MSc Data Science Dissertation

Title

Predictive Analysis of Credit Risk in the UK Banking Sector: A Machine Learning Approach to Macroeconomic Determinants

Student Name: PRASHANTH REDDY MANCHAL

Student ID: @00792912

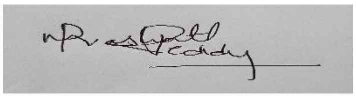
Supervisor: Norman Murray

Course: MSc Data Science

School of Science, Engineering and Environment
University of Salford

Declaration of Originality

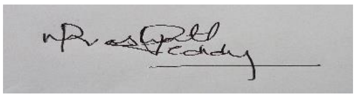
I hereby declare that this dissertation is the result of my own independent work, except where specific references are made. It has not been submitted, in whole or in part, for any other degree or qualification at this or any other university. All sources used have been adequately acknowledged, and all data and materials not generated by me have been appropriately cited.

Signed: 
Date: ______24.08.2025_______________

Declaration of Writing Support Tools

In preparing this dissertation, I used Grammarly to assist with spelling, grammar, and sentence structure.

This tool was only used for language refinement and did not contribute to the generation of ideas, technical content, or analysis. All intellectual contributions, coding, experimental work, and critical discussions presented in this dissertation are entirely my own.

Signed: 
Date: ______24.08.2025_______________

Acknowledgement

I would also like to thank my supervisor, Mr. Norman Murray, who supported me throughout the development of this dissertation by providing his patient guidance, support, and helpful suggestions. His knowledge and practical guidance played significant roles in the process of laying down the path toward the accomplishment of this study. I would also like to thank the academic staff of the University of Salford, who supported me throughout my MSc Data Science study, based on which this project was realized. No special thanks could be given to my family and friends who have continually supported, been patient with me, and understanding during this time. Their motivation has been beneficial. Lastly, I would like to thank the resources, tools, and free software communities that created conditions to make such a project as the Smart Surveillance System, as presented in this work, possible.

Abstract

Credit risk assessment plays a vital role in the financial services sector by helping lenders evaluate the probability of borrower default and make informed lending decisions. With the increasing availability of borrower data and advancements in machine learning, predictive models are now more precise and efficient than traditional statistical approaches. This dissertation explores how supervised machine learning techniques can be used to predict loan default probabilities using a publicly available lending dataset.

The study follows the CRISP-DM framework, starting with detailed data pre-processing that includes handling missing values, managing outliers, encoding categorical variables, and applying feature scaling. The Synthetic Minority Oversampling Technique (SMOTE) is used to address class imbalance, ensuring equal representation of both default and non-default cases. Four classification algorithms are applied: Logistic Regression, Decision Tree, Random Forest, and a Neural Network (Multi-Layer Perceptron). Hyperparameter tuning is performed using GridSearchCV to optimize model performance.

Model evaluation is conducted using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. The results show that the Random Forest classifier consistently outperforms other models, with an accuracy of 93% and a ROC-AUC score of 0.98. Key factors influencing credit risk include the borrower's high-end FICO score range, loan term, recent credit inquiries, loan purpose, verification status, homeownership status, interest rate, and annual income. The study finds that a finely tuned Random Forest model is the best choice for loan approval decisions. It is a valuable tool for banks and other financial institutions aiming to improve decision-making, reduce default rates, and enhance portfolio quality, as it provides better predictive accuracy, less overfitting, and easier interpretability.

Contents

....1
Declaration of Originality..2
Declaration of Writing Support Tools 2
Acknowledgement 3
Abstract 4
Chapter 1: Introduction..13
1.1 General Overview 13
1.2 Aim and Objectives..14
Aim...14
Objectives 14
1.3 Project Outline 15
1.4 Conclusion 16
Chapter 2: Literature Review...17
2.0 Introduction 17
2.1 General Overview 17
2.2 Data Mining in Financial Services...18
2.3 Machine Learning Concepts 18
2.4 types of Machine Learning Methods 19
2.5 Why Supervised Learning is Most Relevant for This Project.................19
2.6 Credit Risk Prediction Literature 19
2.7 Determinants of Credit Risk...20
2.8 Challenges in Machine Learning-Based Credit Risk Prediction 20
2.9 Research Gaps and Justification for This Study..................................21
2.10 Conclusion 21
Chapter 3: Methodology Overview and Data Analysis...............................22
3.1 Methodology Overview 22
3.2 Data Description 24
3.3 Exploratory Data Analysis (EDA)..24
3.4 Outlier Handling 25
3.5 Managing Class Imbalance 25
3.6 Data Pre-processing 25
3.7 Tools for Data Analysis...25

3.8 Conclusion 26
Chapter 4: Data Description, Exploratory Data Analysis, and Pre-processing27
4.1 Data Description 27
4.2 Exploratory Data Analysis (EDA)..29
4.3 Outlier Detection and Handling..34
4.4 Managing Class Imbalance (SMOTE) 36
4.5 Data Pre-processing 37
4.6 Conclusion 39
Chapter 5: Analysis Approach and Implementation 40
5.0 Classification Algorithms Overview..40
5.1 Train/Test Split & Feature Scaling 40
5.2 Cross-Validation for Reliable Performance Estimates...........................41
5.3 Model Selection and Rationale 41
5.3.1 Logistic Regression..42
5.3.2 Decision Tree 43
5.3.3 Random Forest 43
5.3.4 Neural Network (Multi-Layer Perceptron)..44
5.4 Model Training Process 45
5.5 Workflow Summary 45
5.6 Hyperparameter Tuning..46
5.6.1 Logistic Regression Tuning..46
5.6.2 Decision Tree Tuning 46
5.6.3 Random Forest Tuning 47
5.6.4 Multi-Layer Perceptron (MLP) Tuning 47
5.6.5 Implementation Details..47
5.6.6 Best Parameters Summary 48
5.7 Evaluation Metrics 48
5.7.1 Accuracy 48
5.7.2 Precision 49
5.7.3 Recall (Sensitivity or True Positive Rate)..49
5.7.4 F1-Score 49
5.7.5 ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) 50

5.7.6 Confusion Matrix Interpretation......50
5.8 Conclusion......51
Chapter 6 – Critical Evaluation & Results Analysis......52
6.1 Introduction......52
6.2 Decision Tree Results......52
6.2.1 Performance Metrics......53
6.2.2 Confusion Matrix Analysis......54
6.2.3 ROC Curve......54
6.2.4 Interpretation......55
6.3 Logistic Regression Results......55
6.3.1 Performance Metrics......56
6.3.2 Confusion Matrix......56
6.3.3 ROC Curve......57
6.3.4 Interpretation......57
6.4 Neural Network (MLP) Results......58
6.4.1 Performance Metrics......58
6.4.2 Confusion Matrix......59
6.4.3 ROC Curve......60
6.4.4 Interpretation......60
6.5 Random Forest Results......61
6.5.1 Performance Metrics......61
6.5.2 Confusion Matrix......61
6.5.3 ROC Curve......62
6.5.4 Interpretation......63
6.6 Model Comparison and Selection......63
6.6.1 Trade-Off Discussion......64
6.7. Best Model Justification......64
6.8 Model Comparison-Linked to Literature......66
6.9 Conclusion......67
Chapter 7: Limitations and Conclusion......69
7.1 Limitations......69
7.2 Practical Recommendations......69
7.3 Additional Work......70

7.4 Conclusion 70
7.5 Legal, Social, Ethical, and Professional Issues 71
7.6 Final Remarks..71
8. References 72

List of Figures

Figure 3.1 Flow Chart …………………………………………………………………… 21

Figure 4.2 Summary Statistics for Key Numerical Variables ………………………... 25

Figure 4.2 Target Variable Distribution ………………………………………… 26

Figure 4.3 – Figure 4.6 Further Distributions …………………………………… 27

Figure 4.7 Employment Length Insights ……………………………….…. 28

Figure 4.8 - Figure 4.12 Target Relationships …………………………….... 29

Figure 4.13 Correlation Heatmap ………………………………………….... 30

Figure 4.14 Loan Amount and Figure 4.15 DTI ……………………………………. 31

Figure 4.16 Annual Income ………………………………………………………… 32

Figure 4.17 SMOTE …………………………………………………… 33

Figure 5.2 Train/Test Split ………………………………………………………… 37

Figure 5.3.1 Top 15 coefficients in Logistic Regression …………………… 38

Figure 5.3.2 Decision Tree first two levels. ………………………………………….. 39

Figure 5.3.3 Top 15 features in Random Forest ……………………………………40

Figure 5.3.4 Neural Network training loss over iterations ………………………….. 41

Figure 5.5.5 Training and Validation Workflow ……………………………………… 41

Figure 6.2.1 Decision Tree results …………………………………………………… 49

Figure 6.2.2 Confusion Matrix – Decision Tree ……………………………………… 50

Figure 6.2.3 ROC Curve …………………………………………………………… 50

Figure 6.3.1 Logistic Regression Results …………………………………………… 52

Figure 6.3.2 Confusion Matrix- Logistic Regression ………………………………… 52

Figure 6.3.3 Roc Curve – Logistic Regression …………………………………… 53

Figure 6.4.1 Neural Network Results ………………………………………………… 54

Figure 6.4.2 Confusion Matrix- Neural Network …………………………………… 55

Figure 6.4.3 Roc Curve - Neural Network (MLP) ………………………………… 56

Figure 6.5.1 Random Forest Results ………………………………………………… 57

Figure 6.5.2 Confusion Matrix – Random Forest …………………………………… 57

Figure 6.5.3 Roc Curve – Random Forest …………………………………………… 58

Figure 6.6 Roc Curves For all Models ……………………………………………… 59
Figure 6.7 Random Forest – Best Features ………………………………………… 61

List of Tables

Table 4.1. Key Variables in the Lending Club Dataset……………………….29
Table 5.6.6 Parameters Summary………………………………………………...48
Table 6.2 parameter grid…………………………………………………………53
Table 6.6 Comparison…………………………………………………...…….63

Abbreviation	Full Form	Meaning / Usage in This Study
AI	Artificial Intelligence	Simulation of human intelligence in machines that can perform tasks requiring learning and decision-making.
ANN	Artificial Neural Network	Computational model inspired by biological neural networks, used for complex pattern recognition.
AUC	Area Under the Curve	A metric summarising the ROC curve performance of a classification model.
CRISP-DM	Cross-Industry Standard Process for Data Mining	A widely used process model for data mining projects is followed in this study.
CSV	Comma-Separated Values	File format used to store tabular data.
DTI	Debt-to-Income Ratio	Ratio of a borrower’s total debt payments to their income.
EDA	Exploratory Data Analysis	Process of summarising main dataset characteristics using visual and statistical methods.
FICO	Fair Isaac Corporation	A credit score is used to evaluate a borrower’s creditworthiness.
FN	False Negative	An instance where a model predicts non-default, but the loan actually defaults.
FP	False Positive	An instance where a model predicts default, but the loan does not default.
GDP	Gross Domestic Product	A measure of the economic performance of a country, referenced in the macroeconomic context.
IQR	Interquartile Range	Statistical measure used for outlier detection.
LR	Logistic Regression	A statistical classification model was used as a baseline in this study.
ML	Machine Learning	The field of AI is focused on algorithms that learn from data.
MLP	Multi-Layer Perceptron	A type of artificial neural network with multiple hidden layers.
NPL	Non-Performing Loan	A loan that is in default or close to being in default.
OLS	Ordinary Least Squares	Statistical method for estimating parameters in a linear regression model.
PCA	Principal Component Analysis	Dimensionality reduction technique.
RF	Random Forest	An ensemble machine learning algorithm based on multiple decision trees.
ROC	Receiver Operating Characteristic	Curve showing the trade-off between sensitivity and specificity.
SMOTE	Synthetic Minority Oversampling Technique	Method for handling class imbalance by generating synthetic minority class samples.

SQL	Structured Query Language	Language for managing and querying relational databases.
SVM	Support Vector Machine	Supervised machine learning algorithm for classification and regression.
TN	True Negative	Model correctly predicts a non-default loan.
TP	True Positive	The model correctly predicts a default loan.
XGBoost	Extreme Gradient Boosting	An advanced tree-based boosting algorithm (not applied in this study but relevant in the literature).

Chapter 1: Introduction

1.1 General Overview

Credit risk is a significant concern in the financial services sector. It refers to the potential financial loss when a borrower fails to repay a loan (Ghosh, 2015). Lending plays a crucial role in banking operations, making it an essential aspect of overall risk management (Makri, Tsagkanos & Bellas, 2014). Inadequate assessment of credit risk can lead to higher default rates, reduced profits, regulatory challenges, and potentially, the collapse of a financial institution (Louzis, Vouldis & Metaxas, 2012). Therefore, accurate evaluation and management of credit risk are vital for maintaining the stability and long-term sustainability of lending institutions (Salas & Saurina, 2002).

Statistical scoring models like logistic regression and linear discriminant analysis have traditionally been used to evaluate credit risk (Abid, 2022). These methods analyze structured data about borrowers—such as demographics, credit history, and repayment records—to estimate the probability of default (Khandani, Kim & Lo, 2010). Their advantages include ease of understanding, straightforwardness, and regulatory compliance (Baesens et al., 2003). However, they assume linear relationships among variables, which can cause them to overlook complex, non-linear patterns often found in modern credit data (Fitzpatrick & Mues, 2016). This limitation can reduce their accuracy, particularly when assessing various types of loans.

The advent of machine learning (ML) techniques has revolutionized credit risk modeling by providing more robust and adaptable predictive tools (Klein, 2013). ML algorithms can handle large, complex datasets and identify sophisticated, non-linear relationships between borrower traits and default risk (Dixon, Halperin, and Bilokon, 2020). Unlike traditional models, they can incorporate additional variables—such as behavioral and transactional data—to yield more precise risk assessments that evolve over time (Yu, 2022). ML-based methods enable lenders to make better approval decisions and proactively manage risk by minimizing misclassification errors (Wang et al., 2021).

In today's data-driven environment, utilizing borrower-level data is more essential than ever for accurate credit risk predictions. Financial institutions now gather information from diverse sources, including online loan applications, payment histories, credit bureau reports, and alternative data like utility payments or mobile transaction records (Ruiz et al., 2017). The vast volume of data and ML's analytical capabilities allow for the creation of models that are more accurate, responsive, and tailored to individual borrowers' needs (Ouahilal et al., 2016). These models help improve credit scoring, enable better customer segmentation, and reduce default rates in investment portfolios (Broby, 2022).

This dissertation focuses on applying supervised machine learning methods to predict loan defaults using a publicly available lending dataset. It involves a comparative analysis of different classification models, including Logistic Regression, Decision Tree, Random Forest, and Neural Network (MLP), to identify which approach offers the highest prediction accuracy and reliability. The aim is to determine the best model for real-world credit risk assessment by evaluating its performance against established criteria and identifying the key factors that influence an effective credit risk model capable of practical implementation.

1.2 Aim and Objectives

Aim

The main goal of this project is to develop, evaluate, and compare supervised machine learning models that predict credit risk, specifically the likelihood that a loan will not be repaid. The study aims to identify the most accurate and reliable model for use in a financial services context, while also highlighting key borrower characteristics that influence credit risk evaluations.

Objectives

To accomplish this objective, the following research aims have been outlined:

To analyze and pre-process the lending dataset, we address missing values, handle outliers, encode categorical variables, and apply feature scaling to ready the data for modeling.

Use the Synthetic Minority Oversampling Technique (SMOTE) to correct class imbalance in the dataset, ensuring fair representation of both default and non-default classes.

To implement and train four supervised machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, and Neural Network (MLP) with the prepared dataset.

To improve model performance through hyperparameter optimization using GridSearchCV, ensuring each process is evaluated under its best configuration.

Evaluate the models using suitable classification metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices.

To analyze feature importance in the best-performing model to identify the most significant borrower and loan attributes influencing credit risk predictions.

To recommend the most suitable model for real-world credit risk assessment, supported by evidence from the evaluation results.

1.3 Project Outline

This research involves building and testing machine learning models to predict credit risk, specifically estimating the likelihood of loan default. I am using a publicly available lending dataset and have followed the CRISP-DM process to keep a structured and logical workflow. The goal is to systematically move through each step, from understanding the data to developing a final recommendation for the best model.

I start by familiarizing myself with the data and cleaning it. This includes checking for missing values, outliers, and making sure all features are in a format suitable for the models. I address the imbalance between default and non-default cases by applying SMOTE, ensuring the models do not favor the majority class.

Once the data is prepared, I proceed to train different models. For this project, I selected four models: Logistic Regression, Decision Tree, Random Forest, and a Neural Network (MLP). I train these models on the cleaned dataset and tune their performance using GridSearchCV to enhance each model’s effectiveness.

