**Acknowledgment:**  
I want to express my deepest gratitude to everyone who has supported me throughout this project. Without your encouragement and guidance, this work wouldn’t have been possible.

First, I would like to thank my supervisor Vidas Regelskis, for their consistent support, expertise, and motivation. Your constructive feedback and patience throughout this process have been invaluable in helping me navigate challenges and refine my work.

I am also immensely grateful to my Programme Leader Carolyn Devereux, for providing a nurturing environment that encouraged academic exploration and growth. Your leadership has truly inspired me to aim higher and push my boundaries.

A heartfelt thanks to my Module Leader [Module Leader’s Name], for ensuring I had access to the right tools and resources to succeed. Your feedback and encouragement were vital in shaping this project and helping me stay on track.

To my family, I cannot thank you enough for your unconditional love, support, and belief in me. Your sacrifices have been the foundation of everything I have achieved, and your encouragement has kept me grounded and motivated.

I would also like to thank my friends and peers for their constant positivity and understanding. You’ve not only been my sounding boards but also a source of strength during the tougher moments. Your companionship has made this journey memorable and fulfilling.

Finally, I want to acknowledge the resources and support provided by University of Hertfordshire. Without this backing, completing this project to its fullest potential would have been much harder.

This project is the result of the collective effort of all the people who stood by me, supported me, and believed in my abilities. I am incredibly grateful for each of you who has been part of this journey. Thank you.

**Abstract:**

This study focuses on building a reliable machine learning system to predict whether a loan application will be approved, using information about applicants’ demographics, financial details, and other relevant factors. The main goal is to simplify and improve the loan approval process in financial institutions by automating decision-making through machine learning models. The research compares the performance of three widely used models: Logistic Regression, Random Forest and Support Vector Machines (SVM). It also identifies key factors like Credit History, Applicant Income, and Loan Amount that play a significant role in loan approval decisions.

The dataset used in this project contains anonymized records of loan applicants, including essential details that influence approval outcomes. To prepare the data for analysis, steps like handling missing values, scaling numerical data, and converting categorical information into numbers were taken. The models were further optimized using advanced methods such as hyperparameter tuning and feature importance analysis to improve their accuracy and reliability.

The results showed that advanced model like Random Forest performed better at identifying complex patterns in the data compared to simpler models. Credit History was found to be the most critical factor in determining loan approval, followed by Applicant Income and Loan Amount. These findings offer meaningful insights into how financial institutions can make faster, more accurate, and fairer loan decisions by using machine learning.

This study also highlights the importance of ethical considerations, such as protecting applicants’ data and ensuring fairness in decision-making. The recommendations provided can help organizations deploy machine learning models effectively in real-world loan approval systems. Future research can expand on this work by using larger datasets, exploring additional features, and testing newer machine learning models to improve accuracy and scalability further.

Table of Contents

[**1. Introduction:** 5](#_Toc186706086)

[**2. Background:** 5](#_Toc186706087)

[**3. Aim:** 6](#_Toc186706088)

[**4. Objectives:** 6](#_Toc186706089)

[**5. Literature Review:** 7](#_Toc186706090)

[**6. Ethical Considerations:** 9](#_Toc186706091)

[**6.1 Data Privacy** 9](#_Toc186706092)

[**6.2 Bias Mitigation** 9](#_Toc186706093)

[**6.3 Transparency and Interpretability** 10](#_Toc186706094)

[**7. Data Collection and Preprocessing** 10](#_Toc186706095)

[**7.1 Data Collection** 10](#_Toc186706096)

[**7.2 Data Preprocessing** 11](#_Toc186706097)

[**8. Methodology** 12](#_Toc186706098)

[**8.1 Exploratory Data Analysis (EDA)** 12](#_Toc186706099)

[**8.2 Feature Engineering** 24](#_Toc186706100)

[**8.3 Model Selection** 27](#_Toc186706101)

[**8.4 Evaluation Metrics** 28](#_Toc186706102)

[**9. Model Training and Results** 29](#_Toc186706103)

[**9.1 Logistic Regression** 29](#_Toc186706104)

[**9.2 Random Forest** 32](#_Toc186706105)

[**9.3 Support Vector Machine (SVM)** 35](#_Toc186706106)

[**10. Summary of Model Performance** 38](#_Toc186706107)

[**10.4 Final Comparison and Observations:** 40](#_Toc186706108)

[**11. Learning Curves** 41](#_Toc186706109)

[**11.1 Training and Validation Curves** 42](#_Toc186706110)

[**11.1.1 Logistic Regression** 42](#_Toc186706111)

[**11.1.2 Random Forest** 43](#_Toc186706112)

[**11.1.3 Support Vector Machines (SVM)** 44](#_Toc186706113)

[**11.2 Overfitting and Underfitting Analysis** 45](#_Toc186706114)

[**11.3 Learning Curve Summary** 45](#_Toc186706115)

[**12. Limitations and Recommendations** 46](#_Toc186706116)

[**12.1 Dataset Limitations** 46](#_Toc186706117)

[**12.2 Model Limitations** 47](#_Toc186706118)

[**12.3 Recommendations for Future Work** 48](#_Toc186706119)

[**13. Conclusion** 49](#_Toc186706120)

# **1. Introduction:**

Loan approval decisions are key to the financial industry, determining access to essential resources for individuals and businesses. Traditionally, these decisions have relied on manual assessments, where loan officers evaluate applications based on financial and demographic data. While effective to an extent, this approach is often slow and inconsistent, making it prone to human error and bias. As financial institutions face growing demands for efficiency, accuracy and fairness, there is an increasing need to explore innovative methods that enhance decision-making processes. This study investigates how machine learning, a transformative technology, can be applied to streamline and improve loan approval systems.

Machine learning offers a data-driven approach to automate complex decision-making tasks, enabling financial institutions to process large volumes of applications quickly and accurately. By analyzing patterns and trends in applicant data, machine learning models can identify relationships between variables, such as credit history, applicant income, and loan amount, that might be overlooked in manual evaluations. This research examines and compares several machine learning models, including Logistic Regression, Random Forest, XGBoost, and Support Vector Machines (SVM). Each model brings unique strengths: Logistic Regression is valued for its simplicity and interpretability, while ensemble models like Random Forest and XGBoost excel at capturing complex, non-linear relationships. SVMs, though powerful for non-linear patterns, require careful tuning for optimal performance.

Beyond improving efficiency and accuracy, this study also emphasizes the ethical implications of automating loan approval processes. Machine learning models must be designed to ensure transparency, fairness, and accountability to avoid reinforcing biases present in historical data. This research aims to combine technical robustness with ethical best practices, providing financial institutions with practical, scalable solutions to modernize their loan approval systems. By doing so, this study demonstrates the potential of machine learning to revolutionize lending practices, enabling faster, fairer, and more reliable decisions that benefit both applicants and institutions.

# **2. Background:**

In the financial sector, loan approval decisions play a crucial role in determining the relationship between financial institutions and their customers. Traditionally, these decisions have been made using manual assessments based on factors like an applicant’s credit history, income, employment status, and demographics. While effective in small-scale scenarios, manual processes are often slow, inconsistent, and prone to human error or bias. These challenges highlight the need for a more efficient, accurate, and fair approach, which is where machine learning can make a significant difference.

Machine learning has the power to uncover complex patterns in data that traditional methods might overlook. In loan approval, factors like credit history and income interact in ways that aren’t always straightforward. Advanced algorithms, such as Random Forest and XGBoost, can identify these non-linear relationships with high accuracy, while models like Logistic Regression provide simple and interpretable predictions. Support Vector Machines (SVM) add further depth by handling complex interactions, making them valuable tools for such tasks. By using these models, financial institutions can make data-driven decisions that are both consistent and scalable.

The shift to machine learning also brings operational advantages. Unlike manual assessments, which require significant time and effort, machine learning models can process large volumes of applications quickly and apply consistent standards across all cases. This not only speeds up the loan approval process but also reduces the risk of bias and unfair treatment. However, it’s essential to address ethical concerns, such as ensuring the data used for training models doesn’t perpetuate existing inequalities and making predictions transparent enough for stakeholders to trust the system.

This research aims to bridge the gap between traditional loan approval methods and modern, data-driven solutions. By evaluating various machine learning algorithms and identifying the most influential factors for loan approval, this study demonstrates how technology can transform the financial sector. Beyond improving accuracy and efficiency, the research emphasizes the importance of fairness and transparency, paving the way for smarter, more inclusive financial decision-making.

