UNIVERSITY OF PORTSMOUTH

School of Computing

MSc. Data Analytics

Title: LOAN APPROVAL PREDCITION

Project unit: M32616

By

Satyadeep Mohanta

UP2205708

Project Supervisor: Dr. Atefeh Khazaei Ghoozhdi

Word count: 12033

(should not include title page, acknowledgements, abstract, table of contents, table of tables, table of figures, reference list or appendices)

Please tick the following declaration

✓	I hereby declare that this dissertation is substantially my own work

Please read the following and then sign underneath.

I do/do not consent to my dissertation in this attributed format (not anonymous), subject to final approval by the Board of Examiners, being made available electronically in the Library Dissertation Repository and/or Department/School/Subject Group digital repositories. Dissertations will normally be kept for a maximum of ten years
I understand that if I consent, this dissertation will be accessible only to staff and students for reference only
This permission may be revoked at any time by e-mailing data-protection@port.ac.uk

Name: Satyadeep Mohanta

Date: 16 September 2024

PROJECT ABSTRACT

The financial services sector is experiencing a significant transformation with the increasing adoption of artificial intelligence (AI) driven solutions, which are progressively replacing traditional methods of loan approval prediction. The critical role of finance in the global market, influencing both economic conditions and the strategies of financial institutions, underscores the importance of efficient loan approval processes. This project addresses the need to enhance these processes, focusing on improving the speed and accuracy of loan approval predictions to better meet the demands of today's dynamic economy.

This study explores the application of advanced machine learning models to improve loan approval predictions. Traditional manual and heuristic-based methods often result in inefficiencies, such as rejecting creditworthy applicants and approving loans for high-risk individuals. To address these challenges, the CRISP-DM methodology was employed, utilizing machine learning models to analyze and predict loan approval outcomes and applicant behavior.

The results demonstrate that AI/machine learning models significantly enhance prediction accuracy compared to traditional methods. Performance metrics, including precision, recall, F1-score, and AUC, indicate that these models offer a more reliable means of evaluating loan applications. This advancement contributes to reducing default rates and operational inefficiencies while better aligning financial services with economic conditions and customer needs.

By accelerating and refining the loan approval process, this project provides valuable insights into improving decision-making and operational efficiency within the financial sector, ultimately benefiting both financial institutions and their clients.

Keywords: Loan Approval, management, advance machine algorithm, Implementation strategy, performance measurement, data visualisations, data analytics, data analysis.

ACKNOWLEDGEMENT

With profound gratitude and deep appreciation, I dedicate this dissertation to the pillars of my academic journey:

My Parents: Your unwavering support has been my cornerstone. To my father, **Shri Manoj Kumar Mohanta**, I owe an immeasurable debt of gratitude. Your steadfast belief in my abilities, constant encouragement, and sacrifices have driven my success. You've been my rock throughout this master's degree, and words cannot express how much your support means to me.

My Friends, your camaraderie, understanding, and encouragement have been invaluable. In moments of doubt and stress, your presence and support have been a source of strength and motivation. This achievement is as much yours as it is mine. I want to especially thank Mary, Lennard, Naveen, Sumita, Sanskar, Meenakshi, Sanat, Hla Soe Htay and Charit - your friendship has been a beacon of light throughout this journey, making the challenging times bearable and the good times even more joyous. Each of you has contributed uniquely to my growth and success, and I am profoundly grateful for your presence in my life.

The University of Portsmouth Faculty: I sincerely thank all my lecturers. Your dedication to imparting knowledge and skills has equipped me for this dissertation and the challenging journey ahead in my career. The wisdom you've shared will guide me long after leaving university halls.

Dr. Atefeh Khazaei Ghoozhdi: My supervisor, mentor, and guiding light through this academic endeavour. Your expertise, patience, and invaluable insights have shaped this dissertation. Your thoughtful feedback and unwavering support pushed me to exceed my expectations. This work stands as a testament to your exceptional guidance.

To everyone who has been a part of this journey, whether mentioned here or not, knows that your contribution has been significant, however small it may seem. This dissertation is not just a culmination of my efforts but a product of the collective support, love, and wisdom of all those who have touched my life during this academic pursuit.

As I stand at the threshold of a new chapter in my life, I carry a degree and the indelible imprint of your support, belief, and encouragement. For this, I am eternally grateful.
Thank you from the depths of my heart.

TABLE OF CONTENTS:

Abbreviations:

ET: Extra Trees (also known as Extremely Randomised Trees)

RF: Random Forest

CB: CatBoost

LGB: LightGBM (Light Gradient Boosting Machine)

EGB: Extreme Gradient Boosting (often used interchangeably with XGBoost)

DT: Decision Tree

KNN: K-Nearest Neighbors

SVM: Support Vector Machine

DTAB: Decision Tree with AdaBoost

LR: Logistic Regression

NB: Naive Bayes

XGB: XGBoost (Extreme Gradient Boosting)

BN: Bayesian Network

MP: Multilayer Perceptron (a type of Neural Network)

A credit score is a numerical measure of an individual's creditworthiness, typically ranging from 300 to 850. It is calculated based on various factors, including payment history, amounts owed, length of credit history, new credit, and types of credit used. Higher credit scores indicate a lower risk to lenders, facilitating easier access to loans and credit at favorable terms (Investopedia, 2019; Experian, 2019).

Glossary

AUC	Area Under the Curve
ML	Machine Learning
ROC Curve	Receiver Operating Characteristics curve
EBIT	Earnings Before Interest and Taxes
FICO scores	Fair Isaac Corporation Scores
AI	Artificial Intelligence

CHAPTER 1: INTRODUCTION

1.1. Introduction

This chapter outlines the background, goals, and objectives of the study. The financial industry is at a pivotal point where advanced predictive techniques are transforming traditional loan approval processes. Historically, loan approvals were managed manually, relying heavily on human judgments and simple rule based systems. This manual process is time consuming and prone to inconsistencies and biases. Today, financial institutions emphasise optimising their operations, making the prediction of loan approvals increasingly important.

The primary problem addresses the inefficiency and bias in traditional loan approval procedures. This project aims to solve the issue by developing AI/machine learning solutions that includes both quantitative data and the qualitative insights. The main objective is to develop machine learning models that enables more accuracy in predicting loan outcomes while reducing human involvements and automating the process.The focus is on continuous adaptation and improvement of loan approval systems.

The challenges includes while integrating with AI technologies into existing systems, ensuring machine learning models should predict with more accuracy and reliability. Additionally, legal, social and ethical issues for data privacy concerns and also the need of transparency in AI decision making process.

1.2. Background

The loan approval methods began in the mid 20th century using traditional manual procedures that included background checks, police verification, collateral assessment, evaluation for credit score validity and even often income to be verified. These manual procedures are time consuming and prone to inconsistencies and broad biases. Despite of technological advancement, still several companies globally cling to traditional manual procedures.

The financial services sector is now undergoing a substantial revolution driven by predictive analytics and machine learning technologies. Prominent financial firms such as JPMorgan Chase and Wells Fargo use

artificial intelligence (AI)/advanced machine learning to evaluate loans (JPMorgan Chase, 2023; Wells Fargo, 2021). For instance, 'JP Morgan Chase uses artificial intelligence (AI) to evaluate credit scores by analysing vast volumes of data (JPMorgan Chase, 2023).' Wells Fargo also utilises AI to enhance the efficiency of loan pricing and risk management (Wells Fargo, 2021).

1.3. Questions

- Why haven't all banks adopted predictive models for loan approvals yet?
- What are the challenges in integrating AI models with existing systems?
- Is it reliable to trust AI models over human judgment in loan sanctions?
- What are the ethical and regulatory implications of using AI in loan approval systems?
- How can AI models ensure fairness and avoid discrimination in loan approvals?
- How can customer trust be maintained with AI-driven systems?

1.4. Aims and Objectives:

Aims:

The primary aim is to understand the data and the business processes.This project is divided into two sections: one for loan approval prediction and another applicant behaviour (defaulters and non-defaulters). Additionally, the goal is to provide association rules to help banks and loan firms understand loan approvals better.

Objectives:

- Conduct Exploratory Data Analysis (EDA).
- Develop predictive models for applicant behaviour.
- Build a loan approval prediction model.
- Implement feature engineering techniques.
- Integrate quantitative and qualitative data.

- Address imbalanced datasets.
- Enhance model transparency.
- Ensure fair lending practices.
- Adapt to dynamic financial environments.
- Provide actionable recommendations.

1.5. Restrictions and Limitations

- Data quality and completeness.
- Time constraints.
- Regulatory compliance.
- Ethical issues.
- Model interpretability vs. complexity.
- Cross-validation constraints.
- Limitations of feature engineering.

1.6. Evaluation of the Problem Topic and Investigation

This dissertation leverages machine learning models for improved accuracy and efficiency. It stresses on high performance should coexist with ethical loan sanctions practices. By addressing accuracy dynamics and the interpretability of loan approval systems, this project aims to enhance how financial institutions can work more efficiently.

1.7. Structure of Research

The organisation in sections is as follows for this study project:

Chapter 1: Introduction - Explains the problem and objectives and provides the background of the research.

Chapter 2: Literature Review - This section reviews what type of research and methodologies are present in predicting loan approval.

Chapter 3: Project Management and Methodology - This chapter discusses the planning, methodology and management aspects of this project.

Chapter 4: Business Understanding - Description of the business case and specification for loan approval prediction.

Chapter 5: Exploratory Data Analysis - This detailed analysis of the sources from which data is coming and an initial look at it.

Chapter 6: Data Preparation - Provides a guide on how the data is prepared for analysis.

Chapter 7: Classification- Covers the development of predictive models.

Chapter 8: Evaluation - Describe the posteriori examinations of property function and review the findings.

Chapter 9 - Association Rule – Suggests some association rules with loan approval sactions.

Chapter 10: Conclusion - This chapter summaries the research findings and suggests areas for future research.

1.8. Conclusion

This dissertation strives to stimulate the discussion in the financial sector through a novel framework for ethical and reliable loan approval prediction. Considering these challenges and complexities, this study aims to provide novel insights with methods that could help increase efficacy on how financial institutions grant commission loan approvals. The introduction sets up the project, describes the context in which it develops and aims to address the placement scenarios, doing so ethically. It lays the groundwork for the study of this dissertation.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter of literature review covers the development in loan approval prediction in the financial services industry. It traces the transformation from traditional financial measures to advanced machine learning models. This review evaluates a range of methods and considers how they perform in real life scenarios to limit the risk associated with default, while arranging allocated resources efficiently.

This study addresses the nuances of algorithmic lending decisions. This paper dives into the nuances between predictive accuracy and fair lending, discussing ways in which advanced analytics can be used appropriately.

This review adds to a recent analysis of contributions and discusses how these findings could lead to better loan approval. The purpose is to offer insights and practitioners some ideas that might alter financial firms' forecasts on loans.

The review aims to connect dots between new state-of-the-art prediction technologies and the ethics aspect of finance towards more precise, effective yet fairer lending in the financial sector.

2.2 Traditional Methods

As mentioned in the Introduction chapter, Loan approval prediction methods have evolved since the mid 20th century, when statistical techniques were already being applied to financial data. So, this part will touch on core papers that laid a foundation for loan approval and credit risk scoring methods. (as credit scoring plays an important role for loan approval)

In 1966, (William H. Beaver,1966) accomplished seminal work in this field, creating the logistics model using financial ratios to predict business failure but later became an important tool for the crediting process and loan approval process. In a univariate analysis of individual financial ratios, identified that the most successful predictor was cash flow to total debt and net income to total assets. Showing that

financial ratios can predict the commercial gain or loss, the work provided support for the worst ratio of all, i.e., working capital to debt by developing a model of company failure.

Beaver's research showed that financial ratios could accurately predict company bankruptcy. The working capital to debt ratio was the top discriminator, able to identify 90% of failing firms one year prior. This work provided a template for such advanced statistical models and showcased the contribution of cash flow specific ratios in assessing its financial health. The Beaver approach analysed individual ratios rather than collectively.

As Altman (Altman,1968) expanded on Beaver's research, the traditional approach was greatly improved in 1968 from the discriminant analysis based business bankruptcy prediction. At the core of Altman's Z-Score model lie five critical financial ratios. They are sales to total assets, market value of equity to total liabilities, working capital to total assets, retained earnings/total asset ratio and EBIT/Total Assets. This model provides a more precise method of predicting financial difficulties in businesses.

Altman's Z-Score model was extremely predictive; 94% of the initial sample firms that subsequently filed for bankruptcy were correctly identified by this method. Three models defined three zones of financial health: The threshold-based Z-Score model defined low, moderate and high bankruptcy risk zones. This study laid the groundwork for predictive analytics in finance, showing how financial ratios can predict corporate bankruptcy when statistical methods are applied.

(Robert O. Edmister,1972) refined the use of discriminant analysis with financial ratios to predict small business failure in 1972. To accomplish this, certain financial ratios, e.g. liquidity, profitability and leverage ratio, in a discriminant model (Altman, 1968). It was a solution that met an urgent need in lending to score and segment small businesses based on their creditworthiness while contributing a framework for the contemporary quantitative credit scoring models.

Later work examined human judgement in making financial decisions. According to a study conducted in 1980 by (Zimmer, 1980), which examined the prediction of corporate failures through bank loan officers, this research found substantial agreement among these disposers, implying great consistency on their parts. However, in the same year, (Casey,1980) found that there were slip-ups by loan officers when forecasting the outcomes of loans to be approved.

That difference showed the nuance in human decision making in loan origination. Professionals offered some consistency (Zimmer, 1980), but there was also large variability in accuracy among the ones tested (Casey, 1980). These gave way to more advanced predictive models in loan approval for the development of objective data driven approaches.

The contrast between Zimmer's and Casey's work in 1980 indicated the enormous variability in human judgement in loan approval decisions. This inconsistency highlighted the importance of objective and more data oriented strategies in finance.

Further, (Nutt,1989) discovered that the organisational culture in banks influences loan approval and risk assessment; quite often more so than the uncertainty of repayment. The findings of this analysis demonstrated that a loan approval decision is not only determined by purely financial criteria but also depends upon the bank's organisational culture. An example would be a loan application considered unfit for another bank (conservative banks) that might still tick the box of a growth oriented bank even though financial data are the same.

These findings collectively emphasised human judgement limitations and organisational factors' impact on loan approval processes. The traditional methods, while groundbreaking for their time, had significant limitations:

In most cases, they took each financial ratio independently rather than in combination.

They neglected the nuances of human judgement and organisational culture.

However, this led to crucially establishing the foundation for advanced models in loan approval and credit risk assessment. This contributed to modern technology and showed that traditional methods need more standardised tools and a data-driven way of making financial decisions.

2.3 Modern Methods

The late 20th and early 21st centuries marked a shift in loan approval prediction methods, transitioning from traditional statistical approaches to advanced machine learning techniques.

In research, (Dimitras et al., 1999) compared machine learning approaches with traditionally established statistical methods for predicting bankruptcy. The study compares decision trees and neural networks with logistic regression and discriminant analysis. The study discovered that, especially while handling diverse and non-linear data, machine learning approaches frequently surpassed conventional methods in terms of predicted accuracy and resilience. The base for the vast use of machine learning in financial prediction, including loan approval methods, was laid by this study. It showed that these novel approaches were capable of handling the intricacies of financial data more adeptly than conventional statistical approaches.

(Thomas, 2000) conducted a credit scoring and behavioural survey that looked at how well past use of consumer loans can predict future financial risk. This study emphasised the increasing demand for more accurate prediction models in financial markets, but also underscored that new methods to deal with consumer credit risk are necessary. The paper set the groundwork for later advances in loan approval prediction systems.

Zhang's review of Neural Networks in business profoundly expanded the understanding of artificial intelligence's use in financial prediction. (Zhang, 2003) emphasised that neural networks are especially suitable for credit scoring, bankruptcy prediction, and loan approval because of its ability to manage massive datasets and unearth intricate linkages. The study showed how, by identifying patterns and default predictors from historical loan data, neural networks could enhance the accuracy and reliability of credit risk ratings.

(Berger et al, 2005) explored how "small and large banks process soft" information for loan approvals. The study discovered that small banks specialised in loans to firms with inadequate financial information, while big banks standardised their processes and sought a less personal touch. The research showed the impact of organisational structure on credit risk assessment and helped to align small lenders with big banks by implying that technology might create a level field.

(Desai, 2006) formulated predictive models for loan application classification using machine learning algorithms like decision trees, support vector machines and neural networks. The study clearly shows that feature selection and data processing need to be prioritised. Desai's analysis, which proved that machine learning models outperformed traditional approaches in predicting the payment of loans showed how these techniques have a role to play for possibilities they provide towards transforming loan approvals.

(Berger et al., 2005) explored how "small and large banks process soft" information for loan approvals. The study discovered that small banks specialised in loans to firms with inadequate (Huang et al,2007) explored credit scoring using support vector machines (SVM). The research showed that to evaluate credit risk, Support Vector Machine performed better than traditional discriminant analysis and logistic regression models. The research also provided evidence in the support of machine learning approaches to audit financial related decisions, especially loan approval using bank transaction data.

(Khandani, 2010) investigated many machine learning algorithms for credit risk prediction, such as Support Vector Machines, Decision Trees and ensemble methods like Bagging and AdaBoost. To handle these problems, a voting ensemble model to increase accuracy and robustness in credit risk predictions. The research revealed the critical role of feature selection and engineering in improving prediction performance for loan approval models.

(Harris, 2013) also investigated credit support vector machine (SVM) with regard to either a broad or narrow definition of loan default. By analysing data from a Barbados credit union, Harris saw that SVM models using the broader (less than 90 days past due) default definition delivered results superior to those

with narrower definitions. The study indicated that a broader sense of default with predictive models may increase their capabilities, and this realisation can present useful takeaways for financial establishments.

(Malekipirbazari and Aksakalli ,2015) used random forest classification, i.e., the RF method, to predict borrower status in social lending platforms. According to their research, compared with the standard FICO scores achieved in credit scoring, Random Forest delivered substantially better results in identifying dependable debtors and showed that machine learning methods have enormous potential for improving forecasts of loan approval.

(Byanjankar, 2015) implemented credit scoring models in peer-to-peer (P2P) lending. The study promised to manage credit risk in this less regulated sector but did not delve deeply into the ethical considerations of P2P lending using automated decision-making, which was one major future area for research.

(Abellán and Castellano ,2017) investigated various base classifiers within ensemble methods for credit scoring. They evaluated combinations of classifiers (e.g., decision trees, support vector machines, neural networks) and ensemble techniques (e.g., bagging, boosting, stacking) to optimise predictive accuracy, robustness, and computational efficiency. Their research demonstrated that even small improvements in credit scoring models could significantly enhance overall credit risk assessment.

(Hamori ,2018) investigated the predictive performance of ensemble learning methods and deep learning techniques with respect to default risk as well. Boosting in general as an ensemble performed better than deep learning models across the board at both classification and interpretability.

(Zhu ,2019) showed that Random Forest performs well in distinguishing loan default. Although their study showed practical effects of these technologies in loan approval systems, it was mainly focused on technical performance dimensions without providing implications for how such an implementation can take place across different organisations

The modern machine learning algorithms have been used by (Singh, 2021) in loan approval systems and the benefits could be enormous. Nevertheless, that work should have directly considered the interpretability and transparency of these complex models, which are essential for regulatory compliance and stakeholder trust. This gap has highlighted the importance of improving explainable AI techniques to predict loan approval effectively while being interpretable.

Through this evolution of contemporary approaches used in loan approval prediction, the author sees the trend evolve into much more advanced and data driven methods. While these more advanced ensemble methods give better accuracy and efficiency, it also introduces the new challenges faced in interpretability based techniques as well as ethical considerations on practical implementations across a wide spectrum of financial scenarios.

2.4 Summary

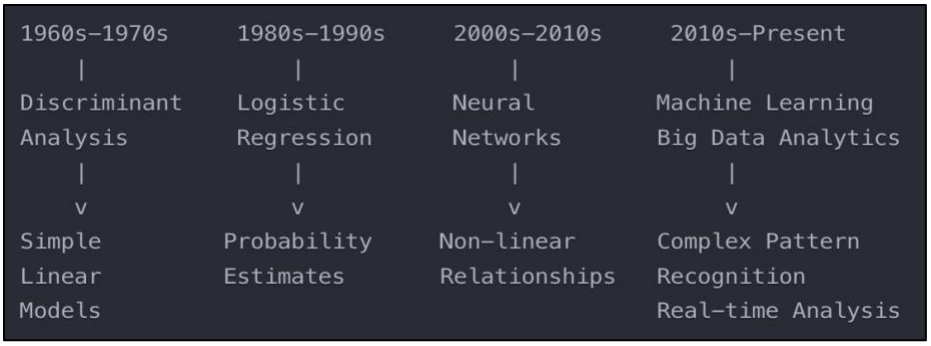
Loan approval prediction has been a huge area in the financial industry over the past few years, and it has shown great advancements in analytical techniques and computational capabilities. This transition has been urged by the rise of complexity in financial markets, as well as greater accuracy and efficiency in loan approval practices. This is one of the key things in the financial sector, also predicting whether to approve or decline loans has evolved over time from simple statistical analyses into more advanced machine learning algorithms but each new system comes with a performance improvements while at the same struggle with different challenges.

