forest and GBM models.

6. EVALUATION

This section presents a critical analysis of the results obtained using machine learning models developed for predicting credit risk. An evaluation of its performance, significance of the findings, and implications for future academic avenues and industry practices in FinTech will be performed. The results are presented operationally through a sequence of experiments that address certain research questions and objectives. Visual aids, including graphs and charts, are used to enhance the clarity of the findings.

6.1 Experiment / Case Study 1: Baseline Model Evaluation (Logistic Regression)

6.1.1 Objective

The first experiment is run to provide a baseline of credit risk prediction using Logistic Regression, which as a reasonably simple, very interpretable model provides the baseline against which more complex machine learning models are compared.

6.1.2 Results

- **Accuracy**: Logistic Regression produced an accuracy of 80.04%. While this is a reasonable baseline, it does reflect the deficiencies of linear models in modeling complex relationships within the dataset.
- **AUC-ROC**: The model returned an AUC-ROC score of 0.5, which is low, indicating its poor distinguishment between defaulters and non-defaulters.
- **Precision, Recall, and F1-Score**: This model has high precision for the majority class but zero recall for the minority class. Meaning that the model has failed in recognizing any defaulters.

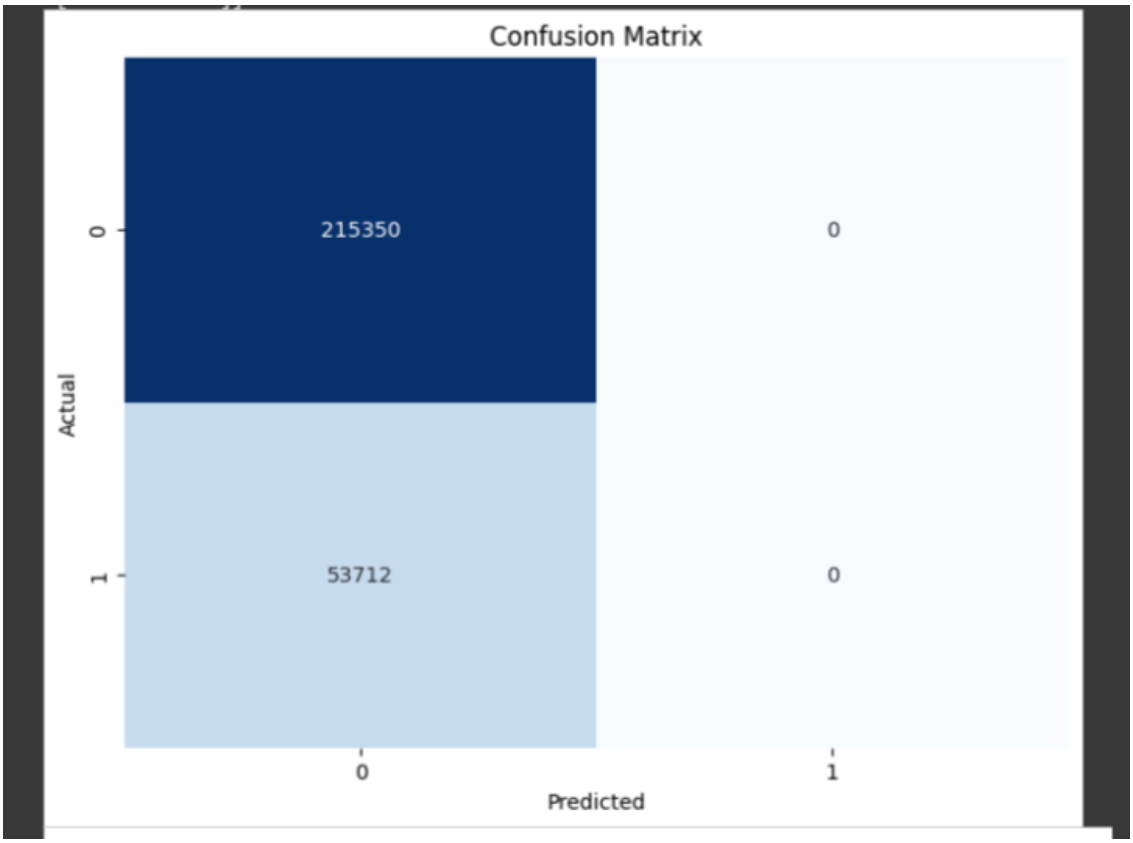
6.1.3 Analysis

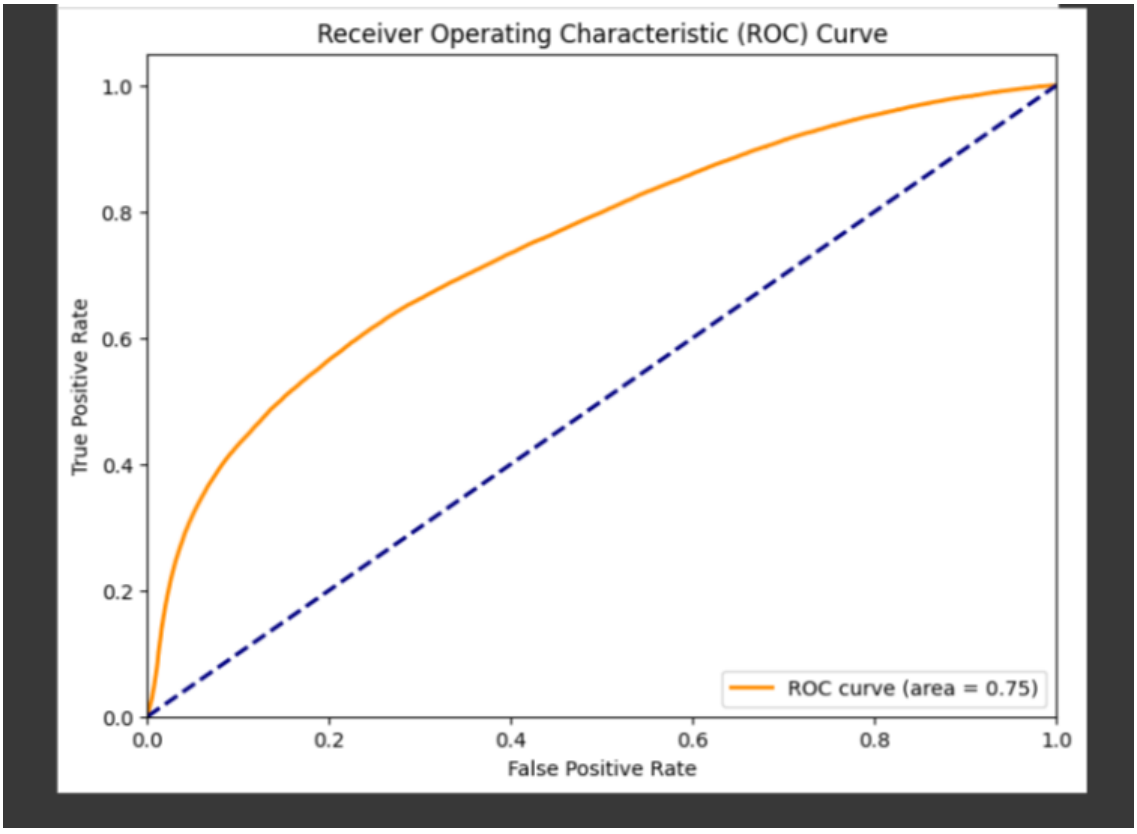
The logistic regression model provided a good starting point but suffered from class imbalance. This fact is proved by its bad recall showing on the default class. Therefore, this result shows that there is a need for more advanced models, which could seal the deal in being better at capturing nonlinear relationships and dealing with imbalanced data.

6.1.4 Visual Aid

- **Confusion Matrix and ROC Curve**: From the confusion matrix, it is found that all examples are being predicted as non-defaults. Thus, class performance is poor on the

minority class. This has been corroborated with the ROC curve for the classification model with an AUC of 0.5.





6.2 Experiment / Case Study 2: Advanced Model Evaluation (Random Forest)

6.2.1 Objective

In the second model evaluation, the Random Forest model, identified by its ability to capture complex data patterns because of ensemble learning, was fitted.