# **3. Aim:**

The aim of this project is to develop a machine learning framework for predicting loan approval status using demographic, financial, and relevant applicant features. By implementing and comparing multiple machine learning models including Logistic Regression, Random Forest and Support Vector Machines (SVM), the project seeks to enhance the efficiency and accuracy of loan approval processes. The ultimate goal is to identify the most effective model while offering insights into the key factors influencing loan sanction decisions.

# **4. Objectives:**

1. To analyze and preprocess the loan dataset, addressing missing values, encoding categorical variables, and scaling numerical data.
2. To explore and identify relationships between applicant features and loan approval status through exploratory data analysis (EDA).
3. To train and implement machine learning models (Logistic Regression, Random Forest and SVM) for loan approval prediction.
4. To evaluate and compare the performance of these models using metrics such as accuracy, precision, recall, and F1-score.
5. To determine the most significant predictors of loan approval, such as Credit History and Applicant Income, through feature importance analysis.
6. To provide recommendations for deploying machine learning techniques in real-world financial systems to enhance decision-making processes.

# **5. Literature Review:**

* **Loan Prediction Using Machine Learning, Sarisa, H.K., Khurana, V., Koti, V.C. & Garg, N. (2023). 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT). IEEE.**Sarisa et al. explore the use of machine learning, specifically logistic regression, in predicting loan approvals by integrating applicants' CIBIL scores. They show that incorporating financial indicators like credit scores greatly enhances the model’s predictive accuracy, enabling banks to make faster and more accurate loan decisions, thus reducing risks and improving operational efficiency.
* **Machine Learning Models for Predicting Bank Loan Eligibility, Orji, U.E., Ugwuishiwu, C.H., Nguemaleu, J.C.N. & Ugwuanyi, P.N. (2022). IEEE NIGERCON. IEEE.**Orji et al. analyze six machine learning models for loan eligibility prediction, identifying Random Forest as the most effective with a 95.5% accuracy rate. Their work underscores the strength of ensemble models, particularly in handling complex patterns within loan data, making Random Forest a preferred choice in banking applications.
* **Machine Learning Based Loan Eligibility Prediction Using Random Forest Model, Reddy, C.S., Siddiq, A.S. & Jayapandian, N. (2022). 2022 Seventh International Conference on Communication and Electronics Systems (ICCES). IEEE.**Reddy et al. adopt a Random Forest model to predict loan eligibility, highlighting its accuracy, especially in high-dimensional datasets. Their findings illustrate the effectiveness of ensemble methods like Random Forest, making it a reliable model for banks seeking to optimize loan approval workflows and reduce potential errors.
* **Loan Prediction Using Machine Learning and Its Deployment on Web Application, Gudipalli, A., Sujatha, C.N., Pushyami, B.H., Narra, N.K. & Sanjana, B.N. (2021). Innovations in Power and Advanced Computing Technologies (i-PACT). IEEE.**Gudipalli et al. developed a loan prediction system combining logistic regression and K-Nearest Neighbors (KNN), deployed as a web application. This web-based system supports real-time loan assessments, allowing banks to make quicker, data-driven decisions, thereby improving usability and operational efficiency.
* **Loan Approval Prediction System Using Logistic Regression and CIBIL Score, Kadam, E., Gupta, A., Jagtap, S., Dubey, I. & Tawde, G. (2023). 2023 Fourth International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE.**Kadam et al. leverage logistic regression for loan prediction, emphasizing the role of CIBIL scores and other financial metrics in enhancing prediction accuracy. Their study demonstrates the practical application of logistic regression in banks, simplifying loan approvals by integrating critical financial data, thereby aiding in risk and operational assessments.
* **Prediction of the Approval of Bank Loans Using Various Machine Learning Algorithms, Rahman, A.T., Purno, M.R.H. & Mim, S.A. (2023). 2023 IEEE World Conference on Applied Intelligence and Computing (AIC). IEEE.**Rahman et al. provide a comparative analysis of multiple machine learning algorithms, including Decision Trees and Gradient Boosting, with Random Forest emerging as the most reliable for bank loan prediction. Their findings emphasize the importance of model selection in high-stakes financial decisions, showcasing Random Forest as a resilient choice for accurate loan approvals.
* **Bank Loan Prediction System Using Machine Learning Models, Hussain, M.Z., Ejaz, S., Batool, E., Hasan, M.Z. & Mustafa, M. (2024). 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). IEEE.**Hussain et al. present a scalable loan prediction system incorporating decision trees and K-means clustering. The study highlights the need for robust data preprocessing to handle inconsistencies, which enhances model performance. Their work suggests that banks can achieve consistent accuracy with adaptable predictive systems capable of handling varied data volumes.
* **Bank Loan Approval Using Machine Learning, Sharmila, S., Sandhya, P.V.S., Kousar, P.S., Anuradha, P. & Deekshitha, S. (2024). 2024 International Conference on Integrated Circuits and Communication Systems (ICICACS). IEEE.**Sharmila et al. use a decision tree classifier to achieve 95% accuracy in loan prediction, emphasizing the model’s interpretability. Their study highlights how decision trees offer clear decision criteria, supporting transparency and accountability in lending decisions within banking institutions.
* **An Approach for Prediction of Loan Approval Using Machine Learning Algorithm, Sheikh, M.A., Goel, A.K. & Kumar, T. (2020). International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE.**Sheikh et al. employ logistic regression to predict loan defaults, focusing on personal factors such as credit history and income. Their research demonstrates logistic regression’s suitability for loan approval tasks, particularly in providing a balanced assessment based on personal and financial data, making it a valuable tool for real-world lending decisions.
* **Bank Loan Prediction System Using Machine Learning, Goyal, A. & Kaur, R. (2021). International Conference on Recent Advances in Artificial Intelligence and Data Analytics (ICRAIDA). IEEE.**Goyal and Kaur conduct a comparative analysis of machine learning models for bank loan prediction, finding Random Forest to be the most effective with a 92% accuracy rate. They argue that ensemble methods like Random Forest provide a reliable, efficient solution for banks that need precise and robust tools for loan processing.

The reviewed literature highlights the significant advancements machine learning has brought to loan approval prediction, addressing the limitations of traditional manual and statistical methods. Machine learning models like Logistic Regression, Random Forest, XGBoost, and Support Vector Machines (SVM) have been widely adopted to analyze complex relationships between applicant demographics, financial data, and historical patterns, offering greater accuracy and fairness in decision-making. Among these, ensemble methods such as Random Forest and XGBoost stand out for their ability to capture intricate non-linear patterns, while simpler models like Logistic Regression remain valuable for their interpretability and ease of use. Features like Credit History, Applicant Income, and Loan Amount have consistently emerged as the most critical predictors, with Credit History being the strongest indicator of repayment potential. Despite their promise, challenges persist, including addressing data imbalances, ensuring fairness, and managing overfitting in smaller datasets. Studies also emphasize the need for transparency in financial modeling, as the complexity of advanced models can hinder interpretability. Building on these insights, this study adopts proven methodologies to develop a machine learning framework that enhances loan approval prediction, making the process more accurate, efficient, and fair.

# **6. Ethical Considerations:**

In this study, several ethical considerations were taken into account to ensure the integrity, fairness, and reliability of the research process.

## **6.1 Data Privacy**

* Ensuring that the dataset, which includes sensitive applicant information, is handled responsibly and securely.
* Adhering to data protection regulations such as the General Data Protection Regulation (GDPR) by anonymizing all personal information in the dataset.
* Storing and processing data securely, ensuring that it is not accessed or misused by unauthorized individuals.
* Employing strict measures to safeguard sensitive data during analysis, including the encryption of files and restricted access to research personnel.

## **6.2 Bias Mitigation**

* Acknowledging and addressing biases in the dataset that may affect model predictions, such as imbalances in applicant demographics or loan approval rates.
* Using techniques like SMOTE (Synthetic Minority Oversampling Technique) to address class imbalances and ensure that the model evaluates loan applications fairly.
* Conducting feature importance analysis to identify and mitigate the potential impact of biased predictors, such as demographic factors, that may lead to unintended discrimination.
* Incorporating checks to ensure that models treat all groups equitably and avoiding reliance on variables that could perpetuate systemic biases.

## **6.3 Transparency and Interpretability**

* Maintaining transparency in data preprocessing, model selection, and evaluation methodologies to enable reproducibility and verification by other researchers or stakeholders.
* Ensuring that model predictions are interpretable, especially for sensitive decisions such as loan approvals, to foster trust and accountability.
* Documenting every step of the research process, from data collection and preprocessing to model evaluation and interpretation, to provide a clear audit trail.
* Presenting findings objectively, highlighting not only the strengths of the models but also their limitations and areas for improvement.