Recent research has used a variety of machine learning techniques, from simple linear models to advanced ensemble methods and deep neural networks. The following table summarises the performance of different models across several studies:

Comparative Analysis of Existing Work (Table 1)

Title of Paper	Models for Loan Prediction used in the Existing Work (Accuracy %)													Author	Year
	ET	RF	CB	LGB	EGB	DT	KNN	SVM	DTAB	LR	NB	XGB	BN	MP	
An ensemble machine learning based bank loan approval predictions system with a smart application	86.6	84.7	84.9			78.9	79.6	68.1		70.6					Nazim Uddin, Md. Ahamed, Md Uddin , Md. Islam , Md. Talukder , Sunil Aryal	2022
Customer loan eligibility prediction using machine learning algorithms in banking sector		72				69	59	70	54						Ch. Naveen Kumar, D. Keerthana, M Kavitha, M Kalyani	2022
Analysis of Loan Availability using Machine Learning Techniques		77.3				66.2	61.9	65		78.5	77.9	77.3			Sharayu Dosalwar1 , Ketki Kinkar , Rahul Sannat , Dr Nitin Pise	2021
Predicting Bank Loan Risks Using Machine Learning Algorithms		78.5				73.5					77.5		75	80	Maan Y. Alsaleem, Safwan O. Hasoon	2020
Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process						71.9		65.3		78.9	80.4				E. Chandra Blessie, R. Rekha	2019
Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms		81					68	73.2		71					Praveen Tumuluru; Lakshmi Ramani Burra; M. Loukya; S. Bhavana; H.M.H. CSaiBaba; N Sunanda	2022
Machine Learning Models for Predicting Bank Loan Eligibility		95.56				91.11	93.33	84.44		80					Ugochukwu. E. Orji; Chikodili. H. Ugwuishiwu,Joseph. C. N. Nguemaleu; Peace. N. Ugwuanyi	2022

The evolution of loan approval prediction methods has been marked by significant advancements in analytical techniques and computational capabilities. This progression can be summarised in several key stages: (Figure 1)



The traditional methods to predict loan approval led by Beaver (1966) and Altman (1968) are considered to be the early stages of predictive analytics within finance. Previously, they mainly relied on financial ratios and discriminant analysis to evaluate credit risk. Although innovative at the time, these approaches have some drawbacks:

- **Reliance on Linear Assumptions:** Most traditional methods often assume linear relationships among variables, which would fail to capture the complexities of financial datasets.
- **Inability to Handle Large Datasets:** As financial datasets have expanded in both scale and complexity; traditional statistical methods have struggled to process and analyse vast data sets.
- **Limited Consideration of Human Judgment:** Zimmer's (1980) and Casey's (1980) studies revealed that humans' judgments vary, but typical approaches do not account for the subtleties of human decision-making or organisational dynamics.

The transition to modern machine learning techniques, as explored by Dimitras et al. (1999) and Singh (2021), substantially improves the precision of loan approval models. These methods such as neural networks, support vector machines and ensemble techniques , incorporate a number of advantages:

- **Handling Non-Linear Data:** Unlike traditional methods, machine learning models have capacity to understand complex and non-linear relationships within the data, which helps in analysing various kinds of financial datasets.
- **Improved Predictive Accuracy:** Recent studies demonstrate that machine learning models can perform better than traditional methods. With the ability of managing large scale frames and detecting detailed oriented patterns in real time.
- **Adaptability:** Modern methods can more easily adapt to changing market conditions and new types of data, making them more flexible for evolving financial landscapes.

Current Challenges and Gaps

Despite these advancements, several key areas that require further attention and improvement:

- **Interpretability vs. Accuracy:** Although machine learning models such as random forests and neural networks have shown better predictive capabilities (Malekipirbazari & Aksakalli, 2015; Byanjankar et al., 2015), they often lack interpretability compared with traditional methods. This 'black box' nature challenges regulatory compliance and stakeholder trust (Singh et al., 2021).
- **Ethical Considerations:** The proliferation of automated decision-making systems, particularly in relatively underregulated sectors such as peer-to-peer lending (Byanjankar et al., 2015), requires models that balance predictive accuracy with fairness and transparency.
- **Adaptability to Dynamic Markets: Most** of the work has been on static data, which could impose constraints as market conditions change rapidly. This also suggests a public policy solution and that future research should look at more dynamic modeling approaches to adjust for the impact of time on changes in economic conditions and borrower behaviour.
- **Integration of Soft Information: While** Berger et al. (2005) highlighted the importance of soft information in loan decisions, many modern models are still predominantly based on hard financial data. Hybrid models combining quantitative and qualitative factors can be widely developed.
- **Ensemble Method Optimization:**Studies like Abellán & Castellano (2017) and Hamori's Ensemble of methods have been shown to be promising (2018). There is still a need for additional research regarding how best to tailor these techniques within the broader realm of loan acceptance.
- **Handling Imbalanced Data:**The theories behind loan defaults are not sustained by evidence. Most default cases are fewer than non-defaults in a credit or lending dataset. Additionally, research will be needed to overcome this issue in order for the models to generalise well with minority classes.

- **Feature Engineering and Selection:**Several studies (Eg. Desai, 2006) have highlighted the importance of feature selection, and advanced methods still need to be developed that can determine and harness the input variables that are most predictive in large financial datasets.
- **Cross-Domain Applicability:**Most studies focused on specific types of loans and geographical regions. Broader models that can work across all types of loans and financial market aspects are needed.
- **Explainable AI:**As highlighted by the limitations in Singh et al. (2021), there is a crucial need for developing methods for ensuring that loan approval decisions can be explained and in a way that does not affect predictive accuracy as it has been shown this will significantly improve understanding of how AI models arrive at their predictions.
- **Real-World Implementation Challenges:**Many studies (eg . Zhu et al. in 2019) explore technical measures of model performance while failing to adequately account for the many real-world problems in applying these models in diverse organisational contexts.

Research Focus:

- Based on the identified gaps, this research aims to explore several potential areas, including developing interpretable machine-learning models that maintain high accuracy.
- Creating adaptive models capable of adjusting to changing market conditions.
- Designing hybrid models that effectively incorporate quantitative (complex) and qualitative (soft) information.
- Exploring novel approaches to handle imbalanced datasets in loan approval contexts.
- Investigating methods to improve the cross-domain applicability of loan prediction models.

CHAPTER 3: PROJECT MANAGEMENT

3.1 Introduction

This chapter outlines the project management aspects of the study, focusing on planning and execution. It addresses the challenges such as selecting appropriate data, defining goals for the data mining process, choosing a project management framework, and establishing a project timeline. It details the role of project client, project cost consideration and ethical issues, methodology, requirements for the project plan.

3.2 Project Client

The project's clients are private banks and loan distribution firms that handle a high volume of loan applications per day. Loan processing is key to the operation of these financial firms. These financial firms' main problem is finding customers they can trust and who will not default on their dues. The challenge can result in financial loss, inefficiencies on the loan approval process very well as for institutional financial stability.

3.3 Project Cost, Ethical Issues

Project Cost:

This project is implemented by using Python programming language through Google Colab tool. Google Colab is available under free licenses, making budget allocations for the completion of this project.

Ethical Issues:

Utilising the Available Dataset and Consent: Banks collect sensitive personal information from applicants, such as financial records, salaries, and family particulars. The present study uses anonymised data obtained from Kaggle, ensuring no personal identifying information is included. Consent for data usage is crucial, and banks must obtain explicit consent from users before collecting and using their data.

Data Privacy and Security: Privacy and security of loan application data are essential. Financial institutions implement data protection through encryption, access controls, and regular audits. The data utilised in this initiative is anonymised to prevent the inclusion of any personal identification information. Regular updates and compliance with data protection regulations, including the General Data Protection Regulation (GDPR), are essential.

Ethical Considerations in Model Development: For this project, features which are user IDs and names are dropped. Important features like credit scores, salary and professions are used. Loan firms can add user data as they take consent, for better accuracy in future predictions. Ensuring the model does not introduce bias is critical.

Transparency and Explainability: To increase transparency, interpretable models or techniques that explain predictions are implemented. Establishing trust with applicants and securing accountability by offering transparent reasons for loan approval or rejection is important.

Societal Impact and Fairness: The project considers the broader implications of automated loan approval systems on financial inclusion and equality. Ensuring that the model does not discriminate against any group and promotes fair access to financial services is vital.

Ethical Checklist: An ethical checklist has been completed in accordance with university requirements, covering data protection, digital rights, and the use of open-source or proprietary software. This checklist ensures that all ethical considerations are systematically addressed throughout the project.

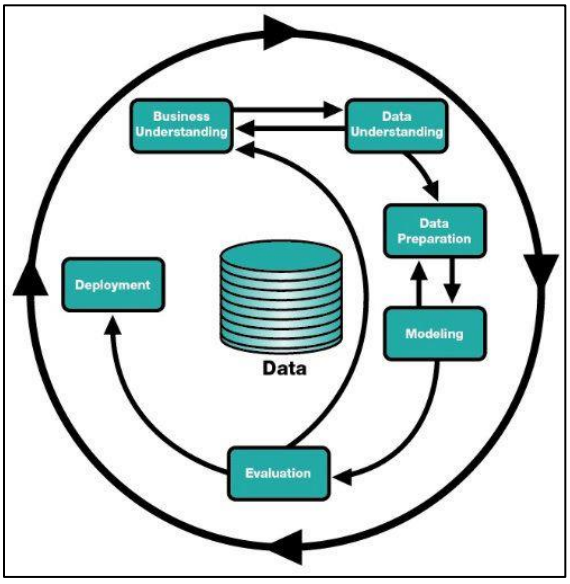
By addressing these ethical considerations, this project ensures responsible, transparent and fair data analytics practices in loan approval prediction.

3.4 Methodology

The dissertation is utilised with the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which is a well recognised methodology for data analytics projects. CRISP-DM is particularly suitable for this project due to its structure, iterative approach which ensures thorough

analysis and model development. This methodology is chosen for this project because of its flexibility and robustness in handling complex data mining tasks, as for predicting loan approval outcomes and applicant behaviour (defaulters and non-defaulters). This method allows for a comprehensive understanding of both the business context and the data, ensuring that the models developed are not only accurate but also aligned with the strategic goals of financial institutions. Additionally, CRISP-DM supports continuous improvement, making it ideal for dynamic environments where models should be able to evolve with new data. For loan approval prediction, where the stakes are high, this method ensures that the models are tested and refined to reduce risks, such as default rates and to optimise loan approval processes.

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a process model that provides a structured approach to planning of data mining projects. It is divided into six major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. (Piatetsky, 2018). The six phases are shown in the Figure 2: (Piatetsky, G., 2018, February 19)



- **Phase 1(Business Understanding):** Provides a depth overview of the project from a business perspective and addresses project requirements.

- **Phase 2(Data Understanding):** Utilise the available dataset and analyse its attributes, structure and quality.
- **Phase 3(Data Preprocessing):** Involves cleaning and transforming the data, addressing the missing values and outliers.
- **Phase 4(Modeling):** Implement machine learning models on the pre-processed data for loan prediction and applicant behaviour.
- **Phase 5(Evaluation):** Evaluates model performance using accuracy, precision and cross validation metrics.
- **Phase 6(Deployment):** This project suggests to utilise the advanced models of predicting loan approval systems for deployment in the real world scenario. It is also recommends for continuous monitoring to be implemented as the models can adapt to future data in order to maintain effectiveness and relevance.

3.6 Project Plan

The initial project plan was created based on the project specification, incorporating all phases of the CRISP-DM framework. However, as the project progressed and gained a deeper understanding from a business perspective, the data understanding and preprocessing phase took longer than expected. The dataset was unbalanced, requiring additional time for proper handling to ensure model accuracy. Furthermore, training the models was time-consuming, taking approximately 5-6 hours for each iteration. This extended training time made the tuning process particularly challenging and time intensive. The Gantt chart below shows the actual timeline that was followed to complete the project.

Table 2:

NO	TASK TITLE	START DATE	DUE DATE	STATUS	W 1	W 2	W 3	W 4	W 5	W 6	W 7	W 8	W 9	W 10	W 11	W 12	W 13	W 14	W 15
1	Using Available Data and Data/Business Understanding	03/06/24	21/06/24	100%															
2	Research about the Project	22/06/24	12/07/24																
3	Data Pre-processing	13/07/24	03/08/24																
4	Model Building and Evaluation	04/08/24	16/08/24																
5	Model Tuning and Finalization	17/08/24	30/08/24																
6	Results Analysis	31/08/24	06/09/24																
7	Conclusion and Recommendations	07/09/24	12/09/24																

3.7 Summary

This project management overview for the loan approval prediction study that follows the CRISP-DM framework, addressing the ethical considerations, costs and the methodology followed. It details six phases, key requirements, and a 15 weeks timeline. The focus is on balancing predictive accuracy with ethics, aiming to develop machine learning models that tackles challenges in data privacy, transparency, and the societal impacts of automated financial decision-making.

CHAPTER 4: BUSINESS UNDERSTANDING

4.1 Loan Products and Services

Financial institutions provide a range of loan solutions to meet the needs of applicants and produce profitability. Different types of loans are personal loans, mortgages, vehicle loans, business loans and credit lines. Due to the inherent risk, personal loans are often unsecured and need higher interest rates. Contrarily, it acquired asset serves as collateral for mortgages and vehicle loans, therefore enabling the use of reduced interest rates and extended payback periods. Business loans often need collateral and have intricate conditions customised to meet the business's specific requirements. Credit lines provide flexible, revolving credit but require careful management to ensure profitability.

4.2 Revenue Generation

These financial institutions mainly depend on the revenue they generate through these loans, which is 'interest'. Interest is calculated using the loan amount, period and rate. Revenue is generated from the applied loan amounts, late instalments and prepayment penalties. These firms, also cross-sell insurances and financial products to generate more revenue. These firms' main challenge is balancing profitability with risk management and ensuring a sustainable loan portfolio.

4.3 Risk Management

Risk is inherent in lending and effective risk management can prevent defaults (credit risk), volatility of market rates, inability to meet mandatory financial obligations and loss from internal processes. Tools like credit scoring models, stress testing and portfolio diversification help manage these risks, aligning with the institution's risk appetite while ensuring profitability and stability.

4.4 Regulatory Compliance

The lending industry is governed by complex regulations designed to protect consumers and ensure financial stability. Financial institutions must meet established anti-money laundering (AML) laws, Know Your Customer (KYC), fair lending practices and data privacy regulations like the General Data

Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA). Compliance is necessary for consumer protection, prevention of financial crimes and trust. Dedicated compliance departments and regular process updates are necessary to meet evolving standards.

4.5 Technology and Innovation

Lending is being moulded by technological advancements, with automated underwriting systems shortening loan approval times, alternative data sources widening credit access and blockchain strengthening security/ transparency. AI/machine learning refines risk assessment and fraud detection. Top priorities are achieving the delicate balance of innovation and regulatory compliance and managing the impact on existing processes and staff.

4.6 Market Competition

The lending market is exceeding competitively with fintech companies, peer-to-peer lending platforms and non-bank lenders posing difficulties for conventional banks. New competitors are offering custom solutions, faster approvals and digital-first experiences. Traditional lenders also need to improve their services and most enter into a partnership with fintech companies, or develop digital platforms. In this competitive landscape, the goal is to balance innovation with stability and trust.

4.7 Project Requirements

This project is focused on creating and suggesting for deploying a robust machine learning based solution. The essential requirements are:

4.7.1 Functional Requirements

This project is focuses on developing and recommends to deploy a machine learning-based solution. The essential requirements are:

Data: Use the available dataset from the Source. (The dataset is downloaded from Kaggle (2022) for this project.)

Data Preprocessing: Perform data cleaning, transformation and feature engineering to prepare the dataset for modeling.

Model Development: Develop two machine learning models for loan approval prediction and another for predicting applicant behaviour.

Model Evaluation: Evaluate the models using accuracy, precision, recall,F1 score and AUC metrics.

Data Visualisation: Create visualisations to understand data patterns, model performance.

Association Rules: Employ association rule mining to identify some specific rules that could assist loan authorities in making critical decisions.

4.7.2 Non-Functional Requirements

Performance: The system should able to handle large datasets efficiently without delays.

Accuracy of the Models: Aims for model accuracy between 80% and 95%.

Scalability: The system should support scaling with more data or complexity.

Usability: Ensure the system is user-friendly and easy to interact with.

Reliability: The system should be robust, handling errors in data or processes without crashing.

Security: The system must ensure data privacy and security if sensitive data is involved.

Maintainability: The system should be easy to get updated , modified and to be maintained as needed.

Ethical Considerations: The system should follow ethical guidelines when using sensitive user data.

4.7.3 New Insights

Client Feedback and Approval: whether requirements are implicitly or explicitly approved by the client.

Challenges and Solutions: Identify obstacles for meeting the requirements, as well of possible solutions to them.

Stakeholder Relevance: Highlight how important the project is to stakeholders.

4.7 Summary

The Business Understanding chapter outlines the key aspects of the lending industry, including various loan products, revenue generation and risk management. It emphasises the importance of regulatory compliance, technological innovations and the competitive landscape. This context is essential for developing machine learning models to predict loan approvals and applicant behaviour. The project aims to use this business knowledge to create practical tools and insights for banks and loan firms, balancing innovation with regulatory requirements and risk management.

CHAPTER 5: EXPLORATORY DATA ANALYSIS

"Exploratory data analysis is an attitude, a flexibility, and a reliance on insight rather than blindness to help solve real world statistical problems." - *John Tukey*

5.1 Introduction

This chapter presents the Exploratory Data Analysis (EDA) conducted for the study, providing insights into the dataset and its attributes. EDA refers to Exploratory Data Analysis for which Python is used in the Google Lab environment. The analysis uses libraries such as Pandas, NumPy, Matplotlib, and Seaborn. This chapter gives a concise and organised background to the EDA steps, laying the foundation for the subsequent loan approval prediction analysis.

5.2 Overview: Dataset & understanding the dataset

5.2.1 Utilise Available Dataset

The dataset for this study includes two CSV files:

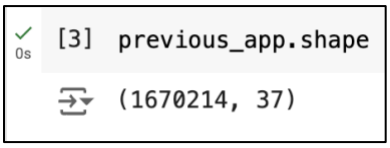
- Current_app dataset (166.13 MB): Information on existing loan applications, including payment difficulties.
- Previous_app dataset (404.97 MB): Details on previous loan applications with statuses like Approved, Cancelled, Refused, or Unused.

These anonymised datasets provide data privacy without the need for individual consent. Data file contributed by Shivam Kapoor (Kaggle, 2021). The currency is INR for analysis.

The dataset under analysis is named 'current_app' and contains 3,07,511 rows and 122 columns.

```
[5]  current_app.shape
0s
     (307511, 122)
```

The dataset under analysis is named 'previous_app' and contains 1,607,214 rows and 37 columns.



5.2.2 Datasets attributes

Loan Applicant Information: Attribute Explanations

5.2.2.1 In the 'current_app' dataset it has a total of 122 attributes.

For this project, 21 key features are selected for visualising applicants and modeling the behaviour analysis (defaulters and non-defaulters). And the remaining features are not important for this project study.

- SK_ID_CURR: Unique identifier for each loan application.
- TARGET: Binary indicator (1 for defaulters, 0 for non-defaulters).
- NAME_CONTRACT_TYPE: Loan type ('Cash loans' or 'Revolving loans').
- CODE_GENDER: Gender ('M', 'F', 'XNA').
- CNT_CHILDREN: Number of children.
- AMT_INCOME_TOTAL: Total income.
- AMT_CREDIT: Credit amount requested/approved.
- AMT_ANNUITY: Annuity amount for the loan.
- AMT_GOODS_PRICE: Price of goods/services requested.
- NAME_TYPE_SUITE: Type of suite/accompaniment.
- NAME_INCOME_TYPE: Source/type of income.

- NAME_EDUCATION_TYPE: Highest education level.
- NAME_FAMILY_STATUS: Marital/family status.
- DAYS_EMPLOYED: Days employed (negative for days before current date).
- OCCUPATION_TYPE: Job category.
- CNT_FAM_MEMBERS: Number of family members.
- REGION_RATING_CLIENT: Region rating (without city).
- REGION_RATING_CLIENT_W_CITY: Region rating (with city).
- REG_REGION_NOT_WORK_REGION: Region difference (registered vs. work).
- LIVE_REGION_NOT_WORK_REGION: Region difference (live vs. work).
- ORGANIZATION_TYPE: Type of organisation.

5.2.2.2 In the previous_app dataset it has a total of 37 attributes.

For this project, 30 key features are selected for visualising loan approval outcomes and modelling loan outcomes (approved, cancelled, refused, unused offer).

- SK_ID_PREV: Unique ID for each previous loan application.
- SK_ID_CURR: Unique ID for each current loan application.
- NAME_CONTRACT_TYPE: Type of loan contract ('Consumer loans', 'Cash loans', etc.).
- AMT_ANNUITY: Annuity amount for regular payments.
- AMT_APPLICATION: Loan amount requested by the borrower.
- AMT_CREDIT: Credit amount granted.
- AMT_DOWN_PAYMENT: Down payment made by the borrower.
- AMT_GOODS_PRICE: Price of goods/services for the loan.

