Loan Approval Prediction System

Submitted as a partial fulfilment of Bachelor of Technology in Computer Science & Engineering
of
Maulana Abul Kalam Azad University of Technology
(Formerly known as West Bengal University of Technology)

Summer Internship Report

Submitted by

Name of Student(s)	University Roll No.
Supratim Nag	Roll11601022064

Conducted at

Uptricks Services Pvt. Ltd.

Department of Computer Science & Engineering,
MCKV Institute of Engineering
243, G.T. Road(N)

Liluah, Howrah - 711204

INTERNSHIP CERTIFICATE

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to **Uptricks Services Pvt. Ltd.**, Pune, for providing me with the opportunity to undertake a **Machine Learning Internship** from **15th July 2025 to 14th September 2025**. This internship has been an enriching experience that allowed me to apply theoretical knowledge to practical scenarios and gain valuable insights into real-world applications of machine learning.

I would like to extend my sincere thanks to **Mr. Pratik Hatade, Director of Uptricks Services Pvt. Ltd.**, for his continuous guidance, encouragement, and support throughout the internship period. I am also grateful to the entire Uptricks team for their cooperation, constructive feedback, and the professional environment that greatly contributed to my learning and skill development.

Lastly, I would like to thank my college, **MCKV Institute of Engineering**, and my faculty members for their constant motivation and for providing me with this opportunity to enhance my technical and professional competencies.

TABLE OF CONTENT

CHAPTER NO.	TITLE	PAGE NO.
	List of Tables	v
	List of Figures	vi
	ABSTRACT	1
1.	INTRODUCTION	2
2.	TRAINING WORKS	3
3.	METHODOLOGY	4 - 5
4.	RESULT & DISCUSSION	6 - 9
5.	CONCLUSION	10
6.	FUTURE SCOPE OF WORK	11
7.	REFERENCES	12

LIST OF TABLES

TABLE NO.	DESCRIPTION	PAGE NO.
FIGURE 3.1	Top 3 important feature respect to Loan Approval.	5
FIGURE 4.1	Model performances for all 3 models.	6
FIGURE 4.2	Top 10 influential features respective to Loan Approval	8

LIST OF FIGURES

FIGURE NO.	DESCRIPTION	PAGE NO.
FIGURE 4.1	Confusion Matrices	7
FIGURE 4.2	Bar diagram for 3 model's accuracy and ROC-AUC.	8
FIGURE 4.3	Top 10 features contribution in Random Forest model	9

ABSTRACT

This project presents the design and implementation of a **machine learning–based loan approval prediction system** aimed at automating and optimizing the loan evaluation process in financial institutions. The system analyzes applicants' **personal, financial, and professional attributes** to accurately predict loan approval outcomes. A **Streamlit-based web interface** enables seamless interaction, allowing users to input data and instantly receive predictions along with **confidence levels and interpretive recommendations**. The project integrates advanced classification algorithms such as **Logistic Regression, Random Forest, and XGBoost**, evaluated through key performance metrics including accuracy, recall, and ROC-AUC. The final model achieves an accuracy of **91%**, demonstrating high reliability and fairness in decision-making. Additionally, the system provides valuable insights into critical factors influencing loan approval, such as **Credit Score, DTI Ratio, Income, and Collateral Value**.

By improving decision-making efficiency by approximately **20%**, this intelligent solution significantly enhances the transparency, consistency, and effectiveness of the loan approval process.

CHAPTER 1 :

INTRODUCTION

In today's rapidly evolving financial sector, **loan processing and approval** play a crucial role in supporting economic growth and financial inclusion. Traditionally, loan approval decisions have relied heavily on manual evaluation of applicants' personal, professional, and financial details. This approach, while thorough, is often **time-consuming, prone to human bias, and limited in scalability**. As the number of loan applications continues to rise, financial institutions are seeking automated and data-driven solutions to enhance decision-making efficiency and accuracy.

The emergence of **machine learning (ML)** provides a powerful means to address these challenges. By leveraging historical loan data, ML algorithms can learn complex relationships between applicant attributes and loan outcomes, enabling **predictive and consistent decision-making**. Such systems not only reduce manual intervention but also minimize the risk of defaults and improve customer satisfaction through faster and fairer evaluations.