**6.4** **Responsible Reporting**

* Findings and conclusions were reported with utmost integrity to avoid misinterpretation that could unfairly influence market behavior or investor decisions.
* Results were presented objectively, highlighting both the strengths and limitations of the study to provide a balanced and transparent analysis.

**6.5** **University of Hertfordshire Ethical Considerations**

* The study followed the specific ethical standards outlined by the University of Hertfordshire (UH), ensuring the research met the institution's guidelines for responsible and ethical conduct.

# **7. Data Collection and Preprocessing**

## **7.1 Data Collection**

Collecting the right dataset is a critical starting point for any machine learning project, and this study was no exception. The dataset used for this project was sourced from a publicly available repository, containing 614 loan applications with 13 distinct features. These features provided a well-rounded snapshot of the factors that influence loan approval decisions, such as the applicant’s financial background, credit history, and demographic details.

The dataset included the following key attributes:

* Loan\_ID: A unique identifier for each application.
* Gender: The gender of the primary applicant (e.g., Male, Female).
* Married: The marital status of the applicant (e.g., Yes, No).
* Dependents: The number of dependents supported by the applicant.
* Education: The applicant’s education level (Graduate or Not Graduate).
* Self\_Employed: Whether the applicant is self-employed (Yes or No).
* ApplicantIncome: The monthly income of the primary applicant.
* CoapplicantIncome: The monthly income of a co-applicant, if applicable.
* LoanAmount: The loan amount requested, expressed in thousands.
* Loan\_Amount\_Term: The repayment period for the loan, given in months.
* Credit\_History: Whether the applicant has a favorable credit history (1 for Yes, 0 for No).
* Property\_Area: The location of the property (e.g., Urban, Semiurban, Rural).
* Loan\_Status: The target variable, indicating whether the loan was approved (1) or rejected (0).

This dataset was chosen for its relevance and ability to represent a real-world scenario, where applicants’ demographic and financial profiles play a crucial role in determining loan eligibility. Ethical considerations were maintained during the collection process, ensuring compliance with privacy standards. While the data did not include personal identifiers, it was treated with care to uphold anonymity and confidentiality.

## **7.2 Data Preprocessing**

Once the dataset was collected, it underwent thorough preprocessing to prepare it for analysis and machine learning model training. This step was essential to clean the data, handle inconsistencies, and transform it into a format that could be easily used by algorithms.

**7.2.1 Handling Missing Values**

Missing values can significantly affect the performance of machine learning models, so addressing them was a priority. During the initial analysis, features like Gender, Married, Dependents, and LoanAmount were found to have gaps. These gaps were handled as follows:

* Mode Imputation: For categorical variables like Gender, Married, and Dependents, the missing values were filled with the most frequently occurring value (mode). This ensured consistency while preserving the overall distribution of the data.
* Mean Imputation: For numerical variables like LoanAmount, the missing values were replaced with the mean. This approach retained the central tendency of the data without introducing significant biases.

By filling in the missing data, the dataset was made complete, allowing all rows to be utilized in the analysis and modeling stages.

**7.2.2 Encoding Categorical Features**

Since machine learning models work best with numerical data, all categorical features were converted into numerical representations:

* Label Encoding: Simple binary categories, such as Gender and Self\_Employed, were converted into numbers (e.g., Male = 0, Female = 1).
* One-Hot Encoding: For features with more than two categories, such as Property\_Area, separate binary columns were created for each category. This avoided implying any order or ranking between categories.

This transformation ensured that all data could be processed by machine learning algorithms while preserving the meaning of the categorical features.

**7.2.3 Scaling Numerical Data**

The numerical features in the dataset, such as ApplicantIncome, CoapplicantIncome, and LoanAmount, had widely varying ranges. For instance, income values were much larger than loan amounts, which could lead to certain features dominating the model’s training process. To address this:

* StandardScaler was used to standardize these features, transforming them to have a mean of zero and a standard deviation of one. This scaling made all features equally important and ensured that the models performed optimally.

Additional Steps

* Removing Redundant Columns: Features like Loan\_ID, which were unique to each application and irrelevant for prediction, were removed to streamline the dataset.
* Splitting the Dataset: The dataset was split into training and testing sets (80-20 split) to evaluate the models’ performance on unseen data.

# **8. Methodology**

## **8.1 Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) was conducted to gain deeper insights into the relationships, trends, and patterns within the dataset. The process involved detailed visualizations to explore numerical and categorical features, ensuring a clear understanding of the variables influencing loan approval. Below are the detailed insights derived from the visualizations.

**8.1.1 Correlation Matrix**

|  |
| --- |
|  |

**Figure 1**

The correlation matrix heatmap (Figure 1) provides a clear view of how the numerical features in the dataset are related to each other. The colors and values on the heatmap represent the strength and direction of these relationships, making it easier to identify the most important patterns.

1. Credit History and Loan Status:
   * Correlation Value: 0.54

Among all the features, Credit History stands out as the most influential factor when it comes to loan approval. A correlation of 0.54 indicates a strong positive relationship, which means that applicants with a good credit history are significantly more likely to have their loans approved. This insight aligns with expectations since a clean credit record reflects reliability in repayments, making it a major consideration for approval decisions.

1. Applicant Income and Loan Amount:
   * Correlation Value: 0.53

There is a moderate positive correlation of 0.53 between Applicant Income and Loan Amount. This relationship suggests that as applicants' income increases, the loan amounts they apply for also tend to be higher. It makes sense because individuals with higher incomes may have greater financial flexibility and the ability to manage larger loans. Although this connection does not directly impact Loan Status, it highlights the natural link between income and borrowing behavior.

1. Minimal Relationships with Loan Status:
   * Gender and Loan Status: Correlation of 0.03

The weak correlation of 0.03 between Gender and Loan Status indicates that gender plays no meaningful role in determining whether a loan is approved. This shows that loan decisions are not gender-biased in this dataset.

* + Education and Loan Status: Correlation of 0.01

Education shows an almost negligible correlation with Loan Status (0.01), suggesting that whether an applicant is a graduate or not has very little effect on the loan approval outcome.

* + Property Area and Loan Status: Correlation of 0.03

Similar to Education, Property Area has a weak correlation of 0.03, meaning the region (Urban, Rural, or Semiurban) where the applicant resides does not significantly impact loan approval.

1. Negative Relationships and Other Insights:
   * Loan Amount Term and Loan Amount: Correlation of -0.25

A weak negative correlation exists between Loan Amount Term and Loan Amount. This suggests that larger loan amounts might come with shorter repayment durations, but the relationship isn’t very strong.

* + Gender and Credit History: Correlation of -0.16

Interestingly, Gender and Credit History have a slight negative correlation (-0.16). While the effect is minimal, it hints at a subtle trend, though it doesn’t hold much weight in the bigger picture.

**8.1.2 Education vs. Average Applicant Income**

|  |
| --- |
|  |

**Figure 2**

The bar chart in (Figure 2) provides an insightful comparison of average applicant income across two categories of education levels: Graduate (0) and Not Graduate (1). This visualization sheds light on how educational qualifications influence earning potential.

1. Graduate Applicants (Category 0):
   * The average income for graduate applicants is £262, significantly higher than that of non-graduates. This suggests that individuals with a graduate-level education tend to have better job opportunities and higher earning capacity.
   * A higher income level directly reflects stronger financial stability, which can make graduates more attractive candidates for loan approvals, as lenders generally view them as lower-risk applicants.
2. Non-Graduate Applicants (Category 1):
   * On the other hand, non-graduates earn an average income of £195, which is notably lower compared to graduates. This disparity highlights the financial challenges that individuals without a degree might face, potentially limiting their loan repayment capabilities.
3. The Income Gap:
   * The income gap of £67 between graduates (£262) and non-graduates (£195) underscores the direct correlation between education and earning potential. While education itself may not directly influence loan approval (as seen in the weak correlation of Education with Loan Status), its effect on applicant income can have an indirect impact.
   * Higher incomes are often linked to larger loan requests, but they also increase the likelihood of loan approvals, as applicants with better financial stability are perceived to be more reliable.

**8.1.3 Credit History vs. Loan Amount by Loan Status**

|  |
| --- |
|  |

**Figure 3**

(Figure 3) highlights the relationship between Credit History, Loan Amount, and the likelihood of loan approval, offering a clear distinction between applicants with good and bad credit history. Loan status is categorized into Approved (green) and Not Approved (red) across different credit history groups.