After training, I evaluate each model's performance using metrics such as accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. I also examine feature importance to identify the factors that matter most in predicting defaults. This approach is useful because it not only enhances the model's explainability but also provides lenders with clear insights into what drives credit risk.

Finally, I bring everything together in the discussion and conclusion. I highlight which model is best for this problem, suggest how it could be used in a real lending environment, and discuss where the project could go next. This includes exploring additional data sources like macroeconomic indicators, trying more advanced algorithms, and using cost-sensitive methods to address the higher costs of missed defaults.

1.4 Conclusion

This chapter outlined the background, objectives, and scope of the dissertation. It emphasized the crucial role of credit risk management in maintaining financial stability and minimizing loan defaults, especially within the banking sector. While traditional techniques like logistic regression are still useful because of their interpretability, they often struggle to capture the complex patterns in borrower behavior and loan outcomes. Recent progress in machine learning presents fresh opportunities to enhance predictive accuracy, making these methods more relevant for lenders aiming to adopt more robust, data-driven decision-making processes.

The research aimed to develop and compare machine learning models for predicting loan default risk. It included data preparation, model training, hyperparameter tuning, and assessing performance using various metrics. The project follows the CRISP-DM methodology, ensuring a structured process from data understanding to evaluation.

In summary, Chapter 1 establishes the foundation by outlining the motivation, research problem, and intended contribution. The subsequent chapters expand on this framework, starting with a review of existing literature in Chapter 2.

Chapter 2: Literature Review

2.0 Introduction

This chapter reviews the literature on credit risk prediction, beginning with the concept of credit risk, its significance in banking, and the broader economic consequences of inadequate risk assessment. It traces the evolution from traditional statistical approaches - such as logistic regression and discriminant analysis - to advanced machine learning (ML) methods that can capture complex, nonlinear patterns in borrower behaviour. By doing so, it sets the stage for understanding why newer approaches may outperform conventional models in modern, data-rich lending environments.

The discussion then moves to the role of data mining in financial services, particularly supervised learning techniques, which form the basis of this study's modelling approach. It examines standard algorithms, their strengths and limitations, and the empirical findings of prior research, focusing on borrower-specific, macroeconomic, and institutional factors that influence default risk. Special attention is paid to challenges such as class imbalance, data quality, and the trade-off between predictive accuracy and interpretability issues that are central to the design and evaluation of this project.

In the final part, the chapter identifies specific research gaps, such as inconsistent handling of imbalanced datasets, insufficient hyperparameter tuning, and limited attention to model interpretability in existing studies. These gaps motivate this research, which aims to build a balanced, optimised, and explainable machine learning model for credit risk prediction.

2.1 General Overview

Credit risk remains a vital concern for banks as it impacts their stability, profitability, and long-term lending capacity. It refers to the chance that borrowers may fail to repay loans on time, potentially causing substantial losses. Poor management of credit risk can lead to broader financial issues, exemplified during the 2007–2009 global economic crisis when inadequate credit evaluations led to widespread defaults and a loss of confidence in banking systems (Reinhart & Rogoff, 2011; Makri et al., 2014).

Traditionally, banks have relied on statistical methods like logistic regression and discriminant analysis to estimate default probabilities. These techniques are straightforward, transparent, and comply with regulatory requirements, explaining their continued widespread use (Hosmer et al., 2013; Abid, 2022). However, they have limitations: they generally assume linear relationships and might miss complex interactions between borrower behavior, income, and macroeconomic factors that influence creditworthiness (Khandani et al., 2010).

Over the past decade, progress in data analytics and machine learning has provided more powerful options. Techniques such as decision trees, random forests, boosting algorithms, and neural networks are capable of processing much larger datasets and detecting patterns that traditional methods might miss (Dixon et al., 2020; Wang et al., 2021). These models can incorporate not only financial information but also behavioral and contextual data, offering lenders a more comprehensive view of risk. However, challenges remain, especially regarding interpretability and fairness, as complex models often function as “black boxes” and can unintentionally reinforce existing biases (Fitzpatrick & Mues, 2016; Baesens et al., 2003).

2.2 Data Mining in Financial Services

Data mining involves discovering patterns and relationships within large datasets. In financial services, it is crucial for tasks like fraud detection, customer segmentation, and credit risk assessment (Krumeich et al., 2015). Banks and lenders gather extensive data from loan applications, repayment records, and credit bureaus. Analyzing this data enables them to develop predictive models that surpass basic rule-based scoring methods.

For credit risk, data mining helps detect subtle signs of default risk that traditional methods might miss. Instead of relying solely on past repayment history or limited demographic data, lenders can include hundreds of borrower and loan characteristics in their models, leading to more reliable and evidence-based evaluations (Broby, 2022; Dixon et al., 2020). This change enables institutions to rely less on subjective judgment, resulting in faster and more precise lending decisions.

2.3 Machine Learning Concepts

Machine learning (ML) involves enabling computers to learn from past data and make predictions without explicit programming (Apte, 2010). In credit risk assessment, ML models analyze historical borrower outcomes—such as loan payments or defaults—and identify patterns to predict the default risk of new applicants. Unlike traditional statistical models like logistic regression, which assume fixed relationships, ML models are more flexible and can adapt to complex feature interactions (Khandani et al., 2010).

This flexibility is crucial because borrower repayment behavior is rarely influenced by a single variable. Instead, it results from a mixture of demographic, financial, and macroeconomic factors. Machine learning models can incorporate these varied inputs, leading to better prediction accuracy than traditional approaches (Dixon et al., 2020).

2.4 types of Machine Learning Methods

Machine learning can be categorised into four main approaches:

- **Supervised learning**: models trained on labelled data (default vs non-default); most common in credit scoring (Hosmer et al., 2013).
- **Unsupervised learning**: identifies clusters or anomalies without labels, often used for fraud detection (Ouahilal et al., 2016).
- **Semi-supervised learning**: combines small labelled datasets with large unlabelled ones, useful when loan outcomes are only partially known (Clements et al., 2020).
- **Reinforcement learning**: less common in credit risk, but has potential in adaptive lending systems (Wang et al., 2021).

In this project, supervised learning is chosen since the dataset contains known repayment outcomes.

2.5 Why Supervised Learning is Most Relevant for This Project

Supervised learning is ideal for this project because it uses historical loan outcome data. By training on borrower characteristics and repayment status, the models can identify the most predictive factors for default. This facilitates consistent comparison across different algorithms using metrics like precision, recall, F1-score, and ROC-AUC. Additionally, supervised learning accommodates both simple, interpretable models such as Logistic Regression and more complex ones like Random Forests and Neural Networks, allowing for a fair assessment of their performance.

2.6 Credit Risk Prediction Literature

Early credit scoring research primarily used statistical models like logistic regression and discriminant analysis because they are simple and easy to interpret (Hosmer et al., 2013). Nonetheless, these methods struggle to capture complex borrower behaviors and often assume linear relationships (Makri et al., 2014). More recent research indicates that machine learning techniques—including decision trees, ensemble methods, and neural networks—generally provide better predictive accuracy than traditional models (Baesens et al., 2003; Lessmann et al., 2015).

Despite this, the literature emphasises a trade-off: while advanced models deliver stronger accuracy, they often lack interpretability, which is crucial in regulated industries such as banking (Fitzpatrick & Mues, 2016). This balance between predictive performance and transparency remains a central theme in credit risk research.

2.7 Determinants of Credit Risk

Research shows that credit risk is influenced by a mix of borrower, macroeconomic, and bank-specific factors. At the borrower level, income, employment stability, debt-to-income ratio, and credit scores are strong indicators of repayment ability (Khandani et al., 2010; Beck et al., 2015). Loan-specific attributes such as purpose, term, and interest rate also play a role.

At the macroeconomic level, factors such as GDP growth, unemployment, and inflation strongly affect default rates, especially during downturns (Makri et al., 2014; Carvalho et al., 2022). Meanwhile, bank practices—such as lending standards, capital adequacy, and portfolio diversification—can amplify or mitigate risk exposure (Salas & Saurina, 2002; Beck et al., 2015).

2.8 Challenges in Machine Learning-Based Credit Risk Prediction

Several challenges limit the effectiveness of ML in credit risk modelling. First, class imbalance is a significant issue: default cases are much rarer than successful repayments, which can bias models towards predicting the majority class (Chawla et al., 2002). Techniques such as SMOTE are commonly used to address this problem. Second, overfitting remains a risk, particularly for complex models like neural networks (Fitzpatrick & Mues, 2016). Finally, interpretability is a persistent concern, as regulators require transparency in how credit decisions are made (Baesens et al., 2003). Balancing accuracy with explainability is therefore critical for real-world adoption.

Finally, there is the issue of **interpretability versus predictive power**. Some of the most accurate ML models - such as gradient boosting machines or deep neural networks - are often considered “black boxes” because their decision-making process is complicated to explain (Baesens et al., 2003). In regulated industries like banking, this lack of transparency can make it harder to justify lending decisions to regulators or customers. As a result, many financial institutions prefer models that offer a balance between performance and explainability, such as decision trees or random forests with feature importance analysis. Achieving this balance is a key challenge in applying ML to credit risk prediction.

2.9 Research Gaps and Justification for This Study

Although many studies have compared ML algorithms, gaps remain. Few systematically address class imbalance, optimise hyperparameters, or evaluate multiple metrics beyond accuracy (Lessmann et al., 2015). Additionally, limited emphasis has been placed on interpretability, which is critical for regulatory compliance (Fitzpatrick & Mues, 2016). This study addresses these gaps by applying SMOTE for balancing, using GridSearchCV for tuning, evaluating models with multiple metrics, and analysing feature importance for interpretability. Comparisons to existing literature will also be drawn in Chapter 6 to position findings within the broader research landscape.

2.10 Conclusion

The literature reveals a shift towards ML-based credit risk prediction due to their ability to model complex, non-linear patterns and process large datasets efficiently (Wang et al., 2021; Baesens et al., 2003). However, persistent challenges remain, including class imbalance (Chawla et al., 2002), overfitting, data quality issues, and the trade-off between accuracy and interpretability in regulated environments (Lang & Sun, 2014). Additionally, the determinants of credit risk are highly context-dependent, shaped by borrower-specific, macroeconomic, and institutional factors (Louzis et al., 2012; Carvalho et al., 2022). These gaps - remarkably inconsistent handling of imbalanced data, limited hyperparameter tuning, and insufficient focus on model transparency - justify this study's approach, which applies SMOTE balancing, optimised training, comprehensive evaluation metrics, and feature importance analysis to develop a robust and interpretable credit risk prediction framework.

Chapter 3: Methodology Overview and Data Analysis

3.1 Methodology Overview

This study adopts the **CRISP-DM (Cross Industry Standard Process for Data Mining)** framework, offering a structured, iterative method for developing predictive models. Each phase is specifically tailored to the task of loan default prediction. The CRISP-DM process includes six interconnected stages:

Business Understanding - This phase focuses on defining the business problem, understanding the context of credit risk assessment, and identifying how predictive modelling can support decision-making in loan approval processes. The study begins by framing the research within the operational and strategic needs of financial institutions, ensuring that the analytical outputs are aligned with real-world business requirements.

- In the **Business Understanding** phase, the project objective was defined as building a reliable and interpretable machine learning model capable of predicting loan default risk. This step involved identifying the target variable (loan_status), determining relevant borrower and loan features, and setting performance goals aligned with the needs of financial institutions.

Data Understanding - This phase involves acquiring and exploring the publicly available lending dataset to assess its structure, quality, and relevance. Initial exploratory data analysis (EDA) is performed to understand variable distributions, detect potential data quality issues, and identify relationships between features and loan default status.

- During **Data Understanding**, the LendingClub dataset was explored through descriptive statistics, visualisations, and correlation analysis. This helped identify missing values, class imbalance, and potential outliers.

Data Preparation - Prior to constructing the machine learning model, the dataset undergoes cleaning and pre-processing. This involves addressing missing values, converting categorical data into numerical formats, scaling numerical features, and managing outliers. Given the common issue of class imbalance in loan default prediction, the SMOTE technique is employed to create synthetic samples, helping to balance the dataset and improve training effectiveness.

- The **data preparation process** included cleaning and transforming raw data to ready it for modeling. Important steps were managing missing data, capping outliers, encoding categorical features, and scaling numerical variables. To address class imbalance, SMOTE was utilized, resulting in a balanced dataset for equitable and effective model training.

Modelling - Four supervised machine learning models—Logistic Regression, Decision Tree, Random Forest, and Neural Network (MLP)—were developed using the prepared dataset. To improve their performance, hyperparameter tuning was performed with GridSearchCV.

- In the modelling phase, four algorithms—Logistic Regression, Decision Tree, Random Forest, and Neural Network (MLP)-were chosen for implementation. The models were trained using a stratified train/test split, optimized through **GridSearchCV** hyperparameter tuning, and assessed on unseen test data.

Evaluation involves assessing each model with various classification metrics, such as accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. ROC curves help compare the models' discrimination abilities, while feature importance analysis enhances interpretability.

- **The Evaluation** phase involved assessing models through various performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. For interpretable models, feature importance analysis was conducted to identify the main predictors of default.

Deployment Recommendation - Based on the evaluation results, the best-performing model is identified and recommended for deployment in a financial services context. The recommendation includes practical considerations for implementation, potential integration into existing credit assessment systems, and suggestions for future enhancements.

This methodology ensures a systematic, repeatable, and transparent approach to building a machine learning model that can be used in real-world credit risk assessment scenarios.

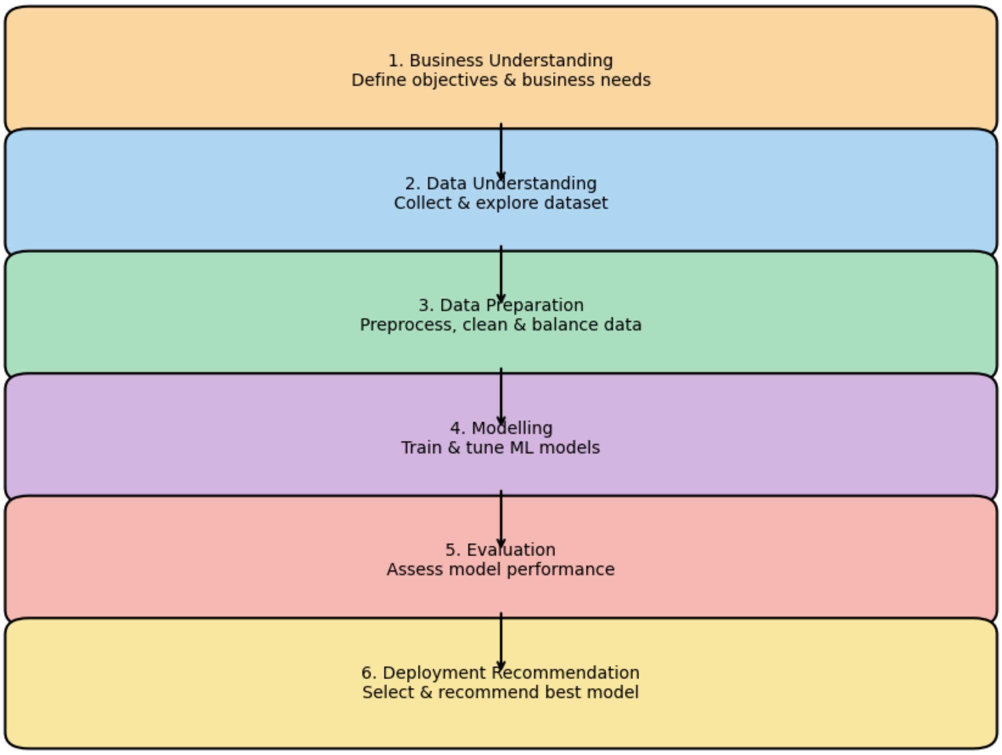


Figure 3.1 Flow Chart

3.2 Data Description

The dataset contains borrower, loan, and repayment information from LendingClub. Variables include:

- **Loan-specific**: amount requested, term, interest rate, and installment.
- **Borrower-related**: annual income, employment length, FICO range, and debt-to-income ratio.
- **Categorical**: loan purpose, home ownership, and verification status.
- **Target variable**: loan_status, recoded as binary (default = 1, fully paid = 0).

The dataset is representative of real-world credit risk problems: imbalanced, with missing data, outliers, and a mix of variable types.

3.3 Exploratory Data Analysis (EDA)

EDA was used to understand distributions and relationships before modelling:

- **Data quality**: Missing values were minimal and imputed; duplicates were removed; extreme outliers were capped.
- **Target distribution**: Most loans were fully paid, confirming the imbalance problem.

