Predicting Bank Loan Approval: A Machine Learning Classification Case Study

Name: REBECCA AGBOLADE
Student ID: D3077427

Abstract

This study explores the application of machine learning techniques for predicting bank loan approval status. With the increasing reliance on data-driven decision-making in the financial sector, the demand for accurate loan approval models has grown substantially. Leveraging historical loan data, we employ various machine learning algorithms to develop predictive models. Through rigorous evaluation and comparison of algorithm performance metrics such as precision, recall, and F-measure, we aim to identify the most effective approach for predicting loan outcomes. Our findings offer valuable insights into financial institutions seeking to optimise their loan approval processes and mitigate risks associated with lending.

Introduction

In the fast-evolving banking and finance landscape, accurately evaluating loan applicants' creditworthiness is paramount. Traditional methods, often manual and error-prone, are being replaced by machine learning (ML) models [1]. These models automate and enhance the loan prediction process, ensuring more accurate assessments and informed lending decisions. Our study focuses on four key ML algorithms: Logistic Regression, Decision Tree, K Nearest Neighbor (KNN), and Random Forest, aiming to evaluate their effectiveness in predicting loan approvals. We seek to assess model reliability, explore integration benefits into traditional practices, and investigate the feasibility of creating a unified model for real-time decision support. By leveraging these ML techniques, we aim to boost efficiency, mitigate default risks, and foster financial inclusion**.**

Related Work

The challenge of accurately predicting loan defaults poses a significant obstacle for financial institutions operating in today's dynamic banking landscape. With evolving consumer behaviors and market dynamics, the need for robust loan prediction models is more critical than ever. This literature review aims to provide an overview of current research in bank loan prediction, highlighting methodologies, key insights, and emerging trends in the field.

A study by [2] introduced a comprehensive set of features tailored to forecast loan defaults effectively. These features encompassed various data categories, including applicant financial history, credit scores, loan details, and socio-economic indicators. Leveraging this extensive feature set, the research evaluated the performance of seven distinct prediction algorithms: logistic regression, decision trees, random forests, gradient boosted machines, support vector machines, neural networks, and ensemble methods [4]. Comparative analyses revealed that the integration of newly added features with advanced modeling techniques yielded superior predictive accuracy compared to conventional methods.

In a separate investigation, [1] examined the efficacy of multiple algorithms for loan prediction using a dataset from a leading banking institution. Notably, the study observed a significant enhancement in model performance, with accuracy soaring from 79.2% to an impressive 86.42% when incorporating borrower credit utilisation ratios and income stability metrics. The evaluation, based on standard metrics such as precision, recall, and F1-score, underscored the importance of incorporating diverse data sources to enhance predictive capabilities.

Building upon existing methodologies, [3][6] proposed an innovative approach to loan prediction by integrating boosting algorithms with clustering techniques. The primary objective was to identify high-risk borrower segments by clustering individuals based on their credit profiles and financial behaviors. Subsequently, logistic regression models were trained for each cluster to predict loan defaults accurately. Experimental results demonstrated that the boosting-enhanced models outperformed traditional logistic regression models, effectively isolating defaulting patterns within specific borrower segments[1].

By synthesizing insights from these studies, this literature review aims to shed light on the evolving landscape of bank loan prediction and provide valuable guidance for financial institutions striving to develop robust predictive models for loan portfolio management.

Methodology

The process involves stages such as gathering data, preparing the data, exploring it, enhancing features, creating models, and assessing their performance[8].

Dataset

The chosen dataset for my work revolves around predicting loan approvals, a critical task in the financial domain. Obtaining accurate predictions relies heavily on various data attributes. These attributes, sourced from Kaggle, encompass 14 columns and 5001 features, provide a comprehensive view of applicant profiles typically scrutinised by financial institutions or banks when evaluating loan applications. The primary focus lies on predicting whether an applicant will be approved for a loan or not, making it a significant undertaking with potential impacts on lending practices and risk assessment strategies.

