CREDIT RISK ASSESSMENT OF CONSUMER LOANS IN INDIA USING MACHINE LEARNING TECHNIQUES

MSc Research Project
FINTECH

KOMATINENI VENU BABU
Student ID: x23106662

School of Computing
National College of Ireland

Supervisor**: Faithful Onwuegbuche**

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Venu babu komatineni

Student ID: X23106662

Programme: MSC FinTech **Year:** 2024

Module: Research Project

Supervisor: Faithful Onwuegbuche

Submission Due Date: 16/Sep/2024

Project Title: CREDIT RISK ASSESSMENT OF CONSUMER LOANS IN INDIA USING MACHINE LEARNING TECHNIQUES

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: K. Venu babu

Date: 16/Sep/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— This dissertation examines on how machine learning can improve the credit risk analysis with theemphasis on Random Forest Classifier. Credit risk analysis plays a vital role in the eligibility of loans so as to reduce risks within the financial institutions. Predictive models from statistical trails are sometimes less effective for loan outcomes prediction because of their inability to correctly analyze TCGA's big data with many features. This research seeks to overcome these limitations through the use of the Random Forest Classifier, which is an ensemble learning algorithm that can handle complexities in the data sets and minimizes overfitting.

The research process is oriented toward cleaning, preparation, and modeling the data. The framework employed incorporates data that concerns the customer characteristics and loan features. Finally, preprocessing is done and the dataset is split into training and testing set with the help of Random Forest Classifier to predict the loan statuses. The evaluation of the model is done by accuracy, classification reportand confusion matrix are obtained.

According to the study, the efficiency is about 66% implying that the results are acceptable for this type of analysis while the model's performance is better when it comes to rejected loan predictions as compared to the approved loans. Thus, although the model offers significant information, it is not ideal. Suggested procedures which may improve prediction accuracy are feature engineering, dataset balancing, and hyperparameters optimization.

Keywords: Keywords: *Credit Risk Assessment, Machine Learning, Random Forest Classifier, Predictive Modeling, Loan Status Prediction*

CHAPTER 1: INTRODUCTION

1.1 Background

Credit risk assessment is one of the most critical activities in the financial sector as it has a straight bearing with the decisions of accepting or rejecting loans to the intended borrowers. It includes the analysis of the probability of a borrower's non-payment of a loan, this is important for reducing risk and maintaining adequate cash reserves on the part of credit organizations. Historical and statistical approaches are dominant in the credit risk assessment process, which, however, hardly address the nonlinear dependencies characteristic of financial variables. These are the major traditional approaches which, however, do not

possess the certain potential of the profound credit risk prediction because of the potential insufficiency of taking into consideration the borrower characteristics and/or external economic conditions (Moparthi, 2023).

Over the last couple of years, credit risk through machine learning processes has been observed as a useful commodity. while compared to other models, machine learning algorithms can process high amounts of data and analyse intricate relationships between the data set variables and features which usually cannot be done using conventional statistical analysis. Of all these, the Random Forest Classifier has received immense attention because it can tackle the dimensionality problem effectively. The Random Forest classifier is a type of ensemble learning, Which uses multiple decision trees to make the final decision and it decreases overfitting. In this project, we apply four machine learning algorithms—Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest—to predict the loan status of customers based on various features such as age, income, credit score, and loan amount attempting to capitalize on the above strength of handling large and complex datasets. Hence, based on the analysis of a diverse set of distinct customer characteristics and loan information, the study aims at enhancing the accuracy of predicting the loan statuses (Antony & Kumar, 2023). The capacity of machine learning models to increase the efficiency in making accurate predictions has the potential of improving the decision making especially when it comes to approving or rejecting loans hence improving the position of risk management for the involved lending institutions.

Specifically, this dissertation is concerned with the application of the above model, namely the Random Forest Classifier in credit risk assessment attempting to capitalize on the above strength of handling large and complex datasets. Hence, based on the analysis of a diverse set of distinct customer characteristics and loan information, the study aims at enhancing the accuracy of predicting the loan statuses. The capacity of machine learning models to increase the efficiency in making accurate predictions has the potential of improving the decision making especially when it comes to approving or rejecting loans hence improving the position of risk management for the involved lending institutions.