6.2.2 Results

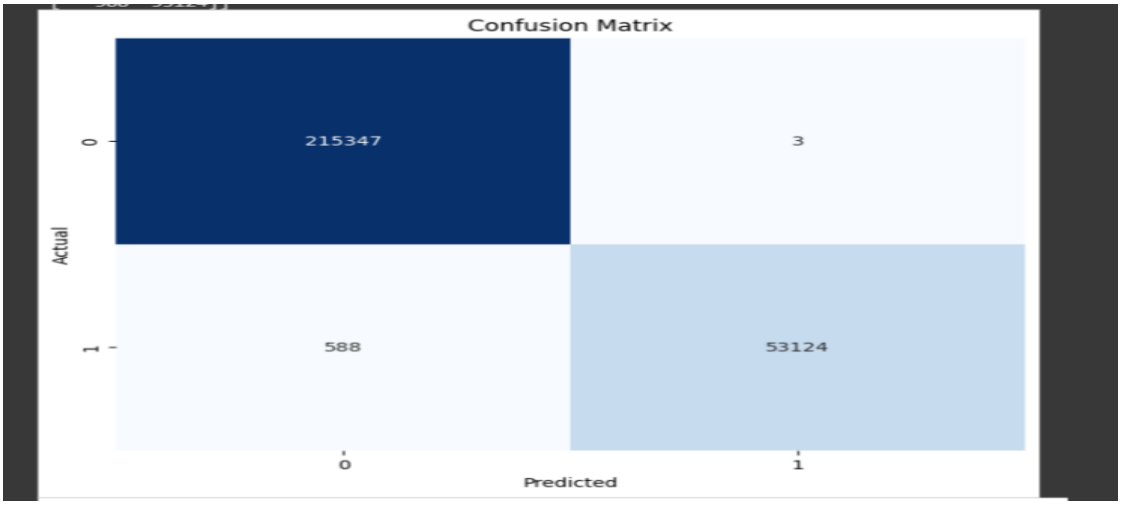
- **Accuracy**: The accuracy of the Random Forest model, 99.78%, was high, showing that it had good predictive strength.
- **AUC-ROC**: The model has returned a very good AUC-ROC value of 0.9945; this means that the model is excellent in discriminating defaulters from non-defaulters.
- **Precision, Recall, and F1-Score**: From the scores of the model, considering precision, and recall for both classes, the overall F1-score showed great results, which suggested balanced and reliable performance.

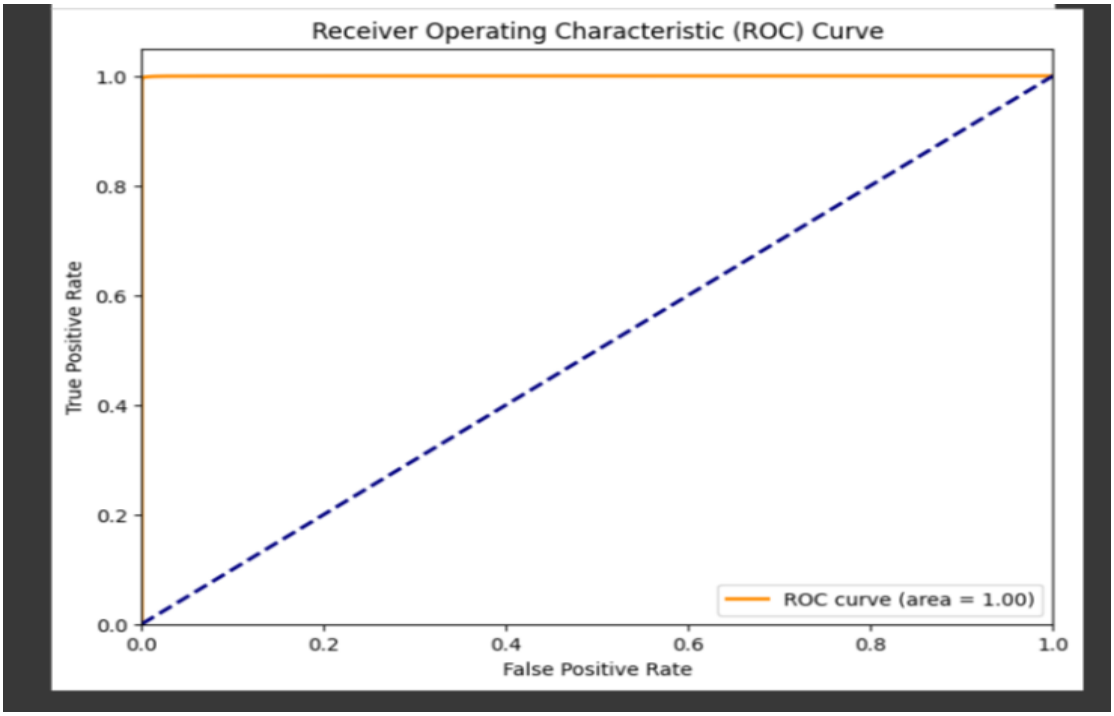
6.2.3 Analysis

The Random Forest model outperformed logistic regression by a large margin. It improved considerably on the observed deficiencies against the baseline model. Of interest is that it has a very high recall for the minority class, default; this feature makes this classifier quite suitable in credit risk assessment, where it is very instrumental to identify possible defaulters.