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
0	1	25	1	49	91107	4	1/60	1	0	0	1	0	0	0
1	2	45	19	34	90089	3	1/50	1	0	0	1	0	0	0
2	3	39	15	11	94720	1	1/00	1	0	0	0	0	0	0
3	4	35	9	100	94112	1	2/70	2	0	0	0	0	0	0
4	5	35	8	45	91330	4	1/00	2	0	0	0	0	0	1
5	6	37	13	29	92121	4	0/40	2	155	0	0	0	1	0
6	7	53	27	72	91711	2	1/50	2	0	0	0	0	1	0
7	8	50	24	22	93943	1	0/30	3	0	0	0	0	0	1
8	9	35	10	81	90089	3	0/60	2	104	0	0	0	1	0
9	10	34	9	180	93023	1	8/90	3	0	1	0	0	0	0

Fig 1: Dataset Overview

Pre-Processing

When obtaining data from Kaggle, it is common to encounter null values or inconsistencies within the dataset, which can diminish the model's efficacy [14]. Therefore, preprocessing the data becomes crucial to rectify such issues. This involves transforming the raw data into a more refined and efficient format, thereby optimizing the performance of the applied algorithm. In essence, data preprocessing entails encoding the Kaggle-acquired data in a manner that is easily comprehensible by the machine, improving its predictive capabilities.

Data Visualisation

Data visualisation is done to quickly show the dataset's features. Count plots and box plots were used to display feature distributions.