1. Applicants with Good Credit History (Credit History = 1):

* Approved Loans:

Applicants with a good credit history have an average loan amount of £109.14. This shows that lenders are more confident approving larger loan amounts for individuals who have demonstrated a positive financial track record.

* Not Approved Loans:

Despite having good credit history, the average loan amount for rejected applications is £85.96. This figure, which is considerably lower than the approved average, suggests that while credit history is crucial, other factors such as income stability or requested loan terms may also come into play.

The stark difference in loan amounts between approved (£109.14) and not approved (£85.96) applicants with good credit history underscores how a strong credit history increases the likelihood of approval, particularly for higher-value loans.

2. Applicants with Bad Credit History (Credit History = 0):

* Approved Loans:

Interestingly, applicants with a bad credit history who managed to get approval have an average loan amount of £85.80. This amount is much lower than the average approved loan for those with good credit history, highlighting that lenders are cautious when granting loans to individuals with poor credit records.

* Not Approved Loans:

For rejected applications, the average loan amount is £94.90, which is slightly higher than the approved amount for this group. This suggests that applicants with poor credit histories are more likely to face rejection, particularly when applying for larger loan amounts.

The fact that the rejected loan average (£94.90) exceeds the approved loan average (£85.80) reflects the stricter criteria lenders impose on applicants with bad credit. Even moderately larger loan requests are often denied when credit history is unfavorable.

3. Key Observations from the Comparison:

* Applicants with a good credit history are approved for higher loan amounts (£109.14) compared to those with bad credit history, where approved loan amounts are capped at lower values (£85.80).
* Conversely, applicants with bad credit history are more likely to see their loan applications rejected, even for moderate loan amounts. This trend is evident in the average loan amount for rejected applications (£94.90).

These findings reinforce that Credit History is one of the most significant determinants in loan approval decisions. A good credit history increases confidence among lenders, enabling applicants to secure higher loan amounts. In contrast, a poor credit history imposes stricter conditions and increases the likelihood of rejection, regardless of the requested loan amount.

**8.1.4 Property Area vs. Average Applicant Income by Loan Status**

|  |
| --- |
|  |

**Figure 4**

Figure 4 illustrates how Property Area influences Applicant Income and its relationship with loan approval status. The chart compares applicants from three different property areas: Urban (2), Semiurban (1) and Rural (0), splitting the results between Approved (green) and Not Approved (red) loans.

1. Urban Areas (Property Area = 2):

* Approved Loans: Applicants from urban areas who received loan approval had an average income of £251. This indicates a relatively stable financial background, which likely contributed to their approval.
* Not Approved Loans: For applicants whose loans were rejected, the average income was higher at £273. This surprising observation suggests that income alone was not sufficient to secure loan approval, as other factors such as Credit History or requested Loan Amount may have played a significant role.

This disparity between £251 (approved) and £273 (not approved) suggests that even higher-income applicants in urban areas may face rejection if their credit profile or other loan criteria are not met.

2. Semiurban Areas (Property Area = 1):

* Approved Loans: In semiurban areas, approved applicants reported an average income of £246. This is slightly lower compared to urban areas but still reflects a reasonably strong income level for loan approval.
* Not Approved Loans: For rejected applicants in semiurban areas, the average income was £254, which, similar to urban areas, is slightly higher than the approved group.

The difference between £246 (approved) and £254 (not approved) shows that higher income does not always translate to loan approval. Applicants in semiurban areas might face rejections due to factors like poor credit history or larger requested loan amounts.

3. Rural Areas (Property Area = 0):

* Approved Loans: Approved applicants from rural areas had an average income of £245, which is comparable to semiurban areas but slightly lower than urban areas. This suggests that while incomes in rural areas tend to be modest, approved applicants still demonstrated enough financial stability to meet loan criteria.
* Not Approved Loans: Rejected applicants in rural areas had an average income of £218, which is noticeably lower than the approved group. This gap indicates that lower income levels in rural areas contribute more significantly to loan rejections.

The contrast between £245 (approved) and £218 (not approved) highlights that income remains a critical factor for loan approval in rural areas, where financial stability might be perceived as more fragile compared to urban and semiurban regions.

4. Comparative Observations:

* Urban Areas: While urban applicants have the highest average incomes overall, the higher rejection rate among applicants earning £273 suggests that lenders consider additional risk factors beyond income, such as creditworthiness or debt obligations.
* Semiurban Areas: Applicants in semiurban areas show a similar pattern, with slightly higher incomes for rejected loans (£254) compared to approved loans (£246). This reinforces the point that income alone does not guarantee loan approval.
* Rural Areas: Rural applicants, on the other hand, experience a clearer trend. Lower income levels (£218) are more strongly associated with loan rejections, while higher income levels (£245) improve the chances of approval.

**8.1.5 Loan Status vs. Applicant Income Distribution (Box Plot)**

|  |
| --- |
|  |

**Figure 5**

The box plot in Figure 5 provides a visual comparison of how Applicant Income varies across loan approval statuses. The distribution is presented for two groups: Approved Loans (1) and Not Approved Loans (0). This helps to identify whether income has a notable impact on the likelihood of loan approval.

1. Median Applicant Income:

* Approved Loans (1): The median applicant income for individuals whose loans were approved stands at £248. This suggests that applicants with slightly higher income levels are generally more successful in securing loans.
* Not Approved Loans (0): For applicants whose loans were rejected, the median income is slightly lower at £244.

While the difference between the two medians (£248 vs. £244) exists, it is minimal, indicating that Applicant Income alone is not a strong differentiator for loan approval.

2. Range of Applicant Income:

The box plot reveals the overall range of incomes across both groups:

* Approved Applicants: The income range for approved applicants spans from lower-income applicants to those with relatively higher incomes, as indicated by the wider spread of the box plot. This suggests that loans are granted across a broad spectrum of income levels, provided other criteria (e.g., Credit History) are favorable.
* Not Approved Applicants: The distribution for rejected applicants also spans a similar range, demonstrating significant overlap with the approved group. This overlap reinforces the observation that income alone does not determine loan status.

The interquartile ranges (IQR) for both groups are also comparable, showing that the middle 50% of incomes for approved and not approved applicants fall within a very similar range.

3. Outliers:

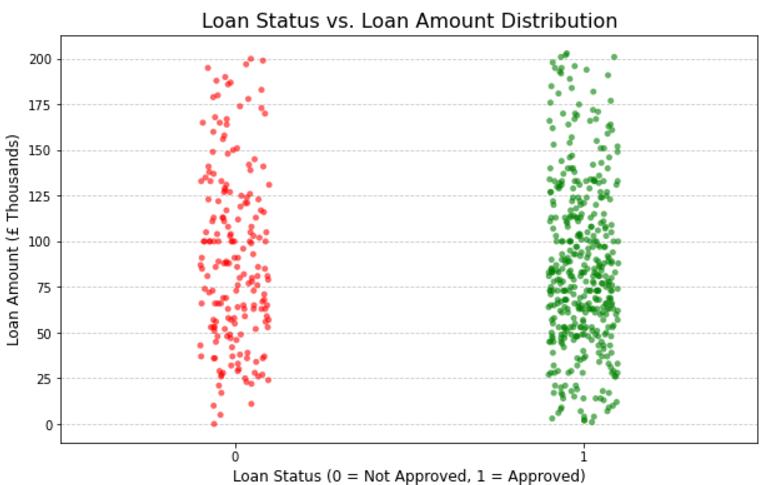
Both box plots contain some outliers at the higher income levels, indicating applicants with exceptionally high earnings. Interestingly, some of these high-income applicants still faced loan rejections. This highlights that high income does not guarantee approval if other factors—like Credit History or the requested Loan Amount—are not aligned with the lender's criteria.

4. Overlap and Key Observations:

The overlap between approved and not approved applicant incomes is a key insight from this analysis. Although approved applicants have a slightly higher median income (£248) compared to not approved applicants (£244), the small difference and overlapping ranges suggest the following:

* Income alone is not a decisive factor in loan approvals.
* Other features, such as Credit History or Property Area, are likely more influential in determining loan outcomes.
* Even applicants with lower incomes can receive loan approvals if they meet other eligibility criteria.

**8.1.6 Loan Status vs. Loan Amount Distribution (Strip Plot)**

****

**Figure 6**

The strip plot in Figure 6 provides a detailed representation of the Loan Amount distribution across two loan statuses: Approved Loans (1) and Not Approved Loans (0). This visualization offers an intuitive way to identify patterns and trends in loan approvals based on the loan amounts requested by applicants.