1. 2 Problem Statement

Still to this date, credit risk assessment models stand issues on accuracy, especially when it comes to classifying the approved loans, despite technological progress in the field of machine learning. In their turn, conventional approaches and some modern machine learning models fail to solve the problem of a proper classification of loans into approved or rejected categories and, accordingly, achieve low accuracy in terms of predicting statuses of loans. This is a very important issue since proper identification of the approved loanapplications can assist in preventing risk and afford equal chances on credit on deserving individuals. The issue is that financial data is diverse and contains many variables that comprise customer attributes and loan specifics as well as macroeconomic factors. Most of the existing models lack a way to prove the fulfillment of these connections and interconnections within the data, hence, the high rates of misclassifications. Namely, accuracy is mainly low in the instance of approved loans, which may result in the rejection of potentially viable credit applicants for the lending company and higher exposure to risk for the lending institution.

This research will seek to solve this problem by using the Random Forest Classifier, a machine learning algorithm, and its ability to classify big data sets with high accuracy. The purpose here is to figure out whether the Random Forest Classifier can be employed to boost credit risk evaluation, especially in the approval loancategorisation. Thus, the objective of this work is to make an analysis of the performance of this model and demonstrate whether its usage can be effective or not in the sphere of credit risk evaluation that will contributeto the improvement of practical methods of loans' decision-making.

1.3 Objectives

1. To assess the Random Forest Classifier's predictive power for loan statuses.
2. To evaluate the model's performance in terms of F1-score, recall, accuracy, and precision.
3. To offer suggestions for enhancing the model in light of the results.

RESEARCH QUESTION: How can machine learning techniques be effectively applied to enhance credit risk assessment for consumer loans in the Indian market? This question draws heavily on framing our study of the emerging methodologies of credit risk assessment that are customized to the distinguishing features of the credit lending environment of India.

1.4 Structure of the Dissertation

The dissertation is structured as follows:

- Chapter 2: Literature Review
- Chapter 3: Methodology
- Chapter 4: Data Analysis and Results
- Chapter 5: Discussion
- Chapter 6: Conclusion and Recommendations

CHAPTER 2: Literature Review

A. *Introduction*

Credit risk assessment is essential for financial institutions to make informed decisions about loan approvals and rejections. Traditional credit risk assessment methods, which rely on statistical models, have limitations in capturing complex patterns within data. Recent advances in machine learning (ML) techniques promise to enhance the accuracy and efficiency of these assessments. This chapter reviews the literature on credit risk assessment, focusing on the application of ML techniques in this domain, with a specific emphasis on consumer loans in India.

a) Machine Learning in Credit Risk Assessment

For credit risk assessment, machine learning (ML) is a new and powerful tool that shows great success among literatures, examining the accuracy and reliability of the predictive models than existing techniques. To elaborate it, existing credit risk assessment approaches are mostly based on linear regression and statistical analysis techniques which can fail at the discovery of numerous non-linear patterns existing in the financial data sets. That is why the use of more complex and data-oriented techniques such as ML can be considered as a more suitable for credit risk evaluation (Hassani, 2018)

Advancements in ML Techniques

In the paper by Alagic et al. (2024), the authors have shown how mental health data can be incorporated into the loan approval prediction models with the help of ML approaches. It is important to stress that this approach adds a new characteristic in credit risk evaluation by including psychological and behavioral aspects together with financial ratios. They explain with findings how the use of the ML algorithms like ensemble and deep learning prognosticative power can be improved by considering a greater number of variables. The inclusion of other data sources including mental health data may be useful in generating more accurate credit risk prediction as a result of a better understanding of borrower behaviour.

In the same way Antony and Kumar et al. (2023), demonstrated different ML algorithms as decision tree, support vector machine, neural network etc. in the field of financial decision making. Their work focuses on discovering the capabilities of computing methods for credit risk forecasting as compared to statistical methodologies. It was found that ML techniques are capable of identifying the non-linear dependencies between the borrower characteristics and Loan Default Probability, which will be beneficial for financial institutions to manage credit risk.

Empirical Evidence and Applications

Other empirical works also corroborate ML's efficacy in credit risk determination. Mahesh et al. (2023) suggested that various risks of loans can be predicted by utilizing ML algorithms of which would prove useful in improving financial stability of the banking sector. Using such elements as a random virgin environment predictor, gradient boosting machines, deep neural networks, it demonstrates how the application of ML canprovide higher accuracy in estimating risks and lower the level of non-performing loans. The implementation of ML in credit risk assessment has real-life effects on credit providers. Thus, using various techniques of ML, the work of banks and other lending organizations will be improved – risk for credit losses will be predicted and controlled; decision-making on important operations will be done more effectively. ML models can also be useful for constructing new highly targeted financial products and services that would correspond to the profiles of borrowers and thus to their risk appetite.