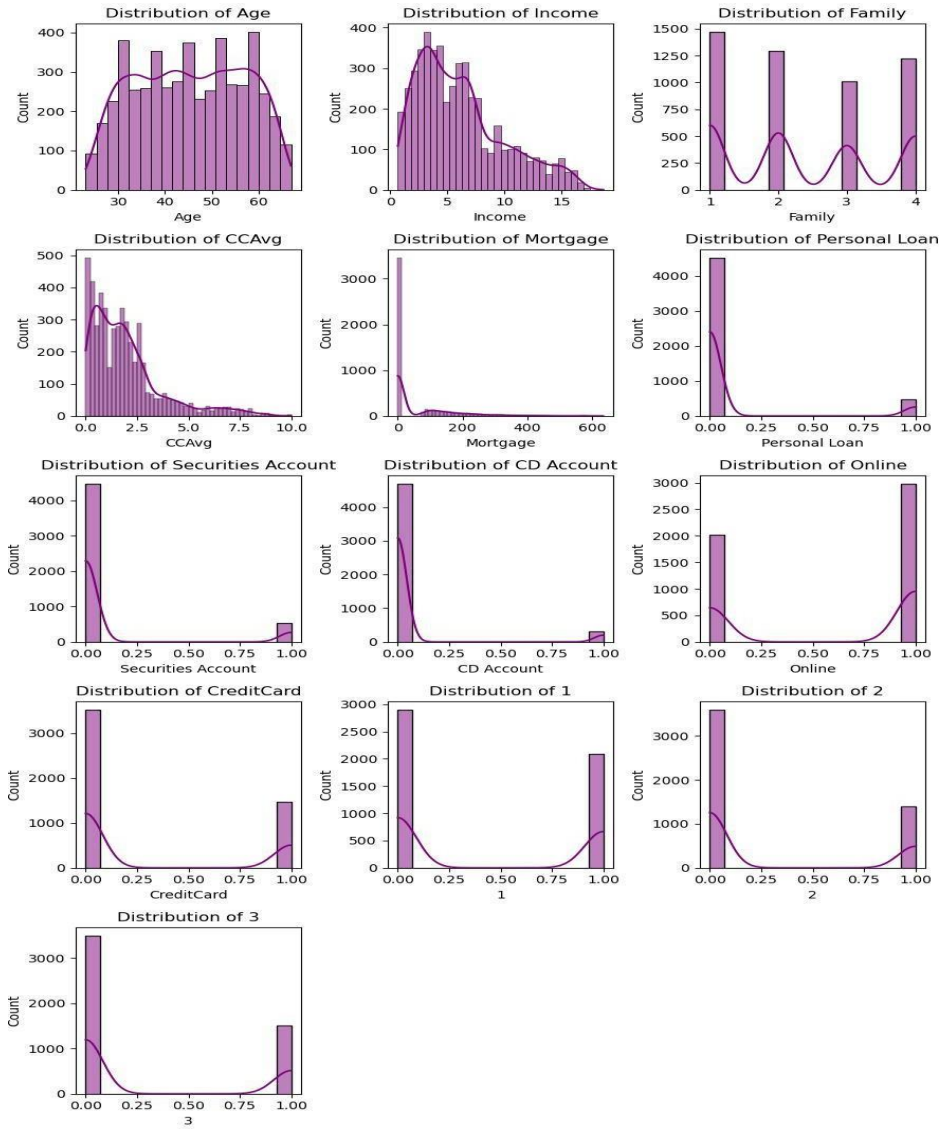

Fig 2: Distribution of features

Data Cleaning

Data cleaning is vital in machine learning, refining datasets to remove irrelevant or inaccurate information [15]. This process ensures that even basic algorithms can deliver accurate results, as illustrated in Fig. 2. By addressing flawed data, data cleaning optimises model performance. Various strategies, like handling null values and converting categorical data, are used systematically to meet the model's accuracy requirements.

Removal of Unnecessary columns

Columns such as ZIP Code and ID will be removed from the dataset as they are deemed unnecessary for this research purpose.

Handling Missing Values

The dataset used for bank loan prediction doesn't have any missing values. However, if there were any, we'd address them using methods like imputation with statistical measures (mean, median, mode), removing incomplete rows/columns, or employing predictive modeling to estimate missing values based on dataset patterns. Even without missing values, it's crucial to acknowledge potential data gaps and discuss strategies to maintain model integrity**.**

Label Encoding

Dummy encoding was utilised to handle categorical data within the 'Education' feature. This method converts categorical variables into binary dummy variables, representing each category with a separate column. Dummy encoding was chosen because it avoids ordinality assumptions and prevents the model from misinterpreting numerical values. This approach enhances the accuracy and robustness of my research findings by effectively incorporating categorical data without introducing bias."

Normalisation

The normalised oversampled training set underwent Min-Max scaling, ensuring feature magnitudes were uniformly scaled within a range of 0 to 1. To maintain consistency, the same scaling parameters from the training set were applied to the testing set.

Following scaling, I transformed the scaled arrays back into data frames for easier interpretation. This facilitated a thorough inspection of the normalised features, confirming the accuracy of the scaling process [13].

Feature Selection

Feature selection is a widely used technique in the initial stages of machine learning [14]. It entails selecting a subset of features from a larger set of available features.

Correlation Analysis for Multi-Collinearity Assessment

Correlation analysis is essential in machine learning for understanding how independent variables interact within a dataset [1]. It helps uncover redundant or highly linked features, aiding in the process of feature selection and guarding against overfitting [5]. Strong correlations can mask individual feature effects in linear models like regression. Thus, addressing multicollinearity through correlation analysis ensures stable coefficient estimates and reduces the likelihood of inflated standard errors, ultimately enhancing both model interpretability and performance [12]. In this analysis, a threshold of 0.7 was used to identify high correlations, and features exceeding this were subsequently removed from the dataset [5].