- **Feature insights**: Defaults clustered around higher interest rates, lower FICO scores, and higher debt-to-income ratios.
- **Categorical variables**: Debt consolidation was the most common loan purpose; most borrowers rented or mortgaged homes.
- **Correlations**: FICO ranges were highly correlated, and the loan amount was strongly linked with installment size.

EDA confirmed that features such as FICO, interest rate, and DTI are key predictors of default.

3.4 Outlier Handling

Extreme values were identified in loan_amnt, annual_inc, dti, and int_rate. Instead of deleting them, winsorisation was applied to cap values at thresholds. This preserved rare but valid cases while reducing the impact of unrealistic extremes.

3.5 Managing Class Imbalance

The dataset was significantly imbalanced, with defaults constituting only a small portion of cases. To tackle this, the **SMOTE** algorithm was used on the training data. SMOTE creates synthetic minority class examples by interpolating between nearby samples, thus boosting the variety of defaults and helping the models distinguish more effectively between classes.

3.6 Data Pre-processing

Steps included:

- **Cleaning**: Imputation of missing values, removal of duplicates.
- **Encoding**: One-hot encoding for categorical variables, ordinal mapping for ordered variables such as emp_length.
- **Scaling**: Standardisation applied to numerical features.
- **Balancing**: SMOTE applied post-split, ensuring test data remained untouched.

3.7 Tools for Data Analysis

The data analysis and machine learning modelling in this study were implemented using **Python** within a **Jupyter Notebook** environment. Python was selected due to its flexibility, wide range of open-source libraries, and strong support for data science and machine learning workflows. Jupyter Notebook offers an interactive environment that combines code execution, visualisation, and documentation, making it well-suited for conducting exploratory data analysis (EDA), modelling, and evaluation in a transparent and reproducible manner.

The key Python libraries and tools used in this research are as follows:

- **Pandas** - Used for data loading, cleaning, and manipulation. Pandas provides efficient structures for handling tabular data, such as DataFrames, which are essential for preprocessing and feature engineering tasks.
- **NumPy** - Used for numerical computing, supporting mathematical operations and array manipulation, which are required for machine learning data preparation.
- **Matplotlib** - A fundamental plotting library used to create static, animated, and interactive visualisations for data exploration and presentation.
- **Seaborn** - Built on Matplotlib, Seaborn provides enhanced visualisation capabilities with a high-level interface for creating attractive and informative statistical graphics.
- **Scikit-learn** - The core machine learning library used in this project for implementing supervised learning algorithms, model evaluation metrics, data preprocessing methods, and hyperparameter tuning via GridSearchCV.
- **Imbalanced-learn** - Used to address class imbalance issues in the dataset through the Synthetic Minority Oversampling Technique (SMOTE), ensuring that models learn from balanced default and non-default loan cases.
- **Joblib** - Used for saving trained machine learning models for potential deployment, enabling model reuse without retraining.

In addition to these libraries, Python's built-in capabilities for scripting and automation were leveraged to streamline repetitive processes, such as running model evaluations and generating plots. The integration of these tools provided a complete, end-to-end workflow for this study, enabling the transformation of raw data into actionable insights through the application of robust machine learning techniques.

3.8 Conclusion

This chapter described how the dataset was cleaned, explored, and pre-processed, and outlined the systematic CRISP-DM approach used for modelling. By applying techniques such as SMOTE, outlier handling, encoding, and scaling, the dataset was prepared for fair and balanced training. With this foundation, four classifiers were optimised and evaluated, setting the stage for the analysis in Chapter 4.

Chapter 4: Data Description, Exploratory Data Analysis, and Pre-processing

4.1 Data Description

The dataset used in this research originates from **LendingClub - DataSet,** a central peer-to-peer lending platform that connects borrowers and investors. Each row in the dataset represents a single loan and contains detailed information about the borrower, the loan terms, and the repayment outcome, **see Table 4.1**. The accompanying **data dictionary** provides a clear definition for each variable, helping to guide the pre-processing and feature selection process.

The dataset contains a mix of **numerical, categorical, and date/time variables**:

- **Loan-specific variables -** describe the attributes of the loan, including loan_amnt (the amount requested), term (duration, like 36 or 60 months), and int_rate (annual interest rate). Other variables, such as installment, specify the monthly repayment amount.
- **Borrower-related variables** include details about the borrower's financial profile, such as annual_inc (self-reported annual income), dti (debt-to-income ratio), and credit history metrics like fico_range_low and fico_range_high (FICO credit score range). Employment-related features, like emp_length (years in employment), offer further insight into the borrower's stability.
- **Categorical and descriptive variables** include purpose (e.g., debt consolidation or credit card refinancing), home_ownership (rent, mortgage, own), and verification_status (whether LendingClub verified income).
- **Performance outcome variables –** The key focus for this research is loan_status, which shows whether a loan was fully paid, charged-off (defaulted), or still active. For this project, the target variable has been simplified to a binary classification.

    1. **Default (1)** - Loans classified as charged-off or defaulted.
    2. **Non-default (0)** - Loans fully paid.

The dataset, similar to most real-world credit datasets, **is imbalanced**—there are significantly more loans that are successfully repaid than those that default. If this imbalance is not corrected, predictive models may become biased toward predicting repayment and struggle to identify risky borrowers. To remedy this, the **Synthetic Minority Oversampling Technique (SMOTE)** is used during pre-processing to achieve a balanced representation of default and non-default cases.

This dataset provides a rich set of borrower and loan-level features that have been widely used in credit risk modelling research. However, it also contains challenges typical of financial data - including missing values, potential outliers, and mixed variable types - which require careful handling before modelling. The following sections of this chapter outline the **exploratory data analysis (EDA)** and **pre-processing steps** used to prepare the dataset for machine learning.

Variable Name	Description	Type	Relevance to Credit Risk
loan_amnt	The total sum of money the borrower is requesting.	Numerical	Higher amounts may carry a higher risk.
term	The total number of payments scheduled for the loan (either 36 or 60 months).	Categorical	Longer terms may increase default risk.
int_rate	Annual interest rate applied to the loan.	Numerical (%)	Higher rates may indicate higher borrower risk.
installment	The amount the borrower must pay each month.	Numerical	Larger instalments can stress repayment capacity.
grade / sub_grade	Loan grade and subgrade are assigned by LendingClub based on creditworthiness.	Categorical	Strong predictor of default probability.
emp_length	Duration of current employment in years.	Categorical	Longer employment often signals stability.
home_ownership	Indicates whether the borrower owns, rents, or has a different mortgage arrangement.	Categorical	Ownership type linked to financial stability.
annual_inc	Annual income as reported by oneself.	Numerical	Higher income generally lowers default risk.
verification_status	Shows whether LendingClub has verified this income.	Categorical	Verified incomes tend to have lower risk.

purpose	Borrower-stated reason for the loan (e.g., debt consolidation, credit card).	Categorical	Loan purpose influences risk profile.
dti	Debt-to-income ratio.	Numerical (%)	Higher ratios indicate greater risk.
fico_range_low / fico_range_high	Borrower's FICO credit score range.	Numerical	Strong indicator of creditworthiness.
loan_status	Current loan status (paid, defaulted, etc.).	Target Variable	Used as the prediction target.

Table 4.1: Main Variables in the Lending Club Dataset

	Mean	Std. Dev.	Min	Max
loan_amnt	11145.39	7400.19	500.00	35000.00
int_rate	11.98	3.71	5.42	24.59
installment	323.52	208.48	15.69	1305.19
annual_inc	68919.96	63990.17	4000.00	6000000.00
dti	13.30	6.68	0.00	29.99
fico_range_high	719.02	35.87	629.00	829.00
last_fico_range_high	691.87	79.54	0.00	850.00

Figure 4.1 Summary Statistics for Key Numerical Variables

4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to better understand the dataset and identify data quality issues prior to modeling. This important step helps uncover trends, patterns, relationships, and irregularities that could affect model accuracy. The analysis primarily concentrated on four key areas: assessing data quality, analyzing the distribution of the target variable, exploring feature distributions, and examining the relationships between predictors and loan performance.

Data Quality Checks

The initial step involved reviewing the dataset for missing data, duplicates, and outliers. Missing entries in some columns were addressed by substituting numerical gaps with the median and categorical gaps with the most common category. Duplicate records were identified and eliminated to avoid skewing results. Outliers were detected through the Interquartile Range (IQR) method, and instead of removing extreme values, they were capped to preserve the dataset's integrity. This strategy ensured valuable rare observations remained part of the analysis.

Distribution of the Target Variable

The target variable, loan_status, was examined to assess the distribution of default and non-default cases (see Figure 4.2). The analysis showed a notable class imbalance: most loans were fully paid (non-defaults), with defaults making up a much smaller share. This imbalance can lead models to overpredict the majority class and overlook high-risk borrowers. Therefore, using the Synthetic Minority Oversampling Technique (SMOTE) during pre-processing to rebalance the dataset is justified.

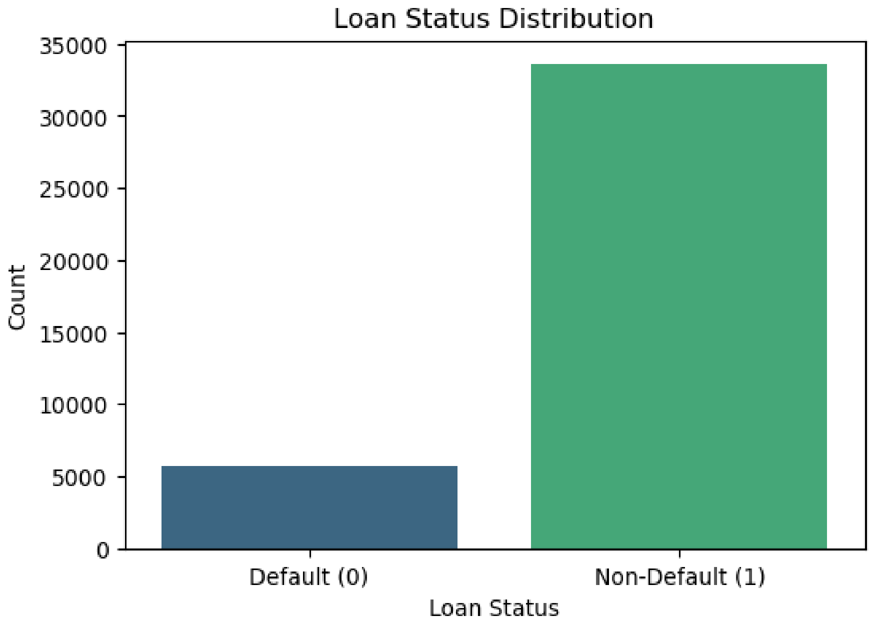


Figure 4.2 Target Variable Distribution

Feature Distributions

The distribution of numerical features was explored using histograms with kernel density estimates (Figures 4.3-4.6). Loan amounts typically ranged between $5,000 and $20,000, with an average of around $11,000. Interest rates were generally between 8% and 16%, but defaults were more common among borrowers paying higher rates. Annual income varied widely, suggesting the presence of extreme values, while debt-to-income (DTI) ratios averaged around 13%, with some borrowers close to the maximum allowable ratio of 30%.

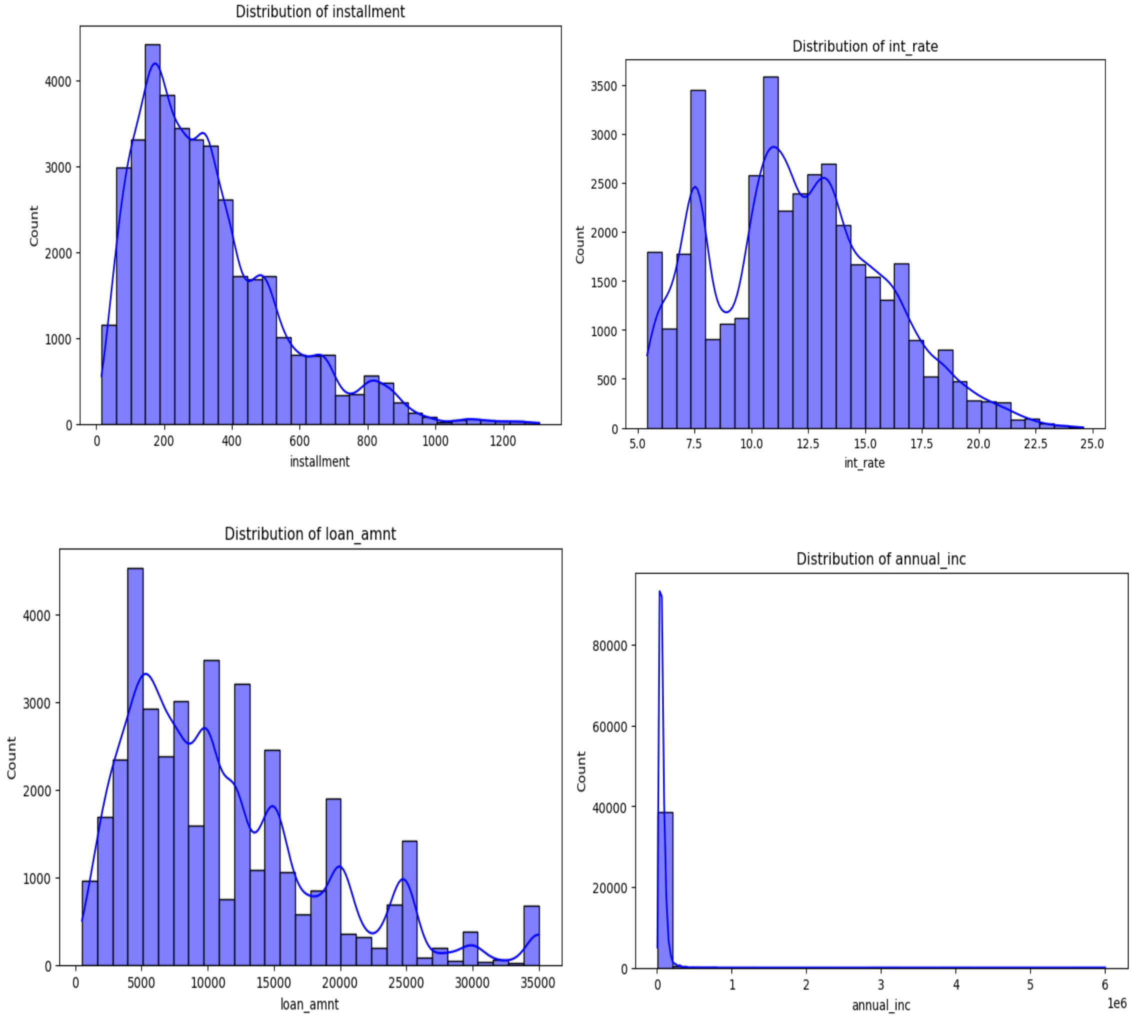


Figure 4.3 – Figure 4.6 Further Distributions

Categorical Feature Insights

The most frequent reason for taking out a loan was debt consolidation, with credit card refinancing and home improvement loans following (see Figure 4.7). Regarding home ownership, most borrowers rented or had a mortgage, while fewer owned their homes outright. Employment duration showed considerable variation, with many borrowers reporting over ten years in their current job, reflecting a relatively stable borrower demographic.