- FLAG_LAST_APPL_PER_CONTRACT: Indicator if it's the last application for the contract.
- NFLAG_LAST_APPL_IN_DAY: Indicator if the application is the last submitted that day.
- RATE_DOWN_PAYMENT: Down payment to loan amount ratio.
- NAME_CASH_LOAN_PURPOSE: Purpose of the cash loan.
- NAME_CONTRACT_STATUS: Status of the loan application.
- DAYS_DECISION: Days since the loan decision was made.
- NAME_PAYMENT_TYPE: Method of loan payment.
- CODE_REJECT_REASON: Reason for loan rejection.
- NAME_TYPE_SUITE: Type of client accompanying the applicant.
- NAME_CLIENT_TYPE: Type of client applying for the loan.
- NAME_GOODS_CATEGORY: Category of goods/services for the loan.
- NAME_PORTFOLIO: Loan portfolio type.
- NAME_PRODUCT_TYPE: Product associated with the loan.
- CHANNEL_TYPE: Channel used for loan submission.
- NAME_SELLER_INDUSTRY: Industry of the seller.
- CNT_PAYMENT: Number of loan payments/instalments.
- NAME_YIELD_GROUP: Profitability category of the loan.
- PRODUCT_COMBINATION: Combination of product features.
- DAYS_FIRST_DRAWING: Days from application to first loan drawing.
- DAYS_LAST_DUE_1ST_VERSION: Days to the last due date in the first loan agreement.
- DAYS_LAST_DUE: Days to the last due date.
- DAYS_TERMINATION: Days to the loan termination date.
- WEEKDAY_APPR_PROCESS_START: Day of the week the loan process started.

5.3.1 Exploratory Data Analysis for numerical attributes (previous_app)

Table 3:

Attribute	Datatype	No. of Missing Values	Percentage of Missing Values
SK_ID_PREV	int64	0	0
SK_ID_CURR	int64	0	0
AMT_ANNUITY	float64	372235	22.29
AMT_APPLICATION	float64	0	0
AMT_CREDIT	float64	1	0
AMT_DOWN_PAYMENT	float64	895844	53.64
AMT_GOODS_PRICE	float64	385515	23.08
NFLAG_LAST_APPL_IN_DAY	int64	0	0
RATE_DOWN_PAYMENT	float64	895844	53.64
DAYS_DECISION	int64	0	0
CNT_PAYMENT	float64	372230	22.29
DAYS_FIRST_DRAWING	float64	673065	40.3
DAYS_LAST_DUE_1ST_VERSION	float64	673065	40.3
DAYS_LAST_DUE	float64	673065	40.3
DAYS_TERMINATION	float64	673065	40.3

Analysis of Missing Values in previous_app Dataset (numerical attribute):

- **Complete Attributes**: SK_ID_PREV, SK_ID_CURR, AMT_APPLICATION, NFLAG_LAST_APPL_IN_DAY, DAYS_DECISION
- **Negligible Missing Value:** AMT_CREDIT (1 missing value)
- **Significant Missing Data:**AMT_DOWN_PAYMENT, RATE_DOWN_PAYMENT: 53.64%

  DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION, NFLAG_INSURED_ON_APPROVAL: 40.3%

  AMT_GOODS_PRICE: 23.08%

  AMT_ANNUITY, CNT_PAYMENT: 22.29%

Some facts need to be focused on this dataset:

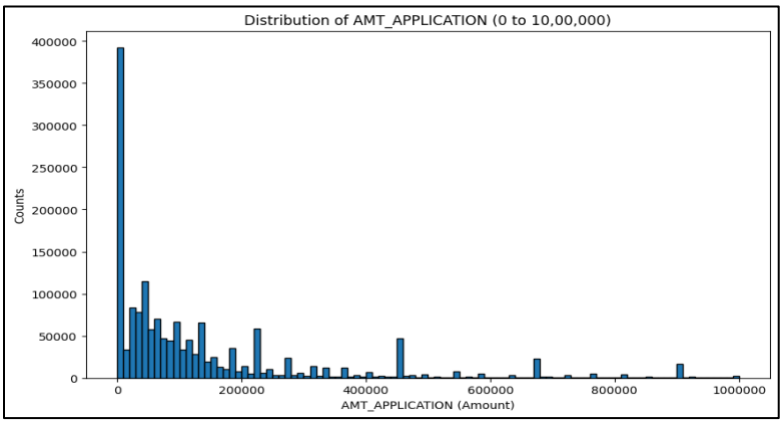
The repeated 'SK_ID_CURR' values show that some applicants have applied for loans multiple times.

This allows us to track the outcomes for the applicant and understand how often they applied for loans.

```
[23] # Get the 'SK_ID_PREV' column
     sk_id_prev = previous_app['SK_ID_CURR']

     # Calculate the number of unique values
     num_unique1 = sk_id_prev.nunique()

     # Calculate the number of repeated values
     num_repeated1 = sk_id_prev.duplicated().sum()

     # Display the summary
     print(f"Number of unique 'SK_ID_CURR' values: {num_unique1}")
     print(f"Number of repeated 'SK_ID_CURR' values: {num_repeated1}")

     Number of unique 'SK_ID_CURR' values: 338857
     Number of repeated 'SK_ID_CURR' values: 1331357
```

5.3.1.1 Amount Application Plot

The below histogram plots the loan applicants' distribution amount application range. (Figure 3)



The above histogram illustrates that most applicants have applied for loans in the range of ₹0 to Rs. 2,00,000. The number of applicants significantly decreases as the loan amount approaches Rs. 10,00,000.

5.3.2 Exploratory Data Analysis for numerical attributes (current_app)

Table 4:

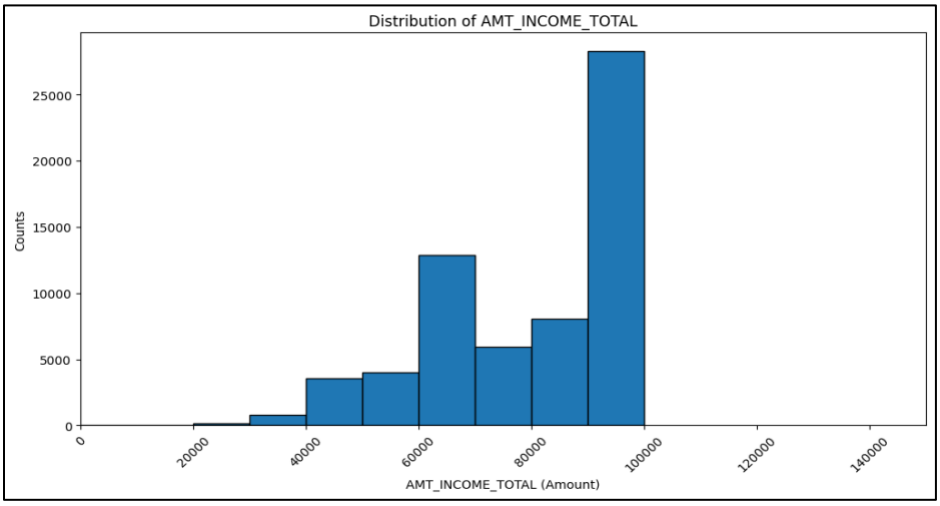
Attribute	Datatype	No. of Missing Values	Percentage of Missing Values (%)
TARGET	int64	0	0
SK_ID_CURR	int64	0	0
CNT_CHILDREN	int64	0	0
AMT_INCOME_TOTAL	float64	0	0
AMT_CREDIT	float64	0	0
AMT_ANNUITY	float64	12	0.004
AMT_GOODS_PRICE	float64	278	0.09
CNT_FAM_MEMBERS	float64	2	0
REGION_RATING_CLIENT	int64	0	0
REGION_RATING_CLIENT_W_CITY	int64	0	0
REG_REGION_NOT_WORK_REGION	int64	0	0
LIVE_REGION_NOT_WORK_REGION	int64	0	0

The dataset generally shows good completeness, with most attributes having no missing values. However, there are a few exceptions:

- AMT_GOODS_PRICE has the highest number of missing values at 278 (0.09% of the data).
- AMT_ANNUITY has 12 missing values (0.004%).
- CNT_FAM_MEMBERS has 2 missing values (0.001%).

5.3.2.1 AMT_INCOME_TOTAL Plot

The below histogram plots the loan applicants' distribution income range. For analyse this visualisation of loan applicants' total income range, it focuses on the range from Rs. 0 to Rs. 12,50,000, with the highest concentration of applications. (Figure 3)



The above histogram illustrates the distribution of total income (AMT_INCOME_TOTAL) of loan applicants. It clearly shows that the largest group of applicants has a total income of Rs. 1,00,000 represented by the tallest bar. The second most common income level is around Rs. 70,000 with approximately 12,000 applicants in this range. There are a few applicants with total incomes between Rs. 20,000 and Rs. 40,000.

5.4.1 Exploratory Data Analysis for categorical attributes(current_app)

Table 5:

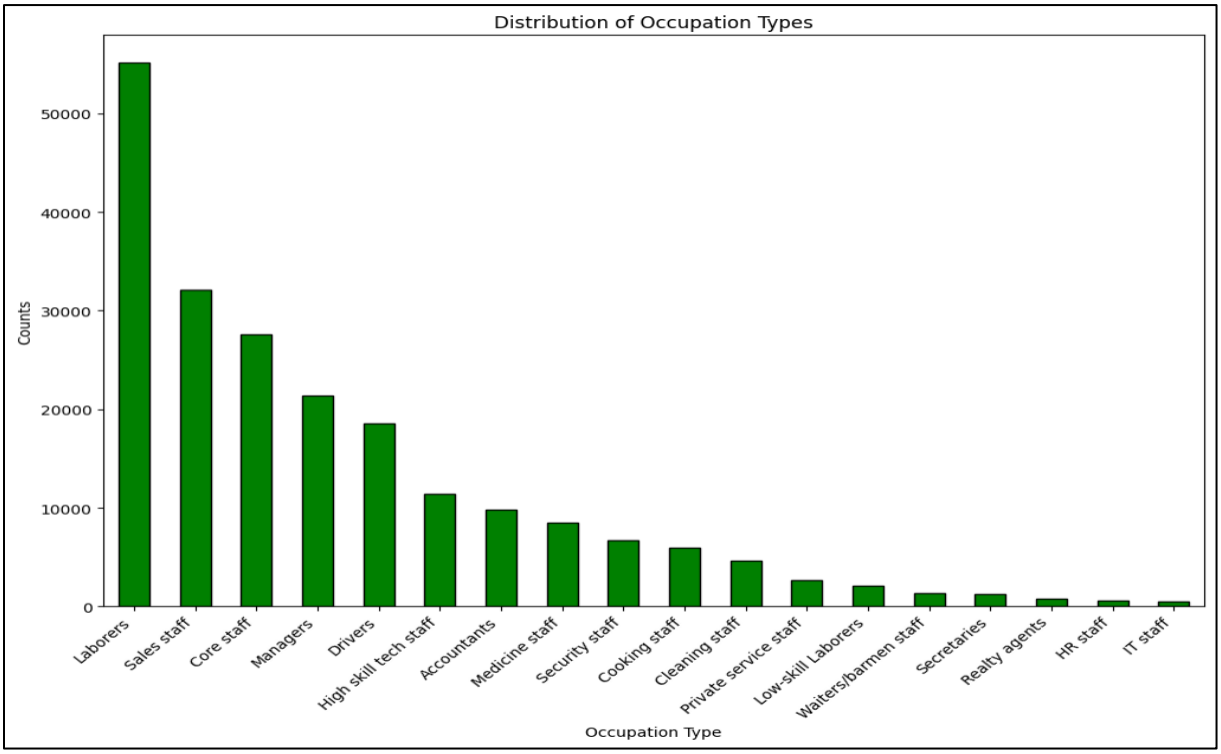
Attribute	Datatype	No. of Unique Variables	List of Unique Variables	No. of Missing Values	Missing Values (%)
NAME_CONTRACT_TYPE	object	2	['Cash loans', 'Revolving loans']	0	0
CODE_GENDER	object	3	['M', 'F', 'XNA']	0	0
NAME_TYPE_SUITE	object	7	['Unaccompanied', 'Family', 'Spouse, partner', 'Children', 'Other_A', 'Other_B', 'Group of people']	1292	0.42
NAME_INCOME_TYPE	object	8	['Working', 'State servant', 'Commercial associate', 'Pensioner', 'Unemployed', 'Student', 'Businessman', 'Maternity leave']	0	0
NAME_EDUCATION_TYPE	object	5	['Secondary / secondary special', 'Higher education', 'Incomplete higher', 'Lower secondary', 'Academic degree']	0	0
NAME_FAMILY_STATUS	object	6	['Single / not married', 'Married', 'Civil marriage', 'Widow', 'Separated', 'Unknown']	0	0
OCCUPATION_TYPE	object	18	['Laborers', 'Core staff', 'Accountants', 'Managers', 'Drivers', 'Sales staff', 'Cleaning staff', 'Cooking staff', 'Private service staff', 'Medicine staff', 'Security staff', 'High skill tech staff', 'Waiters/barmen staff', 'Low-skill Laborers', 'Realty agents', 'Secretaries', 'IT staff', 'HR staff']	96391	31.35
ORGANIZATION_TYPE	object	58	['Business Entity Type 3', 'School', 'Government', 'Religion', 'Other', 'XNA', 'Electricity', 'Medicine', 'Business Entity Type 2', 'Self-employed', 'Transport: type 2', 'Construction', 'Housing', 'Kindergarten', 'Trade: type 7', 'Industry: type 11', 'Military', 'Services', 'Security Ministries', 'Transport: type 4', 'Industry: type 1', 'Emergency', 'Security', 'Trade: type 2', 'University', 'Transport: type 3', 'Police', 'Business Entity Type 1', 'Postal', 'Industry: type 4', 'Agriculture', 'Restaurant', 'Culture', 'Hotel', 'Industry: type 7', 'Trade: type 3', 'Industry: type 3', 'Bank', 'Industry: type 9', 'Insurance', 'Trade: type 6', 'Industry: type 2', 'Transport: type 1', 'Industry: type 12', 'Mobile', 'Trade: type 1', 'Industry: type 5', 'Industry: type 10', 'Legal Services', 'Advertising', 'Trade: type 5', 'Cleaning', 'Industry: type 13', 'Trade: type 4', 'Telecom', 'Industry: type 8', 'Realtor', 'Industry: type 6']	0	0

Several categorical attributes in the dataset have significant proportions of missing values, such as OCCUPATION_TYPE (31.35%) and NAME_TYPE_SUITE (0.42%). The unique values provide

insights into loan applicants, covering demographic and economic factors, including loan types, gender, income sources, education levels, and family statuses. It also captures applicants' occupations and the types of organisations they work for, offering a comprehensive profile of those applying for loans.

5.4.1.1 OCCUPATION_TYPE plot

(Figure 4)



The bar plot indicates that labourers constitute the largest group among loan applicants. Their count exceeds 50,000, making them the most numerous occupation type in this dataset. Other categories, such as sales staff, core staff, managers, and drivers, are significantly lower in comparison. The categories IT staff, HR staff, Realty agents, secretaries, waiters/barman staff, and Low-skill labourers have very minimal applicants.

5.4.2 Exploratory Data Analysis for categorical attributes(previous_app)

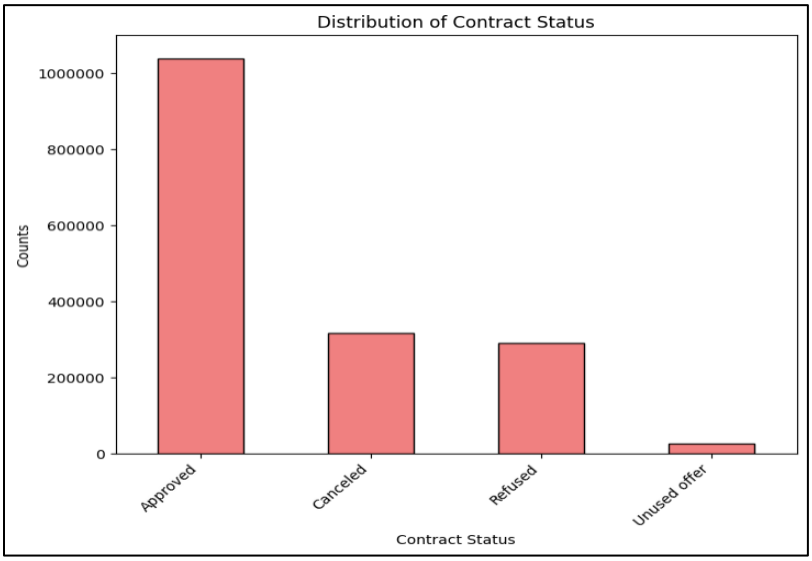
Table 6:

Attribute	Datatype	No. of Unique Variables	Unique Variables	No. of Missing Values	Missing Values(%)
NAME_CONTRACT_TYPE	object	4	['Consumer loans', 'Cash loans', 'Revolving loans', 'XNA']	0	0
WEEKDAY_APPR_PROCESS_START	object	7	['SATURDAY', 'THURSDAY', 'TUESDAY', 'MONDAY', 'FRIDAY', 'SUNDAY', 'WEDNESDAY']	0	0
FLAG_LAST_APPL_PER_CONTRACT	object	2	['Y', 'N']	0	0
NAME_CASH_LOAN_PURPOSE	object	25	['XAP', 'XNA', 'Repairs', 'Everyday expenses', 'Car repairs', 'Building a house or an annex', 'Other', 'Journey', 'Purchase of electronic equipment', 'Medicine', 'Payments on other loans', 'Urgent needs', 'Buying a used car', 'Buying a new car', 'Buying a holiday home / land', 'Education', 'Buying a home', 'Furniture', 'Buying a garage', 'Business development', 'Wedding / gift / holiday', 'Hobby', 'Gasification / water supply', 'Refusal to name the goal', 'Money for a third person']	0	0
NAME_CONTRACT_STATUS	object	4	['Approved', 'Refused', 'Canceled', 'Unused offer']	0	0
NAME_PAYMENT_TYPE	object	4	['Cash through the bank', 'XNA', 'Non-cash from your account', 'Cashless from the account of the employer']	0	0
CODE_REJECT_REASON	object	9	['XAP', 'HC', 'LIMIT', 'CLIENT', 'SCOFR', 'SCO', 'XNA', 'VERIF', 'SYSTEM']	0	0
NAME_TYPE_SUITE	object	7	['Unaccompanied', 'Spouse, partner', 'Family', 'Children', 'Other_B', 'Other_A', 'Group of people']	820405	49.12
NAME_CLIENT_TYPE	object	4	['Repeater', 'New', 'Refreshed', 'XNA']	0	0
NAME_GOODS_CATEGORY	object	28	['Mobile', 'XNA', 'Consumer Electronics', 'Construction Materials', 'Auto Accessories', 'Photo / Cinema Equipment', 'Computers', 'Audio/Video', 'Medicine', 'Clothing and Accessories', 'Furniture', 'Sport and Leisure', 'Homewares', 'Gardening', 'Jewelry', 'Vehicles', 'Education', 'Medical Supplies', 'Other', 'Direct Sales', 'Office Appliances', 'Fitness', 'Tourism', 'Insurance', 'Additional Service', 'Weapon', 'Animals', 'House Construction']	0	0
NAME_PORTFOLIO	object	5	['POS', 'Cash', 'XNA', 'Cards', 'Cars']	0	0
NAME_PRODUCT_TYPE	object	3	['XNA', 'x-sell', 'walk-in']	0	0
CHANNEL_TYPE	object	8	['Country-wide', 'Contact center', 'Credit and cash offices', 'Stone', 'Regional / Local', 'AP+ (Cash loan)', 'Channel of corporate sales', 'Car dealer']	0	0
NAME_SELLER_INDUSTRY	object	11	['Connectivity', 'XNA', 'Consumer electronics', 'Industry', 'Clothing', 'Furniture', 'Construction', 'Jewelry', 'Auto technology', 'MLM partners', 'Tourism']	0	0
NAME_YIELD_GROUP	object	5	['middle', 'low_action', 'high', 'low_normal', 'XNA']	0	0
PRODUCT_COMBINATION	object	17	['POS mobile with interest', 'Cash X-Sell: low', 'Cash X-Sell: high', 'Cash X-Sell: middle', 'Cash Street: high', 'Cash', 'POS household without interest', 'POS household with interest', 'POS other with interest', 'Card X-Sell', 'POS mobile without interest', 'Card Street', 'POS industry with interest', 'Cash Street: low', 'POS industry without interest', 'Cash Street: middle', 'POS others without interest']	346	0.02

Approximately 1.33% of the values in the categorical attributes of the previous_app dataset are missing, with most attributes having no missing values. Notable exceptions are NAME_TYPE_SUITE (49.12%

missing) and PRODUCT_COMBINATION (0.02% missing). The unique values provide insights into applicants' loan types, purposes and buying preferences across goods categories and industries, offering a concise profile of their financial behaviours.

5.4.2.1 NAME_CONTRACT_STATUS plot



(Figure 5)

The bar plot illustrates the loan status of loan applicants. Most of the applicants' loan applications were 'approved', while fewer applicants were either 'cancelled' or 'refused'. The lowest contract status is 'Unused offer.'