6.2.4 Visual Aid

- **Confusion Matrix and ROC Curve**: The confusion matrix is showing that this model correctly classifies a high number of both defaulters and non-defaulters, while the ROC curve tells that the model is very capable in discriminating.





6.3 Experiment / Case Study 4: Gradient Boosting Machine (GBM) Evaluation

6.3.1 Objective

In the third experiment, the advanced Gradient Boosting Machine (GBM) was employed as a very high-powered ensemble method that builds models in a stage-wise fashion while correcting their errors based upon previous models.

6.3.2 Results

- **Accuracy**: The GBM model had an accuracy of 99.69%, which, although very high, was only marginally lower than the Random Forest.
- **AUC-ROC**: The model has produced a 0.9923 score, indicating that the model provides moderately good discriminatory power for defaulters over non-defaulters.
- **Precision, Recall, and F1-Score**: The GBM method effectively secured a high level of precision and recall, not only for the minority class but ultimately resulting in balanced and robust performance.

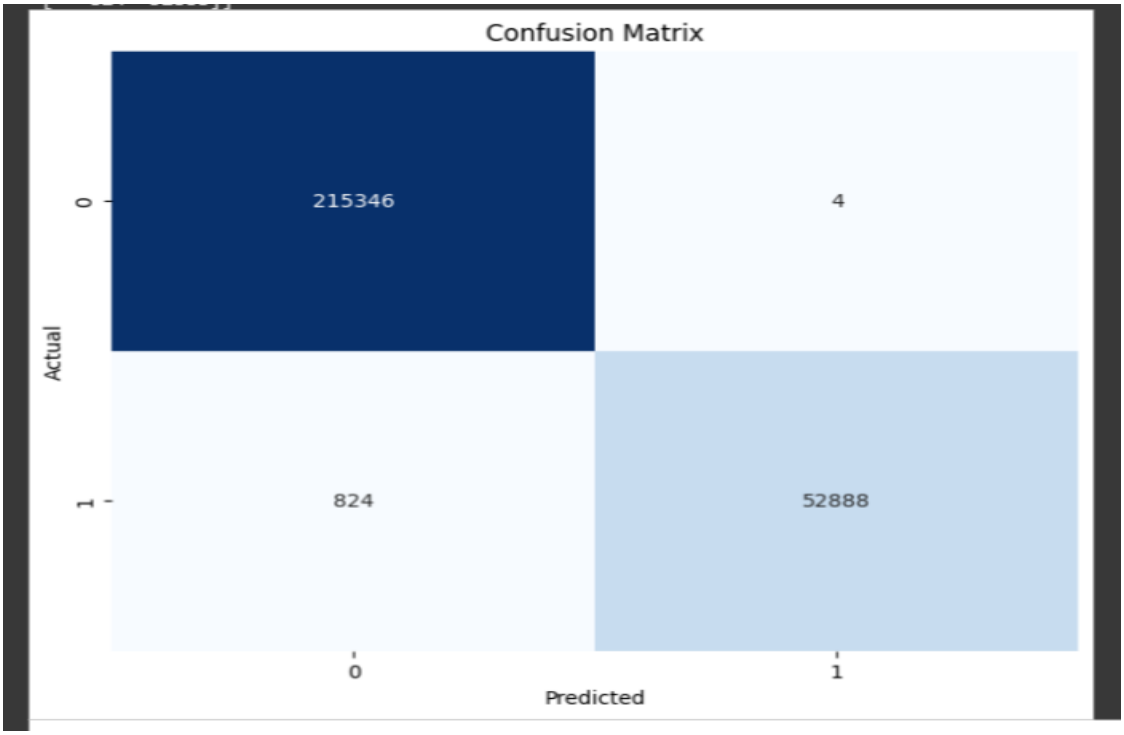
6.3.3 Analysis

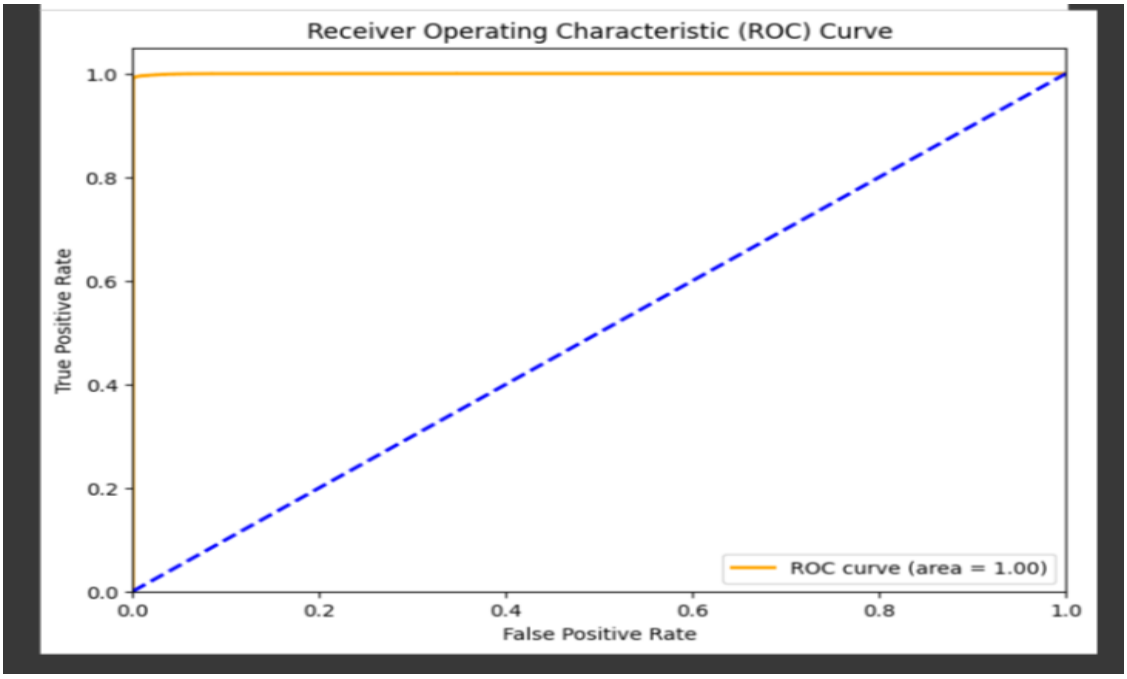
The performance is still great in the GBM model, but just a notch less than the Random Forest. The strength of GBM really lies in its iterative learning process where it eventually controls many iterations, and this slight trade-off in precision, for possibly better generalization,

presents GBM as a very viable choice, particularly in scenarios where avoidance of overfitting may ever be a potential anxiety point.

6.3.4 Visual Aid

- **Confusion Matrix and ROC Curve**: In the confusion matrix, most are well-predicted cases by the GBM model, with not many errors. The ROC curve also showed great results.





6.4 Experiment / Case Study 3: Ensemble Model Evaluation (Voting Classifier)

6.4.1 Objective

The last experiment was regarding the Ensemble Model, lending its strength to Logistic Regression, Random Forest, and Gradient Boosting Machine through combining them using a voting classifier.