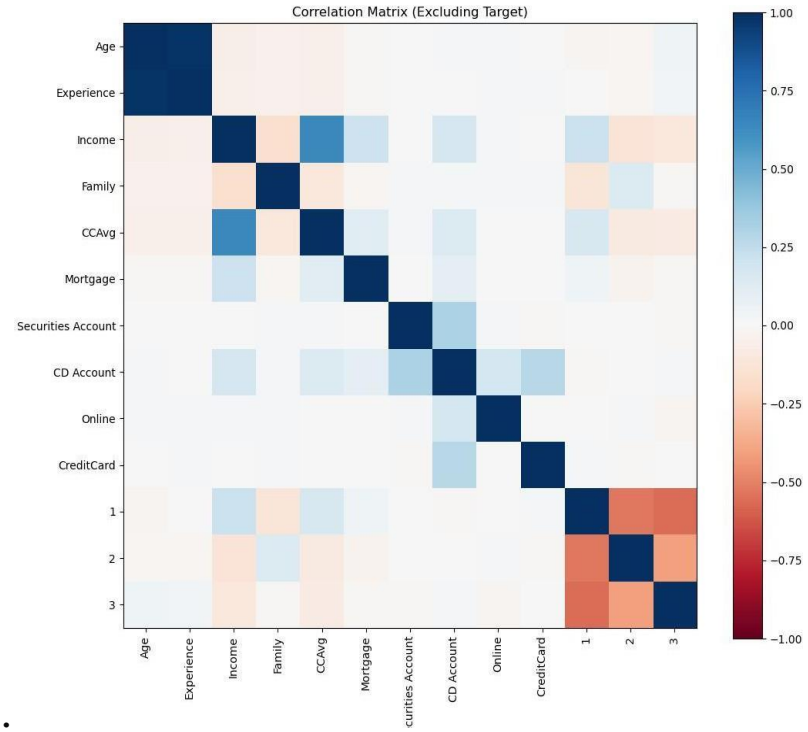

Fig 3: Correlation Plot.

Data Splitting for Machine Learning Algorithm

The dataset underwent a split into training and testing sets using a 70/30 ratio, allocating 70% of the data to the training set and 30% to the test set. This division ensures that the model has sufficient samples to undergo training and enhance its performance [14].

Target Variable Distribution & Up-sampling

Visualizing the distribution of the target variable through count plots is crucial for understanding the balance of classes within a dataset, especially in classification tasks. This visualisation helps assess the percentage of each class, identify potential class imbalances, and evaluate the representativeness of the data. Class imbalances, where one class significantly outnumbers the other, can lead to skewed predictions [8]. To mitigate this issue, data scientists often employ resampling techniques such as upsampling or downsampling based on the observed distribution.

Fig 4: Target Variable Distribution before SMOTE

SMOTE, an upsampling technique, plays a crucial role in addressing class imbalance by creating artificial samples for the minority class. By generating synthetic instances that mimic the features of existing minority class samples, SMOTE helps mitigate the adverse effects of class imbalance during model training. Upsampling ensures a more balanced representation of both classes, enhancing the model's ability to learn from minority class occurrences [16]. This leads to improved classification performance and more reliable predictions across all classes.

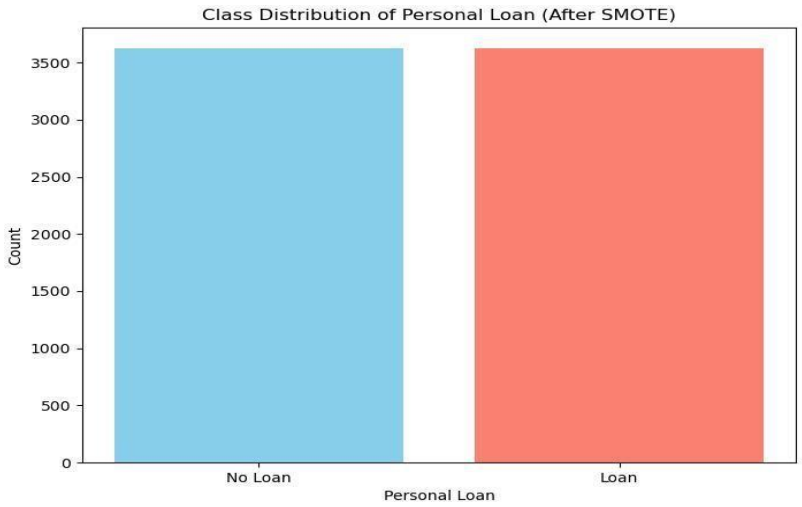

Fig 5: Target Variable Distribution after SMOTE

Machine Learning Algorithm

The choice of classification algorithm for a bank loan prediction problem depends on the type of target feature, which could be categorical (discrete and countable values) or continuous (values that can change over time). Since bank loan prediction typically involves determining whether a loan should be allocated to a customer (a binary outcome), it falls into the category of a two-class classification problem.

Various classification algorithms can be employed for bank loan prediction, considering the nature of the features in the dataset. These features may include categorical data (with distinct categories) or continuous data (values that vary over time).

Modelling Approach

The modeling approach for churn prediction involved considering several machine learning methods, such as K-nearest neighbors (KNN), Gaussian Naive Bayes (GNB), Random Forests, and Logistic Regression. Each algorithm offers distinct advantages and is suited to the churn prediction problem in its own way.