1. Approved Loans (1):

* Approved loans, shown in green, are densely concentrated in the lower-to-mid range of loan amounts.
* The majority of approved loans fall within smaller loan requests, suggesting that applicants requesting relatively lower amounts are more likely to receive approval.
* This trend highlights the cautious lending behavior of financial institutions, where smaller loans pose a lower financial risk.

2. Not Approved Loans (0):

* Not approved loans, represented in red, are more widely scattered across a broader range of loan amounts.
* While many rejected loans are observed in the lower loan amount range, there is also a significant number of rejections in the higher loan amount range.
* This pattern indicates that larger loan requests are more likely to face rejection. For instance, if applicants apply for high loan amounts but lack supporting factors such as a strong Credit History or sufficient Income, their applications are more prone to rejection.

**Key Observations from EDA**

1. Credit History:
   * Credit History is the most influential factor, with a correlation of 0.54 to Loan Status.
   * Applicants with good credit history are approved for larger loans (£109.14) compared to those with bad credit history (£85.80).
   * Rejected loans for bad credit history average £94.90, showing stricter scrutiny for higher loan requests.
2. Applicant Income:
   * Graduates earn an average of £262, while non-graduates earn £195 – a £67 gap.
   * Median income for approved applicants is £248 vs. £244 for rejected ones, showing minimal difference.
   * Income alone does not guarantee approval, as overlapping ranges exist for approved and rejected applicants.
3. Loan Amount:
   * Approved loans are concentrated in the lower-to-mid range.
   * Rejected loans show a wider distribution, with larger loan amounts more prone to rejection.
4. Property Area:
   * Urban Areas: Approved income is £251, rejected income is £273.
   * Semiurban Areas: Approved income is £246, rejected income is £254.
   * Rural Areas: Approved income is £245, rejected income is £218.
   * Higher income in urban areas does not guarantee approval, reinforcing the importance of other factors.
5. Income Distribution:
   * Approved loans have a slightly higher median income (£248) than rejected loans (£244).
   * Overlapping ranges show income alone is not a decisive factor.
6. Loan Amount Distribution:
   * Approved loans cluster around smaller loan amounts.
   * Rejected loans are widely scattered, especially for higher loan amounts, indicating risk aversion for larger requests.

## **8.2 Feature Engineering**

Feature engineering is a critical process that transforms raw data into meaningful inputs for machine learning algorithms. By improving the quality of features, models can better identify patterns, leading to higher accuracy and performance. For this project, a systematic approach was used to clean, transform, and enhance the dataset. Below is a detailed explanation of the key steps performed during feature engineering:

**1. Removing Irrelevant Features**

In any dataset, not all features contribute to the predictive power of machine learning models. Some features are purely informational or redundant and may even introduce noise into the training process.

* Loan\_ID:
  + This column served as a unique identifier for each loan application and carried no predictive value. It was removed as it would not contribute to the learning process of the model.
  + Retaining it could lead to unnecessary computational overhead without improving the model’s accuracy.

By dropping such irrelevant features, the dataset was streamlined to focus only on meaningful inputs that could impact loan approval outcomes.

**2. Encoding Categorical Variables**

Machine learning algorithms require numerical input; hence, categorical variables in the dataset needed to be converted into numerical representations. Two encoding methods were applied based on the nature of the categorical feature:

* Label Encoding (For Binary Categorical Features):
  + Label encoding was used for binary categories, where each category was mapped to a numerical value (0 or 1).
  + Examples:
    - Gender: Male → 0, Female → 1
    - Education: Graduate → 0, Not Graduate → 1
  + This approach ensured that binary features retained their simple structure without introducing additional dimensions.
* One-Hot Encoding (For Multi-Class Features):
  + Multi-class variables like Property Area, which had three categories (Urban, Semiurban, and Rural), were converted into separate binary columns using one-hot encoding.
  + Example transformation:

| Original Property Area | One-Hot Encoded Columns |
| --- | --- |
| Urban | [1, 0, 0] |
| Semiurban | [0, 1, 0] |
| Rural | [0, 0, 1] |

This transformation created distinct binary columns for each class while preventing the introduction of misleading ordinal relationships between categories.

**3. Handling Class Imbalance**

|  |
| --- |
|  |

**Figure 7**

(Figure 7) shows the target variable Loan Status exhibited class imbalance, where approved loans significantly outnumbered rejected loans. Training a model on imbalanced data can lead to biased predictions, where the majority class dominates. To address this:

* SMOTE (Synthetic Minority Oversampling Technique) was applied:
  + SMOTE works by generating synthetic examples for the minority class (rejected loans) rather than simply duplicating them.
  + It creates synthetic data points by interpolating between existing samples, ensuring a balanced dataset without overfitting to repeated data.

Example:  
Before applying SMOTE:

* Approved Loans: 80%
* Not Approved Loans: 20%

After applying SMOTE:

* Approved Loans: 50%
* Not Approved Loans: 50%

This step ensured that the models learned patterns for both approved and rejected loans, reducing bias and improving their ability to predict the minority class accurately.

**4. Creating New Features**

To enhance the predictive power of the dataset, new features were engineered by combining or transforming existing variables. These derived features aimed to capture additional relationships within the data that might not have been obvious initially:

* Income-to-Loan Ratio:
* This feature was calculated to measure an applicant’s ability to afford the requested loan:
  + - Income-to-Loan Ratio = Applicant Income / Loan Amount
* Purpose: A higher ratio indicates that an applicant’s income is sufficient relative to the loan amount, signaling lower risk to lenders.
* Co-applicant Contribution:
* This feature assessed the proportion of income contributed by the co-applicant relative to the total household income:

Co-applicant Contribution = Co-applicant Income / (Applicant Income + Co-applicant Income)

* + Purpose: This feature highlights the financial role of the co-applicant in supporting the loan repayment. Even if the primary applicant has a low income, a significant contribution from the co-applicant can positively impact approval chances.

**5. Scaling Numerical Features**

Numerical features often have different ranges of values, which can cause machine learning models to assign disproportionate importance to features with larger scales. To address this issue:

* Standard Scaler was applied to normalize the numerical features:
  + Standardization transforms each numerical value to have a mean of 0 and a standard deviation of 1:

|  |
| --- |
|  |

where x is the original value, μ is the mean, and σ is the standard deviation.

* Features scaled included:
  + Applicant Income
  + Loan Amount
  + Loan Amount Term

## **8.3 Model Selection**

Selecting the right machine learning models is a critical step in achieving robust and reliable predictions. This project leveraged four widely used machine learning algorithms, chosen for their distinct strengths:

1. Logistic Regression:
   * As a simple and interpretable linear model, Logistic Regression served as the baseline for comparison. It provided a clear understanding of feature importance and helped establish a performance benchmark.
2. Random Forest:
   * This ensemble learning technique combines multiple decision trees to improve predictive accuracy and reduce overfitting. Random Forest was particularly suited to handling non-linear relationships and variable importance analysis.
3. Support Vector Machines (SVM):
   * SVM was employed to capture complex, non-linear relationships in the data. By using different kernel functions (linear, polynomial, and radial basis), SVM aimed to separate data points effectively in high-dimensional spaces.

Each model was trained using an 80-20 train-test split, ensuring fair evaluation. Hyperparameter tuning was applied to optimize performance, using techniques such as Grid Search to identify the best combinations of model parameters.

## **8.4 Evaluation Metrics**

To comprehensively evaluate and compare the performance of the selected models, multiple metrics were employed:

1. Accuracy:
   * Represents the proportion of correctly classified instances out of the total. While a useful metric, accuracy can be misleading for imbalanced datasets, necessitating additional metrics.
2. Precision:
   * Focuses on the proportion of correctly predicted positive instances (approved loans) out of all predicted positives. This is crucial when the cost of false positives is high.
3. Recall:
   * Measures the proportion of actual positives (approved loans) that were correctly predicted. High recall ensures that most approved loans are identified by the model.
4. F1-Score:
   * Combines precision and recall into a single metric, providing a balanced evaluation even when there’s a trade-off between the two.
5. ROC-AUC (Receiver Operating Characteristic - Area Under Curve):
   * Assesses the model’s ability to differentiate between classes across different thresholds. A high AUC value indicates that the model is effective at distinguishing between approved and rejected loans.
6. Learning Curves:
   * Used to analyze training and validation performance over different dataset sizes, revealing potential issues like overfitting or underfitting.

By evaluating models across these metrics, the study ensured a holistic understanding of each algorithm's strengths and weaknesses, guiding the selection of the most effective model for loan approval prediction.