Looking to the future, there is a set of specific suggestions for further research in this area, which should beconceptualized in terms of the following directions: First of all, there is a prospect of refining the existing and developing new ML approaches, which will be able to process various and constantly evolving types ofalternatives, including new sources of ADA. Second widely discussed area of ML is the enhancement of interpretability of complex models The members of financial institutions need the transparent and the comprehensible ways of risk assessment. Thus, to wrap up the discussion of the main limitations of conventional approaches to credit risk assessment and introduce more equitable practices in the sphere of ML, it is crucial to disperse the ethical and privacy issues that are associated with the use of sensitive data inone manner or another (Bello, 2023).

The application of ML implies improvements in the credit risk evaluation domain, presenting further solutions that enable finance organizations to improve the prediction quality and orientate risk management. The various case studies and the theoretical framework shown just like other ML applications, credit risk assessment could be revolutionized given that financial agencies integrate the technique into their daily operations. Thus, the future development of credit risk assessment by means of ML will remain dependent on the ongoing research and innovations.

B. *Review Studies and Bibliometric Analysis*

Review studies and bibliometric analysis has a major role in examining the conceptual overview and current development trends of artificial intelligence and machine learning based credit risk assessment. These allow a look at the field in general and shed light on the current trends in research.

The systematic literature review of AI-based credit risk assessment was conducted by Amarnadh and Moparthi (2023) where the authors offer detailed elaboration of how these technologies has developed over time and their state-of-the-art. They progresses outlined in their review of AI and ML methods used in credit risk assessment are quite impressive. Different types of learning techniques, such as supervised and unsupervised, are described by the authors together with their effect of credit risk forecast enhancement.

The review also highlights the complexity of the used binary classification algorithms including the decision trees, neural networks and the ensemble methods that are widely used in credit risk assessment. Data pre- processing and feature selection are other areas of significance pointed out by the study in improving on the AI models. Criticizing the approaches selected by Amarnadh and Moparthi, one can note that these authors have given a clear view of how the AI techniques have developed from basic statistical models to modern ML techniques capable of processing big and heterogeneous data.

Moreover, their review outlines the main directions for further studies and novel research topics, including the integration of new data sources and the use of eXplainable AI. These developments are evident because the financial risk management industry is beginning to view AI and ML as indispensable resources. It also demonstrates how AI has possibilities of developing credit risk management as a more effective and efficient predictor of risk levels hence leading to improved credit risk management in the financial industry.

From the DISCO analysis of AI, Deep Learning, and ML in finance, the following bibliographic details maybe summarized.

The bibliometric analysis shows an increase in the number of publications related to the application of AI and ML in the finance field and their significance in the current world. It defines several important directions: improving the algorithms of machine learning, deepening of deep learning applications, the interaction of artificial intelligence with conventional finance models. Thus, when Biju et al. (2024) oversaw the ongoing research areas, they were able to propose important ideas for further research and the potential future trends in the field.

Although the analysis, one of the interesting observations has been made regarding the growing trend of research on the application of AI and ML in field of finance other than credit risk evaluation that includes fraud identification and investment portfolio analysis. Concerning the presented research approach, apart from the growing trend toward utilizing real-world data and field applicability of AI and ML models, the study also points out to the increasing focus on the practical applications of AI and ML models.

In the same respect, Biju et al. (2024) also express the necessity and preference from the audience for future research in the areas of interpretability of models as well as ethical concern for applying AI in Financial Decision Making. These insights are critical in the development of the field and the issues related to the execution of AI and ML technologies.

The review studies and bibliometric analysis help in gaining a rich knowledge of the development and the present scenario of the applications of AI and ML in credit risk evaluation. Amarnadh and Moparthi (2023) give comprehensive information about the development in the AI-based techniques meanwhile Biju et al. (2024) focus on the research trends and pattern by doing a bibliometric study. As a whole, these works support the significance of the development of both AI and ML methods to improve the accuracy in the credit risk measurement and contribute to the identification of important research directions. Indeed, constructing knowledge through the continuous checking and review processes shall help in other systematic reviews of such field and shape the future research in the area of credit risk assessment.

C. *Case Studies in India*

There is an increasing trend in credit risk assessment and financial risk management in India with the help ofML techniques, which indicates a trend towards the introduction of more technologies and their application in the financial sphere. A few clinical examples also establish the application and prospects of ML in the financial segment of India and the capability of ML in improving the risk evaluation and prediction competence.