1. **Naive Bayes:** Naive Bayes: Known for its computational speed, Naive Bayes operates on Bayes' theorem, assuming feature independence[9]. It calculates the posterior probability using initial class and feature probabilities [7]. $\boldsymbol{p(c|x) from\ P(c), P(x) and\ p(c|x)}$

$$P(c|x) = \frac{p(x|c)p(c)}{p(x)}$$

- P(c|x) represents the updated probability of a certain class (c, necessary) given a specific indicator (x, feature).
- P(c) - Initial probability of the class.
- P(x|c) denotes the probability of the feature (x) given the class.
- P(x) - Initial probability of the indicator.

2. Logistic Regression: Applied when the outcome variable has a limited number of outcomes, logistic regression is suitable for categorical response variables. It requires predictor variables to have low correlations and estimates log odds of an event, making it ideal for binary classification [1]. Multiple linear regression function defined as:

$$l = \log\left(\frac{p}{1-p}\right) = \beta\theta + \beta 1\chi 1 + \beta 2\chi 2 + \cdots + \beta n\chi n$$

1−p is the probability of the false class. Logistic regression is most appropriate for binary classification problems.

3. **Random Forest**: This algorithm leverages multiple decision trees to improve performance. By introducing randomness into the dataset through methods like bagging and feature subset selection, Random Forests enhance predictive accuracy [8].

4. Decision Tree: A powerful tool for classification and prediction tasks, decision trees adopt a flowchart-like structure. Each node represents a test on a feature, with branches indicating outcomes. The selection of nodes is based on class-feature correlation, evaluated using class entropy and information gain [10].

Model Performance Analysis

The loan prediction model is assessed using various metrics: precision, recall, F1-score, and accuracy from the Classification Report. Precision measures positive prediction reliability, while recall assesses correct positive identification. F1-score combines both for balanced evaluation. The ROC-AUC score evaluates performance across thresholds. These metrics inform implementation and optimisation decisions.

Classification Report

The optimised models' classification reports are displayed below, showing the analysis outcomes. Random Forest and Logistic Regression performed best in terms of accuracy compared to other methods. Additionally, GNB and KNN outcomes are presented, with KNN demonstrating the lowest accuracy.

Logistic Regression Classification Report:

```
Accuracy: 0.962
Precision: 0.935064935064935
Recall: 0.6857142857142857
F1 Score: 0.7912087912087912
Logistic Regression Classification Report before optimisation:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98       895
           1       0.94      0.69      0.79       105

    accuracy                           0.96      1000
   macro avg       0.95      0.84      0.89      1000
weighted avg       0.96      0.96      0.96      1000
```

Fig 6: Classification Report of logistic Regression model

```
Accuracy of best model: 0.967
Precision of best model: 0.9186046511627907
Recall of best model: 0.7523809523809524
F1 of best model: 0.8272251308900523
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       895
           1       0.92      0.75      0.83       105

    accuracy                           0.97      1000
   macro avg       0.95      0.87      0.90      1000
weighted avg       0.97      0.97      0.97      1000
```

Fig 9: Classification Report for Random Forest.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       895
           1       0.98      0.92      0.95       105

    accuracy                           0.99      1000
   macro avg       0.99      0.96      0.97      1000
weighted avg       0.99      0.99      0.99      1000
```

Fig 7: Classification Report for GNB Model

```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.93      0.94       895
           1       0.52      0.62      0.57       105

    accuracy                           0.90      1000
   macro avg       0.74      0.78      0.75      1000
weighted avg       0.91      0.90      0.90      1000
```

Fig 8: Classification Report for KNN.

```
Best K value: 1
Accuracy for best K value: 0.964
Precision for best K value: 0.896551724137931
Recall for best K value: 0.7428571428571429
F1 Score for best K value: 0.8125
KNN Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       895
           1       0.90      0.74      0.81       105