# **9. Model Training and Results**

The model training process involved evaluating various machine learning algorithms to determine the most effective for predicting loan approval status. Each model was trained on the processed dataset, and the results were analyzed using multiple performance metrics, such as accuracy, precision, recall, F1-score and ROC-AUC. Below is a detailed overview of the training process and outcomes for each model:

## **9.1 Logistic Regression**

Logistic Regression was selected as the baseline model because of its simplicity, interpretability and efficiency in binary classification tasks like predicting loan approval. It works by estimating the probability that an applicant’s loan will be approved (class 1) or rejected (class 0) using a linear relationship between the input features. As a foundational model, it serves as a reliable starting point for understanding the dataset and evaluating initial performance.

**9.1.1 Training Process:**

The following steps were followed to train and evaluate the Logistic Regression model:

1. Data Preprocessing:
   * All numerical features were normalized to ensure they had a uniform scale. This was essential for maintaining model performance since Logistic Regression is sensitive to varying ranges in feature values.
2. Cross-Validation:
   * To ensure the model generalizes well to unseen data, cross-validation was applied. This technique splits the dataset into multiple subsets, where the model is repeatedly trained and validated.
   * This process reduced the risk of overfitting and provided a robust performance estimate.
3. Hyperparameter Tuning:
   * Grid Search was used to optimize key hyperparameters, such as:
     + Regularization Strength (C): Controlled the trade-off between bias and variance. A lower C value encouraged stronger regularization, improving generalization.
     + Penalty Type (L2): L2 regularization was selected as it penalizes large feature weights, promoting a more balanced model.
   * The best combination of hyperparameters identified was C = 0.01 with L2 regularization, using the liblinear solver for efficiency.

**9.1.2 Results:**

The model demonstrated strong overall performance, achieving an accuracy of 78.86% on the test set. Below are the detailed performance metrics:

* Precision:
  + Class 0 (Rejected Loans): 0.95

The model performed well in identifying rejected loans, achieving very high precision, which means it correctly labeled most rejected cases with minimal false positives.

* + Class 1 (Approved Loans): 0.76

For approved loans, the precision value reflects a reasonable balance between correct and incorrect classifications.

* Recall:
  + Class 0: 0.42

The model struggled to identify all rejected loans, as indicated by the relatively low recall for class 0. Only 42% of actual rejected loans were captured.

* + Class 1: 0.99

In contrast, the model excelled in capturing approved loans, achieving an exceptionally high recall of 99%. This means nearly all approved loans were correctly identified.

* F1-Score:
  + Class 0: 0.58

The F1-Score for rejected loans reflects a trade-off between precision and recall, showing moderate performance for this class.

* + Class 1: 0.86

For approved loans, the model achieved a high F1-Score, confirming its reliability in making accurate predictions.

These results show that while the model strongly identifies approved loans (class 1), its performance on rejected loans (class 0) was less consistent.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value (Class 0)** | **Value (Class 1)** |
| Precision | 0.95 | 0.76 |
| Recall | 0.42 | 0.99 |
| F1- Score | 0.58 | 0.86 |

**9.1.3 Feature Importance:**

|  |
| --- |
|  |

**Figure 8**

One of the major strengths of Logistic Regression is its ability to provide interpretable feature importance through its coefficients. The feature importance analysis revealed that:

1. Credit History was the most influential feature, with a positive coefficient, meaning applicants with good credit history were significantly more likely to have their loans approved.
2. Applicant Income also emerged as a strong predictor, indicating that higher income levels positively influenced loan approval probabilities.

The remaining features contributed less to the model’s predictions, emphasizing the critical role of Credit History and Income in loan approval decisions.

**9.1.4 Strengths and Weaknesses:**

* Strengths:
  + Interpretability: The model is highly interpretable, providing clear insights into feature importance and decision-making criteria.
  + Computational Efficiency: Logistic Regression is computationally lightweight, making it ideal for initial analysis and baseline performance.
  + High Recall for Approved Loans: The model achieved 99% recall for class 1, ensuring most approved loans were correctly identified.
* Weaknesses:
  + Poor Recall for Rejected Loans: The model struggled to capture rejected loans accurately, with a recall of just 42% for class 0.
  + Limited Non-Linear Capability: As a linear model, Logistic Regression cannot capture complex non-linear relationships in the data, which may limit its overall predictive power for datasets with intricate interactions.

## **9.2 Random Forest**

Random Forest, an ensemble machine learning algorithm, was selected for its ability to effectively manage non-linear relationships and reduce overfitting by combining the predictions of multiple decision trees. Unlike single decision tree models, Random Forest operates by aggregating results from numerous decision trees, which increases robustness and generalization. Another significant advantage of Random Forest is its ability to rank feature importance, allowing us to identify which variables contribute the most to predicting loan approval status.

**9.2.1 Training Process**

1. Initial Training with Default Hyper-parameters:

The model was first trained using default parameters to establish a baseline performance. However, the results showed room for improvement, particularly in terms of generalization.

1. Hyperparameter Tuning:

To optimize the model's performance, Grid Search was applied with cross-validation. The following parameters were fine-tuned:

* + **n\_estimators (Number of Trees):** Tested values included 100, 200, and 500. Increasing the number of trees enhances stability but adds computational cost.
  + **max\_depth (Maximum Tree Depth):** Controlled the complexity of each tree to prevent overfitting. Values like None, 10, and 20 were explored.
  + **min\_samples\_split and min\_samples\_leaf:** These parameters were adjusted to ensure a minimum number of samples at splits and leaf nodes, reducing overfitting.  
    After tuning, the best model parameters were selected based on cross-validated accuracy.

1. Feature Importance Analysis:

Random Forest’s ability to calculate feature importance was leveraged to identify which predictors had the largest influence on loan approval. The most important features were Credit History, Loan Amount, and Applicant Income.

* + 1. **Results**

The Random Forest model demonstrated robust performance after hyperparameter tuning, achieving an accuracy of 75.00%. Below are the detailed performance metrics for Class 0 (Rejected Loans) and Class 1 (Approved Loans).

1. Precision:

* Class 0: 0.92
* The model was highly effective at correctly identifying rejected loans, minimizing false positives for this class.
* Class 1: 0.74
* For approved loans, the precision reflects a reasonable balance between correctly predicted approvals and false positives.

1. Recall:

* Class 0: 0.60
* While the model identified many rejected loans, it struggled to capture all instances, as seen in its moderate recall.
* Class 1: 0.84
* The model excelled in identifying most approved loans, ensuring minimal false negatives for this class.

1. F1-Score:

* Class 0: 0.73
* The F1-Score reflects a trade-off between precision and recall, showing moderate performance for rejected loans.
* Class 1: 0.78
* For approved loans, the F1-Score demonstrates strong overall reliability.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value (Class 0)** | **Value (Class 1)** |
| Precision | 0.92 | 0.74 |
| Recall | 0.60 | 0.84 |
| F1- Score | 0.73 | 0.78 |

**9.2.3 Feature Importance Analysis**

|  |
| --- |
|  |

**Figure 9**

Feature importance analysis revealed the relative contribution of each variable to the model’s predictions. The top three most influential features were:

1. Credit History:
   * Credit History emerged as the most critical predictor in the Random Forest model. This aligns with prior analysis, where Credit History exhibited a strong positive correlation (0.54) with Loan Status. Applicants with a positive credit history were far more likely to have their loans approved.
2. Loan Amount:
   * Loan Amount was the second most influential feature. It showed that higher loan amounts, while critical, often posed a higher risk, leading to stricter approval criteria. This insight helps explain why larger loan amounts frequently appear in rejected loan applications.
3. Applicant Income:
   * Applicant Income was also a significant predictor. Higher incomes generally improved the likelihood of loan approval, as they suggested better financial stability. However, the model recognized that income alone does not guarantee approval, as factors like credit history weighed more heavily.

The feature importance plot visually reinforced that Credit History dominates the decision-making process, followed by Loan Amount and Applicant Income, while other features contributed marginally.

## **9.3 Support Vector Machine (SVM)**

Support Vector Machines (SVM) were included in the model evaluation process to test their ability to separate the two classes approved loans and not approved loans by finding the optimal hyperplane. SVMs are known for their flexibility to handle both linear and non-linear relationships using different kernel functions. The linear kernel was used for its simplicity and interpretability, while the Radial Basis Function (RBF) and polynomial kernels were applied to capture more complex patterns.