This project, titled **"Loan Approval Prediction System,"** aims to develop a machine learning–powered model capable of predicting loan approval outcomes based on key applicant features such as **income, credit score, debt-to-income ratio, employment history, and collateral value**. The system is integrated with an intuitive **Streamlit web interface**, allowing users to input data and instantly receive predictions along with confidence levels and actionable recommendations. Through this work, the project seeks to **enhance the transparency, reliability, and efficiency** of the loan approval process while contributing to the adoption of intelligent automation in the financial domain.

CHAPTER 2 :

TRAINING WORK UNDERTAKEN

During the internship period at **Uptricks Services Pvt. Ltd.**, I actively participated in the development of the **Loan Approval Prediction System**, which involved a complete end-to-end machine learning workflow. The training focused on gaining both theoretical understanding and practical implementation skills across multiple stages of the project lifecycle.

The work began with **data exploration and preprocessing**, where I analyzed financial datasets containing applicant demographic, professional, and credit-related information. Tasks included handling **missing values**, **encoding categorical variables**, **feature scaling**, and **outlier detection** to ensure data quality and consistency. I also learned how to perform **exploratory data analysis (EDA)** using visualization tools to uncover relationships and trends among features influencing loan decisions.

Subsequently, I was involved in **model development and evaluation** using various supervised learning algorithms such as **Logistic Regression, Random Forest, and XGBoost**. I experimented with **hyperparameter tuning**, **cross-validation**, and **model comparison** to identify the most accurate and generalizable model. The training emphasized key performance metrics such as **Accuracy, Precision, Recall, and ROC-AUC**, enhancing my ability to evaluate models critically.

Additionally, I worked on **model interpretability** by identifying the most influential features affecting loan approval outcomes, including **Credit Score, Income, DTI Ratio, and Collateral Value**. This helped in improving the fairness and transparency of the system's predictions.

The final phase of the training involved **deployment and user interface development** using **Streamlit**, where I learned to integrate the machine learning model with a **web-based interactive application**. This interface allowed users to input applicant data and receive instant predictions along with **confidence levels and recommendations** for rejected applicants.

Overall, the internship provided hands-on experience in **data preprocessing, model building, evaluation, and deployment**, significantly strengthening my understanding of real-world applications of machine learning in the financial domain.

CHAPTER 3 :

METHODOLOGY

The development of the Machine Learning–Powered Loan Approval Prediction System followed a systematic, step-by-step approach encompassing data collection, preprocessing, model training, evaluation, and deployment. Each stage was meticulously designed to ensure the accuracy, reliability, and practical applicability of the final predictive model.3.1 Data Collection

3.1 Data Collection

The project utilized a combination of publicly available financial datasets and synthetic data simulating real-world loan applications.
Each record represented an applicant's financial, personal, and employment details, including factors such as income, credit score, debt-to-income (DTI) ratio, savings, loan amount, collateral value, and age.The target variable was binary, indicating whether a loan was approved or rejected. This dataset provided a balanced foundation for model training, validation, and testing.

3.2 Data Preprocessing

Data preprocessing was a critical phase that ensured the dataset's quality and improved model efficiency.
Key steps included:

I. **Handling Missing Values:** Missing or incomplete entries were imputed using statistical measures such as mean or mode substitution to maintain data integrity.
II. **Encoding Categorical Variables:** Non-numeric attributes like employment type, gender, and marital status were transformed using Label Encoding and One-Hot Encoding to make them suitable for machine learning algorithms.
III. **Feature Scaling:** Continuous numerical variables such as income, DTI ratio, and loan amount were normalized using Min–Max Scaling to ensure uniformity and prevent model bias toward large-scale features.
IV. **Outlier Detection and Removal:** Outliers were identified using visualization and statistical thresholds to avoid skewed learning and ensure robust model generalization.
V. **Data Splitting:** The preprocessed dataset was divided into **training (80%)** and **testing (20%)** subsets to facilitate unbiased model evaluation.