    accuracy                           0.96      1000
   macro avg       0.93      0.87      0.90      1000
weighted avg       0.96      0.96      0.96      1000
```

Results and Findings

From the performance metrics below

Model	Accuracy %	Precision %	Recall %	F1-Score
Logistic Regression	0.96%	93.51%	68.57%	0.79
Random Forest	0.99%	98.02%	94.29%	0.96
KNN	0.96%	89.66%	74.29%	0.81
Naive Bayes	0.90%	54.17%	53.97%	0.57

Fig 9: Performance Metric Table.

Logistic Regression achieved a commendable accuracy of 96%, with high precision (93.51%). However, its recall rate (68.57%) indicates that it may have missed identifying some positive cases.

Random Forest outperformed other models with the highest accuracy (99%), precision (98.02%), and recall (94.29%). Its F1-score of 0.96 signifies excellent overall performance.

KNN exhibited a respectable accuracy of 96%, though slightly lower precision (89.66%) and recall (74.29%) compared to Random Forest.

Naive Bayes had the lowest accuracy (90%) and precision (54.17%). While its recall (53.97%) was marginally higher, the F1-score (0.57) indicates a need for improvement.

Conclusion: Based on the analysis, Random Forest emerges as the top-performing model, followed by Logistic Regression and KNN. Naive Bayes, although with lower accuracy and precision, still offers insights but may require further optimisation for better performance. The model's choice should align with the requirements and trade-offs between accuracy, precision, and recall in the classification task.

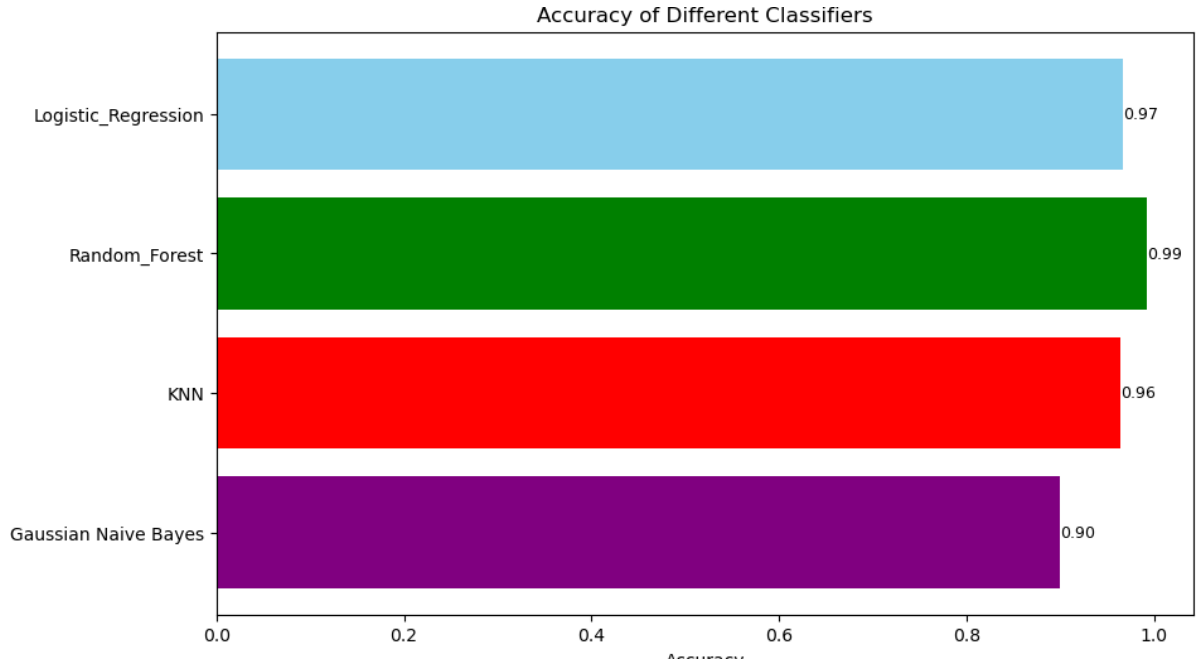

Fig 10: Performance Accuracy.

Precision: Precision reflects how accurately the model identifies individuals predicted to default on their loans among all those predicted to default. For instance, a precision of 93.51% in Logistic Regression means that 93.51% of those predicted to default did.

Recall: Recall shows the model's ability to correctly identify all individuals who truly defaulted on their loans out of all those who did. For example, a recall of 68.57% in Logistic Regression implies that the model captured 68.57% of all true loan defaults.

F1-score: The F1-score, being a balance of precision and recall, provides an overall assessment of the model's performance. It's crucial in loan prediction, where misclassifications carry significant consequences. For instance, an F1-score of 0.79 in Logistic Regression indicates a balanced performance considering both precision and recall.

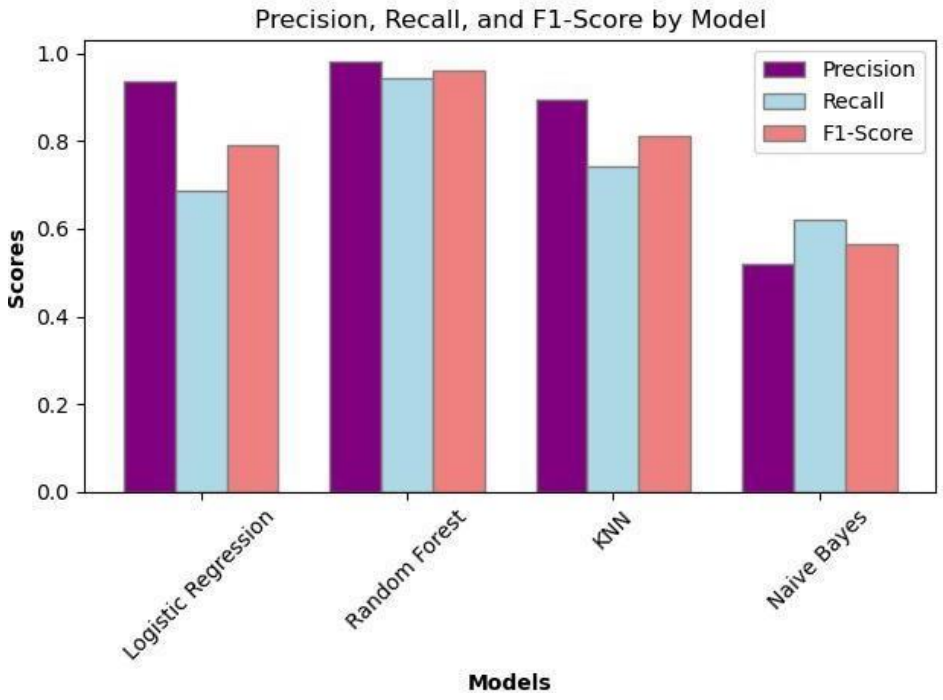

Fig 11: Precision, Recall and F1-score