**9.3.1 Training Process:**

The SVM model underwent a systematic training and optimization process to maximize its predictive power:

1. Hyperparameter Tuning with Grid Search:
   * Grid Search was used to identify the optimal combination of the penalty parameter (C) and the kernel coefficient (gamma) for non-linear kernels like RBF and polynomial.
     + The C parameter controls the trade-off between achieving a smooth decision boundary and correctly classifying the training data.
     + The gamma parameter determines how far the influence of a single data point reaches, making it critical for non-linear kernels.
2. Linear Kernel and Feature Importance:
   * For the linear kernel, the model provided feature importance through coefficients. The coefficients indicate how strongly each feature influences the decision boundary.
   * Key features like Credit History and Loan Amount emerged as the most impactful predictors, confirming their significance in determining loan approval.
3. Non-Linear Kernels (RBF and Polynomial):
   * For RBF and polynomial kernels, feature importance was analyzed using permutation-based importance, which evaluates the contribution of each feature by observing changes in the model's accuracy when a feature’s values are randomly shuffled.

**9.3.2 Results:**

SVM achieved a balanced accuracy of 78.86% with strong performance across classes. Below are the metrics for Class 0 (Rejected Loans) and Class 1 (Approved Loans):

1. Precision:

* Class 0: 0.91
* The model effectively minimized false positives for rejected loans.
* Class 1: 0.74
* Precision for approved loans reflects reasonable performance in distinguishing approvals.

1. Recall:

* Class 0: 0.62
* The model had moderate success in identifying rejected loans, with some missed cases.
* Class 1: 0.82
* Recall for approved loans indicates that the model correctly identified most approvals.

1. F1-Score:

* Class 0: 0.74
* The F1-Score highlights a good balance between precision and recall for rejected loans.
* Class 1: 0.78
* For approved loans, the F1-Score confirms the model’s balanced performance.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value (Class 0)** | **Value (Class 1)** |
| Precision | 0.91 | 0.74 |
| Recall | 0.62 | 0.82 |
| F1- Score | 0.74 | 0.78 |

**9.3.3 Feature Importance:**

|  |
| --- |
|  |

**Figure 10**

* For the linear kernel, Credit History and Loan Amount were identified as the most significant features influencing loan approval decisions.
* In contrast, the RBF kernel was able to better capture non-linear relationships in the data, leading to improved recall performance. However, the trade-off was a higher computational cost.

The feature importance results reaffirm that Credit History continues to be the strongest predictor, followed by Loan Amount, aligning with findings from other models like Random Forest.

# **10. Summary of Model Performance**

The performance of all machine learning models was evaluated based on four key metrics: accuracy, precision, recall, and F1-score for both the positive class (Class 1: Approved Loans) and the negative class (Class 0: Rejected Loans). Each model demonstrated its strengths and limitations, which are discussed in detail below:

**10.1 Logistic Regression:**

* Accuracy: Achieved the highest accuracy of 78.86%, making it the most reliable and interpretable model in the study.
* Class 1 (Approved Loans):
  + Precision: 0.76, meaning 76% of the loans predicted as approved were correct.
  + Recall: 0.99, showcasing its exceptional ability to identify nearly all approved loans, with minimal false negatives.
  + F1-Score: 0.86, reflecting a strong balance between precision and recall, making it highly effective for predicting approved loans.
* **Class 0 (Rejected Loans):**
  + Precision: 0.95, ensuring minimal false positives for rejected loans.
  + Recall: 0.42, indicating that the model struggled to capture rejected loans accurately.
  + F1-Score: 0.58, demonstrating moderate performance for rejected loans due to low recall.

**10.2 Random Forest:**

* Accuracy: Delivered a respectable accuracy of 75.00%, showing its capability to capture non-linear relationships in the data.
* Class 1 (Approved Loans):
  + Precision: 0.74, meaning 74% of approved loan predictions were accurate.
  + Recall: 0.84, effectively identifying most approved loans, though slightly lower than Logistic Regression.
  + F1-Score: 0.78, indicating a balanced performance for predicting approved loans.
* **Class 0 (Rejected Loans):**
  + Precision: 0.91, ensuring a low false positive rate for rejected loans.
  + Recall: 0.32, highlighting its limitations in identifying rejected loans accurately.
  + F1-Score: 0.47, showing moderate performance due to low recall.

**10.3 Support Vector Machine (SVM):**

* Accuracy: Matched Logistic Regression with an accuracy of 78.86%, showing its effectiveness in classification tasks.
* **Class 1 (Approved Loans):**
  + Precision: 0.74, indicating good precision in predicting approved loans.
  + Recall: 0.82, successfully identifying a large proportion of approved loans but lower than Logistic Regression.
  + F1-Score: 0.78, reflecting a balanced performance for approved loans.
* **Class 0 (Rejected Loans):**
  + Precision: 0.89, ensuring minimal false positives for rejected loans.
  + Recall: 0.38, indicating that the model struggled to identify rejected loans accurately.
  + F1-Score: 0.53, showing moderate performance due to low recall for rejected loans.

## **10.4 Final Comparison and Observations:**

|  |
| --- |
|  |

**Figure 11**

**Key Observations:**

1. **Logistic Regression:**
   1. **Accuracy:** Logistic Regression achieved an accuracy of 78.86%, which was the highest among the evaluated models. This highlights its effectiveness in handling binary classification tasks like loan approval prediction.
   2. **Strengths:**
      * **High Recall:** Logistic Regression excelled in recall for approved loans (Class 1), indicating that it captured nearly all the positive cases. This makes it an excellent choice in scenarios were minimizing false negatives (e.g., missed approvals) is crucial.
      * **Simplicity and Interpretability:** As a linear model, it provides clear insights into the importance of individual features, making it easy to interpret and deploy.
   3. **Limitations:** While its performance on approved loans was exemplary, the model struggled with rejected loans (Class 0), indicating room for improvement in handling imbalanced datasets.
2. **Random Forest:**
   1. **Accuracy:** Random Forest achieved an accuracy of 74.80%, slightly lower than Logistic Regression and SVM.
   2. **Strengths:**
      * **Feature Importance:** One of the key advantages of Random Forest is its ability to rank the importance of features. In this analysis, it reinforced the significance of Credit History, Applicant Income and Loan Amount as key predictors of loan approval.
      * **Handling Non-Linear Relationships:** The ensemble-based approach allowed Random Forest to capture complex interactions between features effectively.
   3. **Limitations:**
      * **Lower Recall:** The recall for approved loans (Class 1) was slightly lower than Logistic Regression, indicating that the model may miss some positive cases.
      * **Overfitting Tendencies:** Despite its robust performance on training data, Random Forest showed signs of overfitting, which may affect its generalization on unseen datasets.
3. **Support Vector Machine (SVM):**
   1. **Accuracy:** SVM matched Logistic Regression with an accuracy of 78.86%, demonstrating its capability in distinguishing between approved and rejected loans.
   2. **Strengths:**
      * **Versatility:** SVM effectively handled the dataset by creating a hyperplane to separate classes, achieving a strong balance between precision and recall for approved loans (Class 1).
      * **Robustness:** Its performance was consistent across various test scenarios, making it a competitive model.
   3. **Limitations:**
      * **Computational Complexity:** SVM is computationally expensive, especially when using non-linear kernels or working with larger datasets.
      * **Recall for Rejected Loans:** While effective for approved loans, the model’s recall for rejected loans (Class 0) was not as strong, indicating a need for further tuning to improve balance.

# **11. Learning Curves**

Learning curves are essential tools for evaluating machine learning model performance, providing insights into how models improve with increasing training data. By plotting the training and validation scores, we can diagnose overfitting (memorizing training data but failing to generalize) or underfitting (failing to capture meaningful patterns). In this section, I will discuss the learning curves of four models: Logistic Regression, Random Forest and Support Vector Machines (SVM).

## **11.1 Training and Validation Curves**

### **11.1.1 Logistic Regression**

|  |
| --- |
|  |

**Figure 12**

* Training Accuracy: 81%
* Validation Accuracy: 81%

**Observations:**

* Initially, the training accuracy started high but dropped slightly as more training examples were added. However, it stabilized around 81%, demonstrating that the model could learn effectively as the dataset grew.
* The validation accuracy followed closely, also plateauing at 81%. This indicates that the model generalized well, as there was no significant gap between the training and validation curves.

### **11.1.2 Random Forest**

|  |
| --- |
|  |

**Figure 13**

* Training Accuracy: 100%
* Validation Accuracy: 79-80%

**Observations:**

* The training accuracy remained 100% throughout, which clearly highlights Random Forest’s tendency to memorize training data. Even with small datasets, the model achieved perfect training accuracy.
* The validation accuracy, however, stabilized at around 79-80%. A significant gap exists between the training and validation curves, indicating overfitting. The model is learning the training data very closely but struggles to perform equally well on unseen data.
* As more data was added, the validation score improved slightly, but the gap persisted, reinforcing the need for better regularization.