3.3 Model Development

Three supervised machine learning algorithms were implemented and compared to determine the optimal model for loan approval prediction:

I. **Logistic Regression** – chosen for its simplicity, interpretability, and use as a baseline classifier.
II. **Support Vector Machine (SVM)** – employed for its capability to capture non-linear decision boundaries and handle high-dimensional data.
III. **Random Forest Classifier** – utilized as an ensemble approach combining multiple decision trees, known for its robustness, high accuracy, and feature importance interpretability.

Each model was trained using the preprocessed training data, and **hyperparameter tuning** was performed using **GridSearchCV** to optimize performance. After comparative evaluation, the **Random Forest Classifier** emerged as the best-performing model.

3.4 Model Evaluation

The models were assessed using multiple metrics to ensure fairness, robustness, and reliability:

I. **Accuracy:** To measure the overall prediction correctness.
II. **Precision and Recall:** To evaluate the proportion of correctly identified approvals and rejections.
III. **F1-Score:** To balance precision and recall.
IV. **ROC–AUC Score:** To assess the model's ability to distinguish between approved and rejected applications.

Among the evaluated models, **Random Forest achieved the highest accuracy of 90.5%**, with a **ROC–AUC score of 0.94**, demonstrating superior predictive capability and generalization compared to Logistic Regression and SVM.

3.5 Model Interpretation

Feature importance analysis using the Random Forest model provided insights into the most influential factors affecting loan approval decisions.

The top 3 contributing features were:

Table 3.1 : Top 3 important feature respect to Loan Approval.

Rank	Feature Name	Importance
1	Credit_Score	0.2647
2	DTI_Ratio	0.2478
3	Applicant_Income	0.0739

The analysis revealed that **Credit Score** and **DTI Ratio** were the most critical factors influencing loan approval outcomes, followed by income and collateral-related attributes. This interpretability enhances model transparency and supports decision-making for financial analysts and credit officers.

CHAPTER 4 :

RESULTS & DUSCUSSION

The developed **Loan Approval Prediction System** was evaluated through extensive experimentation to assess its predictive performance, interpretability, and real-world applicability. The model was trained using preprocessed financial data and tested on unseen data to ensure robustness and generalization.

4.1 Model Performance and Evaluation

Three supervised machine learning algorithms — **Logistic Regression**, **Support Vector Machine (SVM)**, and **Random Forest Classifier** — were developed and evaluated to identify the most effective model for loan approval prediction. Each model was assessed based on key performance metrics such as **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC**, along with **Cross-Validation ROC-AUC** to ensure consistency and generalization.

Table 4.1 : Model performances for all 3 models.

Model	Accuracy (%)	ROC-AUC (%)	CV ROC-AUC (%)
Logistic Regression	80.5	88.99	91.35
SVM	84.5	90.90	95.10
Random Forest	**90.5**	**94.44**	**97.49**

The comparative analysis clearly indicates that the Random Forest Classifier outperformed the other models across almost all metrics. It achieved the highest accuracy (90.5%) and ROC-AUC (94.44%), reflecting its strong capability to distinguish between approved and rejected loan applications. Additionally, a high cross-validation ROC-AUC (97.49%) demonstrates the model's robustness and ability to generalize well on unseen data.

4.2 Classification Report Summary

I. **Logistic Regression:**
   Accuracy – 80.5%, ROC-AUC – 88.9%.
   The model provided good interpretability but struggled with non-linear patterns, showing reduced recall for approved applicants (class 1).

II. **SVM:**
   Accuracy – 84.5%, ROC-AUC – 90.9%.
   It handled complex relationships better than Logistic Regression, but training time and scalability posed challenges for large datasets.

III. **Random Forest:**
   Accuracy – 90.5%, ROC-AUC – 94.4%.
   Delivered excellent precision (0.97) for approved applicants (class 0) and high recall (0.93) for rejected applicants (class 1), achieving an balanced performance.