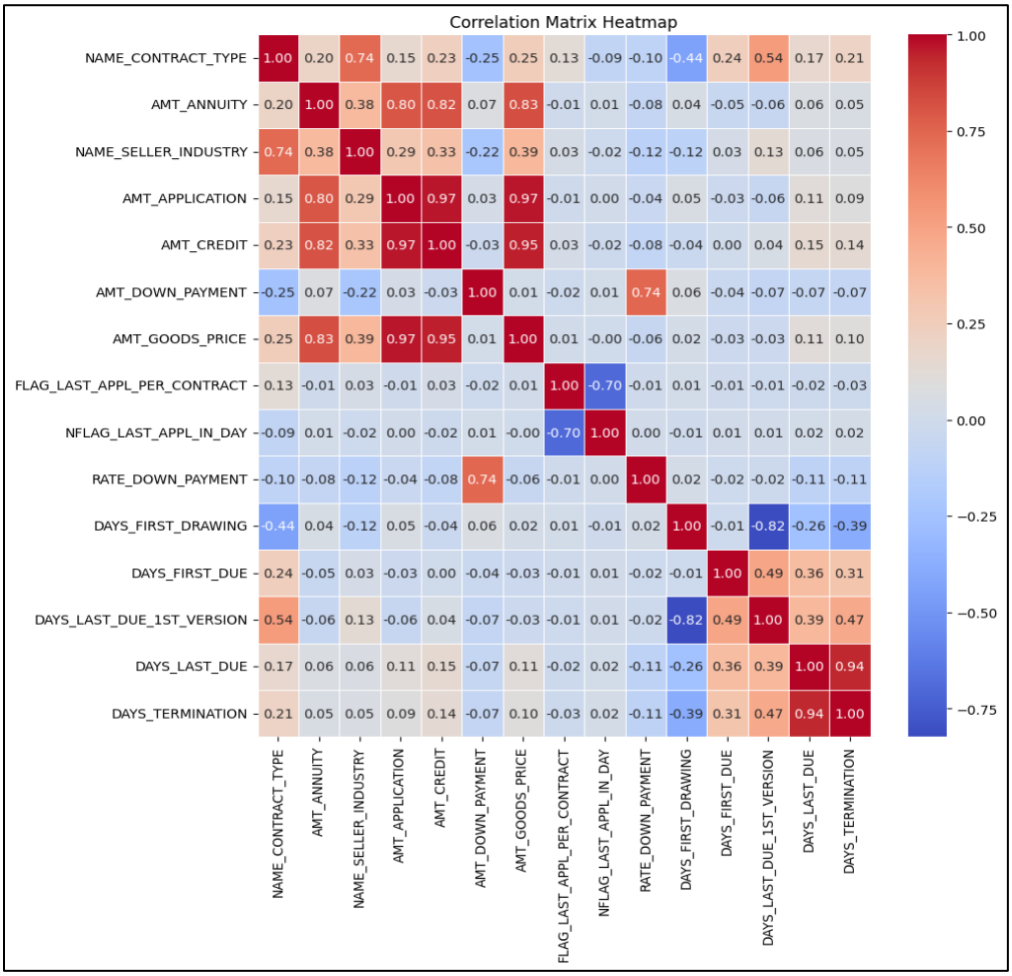
5.5 Correlation

Converted categorical to numerical data(ordinal) .

The heatmap of this correlation matrix presentation . Colour intensity is indicative of the correlation strength and direction where red implies positive correlations, blue imply negative correlations.

5.5.2 Correlation Analysis for 'current_app' Dataset

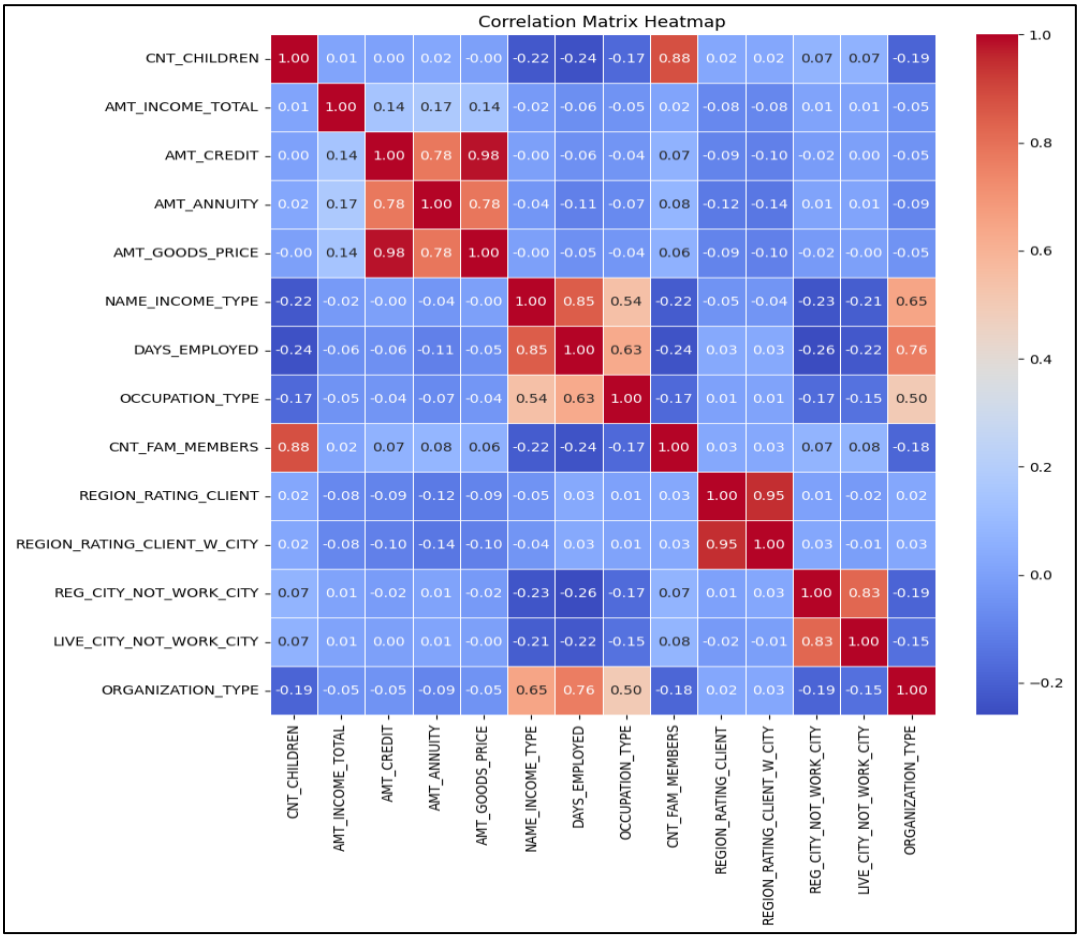
Figure 6:



There is a strong positive correlation between CNT_CHILDREN and CNT_FAM_MEMBERS, showing that more children increase family size. AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE are closely linked, with higher loan amounts leading to higher annuities and goods prices. NAME_INCOME_TYPE, DAYS_EMPLOYED and ORGANIZATION_TYPE indicate job stability, while ORGANIZATION_TYPE and NAME_INCOME_TYPE show income differences across sectors. REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY are positively correlated and REG_CITY_NOT_WORK_CITY and LIVE_CITY_NOT_WORK_CITY reflect commuting patterns

5.5.2 Correlation Analysis for 'previous_app' Dataset

Figure 7:



AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE and related attributes show strong positive correlations with DAYS_LAST_DUE and DAYS_TERMINATION, indicating that higher loan amounts lead to higher annuities and goods prices. In contrast, FLAG_LAST_APPL_PER_CONTRACT, NFLAG_LAST_APPL_IN_DAY, DAYS_FIRST_DRAWING and DAYS_LAST_DUE_1ST_VERSION display negative correlations, suggesting that as one increases, the others tend to decrease. For example, an increase in DAYS_FIRST_DRAWING may lead to a decrease in DAYS_LAST_DUE_1ST_VERSION.

5.6 Summary

Based on the analysis (EDA), there are two datasets. The 'current_app' dataset doesn't have application amounts; it contains all applicant information. The 'previous_app' dataset has multiple records of applicants for applied loans and the status of their loans, whether approved or not. The 'current_app' dataset will be used for visualisation, behaviour analysis and developing machine learning models to identify applicant behaviour as defaulters or non-defaulters. The 'previous_app' dataset will be used for predicting loan approval status (approved, cancelled, refused, or unused) and deriving association rules for loan prediction patterns.

CHAPTER 6: DATA PREPARATION

"The goal is to turn data into information, and information into insights."
-Carly Fiorina

6.1 Introduction

Data preparation is a critical step in any data analytics project, involving tasks such as data cleaning, transformation and organisation to ensure the data is in an appropriate format for analysis.

Key aspects of data preparation include:

- Ensuring data accuracy and validity
- Minimising the possibility of incorrect conclusions
- Enhancing the performance of analytical algorithms
- Reducing uncertainty in results by enabling more accurate data interpretation

This section will examine specific data preparation techniques and outline the methods to convert unorganised data into an organised format for the project's analysis.

6.2 Data Integrity and Alignment:Ensuring Consistency Between Current and Previous Applications

As mentioned in the EDA chapter, the applicant IDs (SK_ID_CURR) have been repeated in the previous_app dataset, suggesting that some applicants have applied for loans multiple times for various objectives and with varied loan amounts. This repetition allows for tracking the outcomes for these applicants and understanding how often they applied for loans.

In order to get precise insights for the same applicants, data preparation entails matching the prior and current application datasets by extracting the common 'SK_ID_CURR' values to remove discrepancies. Implementing this method enhances the data quality and guarantees that all distinct 'SK_ID_CURR' IDs represent the same set of applicants.

To ensure unbiased modeling:

Filter the same set of applicants: This step ensures that the dataset utilised for modeling corresponds to the most recent loan applications by the same applicants, thus providing a consistent basis for the analysis.

Avoid bias: By focusing on the latest applications, the model can avoid biases arising from multiple applications by the same applicants over time.

```
Filtered current_app shape: (291057, 122)
Filtered previous_app shape: (291057, 37)
```

After filtering, the 'current_app' and 'previous_app' datasets each contain 291,057 records. The current_app dataset has 122 features, and the 'previous_app' dataset contains 37 features.

6.3 Handling missing values

6.3.1 Handling missing values for 'current_app' dataset

The output shown represents the dataset features after dropping those with more than 40% missing values and removing irrelevant features.

```
Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
       'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY',
       'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE',
       'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE',
       'DAYS_EMPLOYED', 'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS',
       'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY',
       'WEEKDAY_APPR_PROCESS_START', 'REG_CITY_NOT_WORK_CITY',
       'LIVE_CITY_NOT_WORK_CITY', 'ORGANIZATION_TYPE'],
      dtype='object')
```

The output shows the specific features after addressing missing values by imputing with the median, reducing sensitivity to outliers and providing a reliable estimate for skewed financial data.

```
Missing Values in the Specified Columns:
AMT_ANNUITY        0
AMT_GOODS_PRICE    0
dtype: int64
```

The missing values for categorical features were replaced with 'Unknown', indicating missing information. This approach helps prevent bias and ensures data integrity during analysis.

```
Number of missing values in 'NAME_TYPE_SUITE': 0
Unique values in 'NAME_TYPE_SUITE': ['Unaccompanied' 'Family' 'Spouse, partner' 'Children' 'Other_A'
 'Group of people' 'Other_B' 'Unknown']
Number of missing values in 'OCCUPATION_TYPE': 0
Unique values in 'OCCUPATION_TYPE': ['Laborers' 'Core staff' 'Accountants' 'Managers' 'Unknown' 'Drivers'
 'Sales staff' 'Cleaning staff' 'Private service staff' 'Medicine staff'
 'Security staff' 'Cooking staff' 'High skill tech staff'
 'Waiters/barmen staff' 'Low-skill Laborers' 'Realty agents' 'Secretaries'
 'IT staff' 'HR staff']
```

The mode function is utilised to handle the distribution of family sizes, imputing the most common family size among loan applicants to ensure the data remains consistent and balanced.

```
Number of missing values in 'CNT_FAM_MEMBERS' after imputation: 0
```

6.2.2 Handling missing values for "previous_app" dataset

The output shown represents the dataset features after dropping the irrelevant features.

```
Index(['SK_ID_PREV', 'SK_ID_CURR', 'NAME_CONTRACT_TYPE', 'AMT_ANNUITY',
       'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_DOWN_PAYMENT', 'AMT_GOODS_PRICE',
       'WEEKDAY_APPR_PROCESS_START', 'FLAG_LAST_APPL_PER_CONTRACT',
       'NFLAG_LAST_APPL_IN_DAY', 'RATE_DOWN_PAYMENT', 'NAME_CASH_LOAN_PURPOSE',
       'NAME_CONTRACT_STATUS', 'DAYS_DECISION', 'NAME_PAYMENT_TYPE',
       'CODE_REJECT_REASON', 'NAME_TYPE_SUITE', 'NAME_CLIENT_TYPE',
       'NAME_GOODS_CATEGORY', 'NAME_PORTFOLIO', 'NAME_PRODUCT_TYPE',
       'CHANNEL_TYPE', 'NAME_SELLER_INDUSTRY', 'CNT_PAYMENT',
       'NAME_YIELD_GROUP', 'PRODUCT_COMBINATION', 'DAYS_FIRST_DRAWING',
       'DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE', 'DAYS_TERMINATION'],
      dtype='object')
```

The output shows the specific features after addressing missing values by imputing with the median, reducing sensitivity to outliers and providing a reliable estimate for skewed financial data.

```
Missing Values in AMT_ANNUITY:
0

Missing Values in AMT_DOWN_PAYMENT:
0

Missing Values in AMT_GOODS_PRICE:
0

Missing Values in RATE_DOWN_PAYMENT:
0

Missing Values in CNT_PAYMENT:
0

Missing Values in DAYS_FIRST_DRAWING:
0

Missing Values in DAYS_LAST_DUE_1ST_VERSION:
0

Missing Values in DAYS_LAST_DUE:
0

Missing Values in DAYS_TERMINATION:
0
```

The output shows the dataset after imputing missing values in the specified features using the mode function, which replaces them with the most frequent value for analysis.

```
Number of missing values in 'PRODUCT_COMBINATION' after imputation: 0
```

The missing values for categorical features were replaced with 'Unknown', indicating missing information.

```
Number of missing values in 'NAME_TYPE_SUITE' after replacement: 0
Unique values in 'NAME_TYPE_SUITE': ['Unknown' 'Family' 'Unaccompanied' 'Spouse, partner' 'Other_A' 'Other_B'
 'Children' 'Group of people']
```

6.3 Handling negative values

Negative values represent durations or time periods relative to a reference point, such as the present date. To handle negative values for modeling purposes, these are interpreted as positive durations. Thus, these features are converted to absolute values that align with other positive numerical features in datasets.

```
Before converting to absolute values:
   DAYS_DECISION  DAYS_FIRST_DRAWING  DAYS_LAST_DUE_1ST_VERSION
0           -606            365243.0                      125.0
1           -828            365243.0                     -647.0
2           -815            365243.0                     -694.0
3           -181            365243.0                     -315.0
4           -867            365243.0                     -315.0

   DAYS_LAST_DUE  DAYS_TERMINATION
0          -25.0             -17.0
1         -647.0            -639.0
2         -724.0            -714.0
3         -455.0            -418.0
4         -455.0            -418.0
```

```
After converting to absolute values:
   DAYS_DECISION  DAYS_FIRST_DRAWING  DAYS_LAST_DUE_1ST_VERSION
0            606            365243.0                      125.0
1            828            365243.0                      647.0
2            815            365243.0                      694.0
3            181            365243.0                      315.0
4            867            365243.0                      315.0

   DAYS_LAST_DUE  DAYS_TERMINATION
0           25.0              17.0
1          647.0             639.0
2          724.0             714.0
3          455.0             418.0
4          455.0             418.0
```

6.4 Handling structural error

In the NAME_CONTRACT_TYPE feature, the value 'XNA' appeared 48 times, representing a very small portion of the data. To maintain data consistency and ensure the same number of applicants in both datasets, 'XNA' is replaced with the mode value.

```
NAME_CONTRACT_TYPE
Consumer loans     168120
Cash loans          94068
Revolving loans     28821
XNA                    48
Name: count, dtype: int64
```

```
NAME_CONTRACT_TYPE
Consumer loans     168168
Cash loans          94068
Revolving loans     28821
Name: count, dtype: int64
```

The value 'XNA' is replaced with 'Not Available' to indicate missing or non-applicable information, and 'XAP' is replaced with 'Accepted After Initial Rejection' in the 'CODE_REJECT_REASON' feature to retain insights from cases with initial rejection. 'XNA' in goods categories was replaced with 'Not Specified', and across multiple columns, 'XNA' was replaced with 'Unknown' to ensure clarity and consistency in the dataset.

```
Before imputation:
Number of 'XNA' entries in 'NAME_GOODS_CATEGORY': 124414
Number of 'XNA' entries in 'CODE_REJECT_REASON': 608
Number of 'XAP' entries in 'CODE_REJECT_REASON': 252630
Number of 'XNA' entries in 'NAME_CASH_LOAN_PURPOSE': 84258
Number of 'XAP' entries in 'NAME_CASH_LOAN_PURPOSE': 196989
```

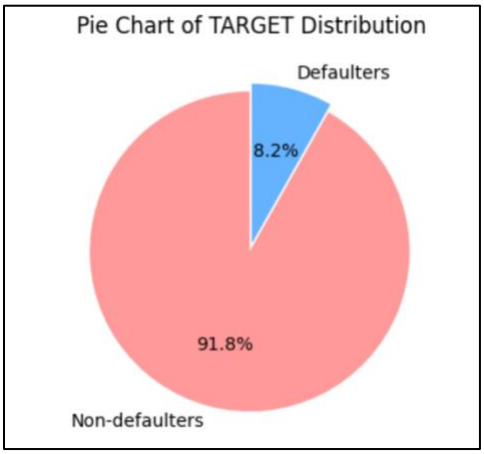
```
After imputation:
Number of 'XNA' entries in 'NAME_GOODS_CATEGORY': 0
Number of 'XNA' entries in 'CODE_REJECT_REASON': 0
Number of 'XAP' entries in 'CODE_REJECT_REASON': 0
Number of 'XNA' entries in 'NAME_CASH_LOAN_PURPOSE': 0
Number of 'XAP' entries in 'NAME_CASH_LOAN_PURPOSE': 0
```

6.4 Features Visualisation

6.4.1 Feature Visualization for Applicants with Existing Loan Applications: Whether Clients Have Payment Difficulties (Dataset: current_app)

Distribution of applicants for 'Defaulters' Vs 'Non-Defaulters'

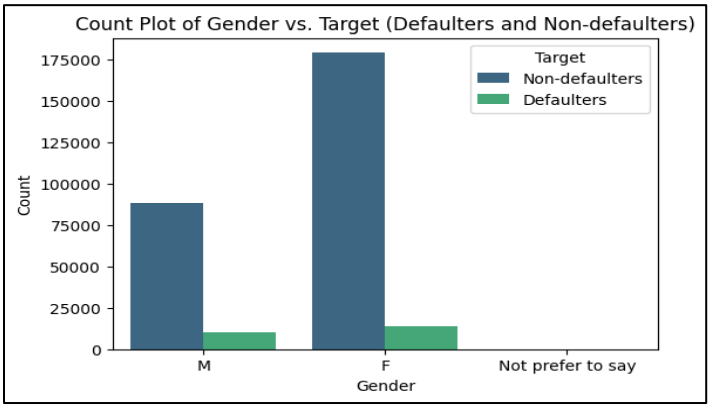
Figure 8:



This pie chart illustrates the distribution of applicants in a loan approval prediction dataset based on their TARGET values, where 91.8% of applicants successfully repay their loans(non-defaulters).The smaller blue portion of 8.2% of the pie chart were unable to meet their loan repayment obligations(defaulters).

Distribution of Loan Defaulters and Non-Defaulters by Gender

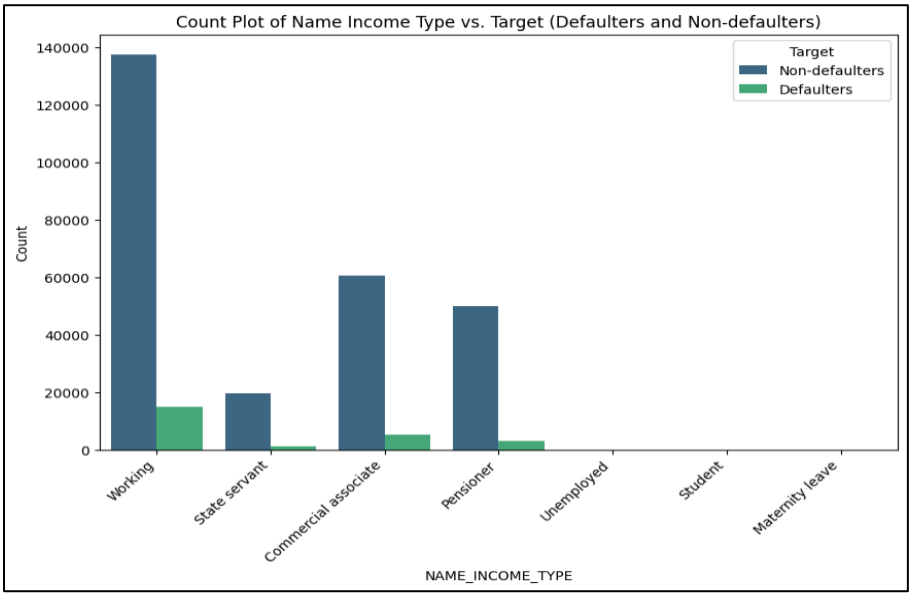
Figure 9:



The count plot compares male (M), female (F), and "Not prefer to say" gender categories against loan default status (TARGET). It shows that both male and female applicants predominantly belong to the non-defaulters category. Female applicants are more represented among both defaulters and non-defaulters, likely due to the larger number of female applicants. The "Not prefer to say" category is insignificant in the data.