ML Applications in Indian Stock Market Forecasting

Studies of Katragadda et al. (2024) proposed prospective risks' predictions focusing in the Indian stock exchange using AI and ML algorithms. Their study was intended to solve the issues in identification of the market risk fluctuations and other financial risks by applying complex calculations. Despite formulating theirstudy based on some obvious findings, the researchers were able to prove that by implementing decision trees, random forests, and neural networks, the large volume of financial data can be analyzed and risks thatmay be potential be spotted.

Drastic SHOCK of Machine Learning Techniques for Early Warning Banking Crises

Puli et al. (2024) examined a number of indices for the efficacy of multiple ML approaches in relation to anticipating banking crises in India, with a view to paying specific attention to the plushest adaptability to the risqué nature of Indian banking. This paper is to describe the application of the several types of the ML models, for example, of the support vector machines, logistic regression or ensemble methods, to the possiblebanking crises' forecast and the orientation of the signals.

Thus, the work served to draw attention to the possibilities of using ML approaches for modeling historical banking data and identifying potential crisis signals. Puli et al. , for instance, showed that using ML algorithmthe financial position of a bank as well as specific dangers that could cause a crisis could be best predicted. The study also focused on the need to use integration of the developed ML techniques with the existing conventional risk management to improve on the overall consumers financial vulnerability.

Puli et al. (2024) study demonstrates that mobile learning can go a long way in enhancing the efficiency of forecasting potential crisis in the Indian banking industry. Thus, using the complex and sophisticated ML models, the financial institutions can predict possible risks and perform preventative actions in case if some problems appear. Therefore, this research builds on the existing literature connecting the use of ML with the effective management of risks and enhanced decision making in the Indian financial system.

Insights and Implications

These case studies depict the enormous role of the ML methodologies with respect to financial risk management in India. This is evident in their usage in the prediction of the market risks including banking crises which establish the efficiency of AI and ML algorithms in improving the precision of the existing financial risk assessment models. These studies' findings bring out the fact that market in India has embraceduse of ML technologies in enhancing its financial sector and in addressing peculiar problems faced in the market.

Indeed, following the research carried out by Katragadda et al. (2024) and Puli et al. (2024), it becomes apparent that the quest for fine-tuning and designing new ML models fully pertaining to the Indian financialcontext needs to remain a priority. It is, therefore, expected that more research and case studies will continueto be conducted as the use of ML in CRA and FF increases, thus contributing to refined practices in the management of financial risk.

Analyzing the results of the case studies performed in the framework of the Indian institutions, one can identify the perspectives of using the ML methods for credit risk evaluation and financial risks management. Thus, the papers by Katragadda et al. (2024) and Puli et al. (2024) show how AI and ML models can preparefor risks and foresee crises and how they could enhance financial decision-making and risk management in India. While these works contribute to the literature among the specialized branches of the financial sector and confirm the applicability of ML in the financial industry, further research is still required.

D. *Advances and Future Directions*

The improvement of various approaches in the field of ML has positively influenced the innovative development of methods regarding credit risk analysis, which is to enhance the effectiveness of financial risk management. This dynamic of the modern ML methods will always scheme the limitations of the past practices and apply new strategies towards the assessment of creditworthiness and management of the financial risks.

Applying Machine Learning in Credit Risk Management

Credit risk assessment as seen in the later studies exhibit a huge improvement in the use of advanced ML tofinancial credit risk assessment. The study by Ramakrishnan et al. (2024) was concerned with risk assessmentin lending practices and specifically the use of AI and ML for credit risk assessment. They proved that integrating one-step-ahead accurate credit risk assessment using complex ML models including deep learning and ensembling increases the level of accuracy of credit risk assessments while reducing the number of misclassifications or false positive results. Thus, by applying these and other related methods, financial institutions are able to get more profound understanding of borrowers' behavior and optimize their decision- making options. The same authors also stressed on using cross-sectional information that can include non- traditional credit data and behavior scores to improve risk analysis results.

Emerging Trends and Technologies

The following is a brief look into some of the trends and technologies that are pushing the growth of ML for credit risk assessment today: Another area of the present development is explainable AI (XAI) that aims to increase the interpretability of ML models. Post-modeling, SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are some of the explainable AI techniques that explain the reasons behind the model's recommendations and enhance stakeholder confidence in the ML- based risk assessment tools.