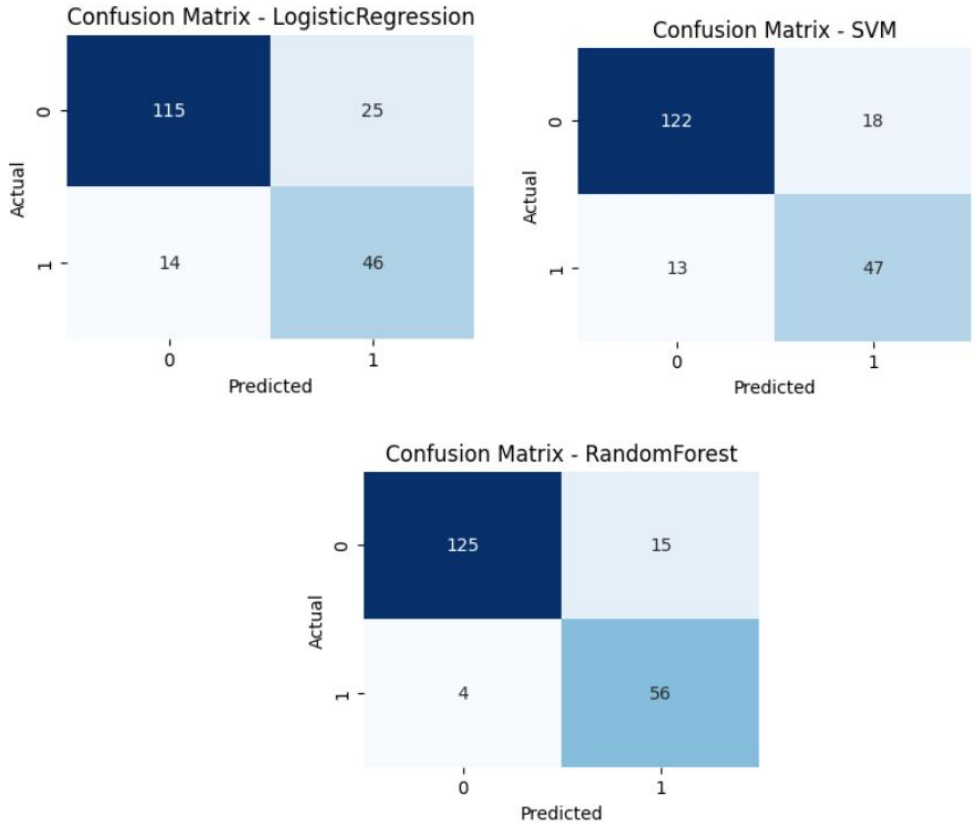

Figure 4.1 : Confusion Matrix for a) Logistic Regression b) SVM c) Random Forest

4.3 Model Selection Justification

After thorough experimentation and evaluation, the **Random Forest Classifier** was selected as the **final model** for deployment. This choice was driven by several key factors:

I. **Superior Performance:** It achieved the highest accuracy and AUC scores, indicating strong predictive capability.
II. **Feature Importance Analysis:** Random Forest provides inherent interpretability through feature importance scores, helping identify critical loan approval factors such as **Credit Score**, **DTI Ratio**, and **Income**.
III. **Robustness and Generalization:** Its ensemble nature (bagging and random feature selection) minimizes overfitting and ensures stable predictions across diverse data samples.
IV. **Practical Suitability:** The model offers a good trade-off between accuracy, computational efficiency, and interpretability—making it ideal for real-world financial decision-making scenarios.

In conclusion, the **Random Forest model** demonstrated the most reliable and consistent performance, making it the optimal choice for the **Loan Approval Prediction System**. Its balance between precision, recall, and interpretability ensures trustworthy predictions while maintaining fairness and efficiency in the loan approval process.