Distribution of Loan Defaulters and Non-Defaulters by Income type

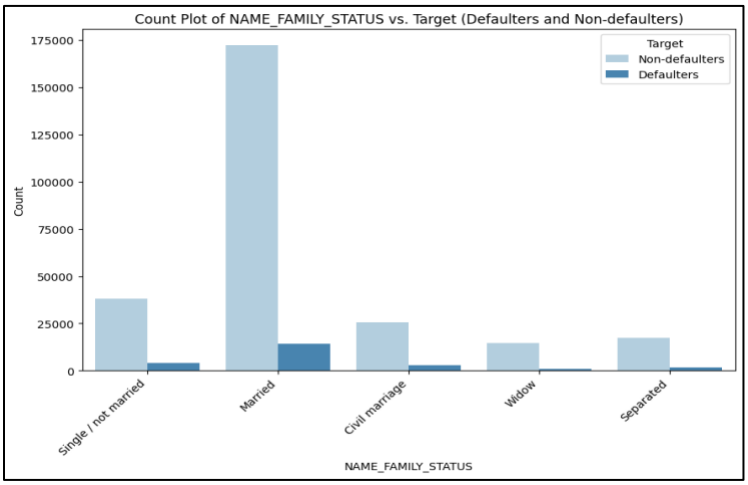
Figure 10:



The bar chart shows the distribution of income types among defaulters and non-defaulters. "Working" individuals are the largest group in both categories, mostly non-defaulters. "State servants," "Commercial associates" and "Pensioners" also have more non-defaulters but in smaller numbers. "Unemployed," "Students" and "Maternity leave" categories have very few individuals in both groups.

Count Plot of NAME_FAMILY_STATUS vs. Target (Defaulters and Non-defaulters)

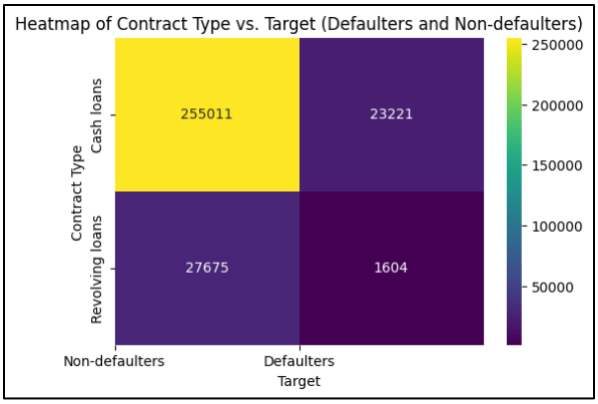
Figure 11:



The bar chart shows the distribution of family status categories among defaulters and non-defaulters. Married individuals are the largest group, mostly non-defaulters. Single/Not Married is the second-largest category, with more non-defaulters but a notable number of defaulters. Civil Marriage, Widow and Separated categories have fewer individuals, mostly non-defaulters. Across all statuses, non-defaulters are the majority but defaulters are present in every category, especially among singles and those in civil marriages.

Heatmap of Contract Type vs. Loan Default Status

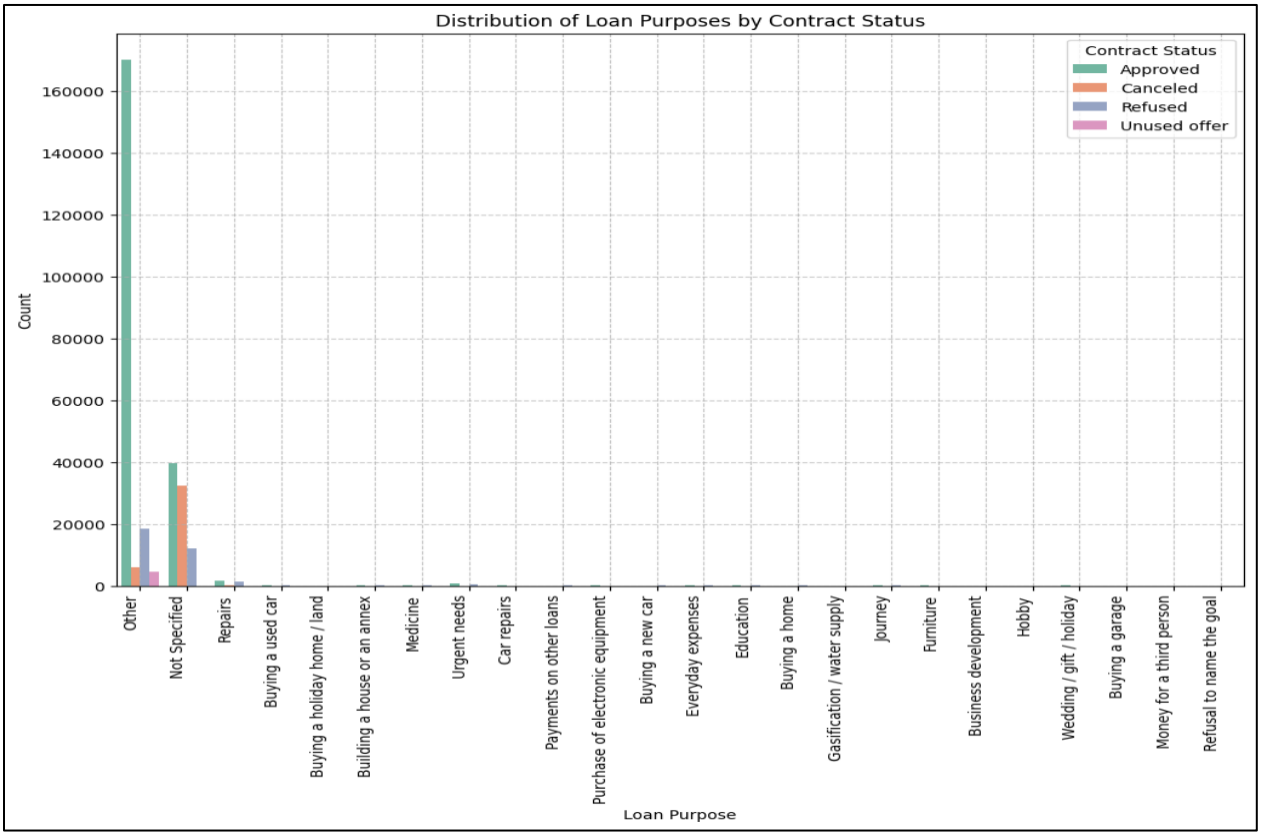
Figure 12:



The heat map shows the relationship between contract type (Cash loans vs. Revolving loans) and loan default status (Defaulters vs. Non-defaulters). Most cash loans are held by non-defaulters (242,302), with 22,366 defaulters. For revolving loans, non-defaulters hold the majority (24,910), with 1,479 defaulters. Cash loans are more common but the proportion of defaulters is higher among cash loan holders compared to revolving loan holders.

6.4.2 Feature Visualization for Applicants with Previous Loan Applications. (Dataset: previous_app)

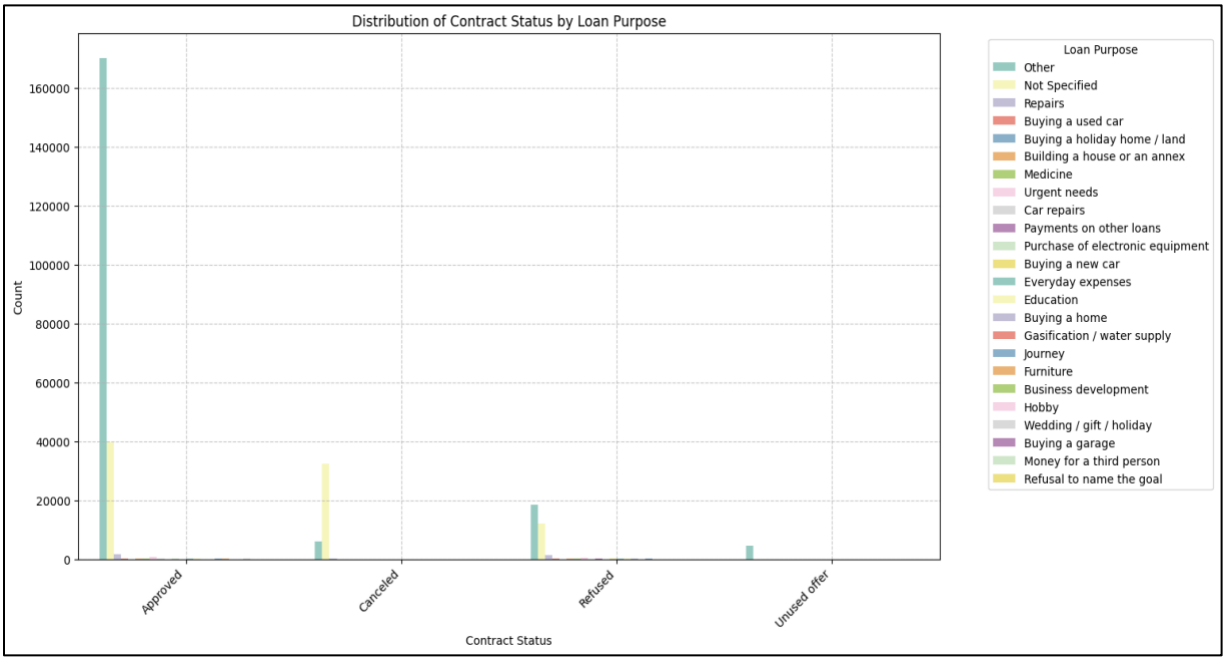
Distribution of Loan Purposes by Contract Status (Figure 12):



The bar chart shows the distribution of loan purposes by contract status. "Other" and "Not Specified" categories have the highest frequencies, indicating many loans lack clearly defined purposes. Approved loans dominate all categories, especially in "Other" and "Not Specified." Cancelled and refused loans are

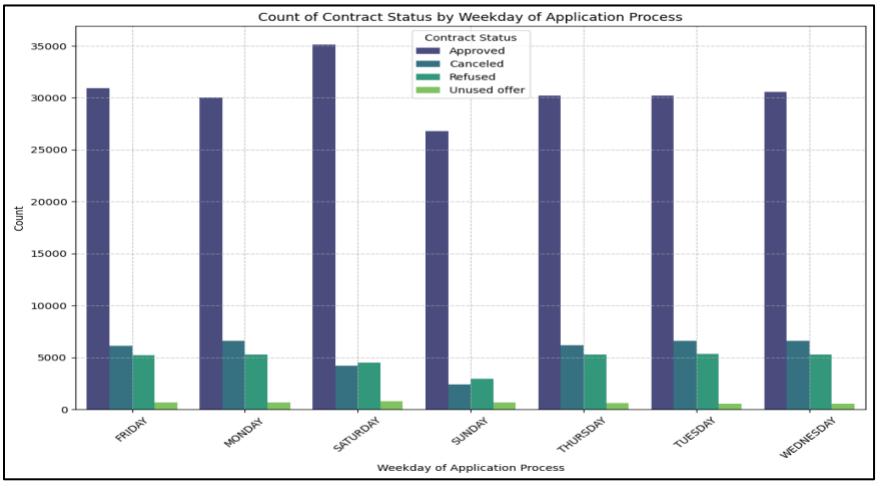
present but much lower in count. "Unused offer" status has the lowest frequency across all loan purposes. Overall, most loans are approved, regardless of purpose.

Distribution of Contract States by Loan Purpose (Figure 13:)



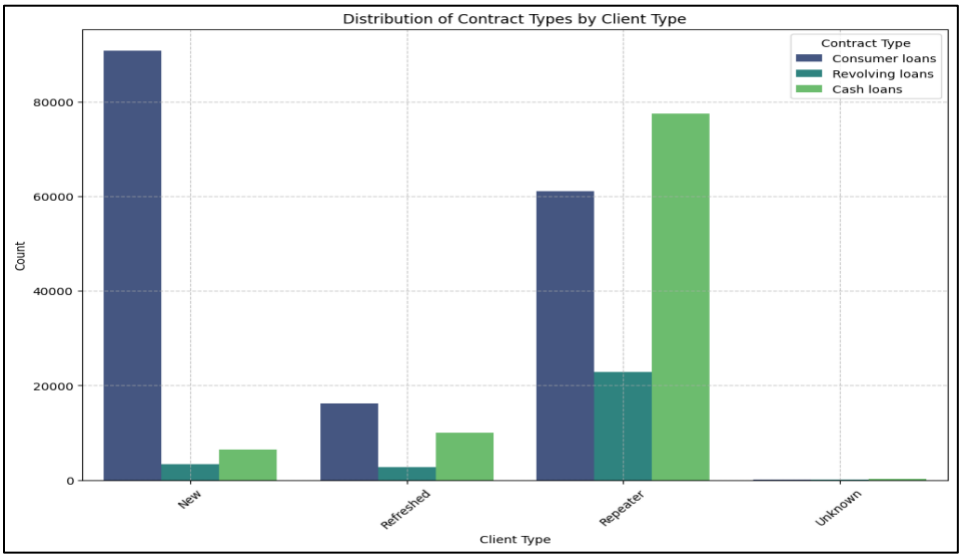
This graph is a bar chart of contract status distribution by loan purpose. In the bar chart, the most of our loan applicants are approved with Buying a new car and Not Specified being top two purposes for acquiring loans. Results there around the other loan purposes are not so significant and everywhere we have very low counts for status — 'Cancelled' or/and status — Unused offer'.

Distribution of Contract States by Weekday of Application Process (Figure 14:)



This bar chart visualises contract statuses across weekdays. Most loan applications are approved consistently, ranging between 30,000 and 35,000. Refused and cancelled applications are significant but lower, decreasing slightly on weekends. Unused offers are the least common outcome. Overall, while total applications drop on weekends, outcome patterns remain stable.

Distribution of Contract Types by Client Type (Figure 15:)



The bar chart shows contract types by client type. New clients prefer consumer loans for their simplicity. Repeat applicants favour cash loans for their flexibility. Refreshed clients, with existing loans, show balanced interest in consumer, revolving and cash loans. The Unknown Client Type has minimal representation across all loan types.

6.5 Converting Categorical Data to Ordinal Format

Categorical data in both the datasets is converted into ordinal format(numerical) to ensure compatibility with machine learning models.

```
current_app.head()
```

TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CRE
1	1	0	1	0	0	202500.0	4065
0	1	1	1	1	0	270000.0	12935
0	2	0	0	0	0	67500.0	1350
0	1	1	1	0	0	135000.0	3126
0	1	0	1	0	0	121500.0	5130

```
[13] previous_app.head()
```

NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_
1	9251.775	179055.0	179055.0	0.0	179055.0	
1	61997.139	337500.0	348637.5	0.0	337500.0	
1	5357.250	24282.0	20106.0	4860.0	24282.0	
3	10125.000	0.0	0.0	1845.0	94320.0	
2	13010.985	225000.0	284400.0	1845.0	225000.0	

6.6 Features selections

By selecting relevant features, computational power is reduced, overfitting is avoided and learning efficiency is enhanced. This project ensures models generalise well and provide reliable predictions for analysing loan application outcomes and applicant behaviour.

6.6.1 Feature Selection for Loan applicant behaviour

For modeling loan applicant behaviour, the target attribute is "TARGET," classifying applicants as defaulters (1) or non-defaulters (0). The selected features include financial attributes such as AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE; demographic information like CNT_CHILDREN, NAME_INCOME_TYPE, AMT_INCOME_TOTAL, OCCUPATION_TYPE, and CNT_FAM_MEMBERS; and regional ratings including REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY, REG_REGION_NOT_WORK_REGION, LIVE_REGION_NOT_WORK_REGION, and ORGANIZATION_TYPE. These features are highly correlated, as seen in the EDA chapter, and are crucial for analysing applicant behaviour.

6.6.2 Feature Selection Loan Approval Prediction

For modeling loan approval prediction, the target attribute is "NAME_CONTRACT_STATUS," which classifies loan applications as 'Cancelled,' 'Refused,' 'Approved,' or 'Unused Offer.' The selected features include contract details such as NAME_CONTRACT_TYPE, AMT_APPLICATION, AMT_CREDIT, and AMT_DOWN_PAYMENT; payment information like AMT_ANNUITY, CNT_PAYMENT, and AMT_GOODS_PRICE; and timing attributes including DAYS_LAST_DUE, DAYS_TERMINATION, DAYS_LAST_DUE_1ST_VERSION, and DAYS_FIRST_DRAWING. Additional features such as FLAG_LAST_APPL_PER_CONTRACT, NFLAG_LAST_APPL_IN_DAY, RATE_DOWN_PAYMENT, NAME_GOODS_CATEGORY, NAME_PORTFOLIO, NAME_PRODUCT_TYPE, and NAME_SELLER_INDUSTRY are also used. These features are highly correlated, as seen in the EDA chapter providing a comprehensive basis for predicting loan outcomes.

6.6.3 Feature Selection Association rule mining

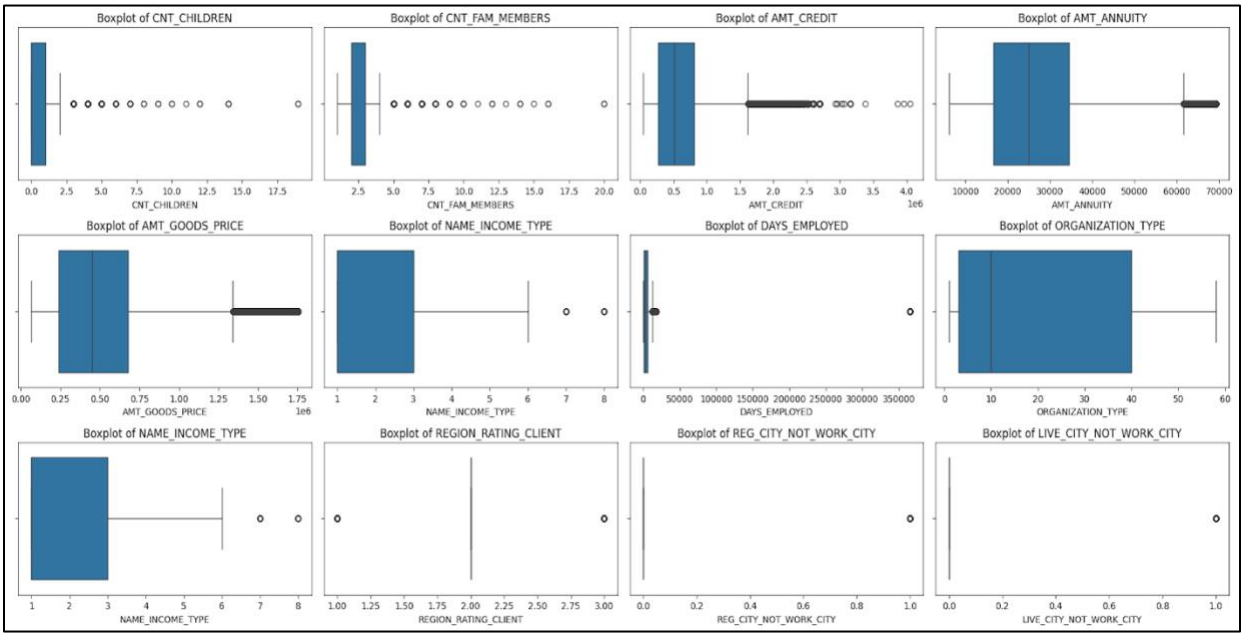
For association rule mining, the selected categorical columns include NAME_CONTRACT_TYPE, NAME_CASH_LOAN_PURPOSE, WEEKDAY_APPR_PROCESS_START, NAME_PAYMENT_TYPE, CODE_REJECT_REASON, NAME_TYPE_SUITE, NAME_CLIENT_TYPE, NAME_GOODS_CATEGORY, NAME_PORTFOLIO,

NAME_PRODUCT_TYPE, CHANNEL_TYPE, NAME_SELLER_INDUSTRY, NAME_YIELD_GROUP, PRODUCT_COMBINATION, and NAME_CONTRACT_STATUS. These features are chosen to uncover interesting relationships between different attributes in the dataset, providing insights into loan applications.