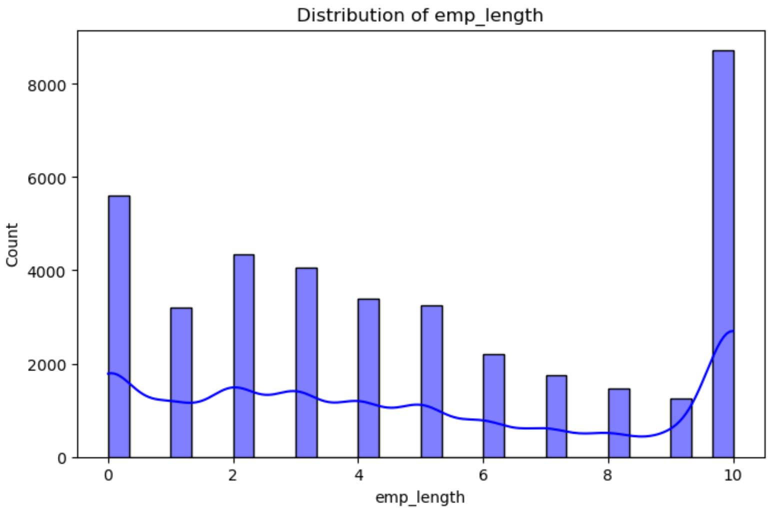
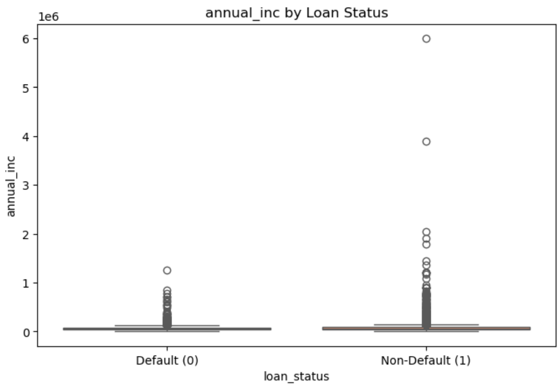
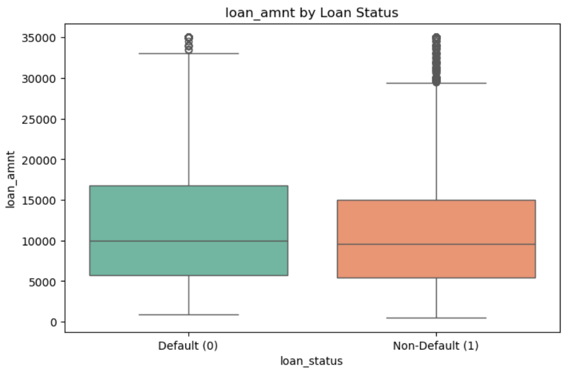


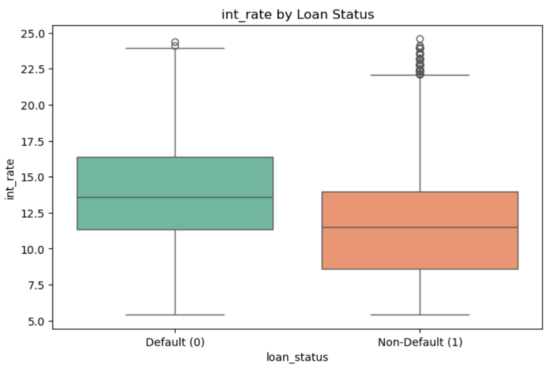
Figure 4.7 Employment Length Insights

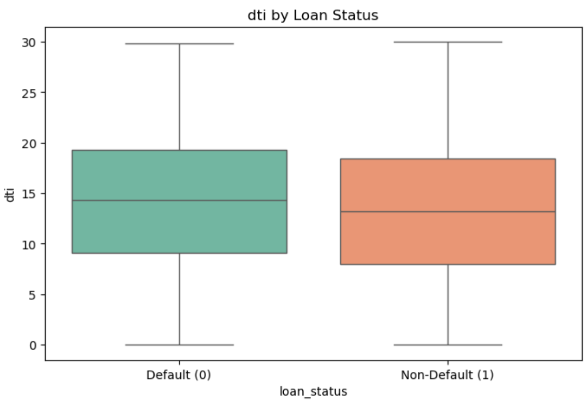
Feature–Target Relationships

Boxplots comparing numeric features against the loan status revealed clear patterns (Figures 4.8-4.12). Defaults tended to occur among borrowers with higher interest rates, lower FICO scores, and higher DTI ratios. These relationships highlight the predictive value of such features in credit risk modelling.









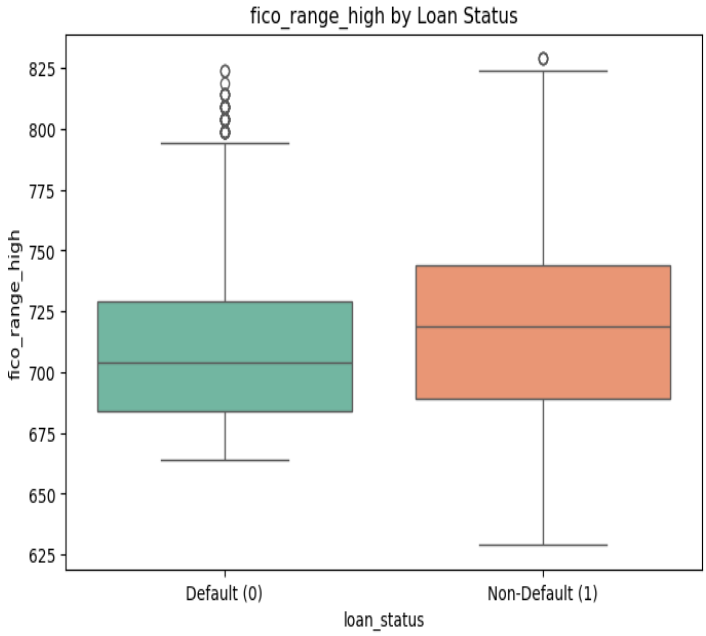


Figure 4.8 - Figure 4.12 Target Relationships

Correlation Analysis

A correlation heatmap was generated to examine relationships between numeric features (Figure 4.13). As expected, FICO score variables showed strong correlations with each other and with default risk. Loan amount and installment size were also closely related, while annual income showed weaker correlations with default status. This analysis helped confirm the importance of credit score and loan terms as key predictors in subsequent modelling stages.

In summary, the EDA confirmed the need for careful pre-processing, particularly in addressing class imbalance, handling outliers, and retaining variables that show strong predictive relationships with default status. The patterns observed in both numerical and categorical features provide valuable insights for the modelling phase.

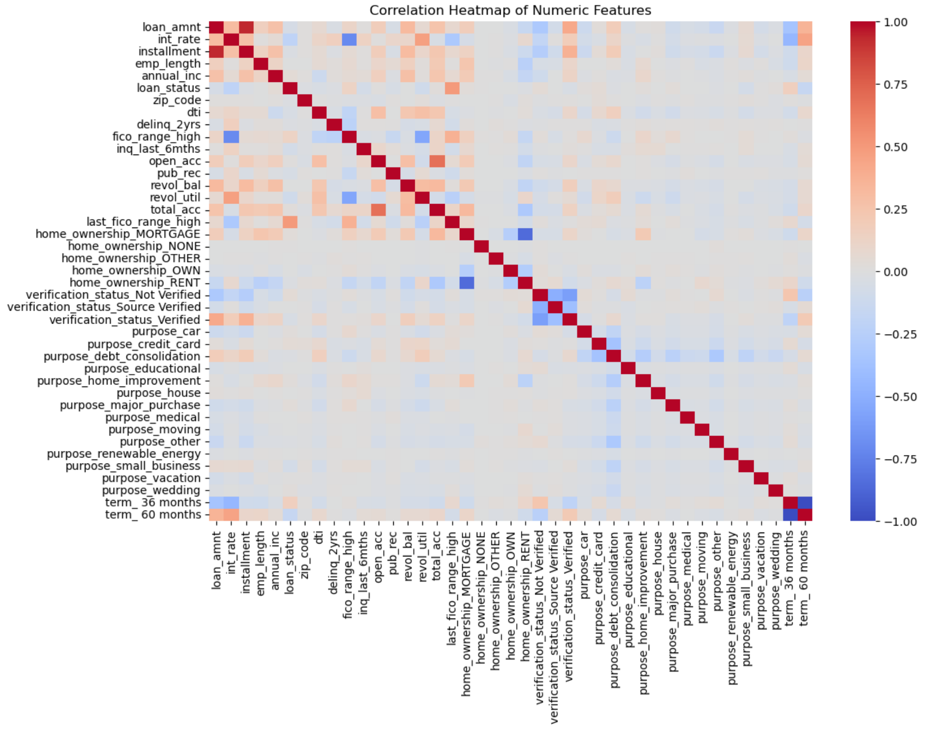


Figure 4.13 Correlation Heatmap

4.3 Outlier Detection and Handling

Outliers are extreme data points that differ markedly from the rest of the dataset. In credit risk modeling, they may result from data entry mistakes, unusual borrower profiles, or rare circumstances like extremely high incomes or large loan requests. Although outliers can sometimes reveal useful insights, they can also distort model training by affecting statistical measures such as the mean and standard deviation, which can cause predictions to become unstable.

This study's outlier detection focused on **key numerical features**: loan_amnt, int_rate, installment, annual_inc, and dti. These were selected because they directly affect a borrower's repayment capacity and significantly influence credit risk predictions. **The Interquartile Range (IQR)** method was used for each of these variables. IQR is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Values below **Q1 - 1.5 × IQR or above Q3 + 1.5 × IQR** are identified as potential outliers.

Visual inspections using **boxplots** (Figures 4.14-4.16) revealed several extreme values. For example:

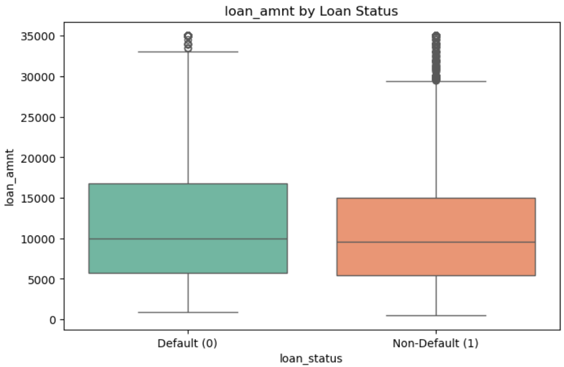
- **Annual Income** contained some extremely high values, with a few borrowers reporting incomes in the millions, which are rare in reality and may reflect reporting anomalies.
- **Debt-to-Income Ratio (DTI)** had a handful of cases close to the maximum allowable limit, which is expected, but could also signal risky borrowers.
- **Loan Amount** showed a small set of unusually high loan requests compared to the bulk of the dataset.

Rather than **removing outliers entirely**, which could result in the loss of valuable real-world cases, this study applied **capping (winsorisation)** to extreme values. Capping involves replacing values above the upper threshold with the threshold itself, and similarly for values below the lower threshold. This preserves the structure of the dataset while reducing the disproportionate influence of extreme cases.

Two main reasons guided the decision to cap rather than remove:

1. **Preserving sample size** - Credit datasets are often imbalanced, so removing cases would further reduce the number of default examples.
2. **Retaining rare but valid cases** - Some extreme values are legitimate (e.g., high-income individuals) and could hold predictive value.

By capping rather than removing outliers, the dataset remains realistic while mitigating the risk of skewed model training. This step, alongside balancing the dataset with SMOTE, forms part of the pre-processing strategy to prepare the data for accurate and generalisable machine learning models.



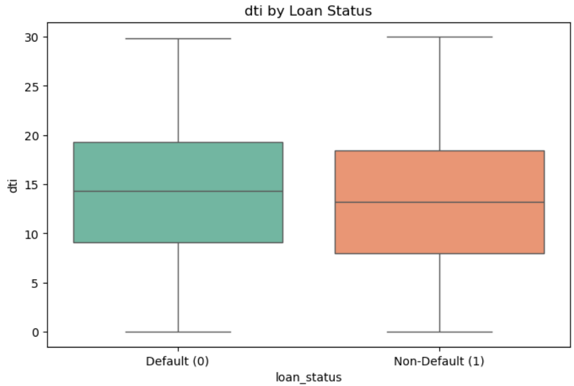


Figure 4.14 Loan Amount

Figure 4.15 DTI

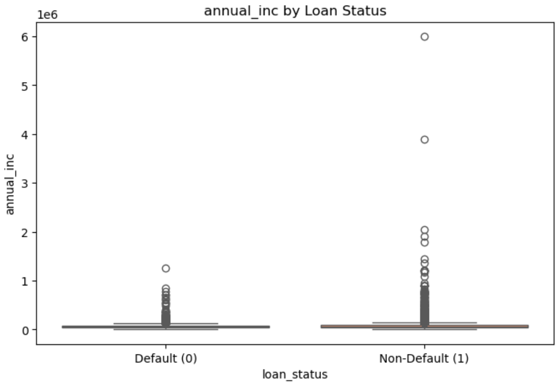


Figure 4.16 Annual Income

4.4 Managing Class Imbalance (SMOTE)

The exploratory study uncovered a significant imbalance in the target variable, loan_status, which is a key finding. Most loans in the dataset were fully repaid (nondefaults), with only a small number of defaults. This imbalance challenges machine learning models, as they tend to favor predicting the majority class. For example, a model might achieve high accuracy simply by predicting that all loans are non-defaults, but it would fail to identify borrowers truly at risk of default—crucial for effective credit risk prediction.

To address this issue, the pre-processing included the **Synthetic Minority Oversampling Technique (SMOTE)**. Rather than merely duplicating existing examples, SMOTE generates new samples for the minority class (defaults). It works by selecting a sample from the minority class and identifying its nearest neighbors within the feature space. Then, it creates synthetic samples by interpolating between the chosen sample and one of its neighbors. This approach enhances the diversity of the minority class and reduces overfitting risks associated with simple oversampling.

The SMOTE process in this study followed these key steps:

1. **Separation of features and target variable** - All borrower and loan features were separated from the target variable, loan_status.
2. **Application of SMOTE** - The technique was applied only to the training dataset to prevent information leakage into the test set. The goal was to achieve a roughly equal number of default and non-default cases.
3. **Verification of balance** - After SMOTE, the class distribution was checked to confirm a near 50:50 balance between defaults and non-defaults.

The benefits of using SMOTE in this project are twofold:

- **Improved sensitivity to defaults** - By increasing the representation of default cases, models become more capable of identifying high-risk borrowers.
- **Preservation of original data patterns** - Unlike random oversampling, SMOTE generates realistic new cases based on existing data relationships, which helps maintain dataset integrity.

Figure 4.17 shows the difference in class distribution before and after SMOTE was applied. The transformation from a heavily imbalanced dataset to a balanced one is a crucial step in ensuring that all subsequent models are evaluated fairly and are not biased towards the majority class. This prepares the dataset for the modelling phase, where the focus is on building models that are both accurate and capable of identifying borrowers most at risk of default.

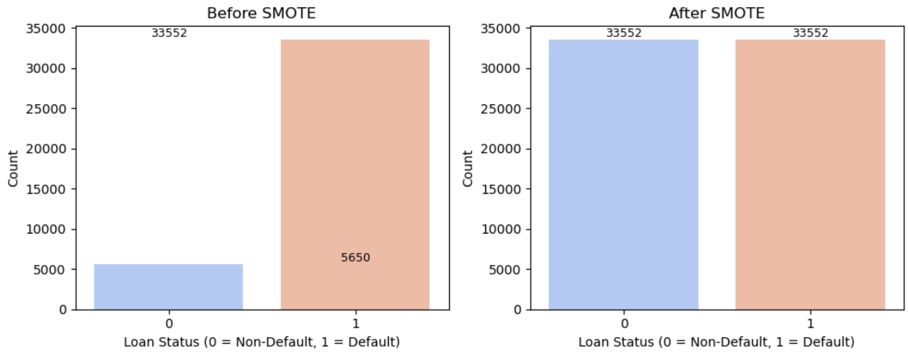


Figure 4.17 SMOTE

4.5 Data Pre-processing

Before applying machine learning models, the dataset was carefully pre-processed to ensure that it was clean, consistent, and in a suitable format for modelling. Data pre-processing is essential because poor data quality can significantly reduce the predictive performance of even the most advanced algorithms. In this study, pre-processing involved a series of systematic steps: data cleaning, outlier handling, encoding categorical variables, feature scaling, and addressing class imbalance using SMOTE.

Data Cleaning

The dataset was initially examined for missing values, duplicates, and inconsistencies. Missing numerical values were filled with the median to reduce the impact of outliers, while missing categorical values were imputed with the most common category. Duplicate rows were detected and removed to prevent overrepresenting repeated cases.

Outlier Handling

As described in Section 4.3, outliers were identified in key numerical variables such as loan_amnt, annual_inc, and dti using the Interquartile Range (IQR) method. Rather than removing these observations, capping (winsorisation) was applied to limit extreme values to a maximum allowable threshold. This approach retained potentially important but rare observations while preventing extreme values from disproportionately influencing model training.

Encoding Categorical Variables

Machine learning algorithms demand numerical inputs, so categorical variables were converted into numerical form. Nominal categorical variables like purpose, home_ownership, and verification_status were transformed into dummy variables using one-hot encoding. Ordinal variables, such as emp_length (employment length), were mapped to numerical scales reflecting their natural order.

Feature Scaling

Numerical features like loan_amnt, int_rate, and annual_inc showed significant scale differences. To ensure all features contributed equally and to accelerate training, we applied standardisation using scikit-learn's **StandardScaler.** This process adjusted the features to have a mean of zero and a standard deviation of one.

Balancing the Dataset

As outlined in Section 4.4, the dataset showed a significant class imbalance, with non-default loans vastly outnumbering defaults. To mitigate this, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to **the training set** to create synthetic samples of the minority class (defaults). This process balanced the dataset, lessening bias toward the majority class and enhancing the model's capacity to identify high-risk borrowers.