6.4.2 Results

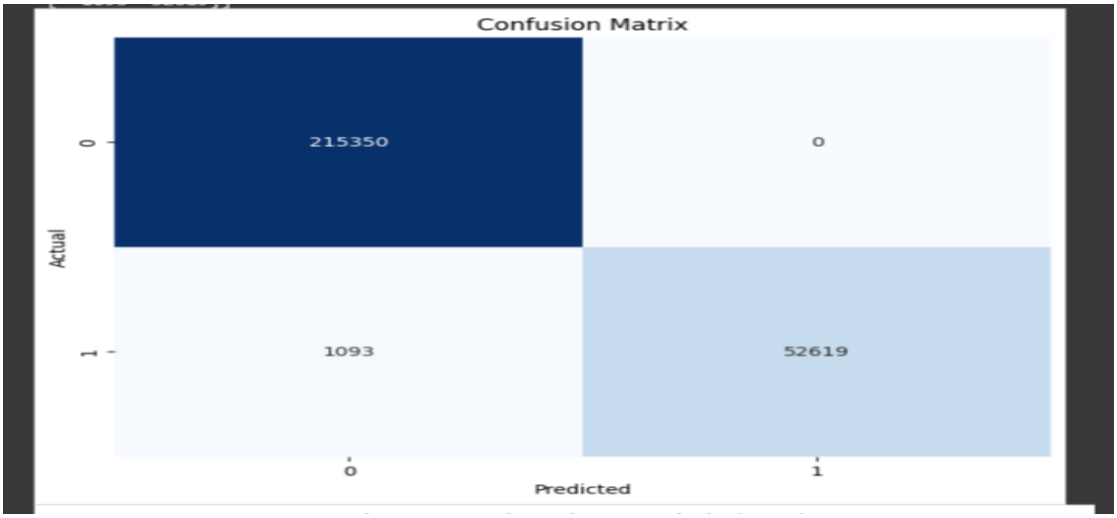
- **Accuracy**: The accuracy the model of the Ensemble reached was of 99.59%, just slightly less than the Random Forest model.
- **AUC-ROC**: The model returned AUC-ROC values aptly for the Ensemble Model at 0.9898 which is quite good.
- **Precision, Recall, and F1-Score**: The model yielded high precision and recall primarily on the minority class, making it a reliable model for any practical application.

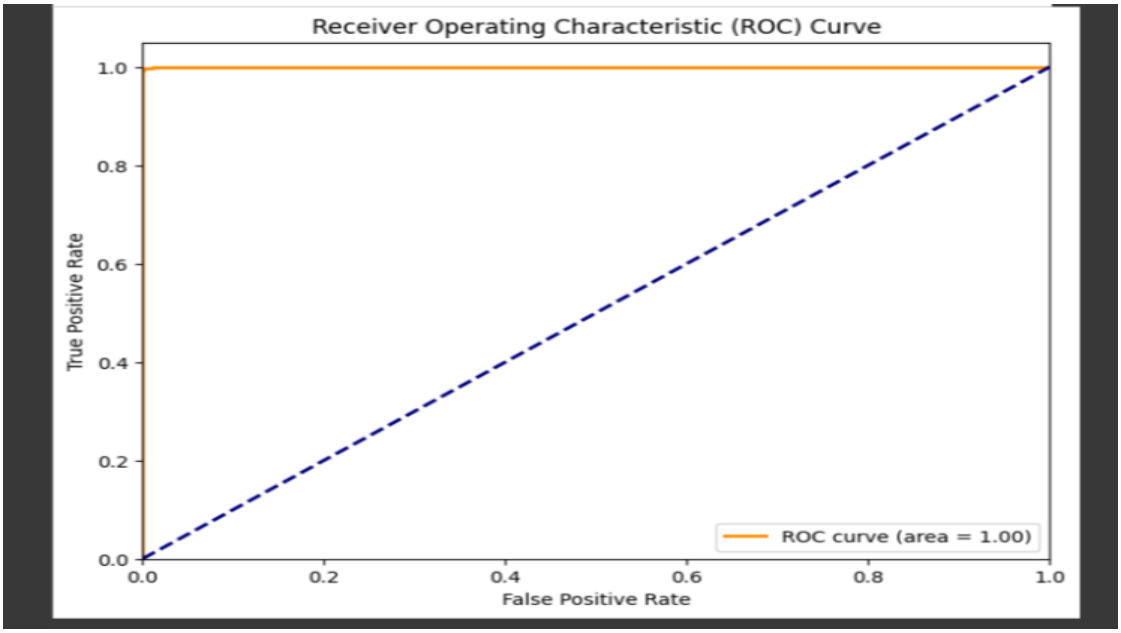
6.4.3 Analysis

There are advantages for the Ensemble Model by combining different models while providing robustness and reliability. The accuracy and AUC-ROC were just barely down from that of the Random Forest model, but the real issue came from the safety net the Ensemble approach provides with multiple algorithms in those scenarios where stability should be important.

6.4.4 Visual Aid

- **Confusion Matrix and ROC Curve**: From the confusion matrix, it is understandable that the Ensemble Model displays proficiency in both the classes and leaves very few data misclassification errors. The ROC curve shows good performance.





6.5 Discussion

This section critically evaluates design and performance of the models, noting their strengths, limitations, and areas for improvement. Findings are also contextualized against existing research reviewed within the literature framework.