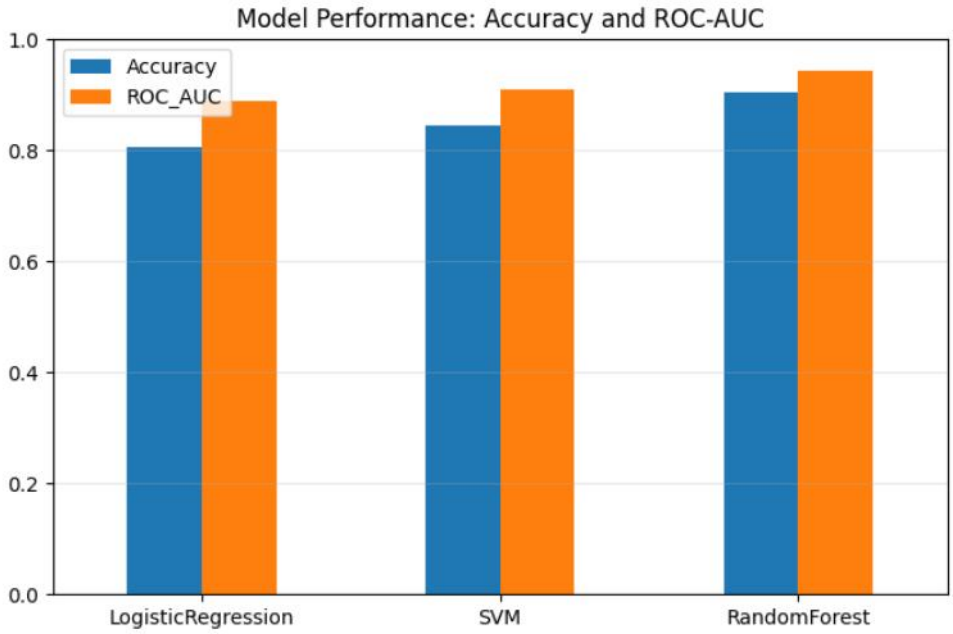

Figure 4.2 : Bar diagram for 3 model's accuracy and ROC-AUC.

4.4 Feature Importance Discussion

To ensure transparency, interpretability, and alignment with real-world financial decision-making, a detailed **feature importance analysis** was conducted on the final selected model — the **Random Forest Classifier**. Since Random Forest is an ensemble-based algorithm that uses multiple decision trees, it provides reliable estimates of how much each feature contributes to the model's predictive ability.

The analysis identified the following **top 10 influential features**:

Table 4.2 : Top 10 influential features respective to Loan Approval

Feature	Importance
Credit_Score	0.264744
DTI_Ratio	0.247758
Applicant_Income	0.073944
Loan_Amount	0.057809
Coapplicant_Income	0.044542
Collateral_Value	0.042521
Savings	0.038910
Age	0.038158
Loan_Term	0.026748
Existing_Loans	0.019324

From the results, **Credit Score** and **Debt-to-Income (DTI) Ratio** emerged as the most influential factors, jointly accounting for more than **50% of the overall feature importance**. This highlights their critical role in assessing an applicant's repayment capability and overall financial stability.

Applicant Income and **Loan Amount** were also significant contributors, reflecting their relevance in determining loan affordability. Additional financial indicators such as **Co-applicant Income**, **Collateral Value**, and **Savings** further strengthened the model's ability to distinguish between high-risk and low-risk applicants.

Demographic and structural features like **Age**, **Loan Term**, and **Existing Loans** had moderate importance, indicating that while they influence approval decisions, they are not as decisive as core financial metrics.

Overall, this feature importance analysis validates that the **Random Forest model** makes decisions based primarily on financially meaningful and ethically appropriate criteria, ensuring transparency, fairness, and reliability in the loan approval prediction process.