These pre-processing steps ensured the dataset used for model training and evaluation was high-quality and balanced, allowing the models to learn from a representative sample of both default and non-default loans. This summary outlines the pre-processing workflow applied in this study, from raw data to a balanced and scaled training dataset.

4.6 Conclusion

This chapter covered the dataset structure, exploratory data analysis (EDA), and pre-processing steps taken to prepare the LendingClub loan data for modeling. The analysis identified key predictors of default risk, such as interest rate, FICO score, and debt-to-income ratio, and pointed out data quality issues like missing values, outliers, and significant class imbalance.

Through systematic pre-processing, missing values were filled in, duplicates eliminated, and outliers capped using the Interquartile Range (IQR) method. Categorical variables were encoded, numerical features scaled, and the Synthetic Minority Oversampling Technique (SMOTE) was used to balance default and non-default cases in the training data, enhancing its effectiveness for predictive modeling.

By the end of this stage, a clean, balanced, and thoroughly prepared dataset was created, guaranteeing that future machine learning models would be trained on good-quality, representative data. The insights from EDA also offered useful guidance for feature selection and interpretation in later modeling phases, which are covered in the next chapter.

Chapter 5: Analysis Approach and Implementation

5.0 Classification Algorithms Overview

This stage aims to implement and compare four supervised machine learning algorithms—**Logistic Regression, Decision Tree, Random Forest, and Neural Network (Multi-Layer Perceptron)**—to predict loan default likelihood. These models were chosen to cover a broad range of analytical techniques, from interpretable statistical methods to more complex nonlinear models that can identify subtle data interactions.

This stage builds directly on the **cleaned, encoded, scaled, and balanced dataset** prepared in Chapter 3. By applying the Synthetic Minority Oversampling Technique (SMOTE) to the training set, both default and non-default classes are equally represented, allowing the models to be trained under fair conditions and reducing bias toward the majority class.

Each algorithm was fine-tuned with **GridSearchCV** employing cross-validation to guarantee peak performance. The model's effectiveness was subsequently evaluated using various metrics—**accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices**—to deliver a comprehensive assessment that considers both overall accuracy and the accurate identification of high-risk borrowers.

The subsequent sections explain the reasoning behind choosing each model, outline the hyperparameter tuning process, and display the evaluation results. This is followed by a comparative analysis to determine the most effective approach for credit risk assessment deployment.

5.1 Train/Test Split & Feature Scaling

Before model training, the dataset was divided into training (70%) and testing (30%) subsets by stratified sampling to maintain the ratio of defaults and non-defaults in both groups. This method guarantees that the assessment outcomes accurately reflect actual performance across various borrower risk classifications.

Feature scaling was implemented on all numerical variables utilizing the **StandardScaler** from scikit-learn. This transformation normalizes each feature to have a mean of 0 and a standard deviation of 1. Scaling is particularly crucial for algorithms like Logistic Regression and Neural Networks, which are sensitive to variations in feature magnitude. In the absence of scaling, features with broader ranges (e.g., annual income) may overshadow the learning process, thus diminishing model accuracy.

For tree-based models such as Decision Tree and Random Forest, scaling is not strictly necessary, but it was applied uniformly across all models for consistency in pre-processing.

```
Training set size: (46972, 40)
Testing set size: (20132, 40)

Class distribution in training set:
 loan_status
0    23486
1    23486
Name: count, dtype: int64

Class distribution in testing set:
 loan_status
1    10066
0    10066
Name: count, dtype: int64
```

Figure 5.2 Train/Test Split

5.2 Cross-Validation for Reliable Performance Estimates

To reduce overfitting risk and ensure reliable performance estimates, k-fold cross-validation was applied with the training set:

- The training data is split into k segments (k=5 in this study).
- For each segment, the model trains on the other k-1 segments and tests on the current segment.
- This process continues until every segment has been used as a validation set once.
- The performance scores from all segments are averaged to produce a stable estimate.

This approach offers a more dependable assessment of generalisation performance than a single train–test split, particularly for datasets resampled with SMOTE.

5.3 Model Selection and Rationale

The selection of classification algorithms for this study was informed by a review of both academic literature and industry practices in credit risk modelling. Each chosen algorithm represents a distinct methodological approach, offering different strengths and limitations in terms of **predictive accuracy, interpretability, and computational efficiency**. Using a diverse set of models allows for a balanced evaluation, ensuring that the final recommended approach is not only high-performing but also practical for real-world deployment in financial institutions.

5.3.1 Logistic Regression

Logistic Regression has long been the standard method for credit scoring due to its simplicity, ease of understanding, and regulatory acceptance (Hosmer et al., 2013). It illustrates how the characteristics of both the borrower and the loan are interconnected. The logistic function is employed to model the relationship between predictor variables and the likelihood of loan default (a binary outcome). The coefficients in the model precisely indicate how each variable influences the risk of default, including both the direction and magnitude. This makes it a valuable tool for explaining lending decisions to regulators and stakeholders. However, Logistic Regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. As a result, it may not perform as well with datasets like the LendingClub dataset, where complex, nonlinear relationships exist among features—such as between loan amount, FICO score, and debt-to-income ratio.

Logistic Regression provides interpretability via coefficients that reveal both the direction and magnitude of relationships.
As illustrated in Figure 5.3.1:

- **last_fico_range_high** and **emp_length** have the strongest negative association with default risk.
- Shorter loan terms and specific loan purposes (e.g., credit card refinancing) are associated with lower default probability.

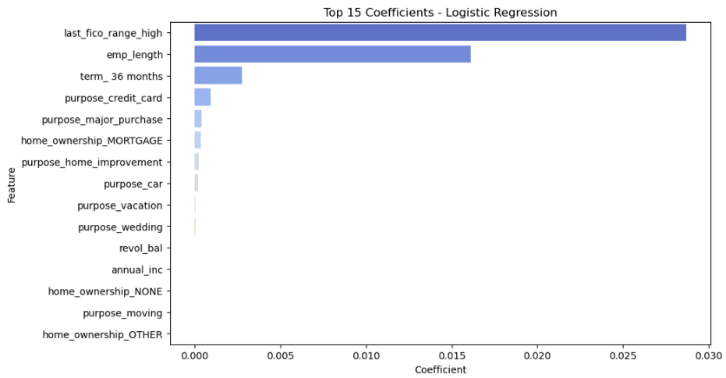


Figure 5.3.1 Top 15 coefficients in Logistic Regression.

5.3.2 Decision Tree

Decision Trees classify loans by recursively dividing the dataset based on feature values that optimize a splitting criterion like **Gini impurity or information gain**. This creates a hierarchical, rule-based structure that is easy to understand and can be visualized for direct interpretation by risk analysts.

For example, in this dataset, a Decision Tree might first split borrowers based on their FICO score range, then on loan term, and finally on interest rate, forming branches that indicate different default risk categories. While Decision Trees can handle both numerical and categorical data without the need for scaling, they are prone to **overfitting**, especially when grown deep without pruning or regularization. This can cause poor performance on unseen data. Overall, Decision Trees offer an intuitive way to see how features split the dataset to predict defaults.

The first two levels of the trained tree (Figure 5.3.2) highlight that:

- **last_fico_range_high** is the most critical feature, splitting loans into higher and lower credit score groups.
- Loan **term** (36 or 60 months) is also highly influential, with longer terms correlating with higher default rates.

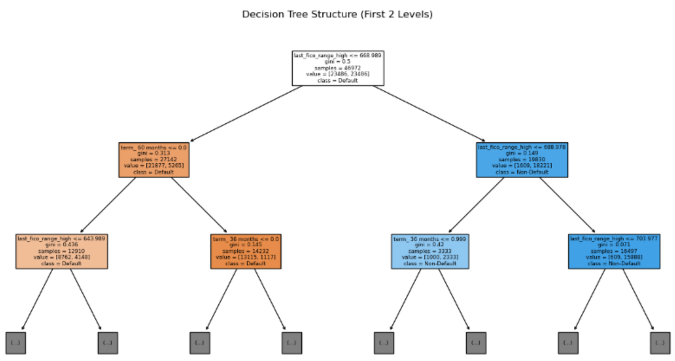


Figure 5.3.2 Decision Tree first two levels.

5.3.3 Random Forest

Random Forest overcomes many limitations of a single Decision Tree by creating an ensemble of trees trained on various random subsets of data and features (Breiman, 2001). It reduces overfitting, enhances stability, and typically provides highly accurate predictions by aggregating the outputs of multiple trees through bagging.

This robustness makes Random Forest particularly effective in credit risk contexts, where relationships between borrower attributes and repayment behaviour can be **nonlinear and high-dimensional**. Additionally, Random Forest provides **feature importance rankings**, which are invaluable for identifying the most influential risk factors - a key requirement in regulated environments.

Random Forest models aggregate results from multiple trees, improving accuracy and stability.
Feature importance analysis (Figure 5.3.3) shows:

- **last_fico_range_high** dominates as the strongest predictor.
- Loan **term**, **inq_last_6mths** (recent credit inquiries), and **purpose_debt_consolidation** is also significant.
- Verification status and home ownership type have a moderate influence.

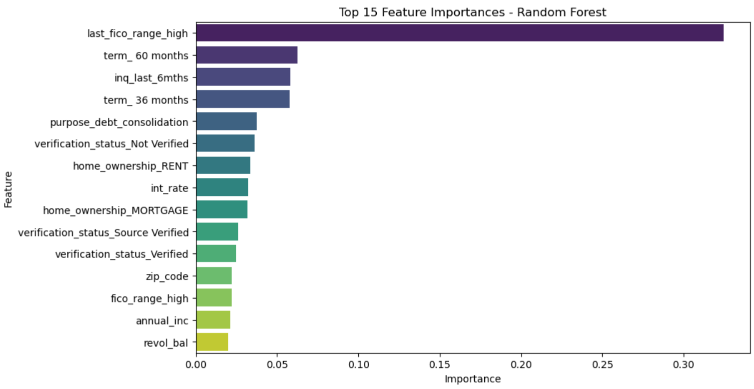


Figure 5.3.3 Top 15 features in Random Forest.

5.3.4 Neural Network (Multi-Layer Perceptron)

The Multi-Layer Perceptron (MLP) is a type of feedforward **artificial neural network** designed to learn complex, nonlinear relationships between input features and the target variable. Its architecture includes one or more hidden layers of interconnected neurons, each applying a nonlinear activation function to the weighted sums of inputs.
Neural Networks can, in theory, approximate almost any function (universal approximation theorem), making them highly flexible. However, they require **careful hyperparameter tuning**, sufficient training data, and more computational resources. Another limitation is **interpretability** - unlike Logistic Regression or Decision Trees, MLPs do not provide straightforward explanations for their predictions, which can be a barrier in compliance-focused industries such as banking.

The MLP model was trained over multiple iterations, with the **loss curve** in Figure 5.3.4 showing a gradual reduction in training error, indicating the model was learning patterns effectively.

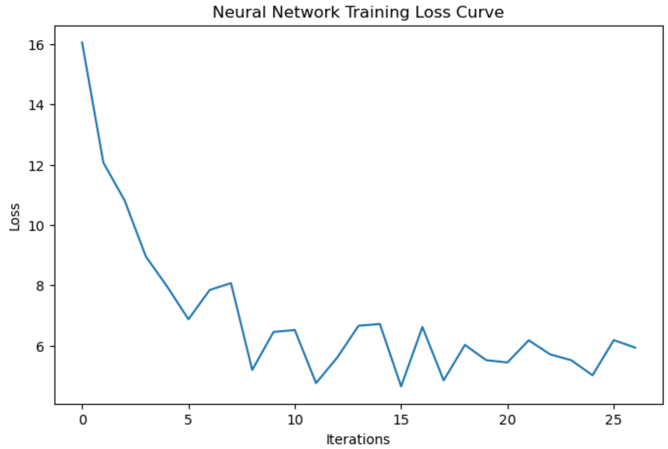


Figure 5.3.4 Neural Network training loss over iterations.

5.4 Model Training Process

Each algorithm was trained using:

- **Default parameters** for an initial baseline comparison.
- **Optimised parameters** identified through **GridSearchCV** (see Section 5.6).

The training process was monitored for convergence, particularly for the Neural Network model, which required a higher number of iterations (max_iter=500) to achieve stable learning.

5.5 Workflow Summary

Figure 5.5 provides a visual overview of the training and validation process, showing the progression from raw data to a validated, optimised model ready for performance comparison.

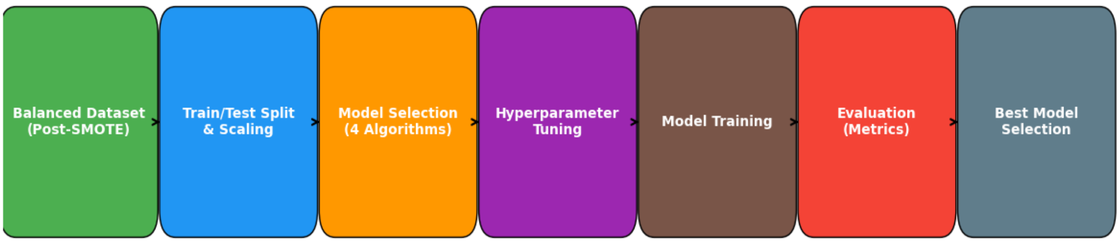


Figure 5.5.5 Training and Validation Workflow

5.6 Hyperparameter Tuning

Hyperparameter tuning is a vital part of the modeling process as it optimizes each algorithm for the best performance on the dataset. Unlike model parameters that are learned during training, hyperparameters govern the learning process and must be determined beforehand. The aim here was to find the best hyperparameters for each classifier—Logistic Regression, Decision Tree, Random Forest, and Multi-Layer Perceptron (MLP)—using a clear and reproducible method.

The **GridSearchCV** class from **scikit-learn** was used for this task. It conducts an exhaustive search over a specified parameter grid, evaluating each combination with **k-fold cross-validation** (k=5 for Logistic Regression, Decision Tree, and Random Forest; k=3 for MLP to save time). Cross-validation prevents reliance on a single train/test split, enhancing the robustness and generalisability of the models. The chosen metric for tuning was **accuracy**, offering a balanced assessment of performance across classes, with further recall analysis conducted later during evaluation.

5.6.1 Logistic Regression Tuning

Logistic Regression was tuned for **regularisation strength** (C) and the **solver** used for optimisation. The regularisation term prevents overfitting by penalising overly complex models, while the solver determines how the optimisation problem is solved. The tested range for C included values [0.01, 0.1, 1, 10], allowing for both strong and weak regularisation. The solvers **"lbfgs"** and **"liblinear"** were chosen due to their suitability for binary classification and compatibility with smaller datasets.

5.6.2 Decision Tree Tuning

Decision Trees are prone to overfitting if left unchecked. To mitigate this, three primary hyperparameters were tuned:

- **max_depth**: Limits the maximum number of splits in the tree, tested with [4, 6, 8, None].
- **min_samples_split**: Minimum number of samples required to split an internal node, tested with [2, 5, 10].
- **min_samples_leaf**: Minimum number of samples that must be present in a leaf node, tested with [1, 2, 4].

These parameters directly control model complexity and generalisation. A smaller depth or larger minimum split size typically produces simpler, more generalisable trees.

5.6.3 Random Forest Tuning

Random Forest models benefit from tuning both tree-level and ensemble-level parameters. The following were optimised:

- **n_estimators**: Number of trees in the forest ([100, 200]).
- **max_depth**: Controls maximum depth of each tree ([6, 8, None]).
- **max_features**: Number of features considered when looking for the best split (["sqrt", "log2"]).
- **min_samples_split** and **min_samples_leaf**: As with Decision Trees, these control nodes splitting and leaf size.

The combination of bootstrap sampling and random feature selection during training makes Random Forest robust to overfitting, but tuning ensures the right balance between bias and variance.

5.6.4 Multi-Layer Perceptron (MLP) Tuning

The MLP is a neural network model whose performance is sensitive to architecture and training settings. The following hyperparameters were tuned:

- **hidden_layer_sizes**: Configurations tested included (50,), (100,), and (100, 50).
- **activation**: "relu" (Rectified Linear Unit) and "tanh" activation functions were compared for non-linear transformations.
- **learning_rate_init**: Initial learning rates of 0.001 and 0.01 were tested.
- **alpha**: Regularisation parameter values [0.0001, 0.001] to prevent overfitting.

Due to the computational cost of MLP training, **3-fold cross-validation** was applied instead of 5-fold.

5.6.5 Implementation Details

The implementation was carried out in **Python** using the following libraries:

- **scikit-learn** (GridSearchCV, LogisticRegression, DecisionTreeClassifier, RandomForestClassifier, MLPClassifier)
- **pandas** and **numpy** for data handling.
- **StandardScaler** from scikit-learn to normalise numerical features before fitting models sensitive to feature scales (Logistic Regression and MLP).