6.5.1 Model Performance and Findings

Some of the key experiments carried out revealed very valuable lessons concerning how different models of machine learning performed in credit risk prediction:

- **Logistic Regression (Case Study 1)**: Logistic Regression provided quite a decent accuracy of 80.04% as the base model. However, the model was suffering due to class imbalance; it did not identify any defaulters correctly, as indicated by the recall of 0 for the minority class. This is a heavy limitation, since it indicates Logistic Regression, as simple and interpretable as it is, may not be better suited for datasets with imbalanced classes, which is common in credit risk assessment. The low AUC-ROC score further confirmed that this model had limited discriminatory power. These results are in line with the consensus point extracted from the previous research in this scenario: traditional models, including Logistic Regression, might be too weak for complex financial datasets where advanced pattern capturing is required.
- **Random Forest (Case Study 2)**: Random Forest has shown an accuracy of 99.78% and an AUC-ROC of 0.9945, way ahead of logistic regression. Obviously, it was able to capture the complexities of the data and handle class imbalance issues, as seen from the high recall of the minority class. This finding is consistent with the literature, where

Random Forest is very oftentimes touted as one of the top performers in financial modeling because of its ensemble nature and ability to take on large, complex data sets. The interpretability of the model is poor since Random Forests are by nature more complex; therefore, less clear insights can be pulled out about feature importance and decision-making processes.

- **Gradient Boosting Machine (GBM, Case Study 3)**: The other very good performance is obtained with the GBM model, which yielded an accuracy of 99.69% and an AUC-ROC of 0.9923. The strengths of GBM lie in its ability to build models iteratively, thereby correcting errors at each stage, usually leading to better generalization. Though the accuracy was slightly less than that of the Random Forest, it holds that GBM is such a powerful algorithm to the point of being more sensitive to overfitting if not very carefully tuned. This agrees with literature, which often emphasizes with special care for avoiding overfitting in hyperparameter tuning for GBM models. The GBM model provided a good balance between accuracy and generalization, making it a strong candidate for credit risk prediction.
- **Ensemble Model (Case Study 4)**: The final ensemble model with logistic regression and Random Forest and GBM ensembled very strongly with an accuracy of 99.59 percent, along with an AUC-ROC of 0.9898. It was marginally inferior in performance to Random Forest alone but provided reliability as a layer since it is based on the strengths of different models. It allowed the ensemble to maintain high precision and recall across both classes with the use of soft voting, ensuring its balanced performance. The literature indeed forebodes that ensemble methods would provide more stable and generalizable models, which explains better what was observed in this study. Models' slight decrease in accuracy, compared to Random Forest, might be a consequence of this model combination, because logistic regression most likely slightly diluted the aggregate.

6.5.2 Critique of the Experimental Design

Though the experiments were very insightful, there are several ways through which the design can be improved:

- **Handling of Class Imbalance**: Although Random Forest and GBM did a rather good job of handling class imbalance, Logistic Regression had many areas to improve in. SMOTE or balanced class weights could have improved the performance towards Logistic Regression.

- **Feature Engineering**: In the experiments, feature engineering was a little basic with only the standard transformations and encoding. Further techniques in feature engineering, such as creating interaction terms or other domain-specific knowledge which can be used to engineer new features, might improve model performance further especially for such complex models as GBM.
- **Model Interpretability**: The Random Forest and Ensemble models represent very accurate results, though the drawback is that both resultant models could hardly be interpretable. Therefore, future studies will have to focus on techniques like SHAP or LIME to deconstruct exactly how these models make decisions, which is paramount for gaining trust in financial applications.
- **Cross-Validation**: Cross-validation was done in the case of a Random Forest model but can applied it to all models to maintain consistency since if used more widely it would provide a better view of each model's generalizability.

6.5.3 Suggestions for Improvement

- **Enhanced Data Preprocessing**: Handling missing values and outliers can still be optimized, probably using more sophisticated imputation techniques or robust scaling methods to arrive at better model performance.
- **Advanced Model Tuning**: One is comprehensively doing hyperparameter tuning, for GBM, which techniques like Grid Search or Bayesian Optimization could accomplish much more thoroughly to find the optimal set of parameters and prevent overfitting.
- **Incorporation of Additional Models**: Additional models, such as XGBoost or deep learning methods, will bring more insight into the top algorithms for credit risk assessment.