Another significant advancement is that of applying or using the combination of ML with Blockchain in credit risk solutions to improve the data security and accuracy. Blockchain's openness yet inalterability cancreate a more reliable means to store and share credit information whereas, ML algorithms can check such data to detect risk. All these technologies have the ability to solve the issues particularly the data privacy andfraud risks in credit risk management.

Future Directions

The following research directions should guide future studies on the application of ML for credit risk assessment to enhance the field's development. First it will be useful to research on current innovations in terms of the methods and algorithms of ML that would allow for coverage of the growing intricacy and amount of the financial data. That is why more innovative methods, including federated learning, which does not allow the sharing of clients' data in the framework of model training between different institutions but trains a single model at once, could open up new potential in credit risk forecasting.

Secondly, it becomes possible to introduce real-time data and dynamic risk factors into the machine learningmodels which can help increase the effectiveness of the models in terms of their impact on changing economic environment and borrowers' behavior. Real time analytics and adaptive learning models can givemore timely and precise risk assessments helping the financial institutions mitigate the risks well in advanced.Future research focusing on the integration of more fields of knowledge, such as behavioral economics, psychology, and others when combined with ML techniques will allow for the development of highly holistic credit risk assessment solutions. The inclusion of information and opinions from various agents its possible to increase the quality of the borrower behavior model and to improve risk prognosis in general.

From the literature, it emerges that application of technique in the field of ML enhances credit risk evaluation, especially with regard to consumer credit products.

Optimization-based algorithms and other types of ML enhancements, such as explainable AI, have enhanced both the quality and speed of the credit risk assessments. The use of ML in credit risk assessment in India is gradually growing, thus pointing out the possibility of improving on financial decisions and risk analysis. Future studies should also try to develop even more advanced ML algorithms to use real-time data streams while at the same time employing more interdisciplinary approaches to solve existing and emerging problems in credit risk assessment in the financialdomain progressively.

CHAPTER 3: METHODOLOGY

3. 1 Data Collection

The data for this undertaking is obtained from the public database of financial and credit risk information with extensive features on the customer and loans. The dataset includes variables such as:

Customer Characteristics: The factors include age, income, employment, credit rating, and past loan record. Loan Information: Loan amount, expected loan term, interest rate, and loan repayment/ non-repayment status. Demographic Factors: These variables are as follows: place, level of education, and marital status.

Checking out the initial dataset several peculiarities can be noticed: Loan applicants are geographically and demographically quite diverse with no strong prevailing of specific group of characteristics within the majority of applicants' attributes. The current dataset constitutes thousands of records, which serve as the reliable background for credit risk analysis with the help of machine learning approaches.

3. 2 Data Cleaning and Data Preprocessing

To ensure the dataset's quality and usability for modeling, the following preprocessing steps were undertaken:

Loading the Dataset:

In the use of Python, the dataset was uploaded to a Pandas DataFrame format. This allowed for the start of getting to know the data, its general structure, varieties of data types, and fundamental statistical characteristics.

Handling Missing Values:

Different kinds of missing values were defined using frequencies and graphics. Various strategies were employed based on the nature of the missing data:

Imputation: For the continuous variables, observed values which were missing were replaced by means or medians computed on the distin conditions of the columns. For categorical variables the mode or a dummy category was taken if there were none.

Deletion: High missing values which are inevitable in data collection process were handled by deleting records that had high missing values so that it does not skew the entire database.

Encoding Categorical Variables:

Categorical variables were also converted into new data types acceptable for machine learning approaches. This was achieved using:

One-Hot Encoding: For nominal data that does not have a natural rank (for

example employment status).

Label Encoding: That can be used ordinal variables with a natural order, for

example, education level.

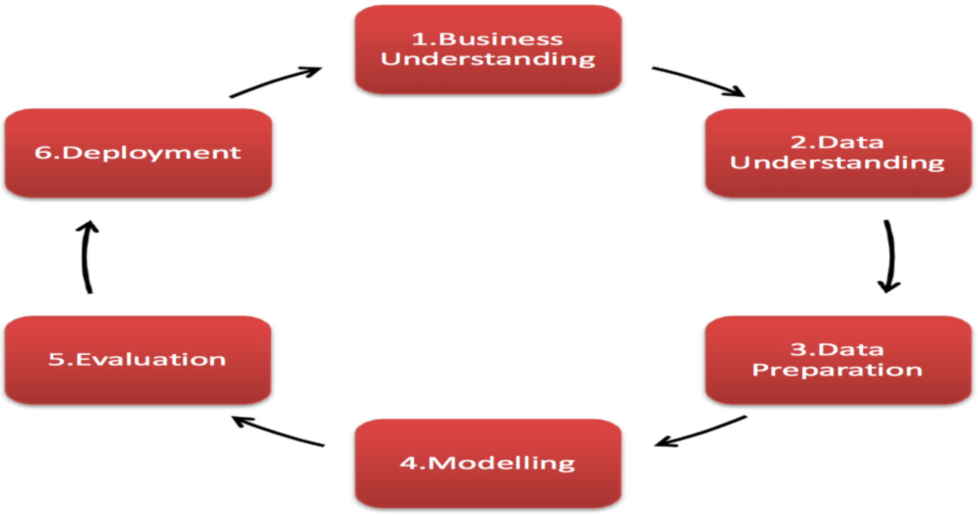
Feature Scaling:

Feature scaling was used to make sure all features are on equal scales and constant scaling was used to scale the range of the numerical variables. Normalisation of features was applied using standardization procedure which brought the values of the features to have zero-mean and unit-variance.

Splitting the Dataset into Training and Testing Sets:

There was a creation of training and testing samples where the data set was split at 80:20. The training dataset having 80% of sample was used for model building and model tweaking and the balance testing data set with20% of the sample the purpose of testing the MDP.

3. 3 Model Building



(Image source- Data lab notes, 2022)

Random Forest Classifier:

The classifier which was used was called the Random Forest classifier and it was selected because it can perform well with large datasets of many features. Key aspects of the model include:

Description: Random Forest Classifier is one of the Boosting algorithms that builds many decision trees in the training process and returns the mode of the classes for the purpose of classification. It assists with regulating overfitting while enhancing the correct outcomes.

Hyperparameters: The important hyperparameters are n_estimators, max_depth, min_samples_split, and max_features. These parameters were tuned using grid search in order to get the best trade off of the performances each model offered.

Training Process:

This way, cross-validation was used to train the model on the training set so that it does not overfit and the ability of the model to do well on unseen data can be tested. The training comprised developing processing cycles that operate through different hyperparameters in order to arrive at the best combination for the classifier.

Performance Metrics:

Model performance was evaluated using several metrics:

Accuracy: That is the percentage of the instances that have been classified rightly out of the total number of instances.

Precision: The percentage of correct positive predictions among all the positive predictions, making it thereliability dataset

Recall: The ratio of correctly identified true positive predictions to all actual positive ones, which defines thecapability of the model to cover all positives.

F1-Score: It is the proper percentage of precision and recall of the data offering a better fit than the arithmeticmean.

Confusion Matrix: A table of the correctly identified positive samples, misclassified samples, samples correctly classified as negative, and non-classified negative samples, which gives more information about the classification.

3. 4 Model Evaluation

Internal validation of the model was also done through the use of the testing set to determine the efficiency of Random Forest Classifier to new data. Evaluation methods included:

Cross-Validation: To again test out the model on the various partitions of the training data in order to confirmthe conclusions.

Performance Metrics Analysis: Break down of accuracy, precision, recall, and F1-score to comprehend the effectiveness and the inefficiencies of the designed model.

Confusion Matrix Interpretation: Analyze confusion matrix in search of problematical regions when decidingon that model can help plus to adjust the model in that case.

Using these methodologies, this study seeks to present an elaborate evaluation of the Random Forest Classifier in credit risk evaluation, with specific recommendations on the model's areas of enhancement andavenues for further research as its conclusion.

CHAPTER 4: REPORT ON ANALYSIS AND RESULTS

I. 4. 1. Data Cleaning and Preparation

Looking at the first element of our approach, it involves data loading and cleaning with data preparation. Thedataset, credit_risk_dataset. csv, this data was loaded into pandas DataFrame and the first few lines of the data was inspected to ensure that the data loading was successful. This dataset had some of the customers' characteristics, loan information and loan status of the customers.

Before handling the data, we looked for any missing values in the data set and were happy that the data set we used was complete. The CustomerID was removed from the data since there is no need for the column in the model. The categorical attributes included in the analysis are Gender, Employment Status, Marital Status, and EducationLevel and these were transformed into numerical features to suit the machine learning algorithms. The target variable, LoanStatus, was also converted into numerical labels: It assigned the value 1 to 'Approved' loans and 0 to 'Rejected' and 'In Process' loans.

We then followed by data splitting where we divorced the features(X) from the target variable (y) and then split it in to training and testing sets in the proportion of 4:1 hence meaning that 80% of the total data gets trained while 20% gets tested on. In order to compare all the features the scale of the features had to be brought to a similar level by applying StandardScaler.