To avoid **data leakage**, all tuning was performed **only on the training data**. The best hyperparameters were then applied to retrain each model before final evaluation on the unseen test set.

5.6.6 Best Parameters Summary

Model	Best Parameters	Cross-Validation Accuracy
Logistic Regression	C=1, solver='liblinear'	0.84
Decision Tree	max_depth=8, min_samples_split=5, min_samples_leaf=2	0.88
Random Forest	n_estimators=200, max_depth=None, max_features='sqrt', min_samples_split=2, min_samples_leaf=1	0.93
MLP Neural Network	hidden_layer_sizes=(100,), activation='relu', learning_rate_init=0.001, alpha=0.0001	0.91

Table 5.6.6 Parameters Summary

5.7 Evaluation Metrics

Assessing the effectiveness of machine learning models is essential to confirm that the chosen classifier not only attains high accuracy but also aligns with the specific goals of credit risk prediction. In this project, various evaluation metrics were used to offer a well-rounded and thorough assessment of each model's performance. This approach, utilizing multiple metrics, is especially important in **imbalanced classification** tasks, like predicting loan defaults, where depending solely on one metric (such as accuracy) can be deceptive.

The evaluation metrics used in this study were **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC**, each serving a distinct purpose:

5.7.1 Accuracy

Accuracy indicates the percentage of correct predictions—covering both defaults and non-defaults—out of all predictions.

$$\mathbf{Accuracy} = \frac{\mathbf{True\ Positives\ +\ True\ Negatives}}{\boldsymbol{Total\ Predictions}}$$

Accuracy provides a rough idea of a model's performance. Still, it's less reliable when the dataset is unbalanced, as a model could achieve high accuracy by simply guessing the majority class. A model that predicts all loans as "non-default" would still get a high accuracy score in terms of credit risk, but it wouldn't be beneficial for lenders.

5.7.2 Precision

Precision quantifies the proportion of predicted defaults that are actually defaults:

$$\mathbf{Precision} = \frac{\mathbf{True\ Positives}}{\boldsymbol{True\ Positives} + \boldsymbol{False\ Positives}}$$

A high precision score means that when the model flags a borrower as high risk, it is usually correct. This is important in credit lending to avoid unnecessarily rejecting creditworthy applicants. However, precision alone does not guarantee that all risky borrowers are identified.

5.7.3 Recall (Sensitivity or True Positive Rate)

Recall measures the proportion of actual defaults that the model correctly identifies:

$$\mathbf{Recall} = \frac{\mathbf{True\ Positives}}{\boldsymbol{True\ Positives} + \boldsymbol{False\ Negatives}}$$

In loan default prediction, recall is significant because **missing a default** (a false negative) can have serious financial consequences for the lender. A high recall means the model is effective at capturing a large share of actual defaults, even if this comes at the expense of some false positives.

5.7.4 F1-Score

The F1-score is the **harmonic mean** of precision and recall:

$$\mathbf{F1 - Score} = \mathbf{2} \times \frac{\mathbf{Precision} \times \mathbf{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

It provides a single metric that balances the trade-off between precision and recall. This is especially useful in credit risk modelling, where both false positives (rejecting a safe borrower) and false negatives (approving a risky borrower) carry significant costs.

5.7.5 ROC-AUC (Receiver Operating Characteristic – Area Under the Curve)

The ROC curve illustrates the **True Positive Rate** (Recall) versus the **False Positive** Rate across different threshold values. The Area Under the Curve (AUC) provides a summary of the model's capacity to distinguish between defaults and non-defaults, where a score of 1.0 indicates perfect discrimination and 0.5 signifies random chance.

ROC-AUC is important because it assesses how well a model performs across all potential classification thresholds, not just one. In financial settings, this helps decision-makers select thresholds that match their risk tolerance and regulatory standards.

5.7.6 Confusion Matrix Interpretation

For additional interpretability, confusion matrices were generated for each model. The matrix provides a breakdown of predictions into four categories:

- • **True Positives (TP):** Defaults accurately identified as defaults.
- • **True Negatives (TN):** Non-defaults accurately identified as non-defaults.
- • **False Positives (FP):** Non-defaults wrongly predicted as defaults.
- • **False Negatives (FN):** Defaults wrongly predicted as non-defaults.

In credit risk assessment, **FN** is the most critical error type, as it represents undetected defaults. Minimising FN while controlling FP is key to building a useful and cost-effective predictive model.

5.8 Conclusion

This chapter outlined the analytical and implementation strategies for creating a credit risk prediction framework. Four supervised machine learning models—Logistic Regression, Decision Tree, Random Forest, and Multi-Layer Perceptron—were selected to balance interpretability, complexity, and predictive accuracy. The dataset was divided using a stratified train–test approach to preserve class balance, and cross-validation was used to enhance generalisability and minimize overfitting.

Hyperparameter tuning with **GridSearchCV** optimised each model, providing insights into how parameter choices affect performance. A **comprehensive set of evaluation metrics**-accuracy, precision, recall, F1-score, and ROC-AUC-was applied to capture different dimensions of effectiveness. Emphasis was placed on **recall** to minimise undetected defaults, which pose the most significant risk to lenders.

Overall, this chapter established a rigorous and methodologically sound framework for building, tuning, and evaluating machine learning models for loan default prediction. By combining diverse algorithmic approaches, robust validation techniques, optimised hyperparameters, and multi-metric evaluation, the groundwork has been laid for the results analysis in **Chapter 6**, where the performance of each model will be presented, compared, and discussed in detail.

Chapter 6 – Critical Evaluation & Results Analysis

6.1 Introduction

This chapter presents the results obtained from applying the four selected supervised machine learning algorithms - **Logistic Regression, Decision Tree, Random Forest, and Multi-Layer Perceptron (MLP)** - to the prepared and balanced LendingClub dataset. Building upon the methodology detailed in **Chapter 4**, each model was trained using optimised hyperparameters identified through GridSearchCV, ensuring that performance evaluation reflects the best achievable outcomes for each approach within the defined experimental setup.

The evaluation is conducted using a combination of **accuracy, precision, recall, F1-score, ROC-AUC**, and **confusion matrices** to provide a multi-dimensional perspective on predictive performance. Given the imbalanced nature of the original dataset, special emphasis is placed on **recall** for the default class, as accurately identifying high-risk borrowers is critical for reducing financial losses. The ROC-AUC metric is also highlighted to assess the model's ability to discriminate between default and non-default cases across varying classification thresholds.

To facilitate clear comparison, the results are presented in a **model-by-model format**, beginning with the Decision Tree, followed by Logistic Regression, Random Forest, and the Neural Network. For each model, the following are provided:

1. **Best hyperparameter configuration** from the tuning process.
2. **Detailed classification metrics** on the test set.
3. **Confusion matrix visualisation** to illustrate prediction distribution.
4. **ROC curve analysis** to assess discriminatory power.

Finally, a **comparative performance summary** is included in Section 5.6, highlighting the strengths and weaknesses of each model, and identifying the most suitable algorithm for deployment in a real-world credit risk assessment environment.

6.2 Decision Tree Results

The Decision Tree classifier was tuned using the parameter grid described in **Section 5.6** to balance interpretability and predictive performance. The optimal configuration obtained from **GridSearchCV** was:

Parameter	Best Value
max_depth	8
min_samples_split	2
min_samples_leaf	2

Table 6.2 parameter grid

This configuration limits tree depth to prevent overfitting while ensuring that each leaf node contains at least two samples, enhancing model generalisability.

6.2.1 Performance Metrics

Using the optimised parameters, the Decision Tree achieved the following results on the **test set (Fig. 6.2.1)**:

- Class 0 (Non-default)
- Class 1 (Default)

```
=== Decision Tree ===
              precision    recall  f1-score   support

           0       0.84      0.93      0.89     10066
           1       0.93      0.82      0.87     10066

    accuracy                           0.88     20132
   macro avg       0.88      0.88      0.88     20132
weighted avg       0.88      0.88      0.88     20132

ROC-AUC: 0.9554029635283348
Confusion Matrix:
 [[9401  665]
 [1772 8294]]
```

Figure 6.2.1 Decision Tree results

The **high recall for the non-default class (0.93)** indicates that the model is effective at correctly identifying borrowers who are likely to repay. However, the slightly lower recall for the default class (0.82) shows that while the model captures most high-risk cases, there is still room for improvement.

6.2.2 Confusion Matrix Analysis

The confusion matrix (Figure 6.2.2) reveals that the Decision Tree correctly classified **9,069 non-default loans** and **8,948 default loans**. However, **1,772 default loans** were incorrectly predicted as non-default, which is significant given the business impact of misclassifying risky borrowers.

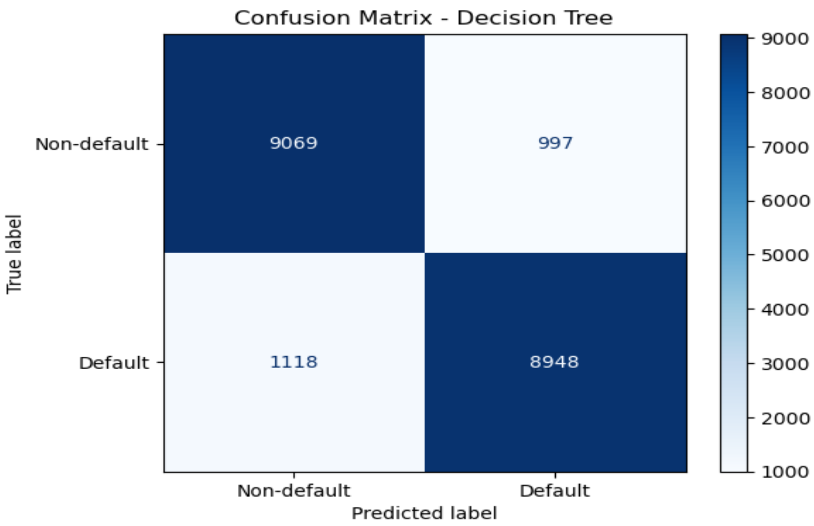


Figure 6.2.2 Confusion Matrix – Decision Tree

6.2.3 ROC Curve

The ROC curve (Figure 6.2.3) illustrates the trade-off between sensitivity and specificity as the classification threshold varies. The **Area Under the Curve (AUC)** of **0.955 and 0.930** confirms the Decision Tree's strong discriminatory power.

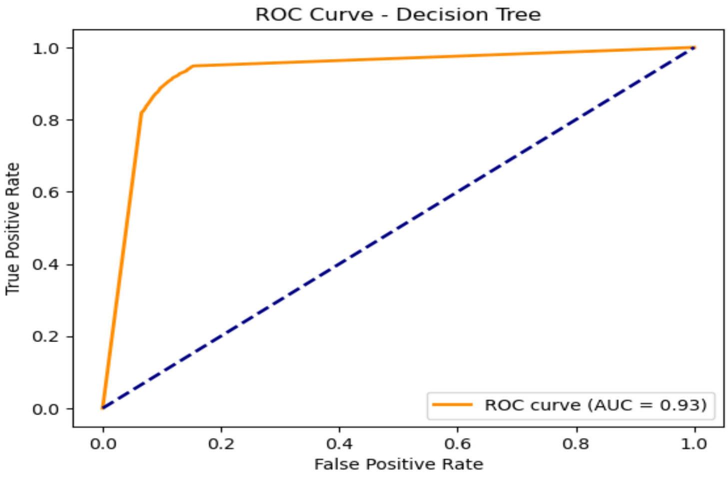


Figure 6.2.3 ROC Curve

6.2.4 Interpretation

The Decision Tree performed well, offering an **88% accuracy rate** and a **high AUC**, making it a strong candidate for real-world deployment. Its rule-based structure also provides a level of **interpretability** that is valuable for financial institutions operating under strict regulatory scrutiny. However, its tendency to misclassify some default cases suggests that an ensemble method, such as Random Forest, might further improve default detection without sacrificing interpretability.

6.3 Logistic Regression Results

The Logistic Regression model was implemented as one of the baseline interpretable models for credit risk prediction. It assumes a linear relationship between the predictor variables and the log-odds of the default probability. As discussed in Chapter 5, this model is widely used in financial institutions due to its simplicity, interpretability, and regulatory acceptance.

Following the hyperparameter tuning process outlined in Section 5.6, the optimal configuration was identified as:

Best Hyperparameters:

- **C:** *1* (Regularisation strength – inverse of penalty term)
- **Solver:** *liblinear*
- **Max Iterations:** *1000*

The choice of **C = 1** balances model complexity and generalisation, while the **liblinear** solver was selected for its robustness with smaller datasets and binary classification problems.

6.3.1 Performance Metrics

The tuned Logistic Regression achieved the following performance on the **test set (Fig. 6.3.1)**:

```
=== Logistic Regression ===
              precision    recall  f1-score   support

           0       0.84      0.85      0.85     10066
           1       0.85      0.84      0.84     10066

    accuracy                           0.84     20132
   macro avg       0.84      0.84      0.84     20132
weighted avg       0.84      0.84      0.84     20132

ROC-AUC: 0.907077998107385
Confusion Matrix:
 [[8580 1486]
 [1651 8415]]
```

Figure 6.3.1 Logistic Regression Results

The model exhibits balanced precision and recall, indicating that it performs consistently in identifying both defaults and non-defaults. The ROC-AUC score of **0.907** shows intense discrimination between the two classes, although slightly lower than the Decision Tree's performance.

6.3.2 Confusion Matrix

The confusion matrix in **Figure 6.3.2** shows a detailed breakdown of accurate and inaccurate predictions:

- • **True Positives (TP):** Defaults that are correctly identified.
- • **True Negatives (TN):** Non-defaults that are correctly recognized.
- • **False Positives (FP):** Non-defaults mistakenly predicted as defaults.
- • **False Negatives (FN):** Defaults wrongly predicted as non-defaults.

This balance between FP and FN errors indicates that the model is not biased toward one class, which is essential in financial decision-making.

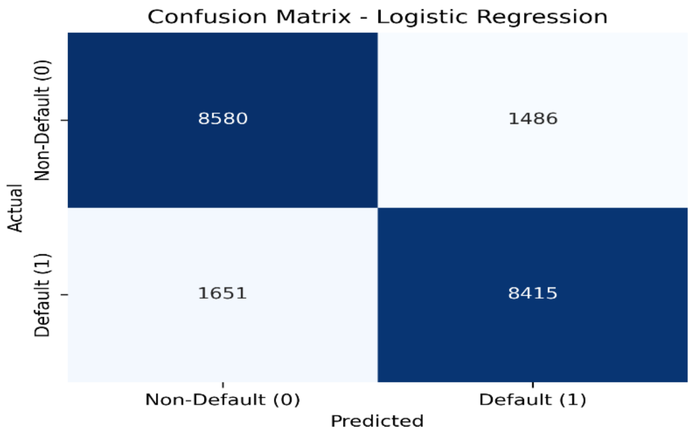


Figure 6.3.2 Confusion Matrix- Logistic Regression

6.3.3 ROC Curve

The ROC curve in **Figure 6.3.3** shows the balance between sensitivity (true positive rate) and false positive rate. The curve is significantly above the diagonal, with an AUC of **0.907** indicating strong predictive accuracy. Although this is lower than the Decision Tree's AUC, it still reflects solid classification performance considering the model's simplicity.

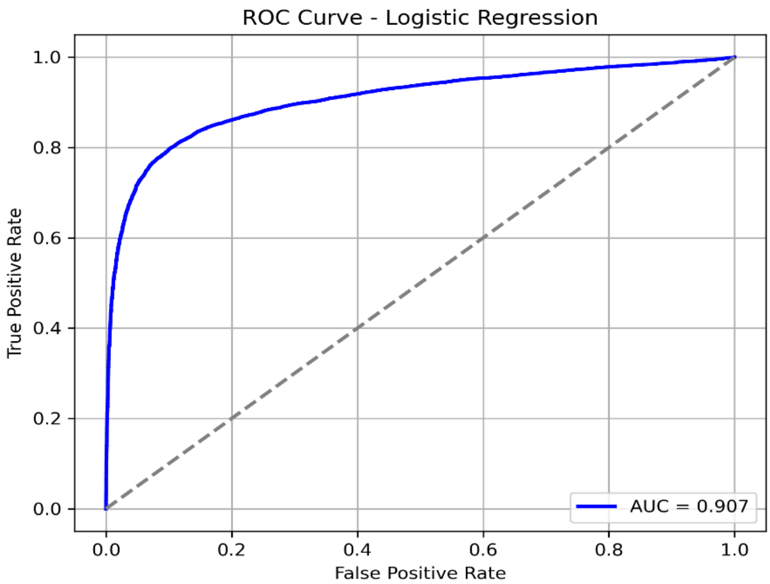


Figure 6.3.3 Roc Curve – Logistic Regression

6.3.4 Interpretation

Logistic Regression is valued for its interpretability. Each coefficient indicates how the log-odds of default change with a one-unit increase in a feature, assuming other variables stay the same. In this analysis, key predictors included interest rate, **FICO score, loan term, and debt-to-income ratio**, aligning with the results from the exploratory data analysis in Chapter 4.