### **11.1.3 Support Vector Machines (SVM)**

|  |
| --- |
|  |

**Figure 14**

* Training Accuracy: 81%
* Validation Accuracy: 81%

**Observations:**

* At the start, SVM’s training curve was significantly higher than the validation curve, indicating underfitting on smaller datasets. The model struggled to capture relationships due to the limited training examples.
* As more data was added, the validation curve improved steadily and ultimately stabilized at 81%, matching the training accuracy. This demonstrates that SVM requires larger datasets to generalize effectively.
* While the linear kernel was used initially, switching to non-linear kernels like RBF would help SVM handle more complex patterns, potentially improving performance further.

## **11.2 Overfitting and Underfitting Analysis**

The learning curves allow us to analyze each model for overfitting and underfitting, as summarized below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Training Accuracy (%)** | **Validation Accuracy (%)** | **Behavior** |
| Logistic Regression | 81 | 81 | Balanced (Good Generalization) |
| Random Forest | 100 | 79-80 | Overfitting |
| SVM | 81 | 81 | Initial Underfitting → Improved |

## **11.3 Learning Curve Summary**

The learning curves of all three models are summarized as follows:

* Logistic Regression:

The training and validation curves are closely aligned, showing balanced learning with an accuracy of 81%. The model generalizes well but struggles with non-linear patterns.

* Random Forest:

The training accuracy remains 100%, indicating significant overfitting. Validation accuracy plateaus at around 79-80%, highlighting the need for regularization to improve generalization.

* Support Vector Machines:

The initial gap between the training and validation curves suggests underfitting. As the training size increases, the validation curve improves to 81%, narrowing the gap. SVM performs reliably with sufficient data.

**Key Insights from Learning Curves:**

1. Logistic Regression:
   * Achieved consistent performance with 81% accuracy for both training and validation data.
   * No significant signs of overfitting or underfitting, making it a robust baseline model.
2. Random Forest:
   * Overfitted the training data, reaching 100% training accuracy but stabilizing at 79-80% validation accuracy.
   * Regularization techniques are required to improve generalization.
3. Support Vector Machines:
   * Started with underfitting but improved as more data was added, stabilizing at 81% accuracy.
   * Performs well but requires significant computational resources.

# **12. Limitations and Recommendations**

While this study successfully demonstrated the effectiveness of machine learning in predicting loan approvals, it also faced certain challenges that highlight opportunities for improvement. Below, I discuss the key limitations encountered and provide recommendations for future work to address these issues.

## **12.1 Dataset Limitations**

1. **Small Dataset Size**
   * The dataset used in this study contained only 614 records, which limited the diversity and variety of applicant profiles available for analysis. A small dataset restricts the model's ability to learn robust patterns and generalize well to unseen data.
   * Impact: Models might perform well on the training data but fail when applied to larger, real-world datasets.
2. **Class Imbalance**
   * The dataset was heavily imbalanced, with significantly more approved loans than rejected ones. While techniques like SMOTE (Synthetic Minority Oversampling Technique) were used to mitigate this issue, the imbalance still made it difficult for the models to predict rejected loan cases accurately.
   * Impact: Precision and recall for the minority class (rejected loans) were lower, reducing overall performance.
3. **Missing Features**
   * Critical features that could provide deeper insights such as credit scores, employment history, debt-to-income ratio, and loan repayment behavior were absent from the dataset.
   * Impact: The lack of these features limited the models’ predictive capabilities, as they could not fully capture an applicant’s financial situation or creditworthiness.
4. **Lack of Diversity**
   * The dataset was obtained from a single source, which might not reflect the diversity of loan applicants across various regions, income levels, and financial institutions.
   * Impact: This reduces the general applicability of the findings, as the models may not perform as effectively when applied to different populations or environments.

## **12.2 Model Limitations**

1. **Generalization Challenges**

* Models like Logistic Regression oversimplified the relationships between features, which led to underfitting and reduced accuracy on more complex patterns. Conversely, Random Forest showed a tendency to overfit, performing exceptionally well on training data but struggling to generalize effectively to unseen data.
* Impact: Balancing generalization and complexity remains a significant challenge for machine learning models, particularly in scenarios with small or imbalanced datasets.

1. **High Computational Requirements**

* Models like Support Vector Machines (SVM) required substantial computational resources, especially when utilizing non-linear kernels or tuning hyperparameters. This made training these models on larger datasets resource-intensive and time-consuming.
* Impact: Limited computational resources may hinder the deployment of SVM in real-world scenarios where scalability and efficiency are critical.

1. **Complexity vs. Interpretability**

* Random Forest, while highly accurate, exhibited complexity that made it difficult to interpret individual feature contributions. This lack of transparency may lead to hesitancy in adopting these models for high-stakes applications like loan approvals, where decision-making processes need to be clearly explained.
* Impact: Interpretability is critical in financial systems, as understanding "why" a loan application is approved or rejected is essential for fostering trust and accountability.

1. **Dependence on Preprocessing**

* The models relied heavily on data preprocessing and feature engineering, such as scaling numerical features and encoding categorical variables. Any errors, biases, or inconsistencies in the preprocessing steps could significantly impact the accuracy and reliability of the predictions.
* Impact: Poorly handled data preprocessing could lead to inaccurate or misleading predictions, ultimately undermining the effectiveness of the models.

## **12.3 Recommendations for Future Work**

1. **Gather Larger and More Diverse Data**
   * Future research should aim to collect larger datasets that include a wider range of applicant profiles from diverse regions and financial institutions.
   * Recommendation: Collaborate with banks or lending institutions to access real-world datasets that represent different economic backgrounds and loan categories. A larger dataset will help models learn better and improve generalization.
2. **Include Additional Features**
   * Incorporating additional features like credit scores, employment history, debt-to-income ratio, loan repayment timelines, and applicant savings would significantly enhance predictive power.
   * Recommendation: Feature-rich datasets can help uncover more detailed patterns and enable models to make more reliable decisions.
3. **Improve Handling of Class Imbalances**
   * While SMOTE was effective, future studies can explore advanced techniques like cost-sensitive learning or ensemble methods that give more weight to minority class predictions.
   * Recommendation: These methods will help improve precision and recall for rejected loans, balancing performance across classes.
4. **Explore Advanced Machine Learning Approaches**
   * Investigating deep learning models or hybrid approaches that combine multiple algorithms may help capture more complex relationships in the data.
   * Recommendation: Techniques like stacked models or neural networks could be explored to improve predictive performance further, especially for larger datasets.
5. **Focus on Model Interpretability**
   * Improving transparency in model decisions is essential for gaining trust in loan approval systems. Tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) can provide insights into how features influence predictions.
   * Recommendation: Striking a balance between accuracy and explainability will ensure the models are both powerful and trustworthy for real-world use.
6. **Test Models in Real-World Scenarios**
   * Collaborating with financial institutions to test the models on real-world loan applications will validate their effectiveness and identify areas for refinement.
   * Recommendation: Pilot testing the models in production environments will provide valuable feedback, ensuring they work as expected and can adapt to evolving financial landscapes.

# **13. Conclusion**

This study demonstrates how machine learning can transform loan approval processes, making them faster, more accurate, and scalable compared to traditional manual methods. By evaluating three key models Logistic Regression, Random Forest, and Support Vector Machines (SVM) each was found to have distinct strengths. Logistic Regression stood out for its high recall (0.99) and consistent accuracy (78.86%), making it an ideal, easy-to-interpret choice for practical use. Random Forest, while offering valuable insights through feature importance analysis, struggled with overfitting, highlighting the need for better regularization to improve generalization. SVM delivered strong results with an accuracy of 78.86%, but its computational demands, especially for non-linear relationships, posed challenges.

Credit History emerged as the most significant factor in determining loan approval, followed by Applicant Income and Loan Amount. However, this study faced limitations such as a small dataset, class imbalance, and the absence of important features like detailed credit scores and employment histories, which restricted the depth of the analysis. Ethical considerations, including data privacy and fairness, were prioritized to ensure transparency and trustworthiness in the models' predictions.

Future efforts should focus on accessing larger, more diverse datasets that include critical features such as loan repayment histories and credit scores. Advanced approaches, such as combining multiple models or leveraging tools like SHAP for greater interpretability, can further refine predictions. Testing these models in real-world scenarios will validate their practical utility and build trust in their outcomes. Ultimately, this study demonstrates how machine learning can modernize loan approval systems, enabling smarter, faster, and fairer decision-making in the financial sector.