6.7 Outlier Detection and Handling

6.7.1 Outlier Detection and Handling for Selected Features in Loan Applicant Behaviour

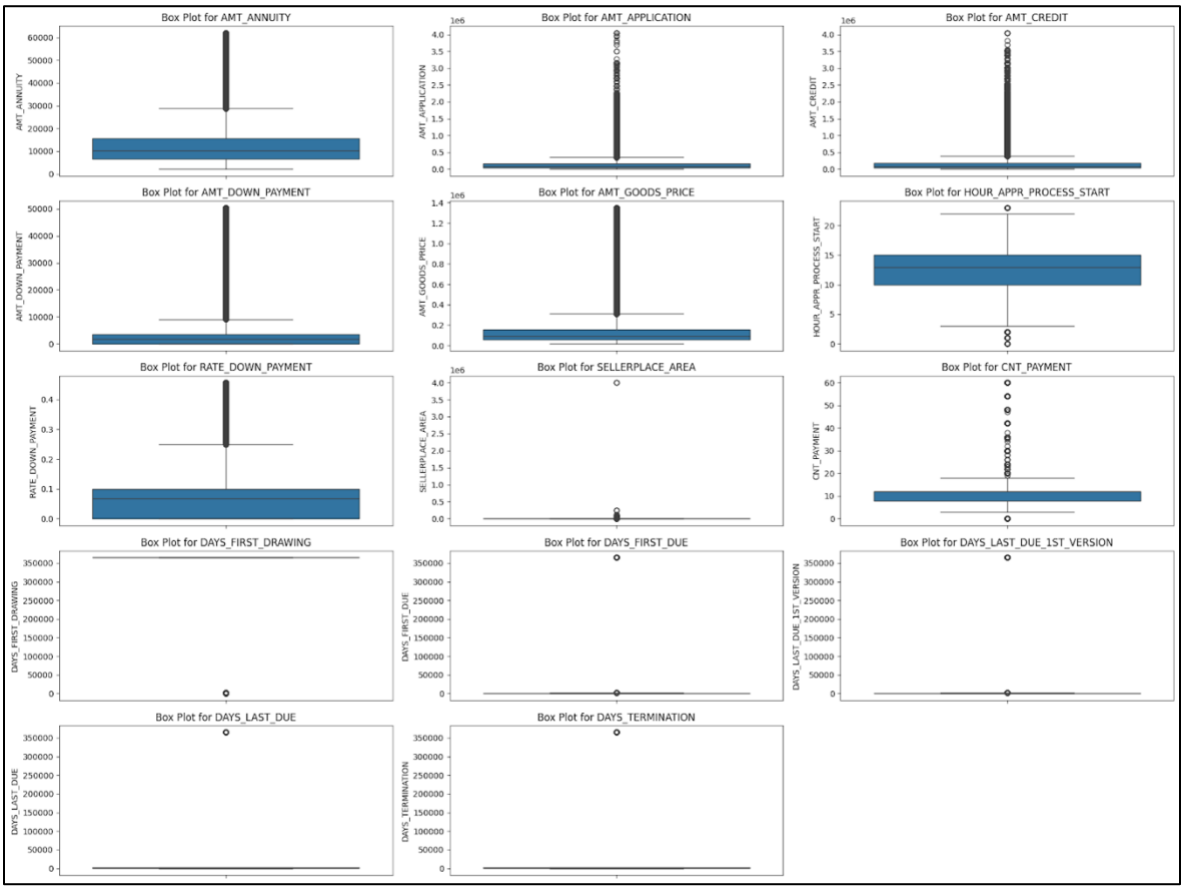
Illustrates the outliers for the selected features for modeling. **(Figure 16:)**



Extreme values in AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, DAYS_EMPLOYED and REGION_RATING_CLIENT can skew results and affect model robustness. To handle these outliers, capping using the Interquartile Range (IQR) method is applied, setting upper and lower bounds based on the IQR. Outliers for CNT_CHILDREN, CNT_FAM_MEMBERS, NAME_INCOME_TYPE, ORGANIZATION_TYPE, REG_CITY_NOT_WORK_CITY and LIVE_CITY_NOT_WORK_CITY are ignored.

6.7.2 Outlier Detection and Handling for Selected Features in Loan Approval Prediction

Illustrates the outliers for the selected features for modeling. **(Figure 17:)**



Extreme values in AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_DOWN_PAYMENT, AMT_GOODS_PRICE, HOUR_APPR_PROCESS_START, RATE_DOWN_PAYMENT, SELLERPLACE_AREA, and CNT_PAYMENT can skew financial assessments, so the Interquartile Range (IQR) method was applied to cap outliers. Additionally, outliers in DAYS_DECISION, DAYS_FIRST_DRAWING, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, and DAYS_TERMINATION exceeding 1000 days were imputed with the median to maintain realistic values and ensure data consistency.

This ensures data consistency and reliability while preserving the natural variance of demographic and categorical features.

6.8 Data Partitioning Strategy: Train, Test, and Validation Sets

For both models, the dataset is divided into three distinct sets:

- **60% for Training:** Used to fit the model.
- **20% for Testing:** Allocated to evaluate the model's performance.
- **20% for Validation:** Reserved for tuning model parameters and preventing overfitting.

6.9 Summary

This chapter focused on preparing the dataset's attributes, cleaning them, aggregating them, and summarising their values. These steps ensured the dataset is ready for building the model and creating the training and test datasets.

CHAPTER 7: CLASSIFICATION

7.1 Classification: Loan Applicant Behaviour(Analysing Defaulters vs. Non-Defaulters)

This classification model, developed from the data, aims to predict whether an applicant is a **defaulter (class 1) or a non-defaulter (class 0).** It will provide banks and loan firms with a behaviour analysis of loan applicants. Feature selection is discussed in the Data Preparation chapter.

Algorithms that have been selected to perform classification on **TARGET** feature.

1. Logistic Regression Classification
2. Decision Tree Classification
3. Gradient Boosting Classification

7.1.1 Logistic Regression Classification

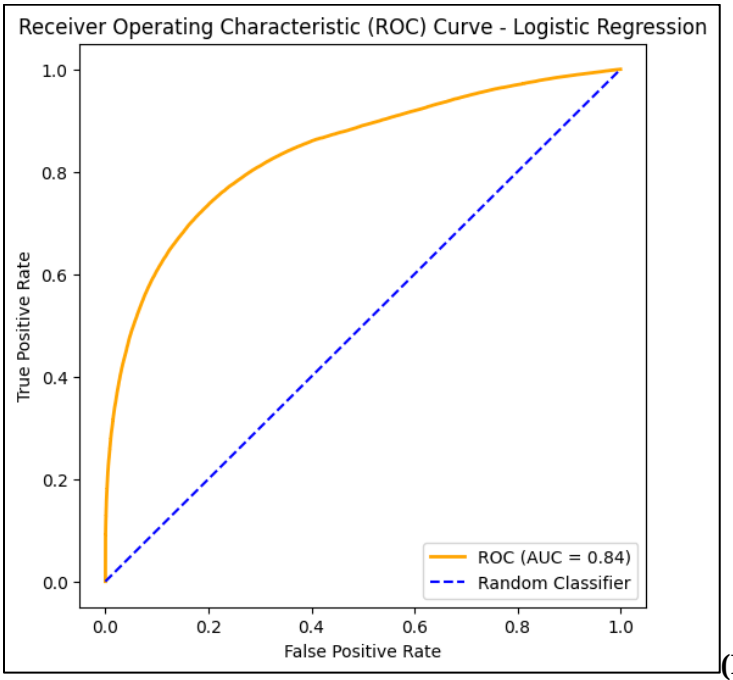
```
Logistic Regression – Training Set Evaluation
Confusion Matrix :
[[126329  33998]
 [ 40561 119766]]
Classification Report :
              precision    recall  f1-score   support

           0       0.76      0.79      0.77    160327
           1       0.78      0.75      0.76    160327

    accuracy                           0.77    320654
   macro avg       0.77      0.77      0.77    320654
weighted avg       0.77      0.77      0.77    320654

Accuracy : 0.7674783411402946
AUC : 0.8393667657804622
```

Confusion Matrix (Table 7)

Predicted / Actual	**Non-Defaulter**	**Defaulter**
Non-Defaulter	126329	33998
Defaulter	40561	119766

This model accurately identified a substantial number of non-defaulters 126,329, making misclassifications of 33,998 as defaulters. The model also correctly identified 119,766 defaulters, misclassified 40,561 as non-defaulters.

The model shows high performance in terms of overall accuracy, with a score of 76.75%. And the AUC score of 0.839 indicates a good ability to distinguish between defaulters and non-defaulters. The performance measures show that class 1 (defaulters) and class 0 (non-defaulters) have similar scores in terms of precision, recall, and F1-score, reflecting a more balanced classification capability.



(Figure 18)

The ROC curve of Logistic Regression shows how well the model separates defaulters from non-defaulters by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds. The closer the curve is to the top-left corner, the better the model. This model does a decent job of identifying defaulters while keeping false positives low.

7.1.2 Decision Tree Classification

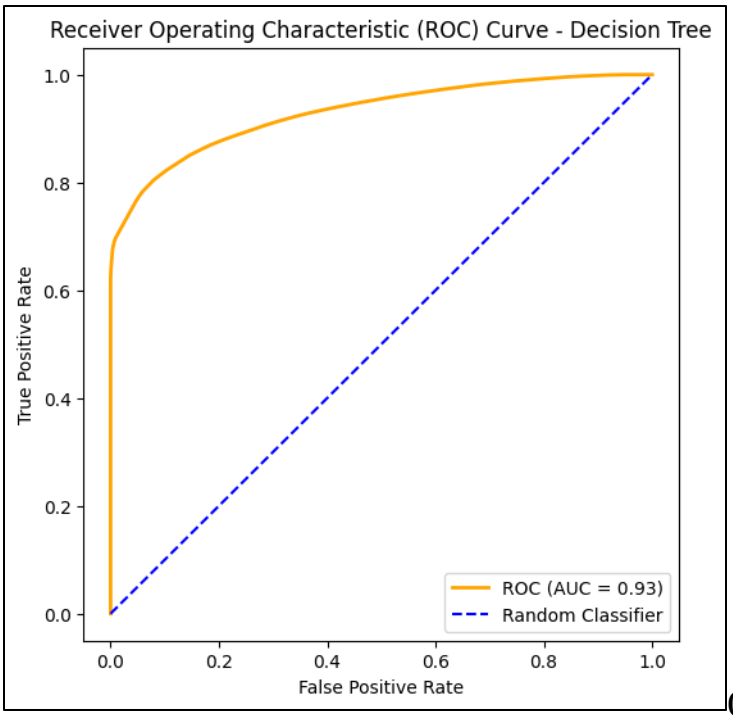
```
Decision Tree - Training Set Evaluation
Confusion Matrix :
[[148243  12084]
 [ 32021 128306]]
Classification Report :
              precision    recall  f1-score   support

           0       0.82      0.92      0.87    160327
           1       0.91      0.80      0.85    160327

    accuracy                           0.86    320654
   macro avg       0.87      0.86      0.86    320654
weighted avg       0.87      0.86      0.86    320654

Accuracy : 0.8624529867084146
AUC : 0.9300415457321152
```

Confusion Matrix (Table 8)

Predicted / Actual	**Non-Defaulter**	**Defaulter**
Non-Defaulter	148243	12084
Defaulter	32021	128306

This model accurately identified a substantial number of non-defaulters 148,243 but misclassified 15,430 instances as defaulters. For defaulters, the model correctly identified 128,306 instances but misclassified 32,021 instances as non-defaulters.

And the model performs well with an accuracy of 86.25%. The AUC score of 0.93 indicates a strong ability to distinguish between defaulters and non-defaulters. Performance measures show that both non-defaulters (class 0) and defaulters (class 1) are reasonably well-classified, with defaulters showing a slightly lower recall but higher precision, indicating balanced performance across both classes.



(Figure 19)

The ROC curve of the Decision tree model shows the True Positive Rate (TPR) against the False Positive Rate (FPR) across various threshold values.The curve signifies a strong model, which is quite capable of correctly identifying both defaulters and non-defaulters.

7.1.3 Gradient Boosting Classification

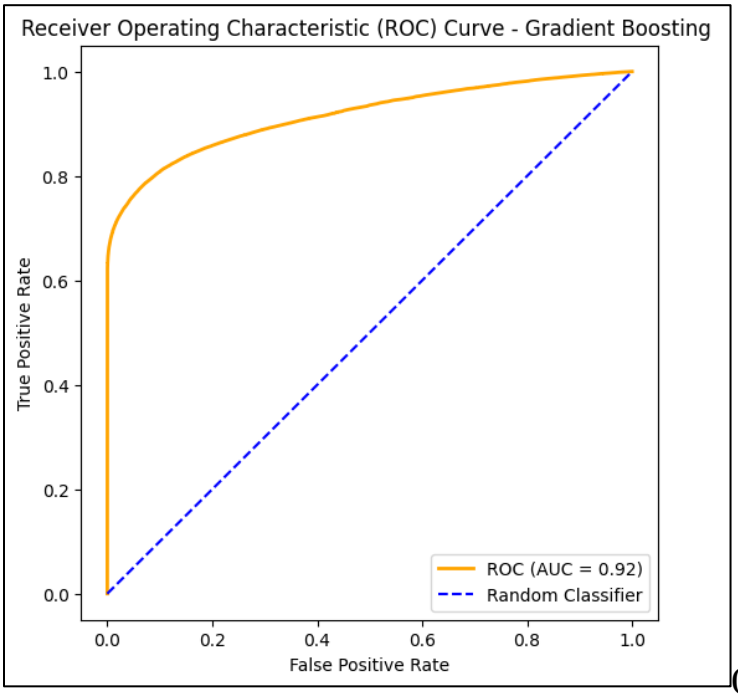
```
Gradient Boosting – Training Set Evaluation
Confusion Matrix :
[[153283   7044]
 [ 38825 121502]]
Classification Report :
              precision    recall  f1-score   support

           0       0.80      0.96      0.87    160327
           1       0.95      0.76      0.84    160327

    accuracy                           0.86    320654
   macro avg       0.87      0.86      0.86    320654
weighted avg       0.87      0.86      0.86    320654

Accuracy : 0.8569517299020127
AUC : 0.9168320421358388
```

Confusion Matrix (Table 9)

Predicted / Actual	**Non-Defaulter**	**Defaulter**
Non-Defaulter	153283	7044
Defaulter	38825	121502

This model accurately identified a substantial number of non-defaulters 153,283 but misclassified 7,044 instances as non-defaulters. For defaulters, the model correctly identified only 121,502 instances but misclassified 38,825 instances as non-defaulters.

The model shows moderate performance, with an overall accuracy of approximately 85.70%. However, the AUC score of 0.92 indicates a strong ability to distinguish between defaulters and non-defaulters. The performance measures reveal that class 1 (defaulters) has slightly lower scores in terms of precision, recall and F1-score compared to class 0 (non-defaulters), indicating a well-balanced classification with room for improvement in reducing misclassifications for defaulters.



(Figure 20)

The ROC curve for the Gradient Boosting model demonstrates a strong performance, closely hugging the top left corner. The curve confirms the model's excellent ability to distinguish between defaulters and non-defaulters, significantly outperforming random guessing.

7.1.4 Summary

Table 10:

Classification model	**Macro-avg: F-1 score**	**Macro-avg: Precision**	**Macro-avg: Recall**	**AUC Score**	**Accuracy Score (train-set)**	**Cross-validation (accuracy)**
Logistic Regression	0.77	0.77	0.77	0.838	0.767	0.593
Decision Tree	0.86	0.87	0.86	0.930	0.862	0.651
Gradient Boosting	**0.86**	**0.87**	**0.86**	**0.916**	**0.857**	**0.917**

The **Gradient Boosting model outperforms the other classifiers with an accuracy of 85.7% and an AUC score of 0.92, along with strong macro-averaged F1 score, precision, and recall values.** Its cross-validation accuracy of 91.7% further confirms its robustness and generalisation ability. In contrast, Decision Tree, while having a slightly higher AUC (0.93), struggles with overfitting, as indicated by its lower cross-validation accuracy (65.1%), and Logistic Regression underperforms with a cross-validation accuracy of 59.3%. Therefore, Gradient Boosting is the best model, showing consistent and strong performance across all metrics, without the need for extensive tuning.

7.2 Classification: Loan Approval Prediction

This classification model aims to predict the results of loan applications based on various applicants and loan characteristics. It will provide banks and loan firms with the outcomes of loan applications. Feature selection is discussed in the Data Preparation chapter.

Algorithms that have been selected to perform classification on **NAME_CONTRACT_STATUS** feature:

1. Logistic Regression Classification
2. Decision Tree Classification
3. Random Forest Classification

7.2.1 Logistic Regression Classification

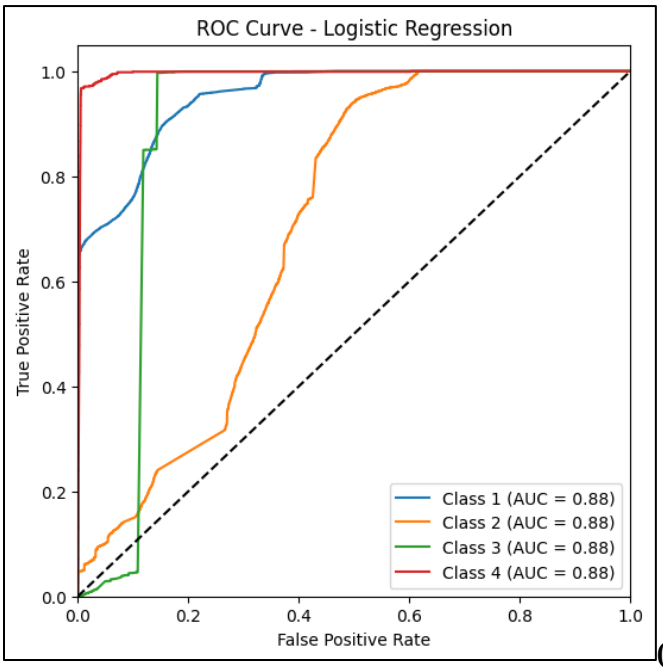
```
Logistic Regression
Confusion Matrix (Test Set):
[[30593  8260  3815    72]
 [  804  2286  3312   320]
 [    2  1159  6661     4]
 [   29     1     0   894]]
Mean Absolute Error: 0.37995602281316565
Mean Squared Error: 0.5325362468219611
Root Mean Squared Error: 0.7297508114568706
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       0.97      0.72      0.82     42740
           2       0.20      0.34      0.25      6722
           3       0.48      0.85      0.62      7826
           4       0.69      0.97      0.81       924

    accuracy                           0.69     58212
   macro avg       0.59      0.72      0.62     58212
weighted avg       0.81      0.69      0.73     58212

Accuracy (Test Set): 0.6945990517419088
AUC (Test Set): 0.8843750988623812
```

Confusion Matrix (Table 11)

Predicted / Actual	**Approved**	**Refused**	**Cancelled**	**Unused offer**
Approved	30593	8260	3815	72
Refused	804	2286	3312	320
Cancelled	2	1159	6661	4
Unused offer	29	0	0	894

The model accurately identified "Approved" cases of 30,593 but it misclassified 12,147 cases into the remaining all other categories. For "Refused" cases, the model correctly identified 2,286 instances but misclassified 804 cases as "Approved" and 3,312 as 'Cancelled' and 325 as "Unused offer.". The model identified "Cancelled" cases correctly 6,661 while it misclassified 1,165 as other categories. And the model correctly identified 894 "Unused offer" cases while only a small number of other categories were misclassified.

This model shows good performance, with an overall accuracy of 69.4%. The AUC score of 0.88 indicates a good ability to distinguish between the various contract statuses. The performance measures reveal that the "Approved" classes have significantly higher scores in terms of precision, recall. In contrast, the "Refused" class has achieved very lower precision and recall, indicating challenges in accurately predicting it, while "Unused offer" cases have high recall but suffer from lower F1 score.



(Figure 21)

The model's ROC curves show a good ability to differentiate between the four contract statuses. The shape of the ROC curve for "Refused" indicates more challenges in distinguishing this class compared to others.

7.2.2 Decision Tree Classification

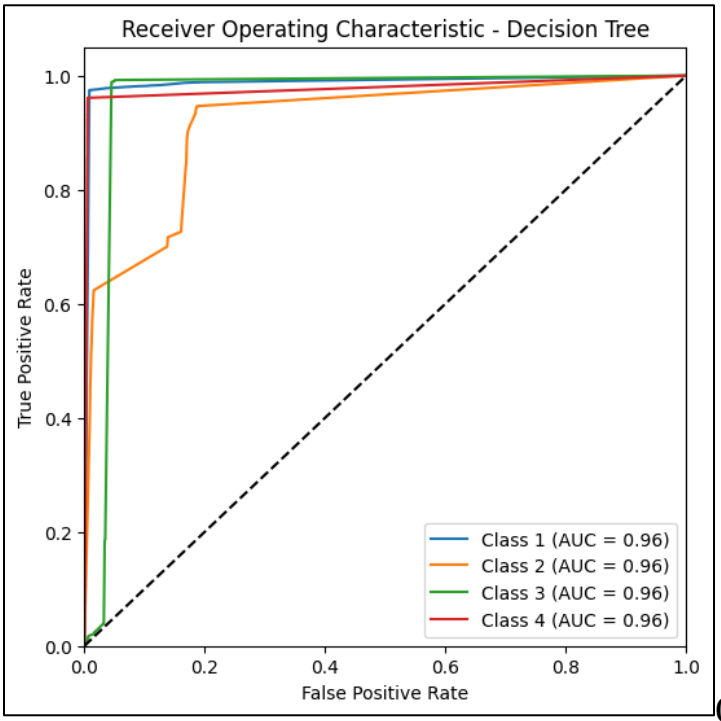
```
Decision Tree
Confusion Matrix (Test Set):
[[41649   688   384    19]
 [  119  4643  1631   329]
 [    2  6337  1486     1]
 [    4    22     2   896]]
Mean Absolute Error: 0.17730021301449872
Mean Squared Error: 0.20499209784924072
Root Mean Squared Error: 0.45276053035709807
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       1.00      0.97      0.99     42740
           2       0.40      0.69      0.50      6722
           3       0.42      0.19      0.26      7826
           4       0.72      0.97      0.83       924

    accuracy                           0.84     58212
   macro avg       0.63      0.71      0.64     58212
weighted avg       0.85      0.84      0.83     58212

Accuracy (Test Set): 0.8361506218649076
AUC (Test Set): 0.964563315068163
```

Confusion Matrix (Table 12)

Predicted / Actual	**Approved**	**Refused**	**Cancelled**	**Unused offer**
Approved	41702	700	324	14
Refused	280	4729	1404	309
Cancelled	10	6363	1452	1
Unused offer	3	39	2	880

This model accurately identified 41,702 "Approved" cases but misclassified 1,038 instances in remaining all other categories. For "Refused" cases, the model correctly identified 4,729 instances but it misclassified 1993 cases as remaining all other categories. The model correctly identified "Cancelled" cases as 1452 cases but misclassified 6374 cases as the categories. And for the "Unused offer" cases, the model correctly identified 880 instances and misclassified only a small number of remaining other categories.