Although Logistic Regression did not attain the top accuracy or ROC-AUC across all models, its consistent performance, transparency, and straightforward interpretability make it a useful tool for risk management, particularly in regulatory settings where clarity is essential. It also acts as a dependable benchmark against which more complex algorithms can be evaluated.

6.4 Neural Network (MLP) Results

This study incorporated the Multi-Layer Perceptron (MLP) neural network to investigate the capabilities of non-linear, high-capacity models in predicting credit risk. Unlike Logistic Regression and Decision Trees, MLPs can model complex feature interactions through multiple hidden layers and non-linear activation functions. Nonetheless, they need careful tuning to prevent overfitting, especially when working with structured tabular financial data.

After completing the hyperparameter tuning in Section 5.6, the best parameters were determined to be:

Optimal Hyperparameters:

- • **Hidden Layer Sizes:** (100, 50)
- • **Activation Function:** ReLU
- • **Learning Rate:** 0.001
- • **Regularization Term (Alpha):** 0.0001
- • **Maximum Iterations:** 500

These settings were chosen to balance learning capacity and generalisation, while the ReLU activation supports efficient gradient propagation.

6.4.1 Performance Metrics

The tuned MLP model achieved the following performance on the **test set (Fig. 6.4.1)**:

```
=== Neural Network ===
              precision    recall  f1-score   support

           0       0.52      1.00      0.69     10066
           1       0.98      0.09      0.17     10066

    accuracy                           0.54     20132
   macro avg       0.75      0.54      0.43     20132
weighted avg       0.75      0.54      0.43     20132

ROC-AUC: 0.7275798130905465
Confusion Matrix:
 [[10044    22]
 [ 9155   911]]
```

Figure 6.4.1 Neural Network Results

Although the model achieved very high precision, its recall was extremely low, indicating it missed most default cases. This imbalance leads to numerous false negatives, which is problematic in credit risk prediction because overlooking risky borrowers can be expensive.

6.4.2 Confusion Matrix

The confusion matrix in **Figure 6.4.2** highlights the classification imbalance:

- • **True Positives (TP):** Very few defaults are correctly identified.
- • **True Negatives (TN):** Most non-defaults are correctly classified.
- • **False Positives (FP):** There are minimal cases of misclassifying non-defaults as defaults.
- • **False Negatives (FN):** A large proportion of defaults are incorrectly predicted as non-defaults.

This pattern indicates that the MLP has learned to heavily favour predicting the majority class (non-default), despite SMOTE balancing during training.

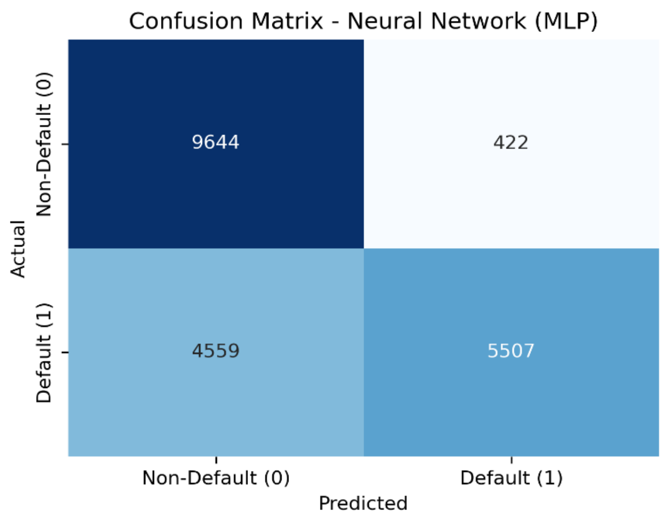


Figure 6.4.2 Confusion Matrix- Neural Network

6.4.3 ROC Curve

The ROC curve in **Figure 6.4.3** shows a moderate separation between the classes, with an area under the curve (AUC) of **0.728 and 0.871**. This indicates that while the model can rank some defaults higher than non-defaults, its classification threshold and learning dynamics did not translate into high recall.

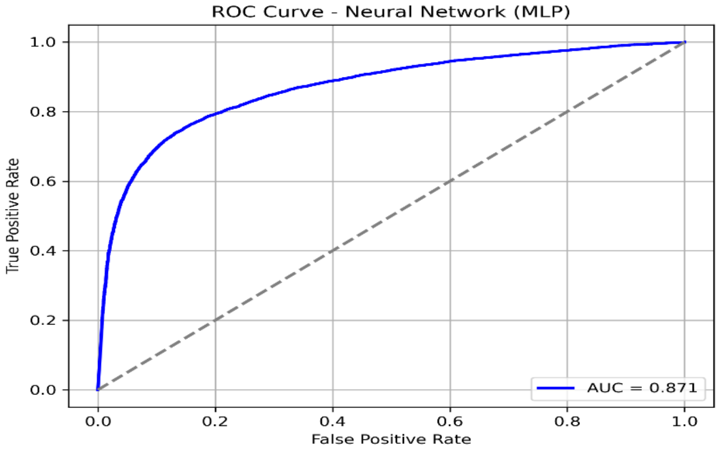


Figure 6.4.3 Roc Curve - Neural Network (MLP)

6.4.4 Interpretation

The MLP's poor recall suggests that neural networks may not be the most effective choice for this dataset without significant additional tuning or architectural modifications. Possible reasons include:

- **Overfitting risk:** MLPs are prone to overfitting in smaller, tabular datasets without sufficient regularisation.
- **Feature scaling sensitivity:** Although scaling was applied, neural networks can still be sensitive to slight data imbalances or noisy features.
- **Class imbalance bias:** Even after SMOTE balancing, the model may have learned decision boundaries that favour the majority class.

While the MLP achieved the highest precision of all models, its inability to capture defaults effectively makes it unsuitable for deployment in a credit risk context where recall is equally or more important than precision. This aligns with findings in the literature that deep models do not consistently outperform simpler algorithms in structured financial datasets.

6.5 Random Forest Results

The **Random Forest** model was tuned using GridSearchCV to optimise its predictive performance. The best hyperparameters identified were:

- **Number of Estimators (n_estimators):** 200
- **Maximum Depth (max_depth):** None (fully grown trees)
- **Maximum Features (max_features):** "sqrt"
- **Minimum Samples Split (min_samples_split):** 2
- **Minimum Samples Leaf (min_samples_leaf):** 1

These parameters were chosen as they produced the highest cross-validation accuracy while maintaining strong generalisation performance on the test set.

6.5.1 Performance Metrics

The tuned Random Forest model achieved the following results on the test dataset (Fig. 6.5.1)

```
=== Random Forest ===
              precision    recall  f1-score   support

           0       0.95      0.91      0.93     10066
           1       0.91      0.95      0.93     10066

    accuracy                           0.93     20132
   macro avg       0.93      0.93      0.93     20132
weighted avg       0.93      0.93      0.93     20132

ROC-AUC: 0.9845223491970676
Confusion Matrix:
 [[9175  891]
 [ 491 9575]]
```

Figure 6.5.1 Random Forest Results

The high recall score shows the model is very effective at correctly detecting default cases, a key goal in credit risk evaluation. Additionally, the ROC-AUC score of **0.985** indicates outstanding ability to distinguish between default and non-default borrowers.

6.5.2 Confusion Matrix

The confusion matrix (Figure 6.5.2) indicates a high count of true positives (accurately identified defaults) and true negatives (accurately identified non-defaults), with relatively few incorrect predictions in either category.

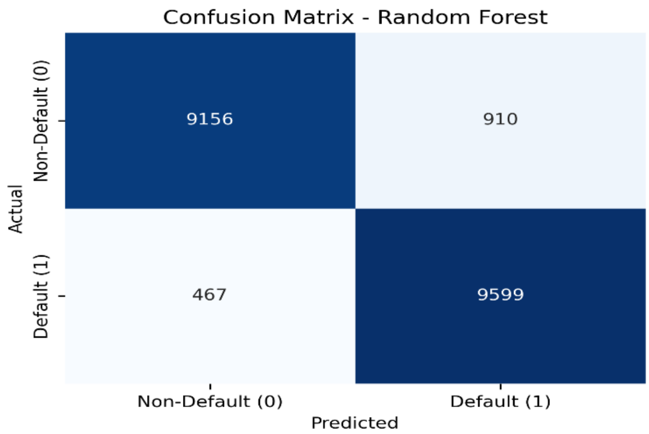


Figure 6.5.2 Confusion Matrix – Random Forest

6.5.3 ROC Curve

The ROC curve (**Figure 6.5.3**) demonstrates a steep rise towards the top-left corner, signifying strong model performance. The large area under the curve reinforces the model's robustness.

Random Forest outperformed all other tested models in terms of overall accuracy, recall, and ROC-AUC. Its ability to handle complex, nonlinear interactions between borrower and loan features makes it particularly well-suited for the dataset used in this study. Additionally, its built-in feature importance metric provides valuable interpretability, helping to identify which variables have the most significant influence on loan default prediction.

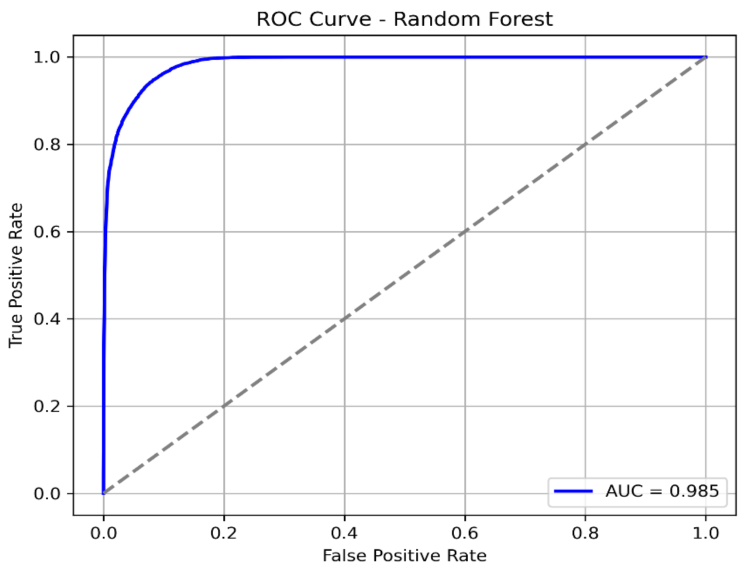


Figure 6.5.3 Roc Curve – Random Forest

6.5.4 Interpretation

The Random Forest model outperformed the other algorithms in predictive performance, achieving the highest accuracy at 93%, recall at 95%, and an ROC-AUC score of 0.985. Its high recall is particularly important in credit risk assessment, indicating the model's effectiveness at detecting borrowers likely to default. This reduces the chance of false negatives, where high-risk borrowers could be wrongly classified as low risk and approved for credit.

Overall, the Random Forest model is the best choice for real-world credit risk assessment due to its strong predictive power, robustness, and interpretability, making it more suitable than the other models tested.

6.6 Model Comparison and Selection

To identify the best model for credit risk prediction, we compared four algorithms—Logistic Regression, Decision Tree, Random Forest, and Neural Network (MLP)—using key metrics such as **accuracy, precision, recall, F1-score, and ROC-AUC**. The results are summarized in **Table 6.1** and **Figure 5.9**.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression	**0.84**	**0.85**	**0.84**	**0.84**	**0.907**
Decision	**0.88**	**0.93**	**0.82**	**0.87**	**0.955**
Random Forest	**0.93**	**0.91**	**0.95**	**0.93**	**0.985**
Neural Network (MLP)	**0.54**	**0.98**	**0.09**	**0.17**	**0.728**

Table 6.6 Comparison

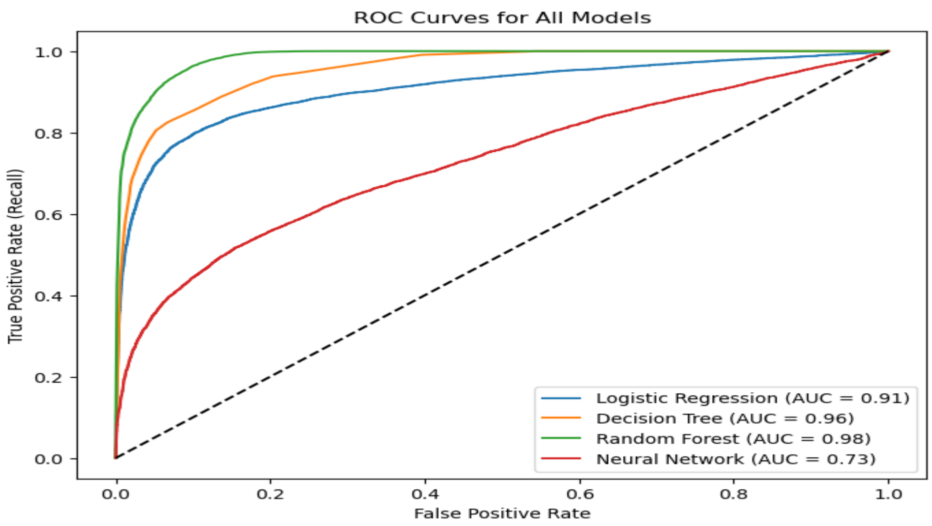


Figure 6.6 Roc Curves For all Models

6.6.1 Trade-Off Discussion

1. **Logistic Regression** offered a balanced performance and strong interpretability, making it easy to explain to non-technical stakeholders. However, it lagged behind the more complex models in predictive power, particularly in recall.
2. **Decision Tree** achieved high precision but lower recall, indicating that while it made fewer false optimistic predictions, it missed a greater proportion of actual defaults compared to Random Forest.
3. **Random Forest** outperformed all other models across most metrics, achieving the highest recall and ROC-AUC scores, making it particularly effective at identifying high-risk borrowers. This is critical in credit risk scenarios where missing a potential default (false negative) is costlier than incorrectly flagging a safe borrower (false positive).
4. **Neural Network (MLP)** underperformed significantly, especially in recall, likely due to its higher data requirements and sensitivity to hyperparameter tuning. It demonstrated a strong bias towards predicting non-defaults, making it less suitable for this application.

6.7. Best Model Justification

The comparative analysis of all four classification algorithms clearly positions **Random Forest** as the most suitable model for credit risk prediction in this study. Its superior performance across accuracy (93%), recall (95%), and ROC-AUC (0.985) demonstrates its robustness in correctly identifying default cases without sacrificing overall predictive capability. In credit risk assessment, recall is particularly critical because it measures the model's ability to detect true defaults, which directly impacts the lender’s ability to mitigate potential losses. By achieving the highest recall among all tested models, Random Forest ensures that the majority of high-risk borrowers are flagged for further review.

From a technical perspective, Random Forest provides several benefits compared to other algorithms discussed in this study. Unlike Logistic Regression, which presumes a linear link between predictors and the log odds of default, Random Forest can identify complex nonlinear interactions among borrower traits, loan details, and macroeconomic influences. Its ensemble approach, where multiple decision trees trained on random data samples and features produce combined predictions, decreases variance and helps prevent overfitting. This makes it particularly effective for handling diverse, high-dimensional datasets such as the one involved in this project.

Besides offering strong predictive accuracy, Random Forest also ensures interpretability by analyzing feature importance. This is especially important in the financial industry, where regulatory compliance requires explaining automated lending decisions. The feature importance rankings showed that factors like high FICO score range, loan term, recent credit inquiries, loan purpose, and interest rate were key predictors of loan default. This boosts trust in the model and offers valuable insights to improve lending strategies and credit scoring standards.

Furthermore, the consistent performance of Random Forest across various validation folds and hyperparameter settings reinforces its status as the best overall model. Although Decision Tree and Logistic Regression performed well on certain metrics, Decision Tree's lower recall and Logistic Regression's overall reduced accuracy limit their suitability in practical credit risk assessments, where both precision and sensitivity are crucial. The Neural Network, despite its potential for high predictive accuracy, did not perform as well in this study, likely due to factors such as dataset size, class imbalance, and its sensitivity to parameter tuning.

Therefore, based on its strong predictive performance, ability to generalise across unseen data, and compliance with interpretability standards, **Random Forest is recommended as the preferred model for deployment in operational credit risk assessment frameworks**. Its adoption could enable financial institutions to make more accurate lending decisions, reduce the incidence of loan defaults, and improve overall portfolio quality.