The classification models' performance was evaluated using ROC-AUC, revealing that both Random Forest and KNN achieved a notable score of 97%. This indicates their strong ability to distinguish between positive and negative cases effectively. Logistic Regression, despite its linear assumptions, demonstrated proficient ranking of positive cases with a ROC-AUC of 91%. Naive Bayes, while competitive in other metrics, exhibited potentially weaker discriminatory power (ROC-AUC of 86%) in identifying positive versus negative events. These observations provide insights into each model's discriminatory capabilities, with Random Forest and KNN excelling, Logistic Regression performing well, and Naive Bayes showing slightly lower discriminatory power.

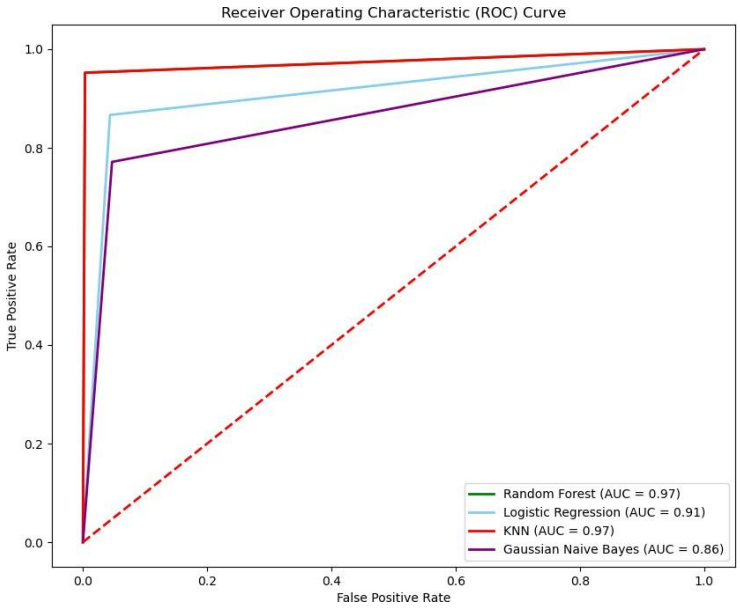

Fig 12: ROC Curve

Recommendation and Conclusion

Based on the evaluation of loan prediction models, it's clear that Random Forest emerges as the top performer, closely followed by Logistic Regression and KNN. Random Forest achieved the highest scores across various metrics, indicating its superior ability to identify loan defaults accurately. While Logistic Regression showed commendable accuracy and precision, it struggled with recall, potentially missing some positive cases. KNN performed reasonably well with good accuracy, but slightly lower precision and recall compared to Random Forest.

Despite its lower accuracy and precision, Naive Bayes still offers valuable insights. However, its lower F1-score and weaker discriminatory power suggest room for improvement, indicating a need for further optimisation.

In summary, for loan prediction tasks, it's advisable to prioritise the use of Random Forest due to its consistently strong performance. Logistic Regression can be considered as a viable alternative, especially when interpretability is crucial, despite its lower recall. KNN may serve as a secondary option, especially in scenarios where model interpretability is less important. Continuous monitoring and refinement of the models are essential to ensure optimal performance, considering the specific requirements and trade-offs between accuracy, precision, and recall in loan prediction.

References:

[1] Vaidya, A., 2017, July. Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In 2017 8th international conference on computing, communication, and networking technologies (ICCCNT) (pp. 1-6). IEEE.

[2] Pandimurugan, parvathi, M. and jenila, A., 2011, November. A survey of software testing in refactoring based software models. In International Conference on Nanoscience, Engineering and Technology (ICONSET 2011) (pp. 571-573). IEEE.