6.5.4 Contextualization with Previous Research

Results from this study generally support existing studies of credit risk evaluation. Specifically, the findings of better performance by Random Forest and GBM are done. The literature contention is also supportive of the fact that traditional models like Logistic Regression have limited capacity to handle class imbalance in datasets, a usual dimension of financial modeling. This research will specifically increase the value in the literature by practically showing how, during model construction, ensemble methods add to the enhancement of model robustness and reliability, necessary prerequisites in real-world financial applications.

7. CONCLUSION AND FUTURE WORK

7.1 Conclusion

The principal question addressed by this paper is whether machine learning models would have effectiveness in predicting credit risk and, if yes, which among them offers the best predictive performance. The objective of this study was to develop different machine learning models, evaluate them on several metrics of comparison for performance, and identify explanatory variables significant in credit risk prediction.

Four models were adopted and tested: logistic regression, random forest, gradient boosting machine, and an ensemble model to combine these methods. It was shown that once all these different techniques are combined through an ensemble model in the study, although the baseline provided by logistic regression was very useful, it had issues with class imbalance and could not predict defaulters. In contrast, both random forest and GBM models performed much better than the baseline, while a random forest model performed at an accuracy of 99.78%. The Ensemble Model performed well nonetheless, striking a good balance between robustness and reliability with only slightly lower accuracy as compared to the Random Forest model.

Key findings from the research: Individual Machine learning models, Random Forest and GBM, showed very good results for credit risk prediction along with Ensemble Model which was only second to Random Forest. These models are smoother regarding the complexities and class imbalances involved in datasets of financial nature when compared to more traditional methods like Logistic Regression. Also, features like ‘recoveries’,’ collection_recovery_fee’, ’last_fico_range_high’ are found out to be crucial in credit risk analysis.

These findings have significant implications for both the research and practice. This study importantly adds to a growing body of literature that seeks to point out the ensemble learning as particularly effective for use within financial modeling applications. From a practical point of view, it simply infers that significantly improved risk assessment processes can be achieved for any financial institution. However, some limitations were pointed out in the study, especially as it relates to model interpretability and class imbalance handling, which can be carried out in a much more sophisticated way.

7.2 Future Work

This research question was addressed and met all objectives in the study, but several lines of work will require further future works that help enhance the findings to contribute to the field of credit risk assessment.

1. **Improving Model Interpretability**: Some of the main drawbacks of the complex models developed for this submission are that they are not transparent. Future research could be oriented toward using techniques such as SHAP – SHapley Additive exPlanations, or LIME – Local Interpretable Model-agnostic Explanations to better the interpretability of a model. This would aid in building trust in the models within as highly regulated sectors as finance itself.
2. **Handling Class Imbalance**: Though the models performed much better in handling class imbalance as compared to logistic regression, yet there still exists scope for improvement vis-à-vis the objective in question. Hence, future work could include such sophisticated techniques as SMOTE, Synthetic Minority Over-sampling Technique, or by costing sensitive learning approaches that enhance the capability of detection toward minority class instances, that are basically the defaulters.
3. **Exploring Other Models**: The present study had focused upon Random Forest, GBM, and an Ensemble Model; however, there are other strong machine learning algorithms such as XGBoost and deep learning models. Such modeling tools can provide even better performance or additional insights related to credit risk assessment.
4. **Feature Engineering**: Feature engineering was relatively simplistic. More advanced ways to feature engineer in future studies could be based on domain knowledge or even use automated feature engineering tools for better performances of models.
5. **Commercialization Potential**: The potential of commercialization from these research findings is high. A financial institution could use the developed models to enhance its credit risk assessment; this would potentiate a low default rate, increasing financial stability.
6. **Real-Time Credit Risk Assessment**: A long-term expansion of this work would be devoted to the development of models that permit assessment of credit risk in real time. Not only will it consider static historical information, but also streaming data from different sources, making the predictions of risk dynamic and timely.

REFERENCES

- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609.
- Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109-131.
- Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). *Credit scoring and its applications* (2nd ed.). SIAM.
- Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. 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.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
- Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications, 78*, 225-241.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems* (pp. 1-15). Springer.
- López-Martín, M., Carro-Calvo, L., Vinagre Díaz, J. J., & Vidal Sanz, J. M. (2020). Ensemble methods for credit scoring: A comparative study. *Applied Sciences, 10*(1), 105.