This model shows strong performance with overall accuracy of 83.6%. The AUC score of 0.96 indicates a high ability to distinguish between the contract statuses. The performance measures reveals that the "Refused" class has moderately lower precision and recall, while the "Approved" and "Unused offer" classes also show strong performance. However, the "Cancelled" class has a lower precision, recall and f1 score values , indicating the model faces some difficulty in accurately predicting this category.



(Figure 22)

The model ROC curve demonstrates robust performance for all contract statuses across all classes. Although the "Refused" category shows a slightly lower curve, the overall model is highly effective in differentiating between statuses, particularly for "Approved" and "Unused offer" contracts.

7.2.3 Random Forest Classification

```
Random Forest
Confusion Matrix (Test Set):
[[41625   678   416    21]
 [   16  4685  1690   331]
 [    1  6329  1495     1]
 [    0    19     0   905]]
Mean Absolute Error: 0.17712842712842713
Mean Squared Error: 0.2056448842163128
Root Mean Squared Error: 0.4534808531970372
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       1.00      0.97      0.99     42740
           2       0.40      0.70      0.51      6722
           3       0.42      0.19      0.26      7826
           4       0.72      0.98      0.83       924

    accuracy                           0.84     58212
   macro avg       0.63      0.71      0.65     58212
weighted avg       0.85      0.84      0.83     58212

Accuracy (Test Set): 0.8367690510547653
AUC (Test Set): 0.971203878094806
```

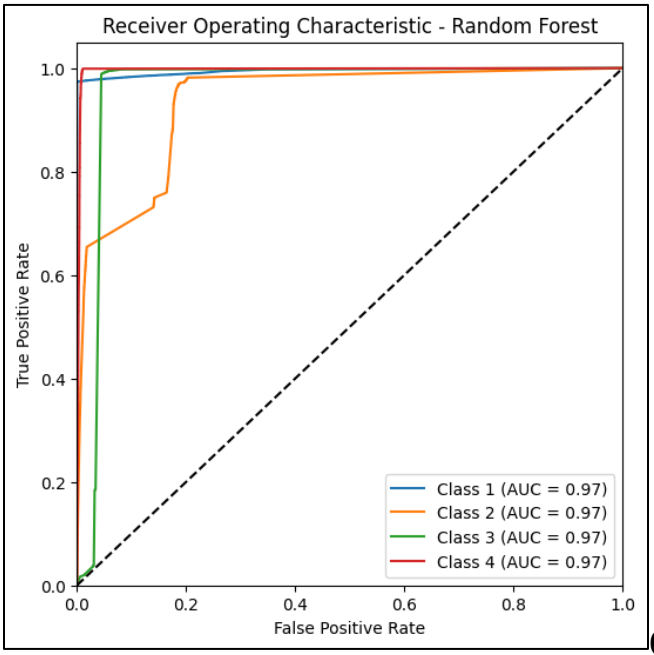
Confusion Matrix (Table 13)

Predicted / Actual	**Approved**	**Refused**	**Cancelled**	**Unused offer**
Approved	41625	678	416	21
Refused	16	4685	1690	331
Cancelled	1	6329	1495	1
Unused offer	0	19	0	905

The model accurately identified as "Approved" cases of 41,625 and misclassified 1,115 instances as remaining all other categories. For "Refused" cases, this model correctly identified 4685 instances but misclassified 2037 into other categories. This model correctly identified "Cancelled" cases of 1,446 and misclassified 6,331 cases as remaining other categories. The "Unused offer", the model correctly predicted 905 instances and misclassified 19 cases as "Refused."

This model's accuracy is 83.6%. AUC 0.97 indicates the capacity to differentiate between contract statuses. The model performs well across all classes, particularly for "Approved" and "Unused offer", which have high precision and recall. The "Refused" class demonstrates good recall, though there are

some misclassifications as "Approved" and "Cancelled." The "Cancelled" class shows strong results but has a lot of misclassification issues.



(Figure 23)

The model's ROC curve excels in classifying all contract statuses. The 'Refused' class curve shows slight fluctuations, but overall, the ROC curves for all classes indicate that this is a strong and reliable model for classification.

7.2.4 Summary (Table 14)

Classification model	**Macro-avg: F-1 score**	**Macro-avg: Precision**	**Macro-avg: Recall**	**AUC Score**	**Accuracy Score(train-set)**	**Cross-validation (accuracy)**
Logistic Regression	0.62	0.59	0.72	0.884	0.695	0.77
Decision Tree	0.64	0.63	0.71	0.965	0.836	0.94
Random Forest	**0.65**	**0.63**	**0.71**	**0.971**	**0.847**	**0.95**

Overall,the Random Forest model outperforms the other classifiers, achieving an accuracy of 84% and an impressive AUC score of 0.972, excellently differentiating between different contract statuses. Additionally, its cross-validation accuracy of 95% confirms its robustness and generalisation capability across different datasets. While the Decision Tree model also shows competitive performance with a high AUC of 0.960 and a cross-validation accuracy of 94%, it slightly lags behind in terms of macro-averaged F1 score and precision. Logistic Regression has the lowest scores across most metrics, particularly in macro-averaged precision (0.57) and cross-validation accuracy (77%). Therefore, Random Forest is the best model, offering a strong balance of high accuracy, precision, recall, and reliable performance across cross-validation.

CHAPTER 8: EVALUATION

8.1 Introduction

In this chapter, we evaluate the performance of various classification models in predicting loan approval based on applicant behaviour

8.2 Applicant Behaviour

8.2.1 Test Evaluation of Logistic Regression Classification Model

```
Logistic Regression - Test Set Evaluation:
Confusion Matrix (Test Set):
[[42098 11368]
 [ 3627  1119]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.92      0.79      0.85     53466
           1       0.09      0.24      0.13      4746

    accuracy                           0.74     58212
   macro avg       0.51      0.51      0.49     58212
weighted avg       0.85      0.74      0.79     58212

Accuracy (Test Set): 0.7424070638356353
AUC (Test Set): 0.52
```

The Logistic Regression model performed modestly on the test set, achieving an accuracy of 74.2%. The confusion matrix shows that the model correctly identified 42,098 non-defaulters but misclassified 11,368 as defaulters. For defaulters, the model correctly identified 1,119 cases but missed 3,627, resulting in a low recall of 24% for class 1. The overall macro-averaged precision, recall, and F1 score were around 0.51, indicating that the model struggles to balance predictions between the two classes, particularly with defaulters (class 1). Additionally, the AUC score of 0.52 suggests poor discrimination ability between defaulters and non-defaulters, as the model's performance is close to random guessing for this task. These results highlight the need for further tuning or alternative approaches to improve the model's ability to detect defaulters effectively.

8.2.2 Test Evaluation of Decision Tree Classification Model

```
Decision Tree - Test Set Evaluation:
Confusion Matrix (Test Set):
[[48926  4540]
 [ 4325   421]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.92      0.92      0.92     53466
           1       0.08      0.09      0.09      4746

    accuracy                           0.85     58212
   macro avg       0.50      0.50      0.50     58212
weighted avg       0.85      0.85      0.85     58212

Accuracy (Test Set): 0.8477118119975263
AUC (Test Set): 0.55
```

The Decision Tree model achieved an accuracy of 84.8% on the test set. It correctly identified 48,926 non-defaulters but misclassified 4,540 as defaulters. For defaulters, the model only correctly identified 421 cases and missed 4,325, resulting in a very low recall of 9% for class 1. This shows that while the model performs well at detecting non-defaulters, it struggles significantly with identifying defaulters. The precision, recall, and F1 score are all around 0.50, and the AUC score of 0.55 indicates a slightly better than random ability to differentiate between the two classes. Overall, the Decision Tree's performance is heavily skewed towards class 0 (non-defaulters), and it needs improvement in handling the minority class (defaulters).

8.2.3 Test Evaluation of Gradient Boosting Classification Model

```
Gradient Boosting - Test Set Evaluation:
Confusion Matrix (Test Set):
[[51093  2373]
 [ 4512   234]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.92      0.96      0.94     53466
           1       0.09      0.05      0.06      4746

    accuracy                           0.88     58212
   macro avg       0.50      0.50      0.50     58212
weighted avg       0.85      0.88      0.87     58212

Accuracy (Test Set): 0.8817254174397031
AUC (Test Set): 0.54
```

The Gradient Boosting model performed better, with an accuracy of 88.2%, correctly identifying 51,093 non-defaulters while misclassifying 2,373 as defaulters. However, like the Decision Tree, it struggled with defaulters, correctly identifying only 234 out of 4,746 and missing 4,512, resulting in a low recall of 5% for class 1. While it shows strong performance for class 0, the macro-averaged precision, recall, and F1 score remain around 0.50, and the AUC score of 0.54 indicates only slight improvement over random guessing. Despite higher accuracy, Gradient Boosting, like the Decision Tree, requires significant improvement in handling the minority class (defaulters).

8.2.4 Summary

Table 15:

Model	**Accuracy (train Set)**	**Cross-Validated Accuracy**	**Accuracy (Test Set)**
Logistic Regression	0.76	0.59	0.74
Decision Tree	0.86	0.65	0.84
Gradient Boosting	**0.85**	**0.91**	**0.88**

Among the three models, Gradient Boosting performed the best, with 88% test accuracy, 91% cross-validated accuracy, and strong generalisation. Despite this, all models had low AUC scores, indicating difficulty in distinguishing between defaulters and non-defaulters. Further tuning and better handling of class imbalance are needed, but Gradient Boosting remains the strongest model.

8.3 Loan Approval Prediction

8.3.1 Evaluation of Logistic Regression Classification

```
Logistic Regression - Test Set Evaluation:
Confusion Matrix (Test Set):
[[30679  8189  3784    88]
 [  841  2226  3333   322]
 [    2  1160  6659     5]
 [   29     1     0   894]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       0.97      0.72      0.83     42740
           2       0.19      0.33      0.24      6722
           3       0.48      0.85      0.62      7826
           4       0.68      0.97      0.80       924

    accuracy                           0.70     58212
   macro avg       0.58      0.72      0.62     58212
weighted avg       0.81      0.70      0.73     58212

Accuracy (Test Set): 0.6950113378684807
AUC (Test Set): 0.89
```

The Logistic Regression model achieved a test set accuracy of 69.5%. The confusion matrix reveals that it correctly identified 30,679 'Approved' cases but misclassified 8,189 as 'Refused' and 3,784 as 'Cancelled 'and 88 instances as 'unused offers'. For 'Refused' cases, only 2,226 were correctly identified, misclassified 4496 into other categories. The model performed well for 'Cancelled' and 'Unused offer' categories, with recall rates of 0.85 and 0.97, respectively. Overall, the macro-averaged precision and recall were 0.58 and 0.72, indicating the model struggled particularly with 'Refused' cases. The AUC score of 0.89 suggests a fair ability to distinguish between classes, but requires improvement.

8.3.2 Evaluation of Decision Tree Classification

```
Decision Tree - Test Set Evaluation:
Confusion Matrix (Test Set):
[[41739   674   315    12]
 [  383  4739  1326   274]
 [   22  6365  1438     1]
 [    4   185     2   733]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       0.99      0.98      0.98     42740
           2       0.40      0.70      0.51      6722
           3       0.47      0.18      0.26      7826
           4       0.72      0.79      0.75       924

    accuracy                           0.84     58212
   macro avg       0.64      0.66      0.63     58212
weighted avg       0.85      0.84      0.83     58212

Accuracy (Test Set): 0.8357211571497286
AUC (Test Set): 0.93
```

The Decision Tree model achieved a test set accuracy of 83.6%. The confusion matrix shows it correctly identified 41,739 'Approved' cases but misclassified 1001 into other remaining categories. For 'Refused' cases, the model correctly identified 4,739 but misclassified 1983 into other remaining categories. The 'Cancelled' class had 1,438 correct predictions, with 6,388 misclassified in other remaining categories. The model demonstrated strong performance for 'Approved' cases, with a precision of 0.99 and recall of 0.98. However, it moderately struggled with 'Refused' cases, achieving a recall of 0.70 but with a lower precision of 0.40. The model strongly struggled with 'cancelled' class with low recall and f1 score. The AUC score of 0.93 indicates a strong ability to discriminate between classes, reflecting the model's good overall performance with some challenges in specific categories.

8.3.3 Evaluation of Random Forest Classification

```
Random Forest - Test Set Evaluation:
Confusion Matrix (Test Set):
[[41676   737   315    12]
 [  117  5014  1313   278]
 [   18  6372  1435     1]
 [    0   186     0   738]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       1.00      0.98      0.99     42740
           2       0.41      0.75      0.53      6722
           3       0.47      0.18      0.26      7826
           4       0.72      0.80      0.76       924

    accuracy                           0.84     58212
   macro avg       0.65      0.68      0.63     58212
weighted avg       0.85      0.84      0.83     58212

Accuracy (Test Set): 0.8393973751116608
AUC (Test Set): 0.97
```

The Random Forest model achieved a test set accuracy of 84.0%. The confusion matrix indicates that it correctly identified 41,676 'Approved' cases but misclassified 1064 into other remaining categories. For 'Refused' cases, the model correctly identified 5,014 but misclassified 1708 into other remaining categories. The 'Cancelled' class had 1,435 correct predictions with 6,391 misclassified to remaining categories. The 'Unused offer' class had 738 correct predictions and minimal misclassifications. The model demonstrated excellent performance for 'Approved' cases with a precision of 1.00 and recall of 0.98. The 'Refused' class showed improved performance with a recall of 0.75 but had lower precision at

0.41. The model mainly struggled with 'Cancelled' class with low recall and f1 score. The AUC score of 0.97 reflects the model's strong discriminative ability across all classes, highlighting its robust performance overall.

8.3.4 Summary

Table 16:

Model	Accuracy (train Set)	Cross-Validated Accuracy	Accuracy (Test Set)
Logistic Regression	0.695	0.77	0.69
Decision Tree	0.836	0.94	0.83
Random Forest	**0.837**	**0.95**	**0.83**

Random Forest performed the most effective model for loan approval prediction. It achieved the highest accuracy on the training set (83.7%) and cross-validated accuracy (95%), reflecting its strong performance and robustness in generalising to unseen data. Although its test set accuracy (83%) was on par with the Decision Tree, its superior cross-validation accuracy suggests better overall performance and reduced risk of overfitting. The Logistic Regression model, while useful, lagged behind with lower accuracy metrics across the board. Therefore, Random Forest is recommended for its consistent and reliable performance across various evaluation metrics.

CHAPTER 9: ASSOCIATION RULE

9.1 Introduction

This chapter investigates the utilisation of association rule mining to forecast loan approval results. Loan firms can improve their decision making processes and risk assessments by identifying patterns in loan applications.

9.2 Association Rules

Association Rule Mining - Row 35354 ((Figure 24)

Row - 35354	Row - 35354 - Items : frozenset({'Consumer loans', 'high', 'POS'}) => frozenset({'Approved'})
	Row - 35354 - Antecedent :Consumer loans, high, POS
	Row - 35354 - Consequent :Approved
	Row - 35354 - Support :0.168355953
	Row - 35354 - Confidence :0.944123314
	Row - 35354 - Lift :1.288567834

People who applied for 'consumer loans', are associated with 'high yield' (indicating a high-interest rate or high-risk loan), and used the 'POS (Point of Sale)' system for financing are highly associated with having their loan application status marked as "Approved." The support for this rule is 0.1684, indicating that 16.84% of the transactions in the dataset follow this pattern. The confidence is 0.9441, meaning there is a 94.41% chance that clients fitting this profile will have their loans approved. The lift is 1.2886, showing that the occurrence of the antecedent increases the likelihood of loan approval by approximately 28.86% compared to random chance.

Association Rule Mining - Row 34451 (Figure 25)

Row - 34451	Row - 34451 - Items : frozenset({'POS household with interest', 'Consumer electronics'}) => frozenset({'Cash through the bank', 'Unknown'})
	Row - 34451 - Antecedent :POS household with interest, Consumer electronics
	Row - 34451 - Consequent :Cash through the bank, Unknown
	Row - 34451 - Support :0.176601958
	Row - 34451 - Confidence :0.874893617
	Row - 34451 - Lift :1.307846878

People who made purchases of 'Consumer Electronics' and financed them through a POS household (Point of Sale: Household) with interest plan are highly associated with using 'Cash' through the bank as the payment type, while the name of industry is 'unknown'. The support for this rule is 0.1766, indicating

that 17.66% of the transactions in the dataset follow this pattern. The confidence is 0.8749, meaning there is an 87.49% chance that clients fitting this profile will use cash through the bank as their payment method, with the name of the industry as unknown. The lift is 1.3078, showing that the occurrence of the antecedent increases the likelihood of this payment method and status by about 30.78% compared to random chance.

Loan Approval Prediction - Association Rule Mining - Row 29105 (Figure 26)

Row - 29105	Row - 29105 - Items :frozenset({'Not Specified', 'Cash loans', 'Credit and cash offices'}) => frozenset({'Repeater'})
	Row - 29105 - Antecedent :Not Specified, Cash loans, Credit and cash offices
	Row - 29105 - Consequent :Repeater
	Row - 29105 - Support :0.206150146
	Row - 29105 - Confidence :0.842696629
	Row - 29105 - Lift :1.496442062

People who applied for 'Cash Loans', had their loan purpose marked as "Not Specified," and obtained the loan through 'credit and cash offices' are highly associated with being 'Repeater' clients (returning customers). The support for this rule is 0.2062, indicating that 20.62% of all transactions in the dataset follow this pattern. The confidence is 0.8427, meaning there is an 84.27% chance that clients fitting this profile will be repeater clients. The lift is 1.4964, showing that the occurrence of the antecedent increases the likelihood of being a repeater by approximately 49.64% compared to random chance.

Loan Approval Prediction - Association Rule Mining - Row 5565 (Figure 27)

Row - 5565	Row - 5565 - Items : frozenset({'Repeater', 'Credit and cash offices', 'Accepted After Initial Rejection'}) => frozenset({'Cash loans', 'Cash'})
	Row - 5565 - Antecedent :Repeater, Credit and cash offices, Accepted After Initial Rejection
	Row - 5565 - Consequent :Cash loans, Cash
	Row - 5565 - Support :0.167840577
	Row - 5565 - Confidence :0.803453947
	Row - 5565 - Lift :2.478487243

People who are 'Repeater' clients (returning customers), obtained their loan through 'credit and cash offices', and were 'accepted after initial rejection' are highly associated with taking out 'cash loans' and receiving 'cash' as the disbursement method. The support for this rule is 0.1678, indicating that 16.78% of all transactions in the dataset follow this pattern. The confidence is 0.8035, meaning there is an 80.35% chance that clients fitting this profile will take out cash loans and receive cash. The lift is 2.4785, showing

that the occurrence of the antecedent increases the likelihood of taking cash loans and receiving cash by approximately 147.85% compared to random chance.

9.3 Summary

This chapter identifies the association rules, highlighting the essential patterns loan firms can utilise to enhance their understanding of client profiles and optimise their loan approval processes. These rules allow firms to enhance customer targeting, refine risk management strategies and predict the likelihood of loan approvals by considering factors such as loan types, payment methods and client history.

CHAPTER 10: CONCLUSION

10.1 Introduction

This chapter concludes the study of the project "Loan Approval Prediction." It summarises the challenges faced, the learnings, the limitations and the potential future developments of the project.

10.2 Summary of the Study

The project began with selecting and understanding the dataset, followed by research and literature review on the significance of loan firms in today's economy. Loans have an important part to play in stabilising the financial status of individuals as well businesses, boosting economy and catering to various needs. The loan approval process is crucial, as it can determine individuals' financial health and even affect national economic power.

The challenges included understanding the importance of each dataset feature and it was also time-consuming process of data preprocessing, which took up 70% of the project timeline. This reinforced the importance of domain knowledge in data analytics projects. Modeling efforts highlighted the need for balanced data and tuning for better prediction accuracy.

The project successfully achieved over 80% accuracy in predicting loan approvals and analyzing applicant behavior. However, there is still room for improvement through further tuning and parameter optimization, as time constraints limits the extent of these adjustments. Additionally, implementing association rules for mining could identify specific patterns to assist loan authorities in making more informed decisions. This study lays the groundwork for future research, emphasizing the need for more balanced data and refined modeling techniques to enhance predictive performance.