[3] Hussain, A., Raja, M., Vellaisamy, P., Krishnan, S. and Rajendran, L., 2021. Enhanced framework for ensemble effort estimation by using recursive-based classification. IET Software, 15(3), pp.230-238.

[4] Pandimurugan, V., Jain, A. and Sinha, Y., 2020, December. IoT based face recognition for smart applications using machine learning. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 1263-1266). IEEE.

[5] Pandimurugan, V., Jayaprakash, R., Rajashekar, V. and Singh, Y., 2019, April. Smart Buspass System Using Android. In *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)* (pp. 1-4). IEEE.

[6] Bayraktar, M., Aktaş, M.S., Kalıpsız, O., Susuz, O. and Bayracı, S., 2018, May. Credit risk analysis with classification Restricted Boltzmann Machine. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[7] Hussain, A., Raja, M., Vellaisamy, P., Krishnan, S. and Rajendran, L., 2021. Enhanced framework for ensemble effort estimation by using recursive-based classification. *IET Software*, *15*(3), pp.230-238.

[8] Marsland, S., 2011. Machine learning: an algorithmic perspective. Chapman and Hall/CRC.

[9] Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. eds., 2013. Machine learning: An artificial intelligence approach. Springer Science & Business Media.

[10] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R. and Suraj, R.S., 2021, January. Heart disease prediction using hybrid machine learning model. In 2021 6th international conference on inventive computation technologies (ICICT) (pp. 1329-1333). IEEE.

[11] Oza, K.S. and Kamat, R.K., 2021. Intelligent Prediction of Loan Eligibility using Soft Computing Towards Digital Banking Sector. SPAST Abstracts, 1(01).

[12] Kavitha, M., Srinivas, P.V.V.S., Kalyampudi, P.L. and Srinivasulu, S., 2021, September. Machine learning techniques for anomaly detection in smart healthcare. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1350-1356). IEEE.

[13] Gupta, A., Pant, V., Kumar, S. and Bansal, P.K., 2020, December. Bank loan prediction system using machine learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE.

[14] Arutjothi, G. and Senthamarai, C., 2017. Comparison of feature selection methods for credit risk assessment. International Journal of Computer Science, 5(5), p.492.

[15] Henley, W., and Hand, D.J., 1996. Ak-nearest-neighbour classifier for assessing consumer credit risk. Journal of the Royal Statistical Society Series D: The Statistician, 45(1), pp.77-95.

[16] Sudhamathy, G. and Venkateswaran, C.J., 2016, October. Analytics use R for predicting credit defaulters. In 2016 IEEE International Conference on Advances in Computer Applications (ICACA) (pp. 66-71). IEEE.