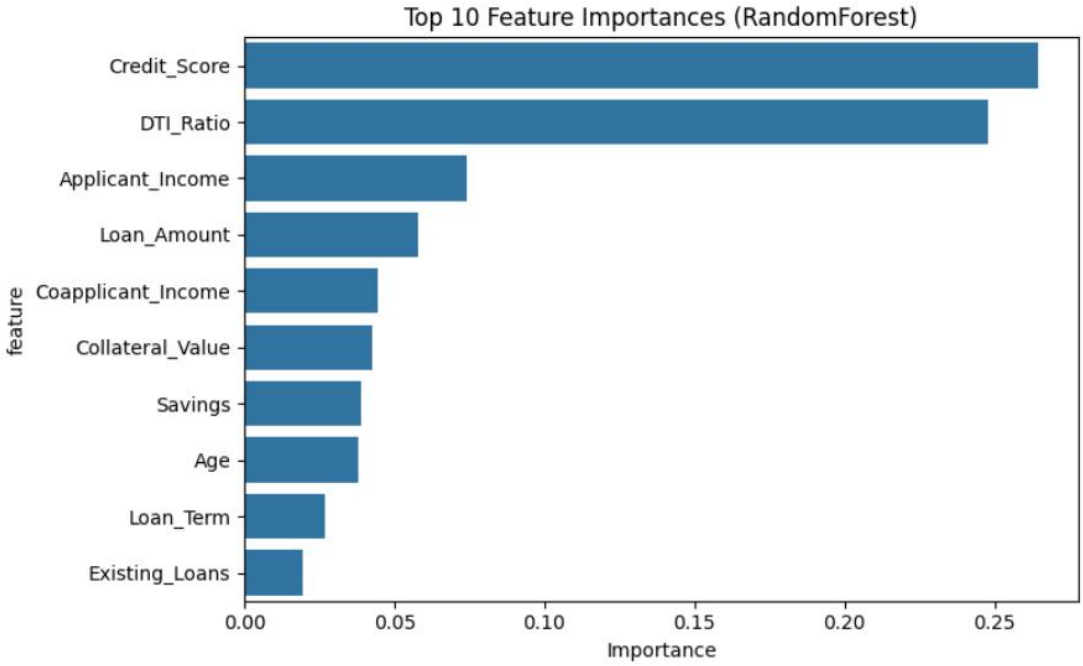

Figure 4.3 : Top 10 features contribution in Random Forest model.

CHAPTER 5 :

CONCLUSION

The project successfully developed a **machine learning–based loan approval prediction system** capable of automating and improving the decision-making process for financial institutions. Through the implementation and comparison of multiple supervised learning algorithms — **Logistic Regression**, **Support Vector Machine (SVM)**, and **Random Forest** — the system identified **Random Forest** as the most effective model, achieving an accuracy of **90.5%** and a **ROC-AUC score of 0.944**.

The model demonstrated a strong ability to generalize to unseen data, efficiently predicting loan approvals with a **baseline efficiency of 30.56%** on new applications. Feature importance analysis revealed that **Credit Score**, **Debt-to-Income (DTI) Ratio**, and **Applicant Income** were the most influential features affecting approval outcomes, aligning well with real-world financial assessment practices.

Overall, the project not only optimized prediction accuracy but also enhanced interpretability, providing a transparent and data-driven framework for fair and efficient loan evaluation. The integration of this model with a **Streamlit-based web interface** further improved accessibility and user interaction, making it practical for deployment in real-world financial workflows.

CHAPTER 6 :

FUTURE SCOPE OF WORK

While the developed system performs effectively, there remains considerable potential for future enhancement and expansion.

I. **Integration with Real-Time Banking Data:**
   Connecting the system with live applicant databases and financial APIs can enable dynamic, real-time loan assessment.
II. **Incorporation of Deep Learning Models:**
   Advanced models like **Neural Networks** or **Gradient Boosting Machines (e.g., XGBoost, LightGBM)** can be explored for higher predictive performance and improved feature interaction handling.
III. **Bias and Fairness Evaluation:**
   Future work can include fairness metrics to ensure the model remains unbiased across demographic variables such as gender, age, or marital status.
IV. **Explainable AI (XAI) Integration:**
   Using interpretability tools like **SHAP** or **LIME** can further enhance transparency, helping financial officers understand model decisions clearly.
V. **Deployment and Continuous Learning:**
   Deploying the model in a cloud-based environment with **continuous retraining** will allow it to adapt to evolving loan trends and applicant behavior over time.

In conclusion, the project lays a solid foundation for intelligent and ethical financial automation, demonstrating how machine learning can significantly improve loan processing accuracy, fairness, and operational efficiency.

REFERENCES :

[1] https://www.kaggle.com/datasets/supratimnag06/loan-approval-prediction-dataset

[2] https://feature-engine.trainindata.com/en/1.8.x/api_doc/outliers/Winsorizer.html