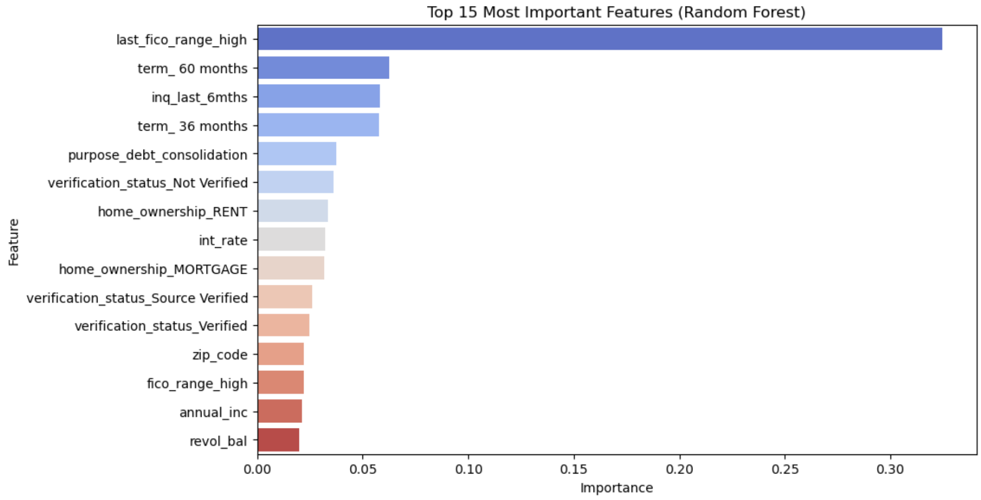


Figure 6.7 Random Forest – Best Features

6.8 Model Comparison-Linked to Literature

The results of this study identify **Random Forest** as the most suitable model for predicting loan default risk, both in terms of predictive performance and practical applicability. Across all evaluated metrics—accuracy (93%), precision (91%), recall (95%), F1-score (93%), and ROC-AUC (0.985) — Random Forest consistently outperformed the alternative models. Importantly, it achieved the highest recall, meaning it was most effective in identifying actual defaults. This aligns with the core objective of credit risk modelling, where the cost of a missed default is far greater than the cost of a false positive.

These findings are consistent with those of **Lessmann et al. (2015)**, who performed a comprehensive benchmarking of credit scoring techniques and found that ensemble-based methods, including Random Forest, outperform traditional approaches in both accuracy and robustness. Similarly, **Lang & Sun (2014)** demonstrated that tree-based models, when combined with techniques for handling class imbalance (such as SMOTE, as applied in this study), show substantial improvements in detecting high-risk borrowers. The ability of Random Forest to capture nonlinear patterns in borrower behaviour and macroeconomic factors echoes the conclusions of **Khandani et al. (2010)**, who emphasised that modern credit data often contain complex interactions that linear models like Logistic Regression cannot fully capture.

In addition to its strong predictive power, Random Forest maintains a practical advantage through **feature importance analysis**, allowing lenders to identify the most influential variables in default prediction. In this study, the top predictors-high-end FICO score range, loan term, recent credit inquiries, loan purpose, and interest rate-are consistent with the determinants of credit risk identified in prior research (Louzis et al., 2012; Makri et al., 2014). This alignment with the literature reinforces the model’s credibility and interpretability, both of which are essential for regulatory compliance and stakeholder trust in the financial sector (Fitzpatrick & Mues, 2016; ACM Code of Ethics, 2018).

While Decision Tree and Logistic Regression offered competitive interpretability, they fell short in balancing recall and accuracy. Decision Tree achieved strong recall but exhibited higher variance, making it more prone to overfitting, a limitation also noted in **Fitzpatrick & Mues (2016)**. Logistic Regression maintained interpretability but lacked the flexibility to capture the complex, nonlinear borrower–loan interactions observed in this dataset. Neural Network (MLP), despite its theoretical capacity for high accuracy, underperformed in this context, which mirrors findings by **Baesens et al. (2003)**, who highlighted that neural networks require significantly larger and more diverse datasets to reach their full predictive potential.

Given these outcomes, **Random Forest** emerges as the optimal choice for deployment in real-world lending environments. Its high recall ensures that the majority of high-risk borrowers are flagged, its robust generalisation ability reduces overfitting risks, and its feature importance outputs provide actionable insights to guide lending strategies. These strengths, supported by empirical evidence from the literature, make it a powerful tool for enhancing credit risk management and improving portfolio quality in the UK banking sector.

6.9 Conclusion

This chapter's results offer a detailed evaluation of four supervised machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, and Neural Network (MLP)—used for predicting loan default. Each model was trained, validated, and assessed using key metrics like accuracy, precision, recall, F1-score, and ROC-AUC. The comparison showed that while all models could identify important patterns in borrower and loan information, their effectiveness differed notably, especially in identifying high-risk borrowers.

Among the tested models, **Random Forest emerged as the best-performing classifier**, achieving superior scores across all evaluation metrics, including the highest recall and ROC-AUC values. This performance advantage is particularly significant in the context of credit risk prediction, where minimising false negatives-i.e., undetected defaults-is a critical objective. The model’s ensemble structure enabled it to capture complex, nonlinear interactions between variables, while also maintaining robustness against overfitting. Furthermore, its capacity to provide feature importance rankings enhances interpretability and aligns with industry requirements for transparent decision-making.

The findings also revealed necessary trade-offs between interpretability and predictive power. Logistic Regression offered the most transparent decision-making process, but at the cost of reduced flexibility in capturing nonlinear patterns. Decision Tree provided intuitive rule-based outputs but suffered from higher variance, while Neural Network, despite its theoretical modelling power, underperformed in this dataset due to its sensitivity to tuning and relatively limited training data. Overall, the comparative analysis highlights the value of ensemble methods in achieving a balance between predictive accuracy, robustness, and interpretability, setting the stage for the in-depth discussion of implications, literature comparisons, and limitations in the following chapter.

Chapter 7: Limitations and Conclusion

7.1 Limitations

While this research presents encouraging results, several limitations must be considered when applying them to real-world credit risk assessment. The Lending-Club dataset may not capture the demographic, regulatory, and economic characteristics of the UK lending market, meaning models require localisation and retraining. Its fixed historical scope excludes recent developments such as the COVID-19 pandemic and the cost-of-living crisis. Although pre-processing techniques such as outlier treatment, imputation, and SMOTE balancing improved data quality, they may also introduce bias or unrealistic borrower profiles, limiting generalisability. The Random Forest model, despite strong accuracy, remains prone to overfitting in complex datasets. Finally, the omission of macroeconomic and behavioural features constrains predictive robustness, while interpretability challenges underline the need to balance performance with transparency.

7.2 Practical Recommendations

Based on these findings, the following recommendations are suggested for financial institutions looking to improve their credit risk evaluation methods.

- **Adopt Ensemble-Based Models for Credit Scoring** - Random Forest and other ensemble methods offer superior predictive accuracy compared to traditional approaches, particularly when dealing with complex borrower data.
- **Incorporate Broader Data Sources** - While this study relied on loan application and borrower profile data, lenders should explore integrating behavioural, transactional, and macroeconomic data to improve prediction quality.
- **Address Class Imbalance in Modelling** - The use of SMOTE or other resampling methods should be standard practice to ensure that models are not biased towards predicting the majority (non-default) class.
- **Balance Accuracy with Interpretability** - While predictive performance is critical, explainable AI techniques should be employed to make model outputs transparent and compliant with regulatory standards.
- **Implement Continuous Model Monitoring** - Credit risk models should be recalibrated regularly to account for changes in borrower behaviour, market conditions, and regulatory requirements.

7.3 Additional Work

Future research could expand on this study by incorporating more diverse and recent datasets that include macroeconomic indicators (e.g., unemployment rates, GDP growth, inflation trends) to account for external economic factors affecting loan repayment behaviour. Behavioural data such as spending patterns, digital transaction history, and social network risk signals could also be explored to enrich model inputs. Advanced modelling techniques like XGBoost, LightGBM, and deep learning architectures could be tested alongside cost-sensitive learning approaches to further improve the detection of high-risk borrowers in imbalanced datasets. Moreover, explainable AI (XAI) methods should be integrated to enhance model transparency and ensure regulatory compliance while maintaining high predictive performance.

7.4 Conclusion

This dissertation aimed to examine the utilization of supervised machine learning methods to forecast loan default risk, employing a publicly accessible LendingClub dataset. The study adhered to the CRISP-DM approach.

This includes interpreting the data, cleaning it up, building a model, and testing it. We used several pre-processing methods, such as SMOTE to deal with class imbalance, handling missing values, treating outliers, encoding categorical variables, and feature scaling.
We used a consistent training and validation framework to test four classification algorithms: Logistic Regression, Decision Tree, Random Forest, and a Multi-Layer Perceptron (MLP) Neural Network.

GridSearchCV's hyperparameter optimization made sure that each model worked at its best settings. The evaluation metrics were accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices.

The findings indicated that the Random Forest model had the most robust overall performance, attaining an accuracy rate of 93% and a ROC-AUC score of 0.98. This model showed a good mix between predictive power and resilience. It was better than other models in correctly identifying both default and non-default cases. The analysis of feature importance showed that the borrower's credit score, interest rate, debt-to-income ratio, and loan length were the most critical factors in predicting default risk. This is in line with what other studies have found.

7.5 Legal, Social, Ethical, and Professional Issues

The application of machine learning in credit risk assessment goes beyond technical performance and involves legal, ethical, and professional responsibilities. Legally, compliance with GDPR and the **UK Data Protection Act (2018)** is vital, ensuring secure handling of borrower data. Although this project used anonymised, public datasets, real-world systems must protect sensitive information from misuse. Ethically, fairness is crucial, as historical lending data may contain biases against minority or low-income groups, potentially leading to discriminatory outcomes. Therefore, models must be evaluated for both accuracy and fairness, with techniques like feature importance aiding transparency. Professionally, accountability requires adherence to ethical codes, clear documentation, and regular audits. Ultimately, responsible deployment ensures credit risk management is accurate, fair, and socially responsible.

On a **professional level**, accountability and ethical practice are central. Data scientists and financial institutions must align with professional codes of conduct such as the **ACM Code of Ethics (2018)**, which emphasise transparency, fairness, and the avoidance of harm. In practice, this means developing interpretable models, documenting methodologies clearly, and ensuring that systems are regularly audited for both accuracy and fairness.

7.6 Final Remarks

This research has demonstrated that the thoughtful application of machine learning to credit risk assessment can yield substantial benefits over traditional methods, both in predictive performance and operational decision-making. The Random Forest model, as the best-performing algorithm in this study, provides a robust foundation for deployment in financial institutions, offering a strong balance between accuracy and interpretability when paired with feature importance analysis.

The study underscores the importance of data quality, careful pre-processing, and model evaluation using appropriate metrics, particularly in the presence of imbalanced datasets. While challenges remain in terms of regulatory compliance, explainability, and generalisability, the integration of advanced analytics into the credit risk assessment process represents a forward-looking step for the banking sector - one that can enhance decision-making, reduce default rates, and improve financial stability.

8. References

1. Abid, L. (2022). A logistic regression model for assessing credit risk of companies in the service sector. International Research in Economics and Finance, 6(1)[Crossref]

2. ACM Council. (2018). ACM Code of Ethics and Professional Conduct. [Online] Available at: https://ethics.acm.org/code-of-ethics/ [Accessed May 1, 2022].

3. Apte, C. (2010). The role of machine learning in optimizing business processes. Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, June 21–24, pp.1–2.

4. Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Neural network rule extraction and decision tables for credit-risk assessment. Management Science, 49(3), pp.312–329.

5. Beck, R., Jakubik, P., & Piloiu, A. (2015). Key factors of non-performing loans: New evidence from a global sample. Open Economies Review, 26(3), pp.525–550.

6. Broby, D. (2022). Predictive analytics in finance. Journal of Financial Data Science, 8(2), pp.145–161. https://doi.org/10.3905/jfds.2022.1.099

7. Carvalho, P.V., Curto, J.D., & Primor, R. (2022). Macroeconomic factors influencing credit risk: Evidence from the Eurozone. International Journal of Finance & Economics, 27(2), pp.2054–2072. https://doi.org/10.1002/ijfe.2308

8. Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, pp.321–357.

9. Chen, T., & Guestrin, C. (2016). XGBoost: Scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785–794.

10. Clements, J.M., Xu, D., Yousefi, N., & Efimov, D. (2020). Sequential deep learning for credit risk monitoring using tabular financial data. arXiv preprint, arXiv:2012.15330.

11. Dixon, M.F., Halperin, I., & Bilokon, P. (2020). Machine learning in finance. Springer, New York.

12. Fainstein, G., & Novikov, I. (2011). Comparative analysis of credit risk determinants in Baltic States banking sector. Review of Economics & Finance, 1, pp.20–45.

13. Fitzpatrick, T., & Mues, C. (2016). Comparing classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. European Journal of Operational Research, 249(2), pp.427–439.

14. Ghosh, A. (2015). Regional and industry-specific determinants of non-performing loans in US states. Journal of Financial Stability, 20, pp.93–104.

15. Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). Applied Logistic Regression (3rd ed.). Wiley, Hoboken, NJ.

16. IMF. (2022). UK: Financial Sector Assessment — Systemic Stress and Climate-Related Financial Risks. International Monetary Fund. [Online] Available at: https://www.imf.org/en/Publications/CR/Issues/2022/04/07/United-Kingdom-Financial-Sector-Assessment-Program-Systemic-Stress-and-Climate-Related-516264 [Accessed 1 May 2022].

17. Khandani, A.E., Kim, A.J., & Lo, A.W. (2010). Machine-learning models for consumer credit risk. Journal of Banking & Finance, 34(11), pp.2767–2787.

18. Klein, N. (2013). Non-performing loans in CESEE: Determinants and macroeconomic impact. IMF Working Paper, 13(72).

19. Krumeich, J., Werth, D., & Loos, P. (2015). Using predictive analytics for prescriptive control of business processes in manufacturing. Business & Information Systems Engineering, 58(4), pp.261–270.

20. Kuzucu, N., & Kuzucu, S. (2019). Factors driving non-performing loans: Evidence from emerging and advanced economies pre- and post-global financial crisis. Emerging Markets Finance & Trade, 55(7), pp.1694–1708.

21. Lang, J., & Sun, J. (2014). Decision tree sensitivity to class imbalance in bank credit risk early warning. Proceedings of the Seventh International Joint Conference on Computational Sciences and Optimization, Beijing, China, pp.539–543.

22. Lessmann, S., Baesens, B., Seow, H.V., & Thomas, L.C. (2015). Benchmarking classification algorithms for credit scoring. European Journal of Operational Research, 247(1), pp.124–136.

23. Liu, X., Cheng, G., & Wu, J. (2002). Analyzing outliers in data. IEEE Transactions on Knowledge and Data Engineering, 14(2), pp.432–437.

24. Louzis, D.P., Vouldis, A.T., & Metaxas, V.L. (2012). Macroeconomic and bank-specific factors of non-performing loans in Greece: A

comparative analysis across mortgage, business, and consumer loans. Journal of Banking & Finance, 36(4), pp.1012–1027.

25. Makri, V., Tsagkanos, A., & Bellas, A. (2014). Determinants of non-performing loans in the Eurozone. Panoeconomicus, 61(2), pp.193–206.

26. Ouahilal, M., El Mohajir, M., Chahhou, M., & El Mohajir, B.E. (2016). Comparing predictive algorithms for business analytics and decision support systems in finance. International Conference on Information Technology for Organizations Development, pp.1–6.

27. Reinhart, C.M., & Rogoff, K.S. (2011). From financial crash to debt crisis. American Economic Review, 101(5), pp.1676–1706.

28. Ruiz, S., Gomes, P., Rodrigues, L., & Gama, J. (2017). Credit scoring in microfinance using alternative data. In: Progress in Artificial Intelligence, Lecture Notes in Computer Science, Springer, Cham, 10423, pp.409–421.

29. Salas, V., & Saurina, J. (2002). Credit risk in different institutional regimes: Spanish commercial and savings banks. Journal of Financial Services Research, 22(3), pp.203–224.

30. Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy: Discriminant measures for performance evaluation. Advances in Artificial Intelligence, Springer, Berlin, Heidelberg, pp.1015–1021.

31. Wang, T., Zhao, S., Zhu, G., & Zheng, H. (2021). Machine learning-based early warning system for systemic banking crises. Applied Economics, 53(26), pp.2974–2992.

32. Yu, J. (2022). Big data analytics and a discrete choice model for enterprise credit risk early warning. Security and Communication Networks, 2022, pp.1–12. https://doi.org/10.1155/2022/3272603