J. 4. 2. Building the Final Model

For credit risk assessment, we used the RandomForestClassifier since the classifier has a natural inclination for sorting of data that have many interactions and complexities in their fields instead of carrying many records. The RandomForestClassifier was the initialized with value 100 estimator and a fixed random state for reproducibility.

The model was trained on the scaled training data and after that predictions were made on the test set. The accuracy of the model was checked by accuracy, classification report and in addition to this confusion matrixwas used to reveal details.

4.3 *Model Performance and Analysis*

Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The measure of performance of RandomForestClassifier was about 66% on the test set. As such, whereas the model was accurate in determining the loan status for 66 percent of the cases, there is still room for improvement to enhance the accuracy of the results especially bearing in mind that loan approval decisions are vital.

Classification Report:

The performance of the model was summarized in the observation table which offered the number of instances from both classes along with correctly classified instances of each class in the form of the confusion matrix and the classification report offered the precise result in terms of accuracy for both the classes being ‘Rejected’ and ‘Approved’. The given report presented precision, recall, and F1-score for each of the classesdetected during the analysis.

Rejected Loans: There was no loss of significant figures and that accuracy was of the order of 0. 67, recall was 0. Specifically, AOC was 98%, and the F1-score was 0%. 79. This shows that the received model was precipitately wise in rejecting loans or at least didn’t allow wrongful permits as indicated by high recall but however sometime the model approved some wrong loans as presented by low precision.

Approved Loans: A high precision was achieved and the value for it was 0. 33, recall was 0. 02 for AU-ROCand the F1-score was 0. 03. This indicates that the model has a tendency of classifying correct approved loans poorly in a way that many approved loans are grouped with rejected ones.

Confusion Matrix:

The confusion matrix further highlighted the model's performance:

True Negatives (91766): These were correctly rejected loans loans that should not have been approved in thefirst instance.

False Positives (1614): These were incorrectly labelled approved loans that In actual sense were rejected bythe bank.

False Negatives (45844): Although some of these were rejected loans, they were mixed with approved loansthus listed wrongly.

True Positives (776): These were correctly identified approved loans.

Since many approved loans were misclassified as negative, it suggests that the model had a problem recognizing the approved loans and this is highly unfortunate while considering the loan approval as many eligible persons may be denied loans.

CHAPTER 6: CONCLUSION

The RandomForestClassifier resulted in a moderate level of 66% accuracy as far as loan status was concerned. However, the model’s ability to correctly identify approved loans was significantly low based on the results presented in the form of recall of ‘Approved’ category and F1-score and high number of false negatives. This implies that the there is need to further enhance on the presented model. Possible steps for improvement include:

Feature Engineering: Further enhancing more features or, map existing features in ways that it’s able toextract more patterns from the obtained data.

Balancing the Dataset: Balancing the classes in case they are imbalanced and this can be done through variousways like oversampling the minority class or undersampling the majority class.

Hyperparameter Tuning: Tweak the RandomForestClassifier and adjusting hyperparameters as applicable tofor better result.

All in all it can be stated that the discussed model could serve as a base for credit risk analysis but furtherrefinements are needed to make the credit risk assessment more accurate and close to real life conditions.

Recommendations

The following are the areas of the credit risk assessment models that need some improvements and future research work:Despite the models offering a benchmark for loan status predictions, there are ways and measures that could be put in place to further improve the model accuracy, make it more resistant to operatingconditions and align it to the real business environment.

The following are some recommendations that are useful in enhancing some of the models and enhancing the forecast performance.

a) Advanced Feature Engineering: There is always room for perfecting models, including the ability tocorrectly classify approved loans, so it will be pertinent to discuss the topic further with regard to feature engineering. Introducing new features that may define previously unused interdependencies between the original or deriving new features that can represent latent dependencies with higher accuracy can improve the predictive capability of the models greatly. For example, using income, loan amount jointly as a debt-to-income ratio, or categorizing age differently may be beneficial for the models themselves to distill more information from.

b) Addressing Class Imbalance: This can be evidenced by the fact that the specific loan statuses are notfully represented within the dataset, making bias within the models the result. In addressing the problem of imbalance in the dataset, one must consider using methods such as Synthetic Minority Over-sampling Technique (SMOTE), undersampling the majority class, or even incorporating cost- sensitive algorithms. These approaches can help in decreasing the number of misclassified cases andincreasing the recall of the lesser frequent class and produce a more accurate model.