10.4 Future development

Future research should focus on improving the reliability and adaptability of loan approval models by:

- Developing models that can evolve in changing market conditions.
- Enhancing techniques for handling imbalanced datasets.
- Incorporating real time streaming, customer feedback and global market insights to enhance model outputs.
- Ensure reliable background checks and continuously update applicants' data to assess risk management.
- Applying advanced tuning methods to optimise machine learning models for better predictive performance.
- Providing transparent decision-making processes and addressing the ethical implications of advanced AI technologies.

10.5 Summary

This chapter concludes the study by summarising the accomplishments in meeting the research objectives and addressing the research questions. It also highlights the project’s drawbacks and provides insights into potential future work. As a master’s student of Data Analytics, I have understood that it doesn’t matter where the data comes from. For tasks such as classification, regression, clustering, and association rule mining for prediction, it is crucial to have a good understanding of the data and the domain knowledge/business process. This understanding is essential to decide what can be done with the data.

REFERENCES:

AI Boosting Payments Efficiency & Cutting Fraud | J.P. Morgan. (n.d.). Www.jpmorgan.com. https://www.jpmorgan.com/insights/payments/payments-optimization/ai-payments-efficiency-fraud-reduction#:~:text=J.P.%20Morgan%20has%20been%20using

Martin, S. (2021, April 13). XAI Explained at GTC: Wells Fargo Examines Explainable AI for Modeling Lending Risk. NVIDIA Blog. https://blogs.nvidia.com/blog/wells-fargo-examines-explainable-ai-for-modeling-lending-risk/

Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems with Applications, 73(0957-4174), 1–10. https://doi.org/10.1016/j.eswa.2016.12.020

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4), 589–609.

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. Management Science, 49(3), 312–329. https://doi.org/10.1287/mnsc.49.3.312.12739

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. Journal of Accounting Research, 4(3), 71–111. https://doi.org/10.2307/2490171

Berger, A. N., Miller, N. H., Petersen, M. A., Rajan, R. G., & Stein, J. C. (2005). Does function follow organizational form? Evidence from the lending practices of large and small banks. Journal of Financial Economics, 76(2), 237–269. https://doi.org/10.1016/j.jfineco.2004.06.003

Blum, M. (1974). Failing Company Discriminant Analysis. Journal of Accounting Research, 12(1), 1. https://doi.org/10.2307/2490525

Byanjankar, A., Heikkilä, M., & Mezei, J. (2015, December 1). Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach. IEEE Xplore. https://doi.org/10.1109/SSCI.2015.109

Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C. (2018). Ensemble Learning or Deep Learning? Application to Default Risk Analysis. Journal of Risk and Financial Management, 11(1), 12. https://doi.org/10.3390/jrfm11010012

Harris, T. (2013). Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions. Expert Systems with Applications, 40(11), 4404–4413. https://doi.org/10.1016/j.eswa.2013.01.044

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 33(4), 847–856. https://doi.org/10.1016/j.eswa.2006.07.007

Kemalbay, G., & Korkmazoğlu, Ö. B. (2014). Categorical Principal Component Logistic Regression: A Case Study for Housing Loan Approval. Procedia - Social and Behavioral Sciences, 109(1877-0428), 730–736. https://doi.org/10.1016/j.sbspro.2013.12.537

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. Expert Systems with Applications, 42(10), 4621–4631. https://doi.org/10.1016/j.eswa.2015.02.001

Nutt, P. (1989). Uncertainty and culture in bank loan decisions. Omega, 17(3), 297–308. https://doi.org/10.1016/0305-0483(89)90035-2

Orji, Ugochukwu. E., Ugwuishiwu, Chikodili. H., Nguemaleu, Joseph. C. N., & Ugwuanyi, Peace. N. (2022). Machine Learning Models for Predicting Bank Loan Eligibility. 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), 978-1-6654-7979-0(978-1-6654-7979-0). https://doi.org/10.1109/nigercon54645.2022.9803172

Singh, V., Yadav, A., Awasthi, R., & Partheeban, G. N. (2021, June 1). Prediction of Modernized Loan Approval System Based on Machine Learning Approach. IEEE Xplore. https://doi.org/10.1109/CONIT51480.2021.9498475

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, 16(2), 149–172. https://doi.org/10.1016/s0169-2070(00)00034-0

Tumuluru, P., Burra, L. R., Loukya, M., Bhavana, S., CSaiBaba, H. M. H., & Sunanda, N. (2022). Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms. 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 978-1-6654-0052-7(978-1-6654-0051-0). https://doi.org/10.1109/icais53314.2022.9742800

Uddin, N., Uddin Ahamed, Md. K., Uddin, M. A., Islam, Md. M., Talukder, Md. A., & Aryal, S. (2023). An ensemble machine learning based bank loan approval predictions system with a smart application. International Journal of Cognitive Computing in Engineering, 4(ISSN 2666-3074), 327–339. https://doi.org/10.1016/j.ijcce.2023.09.001

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162(1877-0509), 503–513. https://doi.org/10.1016/j.procs.2019.12.017

Zimmer, I. (1980). A Lens Study of the Prediction of Corporate Failure by Bank Loan Officers. Journal of Accounting Research, 18(2), 629. https://doi.org/10.2307/2490599

Alsaleem, M., & Hasoon, S. (2020). Predicting Bank Loan Risks Using Machine Learning Algorithms. AL-Rafidain Journal of Computer Sciences and Mathematics, 14(1), 159–168. https://doi.org/10.33899/csmj.2020.164686

Naveen Kumar, Ch., Keerthana, D., Kavitha, M., & Kalyani, M. (2022). Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector. *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, *978-1-6654-9634-6*(978-1-6654-9633-9). https://doi.org/10.1109/icces54183.2022.9835725

Blessie, E. C., & Rekha , R. (2019). Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process. *International Journal of Innovative Technology and Exploring Engineering*, *9*(1), 2714–2719. https://doi.org/10.35940/ijitee.a4881.119119

Piatetsky, G. (2018, February 19). 5 Things You Need To Know About Data Science. KDnuggets. https://www.kdnuggets.com/2018/02/5-things-about-data-science.html

Kagan, J. (2023, July 31). *Loan – Definition*. Investopedia. https://www.investopedia.com/terms/l/loan.asp

kapoorshivam. (2021, February 20). *Credit Analysis using EDA*. Kaggle.com; Kaggle. https://www.kaggle.com/code/kapoorshivam/credit-analysis-using-eda/input

What is AML? Anti-Money Laundering Explained. (n.d.). Www.lexisnexis.com. https://www.lexisnexis.com/en-gb/glossary/aml

Sewell, T. (2024, May 3). *What is KYC?* Experian UK; Experian PLC. https://www.experian.co.uk/blogs/latest-thinking/guide/what-is-kyc/

Fair Lending | FDIC. (2024). Fdic.gov. https://www.fdic.gov/banker-resource-center/fair-lending

Kucera, D. (2021, April 13). *CCPA vs. GDPR: Similarities and Differences Explained*. Www.okta.com. https://www.okta.com/blog/2021/04/ccpa-vs-gdpr/

Kaggle. (2022). *Kaggle: Your Home for Data Science*. Kaggle.com. https://www.kaggle.com/

The Investopedia Team. (2019). *What is a credit score? Definition, factors, and ways to raise it*. Investopedia. https://www.investopedia.com/terms/c/credit_score.asp

APPENDICES

APPENDIX A: Project Specification and Gantt chart(s)

Project Specification

Basic details

Student name:	Satyadeep Mohanta
Draft project title:	Loan Approval Prediction
Course and year:	MSc in Data Analytics, (2023-2024)
Client organisation:	
Client contact name:	
Project supervisor:	Dr. Atefeh Khazaei Ghoozhdi

Outline of the project environment

Potential Clients: Our clients can be private banks and loan distribution firms that handle a high volume of loan applications on a daily basis. These institutions are key players in the financial sector and rely on efficient loan processing to maintain their operations.

Client's Business: The client's core business revolves around processing a large number of loan applications daily. Their primary function is to assess the creditworthiness of applicants, determine loan eligibility, and make decisions on loan approvals.

Client's Business Problem: The primary challenge faced by these institutions is the inability to accurately identify reliable clients and distinguish them from individuals who may default on their loans. This results in financial losses, inefficiencies in the loan approval process, and potential risks to the institution's financial stability.

Solution Justification: The proposed solution aims to address the client's business problem by leveraging predictive analytics to assess the likelihood of loan approval. By predicting the outcome of loan applications, these institutions can make more informed decisions, reduce the risk of defaults, enhance the quality of loan approvals, and ultimately improve financial stability outcomes and customer satisfaction.

The problem to be solved.

Detailed Problem Description: The bank faces challenges in accurately assessing loan applications, leading to potential losses when valid applications are mistakenly rejected. The manual and heuristic-based processes currently in place are insufficient for accurately estimating the risk associated with each loan application. This results in wasted time and resources spent on reviewing and documenting applications that could have been approved, causing financial losses for the bank.

Aims and Objectives: The primary aim of this project is to develop a predictive model that can effectively determine the eligibility of loan applications, thereby reducing the number of valid applications that are incorrectly rejected. The model will utilize historical data from previous loan applications as well as current data to make informed decisions on loan approvals.

Constraints: It is imperative to ensure the confidentiality of the data used in the predictive model and to comply with all relevant laws and regulations regarding data privacy.

Computational Resources: Sufficient computational resources must be available for training and deploying the predictive models effectively.

Data Quality: Addressing missing values (NA values) in the dataset is crucial to ensure the accuracy and reliability of the predictive model. The number of rows affected by missing data should be identified, and appropriate strategies for handling missing data should be implemented to restore the dataset's integrity.

Breakdown of tasks

Approach:

Data Collection: Gather and comprehend the datasets, including the current loan application data (Current_app) and historical loan application data (Previous_app).

Data Pre-processing: Clean the datasets by addressing missing values and handling categorical variables through encoding techniques to prepare the data for analysis.

Feature Engineering: Create informative and relevant features by extracting insights from both the current and previous loan application datasets. This step enhances the predictive power of the model.

Building Model: Train various machine learning models using the pre-processed data to predict loan approval outcomes accurately.

Model Evaluation: Perform cross-validation techniques to assess the performance of the trained models and select the most effective model based on predefined evaluation metrics.

Deployment: Implement the predictive model for loan approval predictions in a standalone application or as a command-line tool for easy access and usage within the bank's operational workflow.

Background Research: Conduct in-depth research on machine learning algorithms suitable for handling imbalanced data to address the challenges in loan approval prediction.

Explore best practices in feature engineering and model evaluation to enhance the accuracy and reliability of the predictive model.

Tools and Skills:

Programming Languages: Proficiency in Python, with knowledge of libraries such as pandas and scikit-

learn for data manipulation and machine learning tasks.

Data Processing: Understanding of data cleaning techniques, pre-processing methods (e.g., scaling, normalization), and feature engineering strategies to optimize the data for modelling.

Machine Learning: Familiarity with model training, validation procedures, and hyper-parameter tuning to improve model performance.

Deployment: Ability to package the predictive model into a standalone application or command-line tool for easy integration and usage.

Design and Build: Design the data pipeline for pre-processing tasks, develop the predictive model, and ensure seamless integration into the bank's operational workflow.

Project deliverables

a. Data Pre-processing: An efficient data pre-processing pipeline will be developed to clean, transform, and prepare the historical loan application dataset for model training. This pipeline will handle missing values, feature engineering, and normalization to ensure data quality.

b. Machine Learning Models: The project will develop and fine-tune machine learning models for predicting loan approval outcomes based on historical loan application data. These models will be trained using algorithms such as logistic regression, decision trees, and neural networks to enhance accuracy and reliability.

c. Model Evaluation Framework: A comprehensive model evaluation framework will be implemented to assess the performance of the trained machine learning models. Metrics such as accuracy, precision, recall, and F1 score will be utilized to evaluate the models' predictive capabilities.

d. Code:

Python Code: The Python code scripts used for data pre-processing, model training, and evaluation will be provided. These scripts will be well-documented and organized for reproducibility and future reference.

e. Models and Visualizations:

Trained Machine Learning Models: The trained machine learning models, including the final selected model for loan approval prediction, will be saved and shared. These models will be available for deployment and further analysis.

Visualizations: Informative visualizations such as ROC curves, feature importance plots, and prediction distributions will be generated to enhance the understanding of the model performance and insights gained from the data.

f. A dissertation report consisting of below points:

Project Proposal: Provide a detailed overview of the research objectives, including the problem statement and research questions. Describe the methodology that will be used to address the research objectives. Outline the scope of work, including the specific tasks and activities to be undertaken. Clearly state the project's goals and expected outcomes to set the context for the research.

Literature Review: Summarize existing research on loan approval prediction models and related machine learning techniques. Explain how the literature review will serve as the foundation for the project's theoretical framework. Identify key studies, methodologies, and findings that will inform your research.

Methodology Report: Detail the data collection process, including sources of data and data pre-processing steps. Explain the feature selection criteria and methods used to select relevant features. Describe the model development process, including the algorithms and techniques employed. Outline the evaluation techniques used to assess the performance of the machine learning models.

Results Analysis: Provide an in-depth analysis of the model performance metrics such as accuracy rates, confusion matrices, and feature importance.

Interpret the results to gain insights into the effectiveness of the developed models in predicting loan approval outcomes. Discuss any limitations or challenges encountered during the analysis.

Conclusion and Recommendations: Summarize the project findings and their implications for the research field. recommendations for future research directions based on the project outcomes. Highlight the significance of the project outcomes and suggest areas for further exploration or practical applications.

Requirements

Client Requirements:

Accuracy: The primary focus is on achieving high accuracy in predicting loan approval outcomes to minimize errors and enhance decision-making processes.

Ensuring the model's precision is crucial to avoid misclassifications and provide reliable insights to loan officers.

User Interface: Develop an intuitive and user-friendly interface tailored for loan officers to interact seamlessly with the predictive model.

A well-designed interface will enhance user experience, streamline operations, and facilitate efficient decision-making.

Real-Time Processing: Prioritize quick response times for predictions to support real-time decision-making in loan processing.

Swift processing ensures timely actions, improves operational efficiency, and enhances customer satisfaction.

Eliciting Requirements:

Literature Review: Start by conducting a comprehensive literature review on loan approval prediction systems, machine learning models, and financial risk assessment in the banking sector. This will help you understand the current trends, challenges, and best practices in the field.

Case Studies and Research Papers: Explore case studies and research papers on loan approval prediction systems in the banking industry. Look for successful implementations, key features, and methodologies used in developing predictive models for loan approvals.

Online Surveys and Questionnaires: Create online surveys or questionnaires using platforms like Google Forms or SurveyMonkey and distribute them to professionals in the banking industry through social media groups, professional networks, or university mailing lists.

Collaborate with Academic Advisors: Seek guidance from your academic advisors or professors specializing in data science, finance, or banking. They can provide valuable insights, suggest relevant resources, and help you refine your research focus.

Utilize Online Resources: Explore online platforms such as research databases, academic journals, and industry reports to gather information on loan approval prediction systems, credit risk assessment models, and industry trends.

Legal, ethical, professional, social issues

Ethical Considerations in Data Analytics Research Abstract: This dissertation project focuses on exploring the ethical implications of data analytics practices in research using a dataset obtained from Kaggle. The study aims to investigate the ethical handling of data throughout the data life cycle and machine learning research practices using real-world data. By integrating data science ethics into the analysis of the Kaggle dataset, the project seeks to address privacy concerns, data security, transparency in data usage, and the societal impacts of data science practices. The project will critically examine the ethical dilemmas surrounding big data analytics and machine learning algorithms, emphasizing responsible and ethical decision-making in data analysis research.

Key Objectives:

- Investigate the ethical considerations in data analytics practices using the Kaggle dataset.
- Analyze the ethical handling of data throughout the data life cycle in the context of the Kaggle dataset.

- Examine the societal impacts of data analytics practices based on the analysis of the Kaggle dataset.
- Critically evaluate the ethical dilemmas in big data analytics and machine learning algorithms using the Kaggle dataset.
- Integrate ethical considerations into the analysis of the Kaggle dataset for responsible data analytics practices.

Facilities and resources

Facilities and Resources Computing/IT Facilities:

High-Performance Computing Resources: The project will utilize Google Colab, a free cloud-based platform, for model training. Google Colab provides high-performance computing resources without the need for additional funds, enabling efficient training of machine learning models.

Other Facilities/Resources:

Access to Historical Loan Application Data: Access to historical loan application data is essential for model training and validation. Utilizing Google Colab allows seamless access to and analysis of the required data for building predictive models.

Software Licenses for Necessary Tools: The project will leverage Python libraries, machine learning algorithms and tools available within Google Colab, eliminating the need for separate software licenses. Google Colab provides a comprehensive environment with pre-installed libraries for machine learning tasks.

Availability Constraints:

Ensuring Access to Necessary Computing Resources: Continuous access to Google Colab is crucial to maintain workflow efficiency. As Google Colab is a free platform, ensuring uninterrupted access to the necessary computing resources is essential for the successful execution of the project.

Timely Availability:

Availability of Resources: Since Google Colab is readily available and does not require funds for usage, there are no constraints related to budget approval or resource acquisition. The project can commence immediately, utilizing the computing resources provided by Google Colab.

Project plan

Project Plan: Please find the below project plan**.**

NO	TASK TITLE	START DATE	DUE DATE	STATUS	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11
1	Data Collection and Understanding	03/06/24	05/06/24	100%											
2	Data Pre-processing	06/06/24	14/06/24												
3	Research about the Project	15/06/24	25/06/24												
4	Model Building and Evaluation	26/06/24	12/07/24												
5	Model Tuning and Finalization	13/07/24	20/07/24												
6	Results Analysis	21/07/24	31/07/24												
7	Conclusion and Recommendations	01/08/24	15/08/24												

Risks and Mitigation:

Data Quality Issues:

Description: Address data quality concerns by implementing rigorous data cleaning procedures.

Mitigation: Ensure data accuracy through thorough pre-processing steps.

Model Performance:

Description: Focus on improving model performance through iterative tuning and feature

engineering.

Mitigation: Enhance model accuracy and generalization by optimizing parameters.

Resource Availability:

Description: Secure necessary resources in advance to prevent project delays.

Mitigation: Plan resource allocation to support project timelines and deliverables.

Technical Challenges:

Description: Anticipate technical hurdles and seek expert advice when needed.

Mitigation: Conduct regular progress reviews and consult experts for complex issues.

Backup Plans:

Description: Prepare alternative models in case the primary model underperforms.

Mitigation: Develop and validate backup models to ensure project continuity.

Supervision meetings

The general plan agreed with the supervisor for supervision meetings involves conducting regular meetings once every two weeks via a 30-minute Google Meet session. The mode of meeting will primarily be virtual to facilitate efficient communication and collaboration. In the event of extended absences of the supervisor due to reasons such as annual leave, the project plan includes provisions for continuity and progress tracking. Contingency measures will be in place to ensure that project milestones are met and any necessary adjustments are made to accommodate the supervisor's absence without significant disruption to the project timeline.

Project mode

If there are two possibilities for your project mode, after negotiation, please record your planned duration and submission date. It is also helpful to record your initial registration mode (i.e. are you a full time or a part time student). Remember, the exact dates will be announced through Moodle − these represent a generic guideline.

Registration mode	Full Time
Project mode	Full Time
Planned submission deadline	16/9/2024

12. Signatures

	Signature:	Date:
Student	Satyadeep Mohanta	31/05/2024
Client		
Project supervisor	Dr. Atefeh Khazaei Ghoozhdi	31/05/2024

APPENDIX B: Ethics certificate, generated from ethics link and signed by your supervisor.

UNIVERSITY OF PORTSMOUTH

Certificate of Ethics Review

Project title: Loan Approval Prediction

Name:	Satyadeep Mohanta	**User ID**:	2205708	**Application date**:	30/05/2024 18:03:16	**ER Number**:	TETHIC-2024-108824

You must download your referral certificate, print a copy and keep it as a record of this review.

The FEC representative(s) for the **School of Computing** is/are **Elisavet Andrikopoulou, Kirsten Smith**

It is your responsibility to follow the University Code of Practice on Ethical Standards and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers including the following:

- University Policy
- Safety on Geological Fieldwork

It is also your responsibility to follow University guidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

Which school/department do you belong to?: **School of Computing**
What is your primary role at the University?: **Postgraduate Student**
What is the name of the member of staff who is responsible for supervising your project?: **Atefeh Khazaei Ghoozhdi**
Is the study likely to involve human subjects (observation) or participants?: No
Will financial inducements (other than reasonable expenses and compensation for time) be offered to participants?: No
Are there risks of significant damage to physical and/or ecological environmental features?: No
Are there risks of significant damage to features of historical or cultural heritage (e.g. impacts of study techniques, taking of samples)?: No
Does the project involve animals in any way?: No
Could the research outputs potentially be harmful to third parties?: No
Could your research/artefact be adapted and be misused?: No
Will your project or project deliverables be relevant to defence, the military, police or other security organisations and/or in addition, could it be used by others to threaten UK security?: No
Please read and confirm that you agree with the following statements: I confirm that I have considered the implications for data collection and use, taking into consideration legal requirements (UK GDPR, Data Protection Act 2018 etc.), I confirm that I have considered the impact of this work and and taken any reasonable action to mitigate potential misuse of the project outputs, I confirm that I will act ethically and honestly throughout this project

Supervisor Review

As supervisor, I will ensure that this work will be conducted in an ethical manner in line with the University Ethics Policy.

Supervisor comments:

Supervisor's Digital Signature: **atefeh.khazaei@port.ac.uk** Date: **12/08/2024**