Algorithmic Enhancements and Ensemble Techniques: Although RandomForestClassifier gave a decent firstlead, it might improve results by trying out other sophisticated models like GBM, XGBoost or even deep learning. Other approaches such as ensemble techniques where several algorithms are combined with the intention of minimizing their weaknesses while exploiting their strengths could also be utilized in the development of a more reliable model of prediction. Tuning the hyperparameters of all the algorithms needsto be stringently followed to enhance the performance of these algorithms further.

REFERENCES

Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M., & Selmanovic, E. (2024). Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data. *Machine Learning and Knowledge Extraction*, *6*(1), 53-77.

Al Ayub Ahmed, A., Rajesh, S., Lohana, S., Ray, S., Maroor, J. P., & Naved, M. (2022, June). Using MachineLearning and Data Mining to Evaluate Modern Financial Management Techniques. In *Proceedings of SecondInternational Conference in Mechanical and Energy Technology: ICMET 2021, India* (pp. 249-257). Singapore: Springer Nature Singapore.

Amarnadh, V., & Moparthi, N. R. (2023). Comprehensive review of different artificial intelligence-based methods for credit risk assessment in data science. *Intelligent Decision Technologies*, *17*(4), 1265-1282.

Antony, T. M., & Kumar, B. S. (2023, August). Predicting of Credit Risk Using Machine Learning Algorithms. In *International Conference on Artificial Intelligence on Textile and Apparel* (pp. 99-114). Singapore: Springer Nature Singapore.

Addo, P.M., Guegan, D. and Hassani, B., 2018. Credit risk analysis using machine and deep learning models. *Risks*, *6*(2), p.38.

Bello, O.A., 2023. Machine learning algorithms for credit risk assessment: an economic and financial analysis. *International Journal of Management*, *10*(1), pp.109-133.

Biju, A. K. V. N., Thomas, A. S., & Thasneem, J. (2024). Examining the research taxonomy of artificial intelligence, deep learning & machine learning in the financial sphere—a bibliometric analysis. *Quality & Quantity*, *58*(1), 849-878.

Jayaram, E. S., Balachandar, G., & Kumar, K. S. (2024). Leveraging Machine Learning Techniques For Developing Robust Credit Scores For Peer-To-Peer Lending Platforms. *Educational Administration: Theoryand Practice*, *30*(5), 12958-12966.

Katragadda, R., Bathini, H. B., & Atluri, S. R. (2024). Application of Artificial Intelligence and Machine- Learning Algorithms for Forecasting Risk: The Case of the Indian Stock Market. *Artificial Intelligence Enabled Management: An Emerging Economy Perspective*, 249.

Mahesh, T. R., Vinoth Kumar, V., Shashikala, H. K., & Roopashree, S. (2023). ML algorithms for providingfinancial security in banking sectors with the prediction of loan risks. In *Artificial Intelligence and Cyber Security in Industry 4.0* (pp. 315-327). Singapore: Springer Nature Singapore.

Pattnaik, D., Ray, S., & Raman, R. (2024). Applications of artificial intelligence and machine learning in thefinancial services industry: A bibliometric review. *Heliyon*.

Puli, S., Thota, N., & Subrahmanyam, A. C. V. (2024). Assessing Machine Learning Techniques for Predicting Banking Crises in India. *Journal of Risk and Financial Management*, *17*(4), 141.

Ramakrishnan, R., Rohella, P., Mimani, S., Jiwani, N., & Logeshwaran, J. (2024, March). Employing AI andML in Risk Assessment for Lending for Assessing Credit Worthiness. In *2024 2nd International Conferenceon Disruptive Technologies (ICDT)* (pp. 561-566). IEEE.

Rao, M. K., Haralayya, B., Mishra, A., & Muda, I. (2024, March). Credit Risk Assessment in Banking Industry Using Optimization Based ML Algorithm. In *Advancements in Business for Integrating Diversity, and Sustainability: International Analytics Conference 2023| IAC 2023 February 2& 3, 2023| Virtual Conference* (p. 93). Taylor & Francis.

Singh, H., & Arora, S. (2023). Artificial Intelligence & Machine Learning Models for Credit Scoring and Risk Management. *IITM Journal of Management and IT*, *14*(1and2), 7-13.

Sharma, A., & Kumar, V. (2022). An exploratory study-based analysis on loan prediction. *InventiveCommunication and Computational Technologies: Proceedings of ICICCT 2022*, 423-433