Prediction of Loan Approval in Banks using Machine Learning Approach

14.03.2025

—

Mid-Term Project

Group 1: Carlos, Maria, Suguru

Summary

The growing demand for loans has created the need for faster and more reliable ways to evaluate applications while minimizing risks. This project focuses on machine learning models to enhance the accuracy and efficiency of loan approval predictions. By replicating the method from a research paper and introducing significant enhancements, our primary goals included:

- Data preprocessing and feature engineering.
- Training and optimizing machine learning models.
- Introducing alternative classification models for comparison.

The results demonstrated improved predictive accuracy, benefiting both applicants and financial institutions.

Motivation

This project addresses the growing complexity of loan approval processes as banks face increasing demand and the need to minimize risks.

We selected the referenced paper because it effectively applies machine learning models to tackle these challenges, providing a solid foundation for further improvements. By replicating and enhancing its methods, we aimed to refine predictive accuracy, explore additional algorithms, and contribute to a more reliable model.

Research paper details

With advancements in technology and the expansion of the banking industry, the demand for loans has increased significantly. This has made the loan approval process more complex, as banks must carefully assess each applicant's eligibility while minimizing the risks. Evaluating numerous applications thoroughly and accurately presents a major challenge for financial institutions.

This research proposes the use of machine learning (ML) models combined with ensemble learning techniques to predict the possibility of approving individual loan applications. By improving the accuracy of identifying qualified applicants, this approach addresses the challenges in the loan approval process. Moreover, it reduces processing time significantly, benefiting both loan applicants and bank employees.

To predict loan approval status, they evaluated four machine learning algorithms: Random Forest, Naive Bayes, Decision Tree, and KNN. Among these, the Naive Bayes algorithm achieved the highest accuracy of 83.73%, proving to be the most effective in this study.

Link:
https://www.researchgate.net/publication/372909313_Prediction_of_Loan_Approval_in_Banks_using_Machine_Learning_Approach

Dataset details

Loan Prediction Problem Dataset

Contains information about four different loan applicants. Each entry includes details such as Loan_ID, Gender, Marital status, Dependents, Education, Self-Employment status, Applicant Income, Co Applicant Income, Loan Amount, Loan Amount Term, Credit History, and Property Area.

Link:
https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset/data

Data head

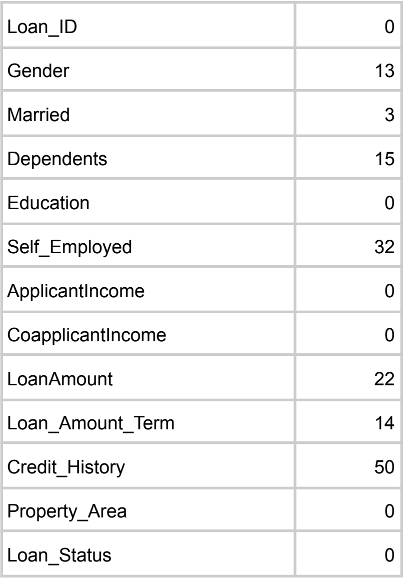
Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0	NaN	360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y

Data Shape

```
(614, 13)
```

Steps reproduced from the paper

- Data Preprocessing
    - Find and fill missing values



Loan_ID	0
Gender	13
Married	3
Dependents	15
Education	0
Self_Employed	32
ApplicantIncome	0
CoapplicantIncome	0
LoanAmount	22
Loan_Amount_Term	14
Credit_History	50
Property_Area	0
Loan_Status	0

    - Fill the missing values with `mode()`
- Feature Engineering
    - Performed Label encoding for categorical columns

      *In the paper, they performed label encoding for all columns, but we here applied it in a more proper way we believe.
    - Split the data into train and test set 80/20
    - Performed Standard Scaler
- Model implementation

  Accuracy of different Algorithms

Algorithms	**Reproduction**	Research Paper
Random Forest	77.23%	77.23%
Naive Bayes	Skipped	83.73%
Decision Tree	70.73%	63.41%
KNN Algorithm	77.23%	77.23%

Contributions

To make the loan approval prediction more accurate and reliable, we introduced some key enhancements:

- **Smarter Data Preprocessing**: Proper data cleaning and preprocessing were conducted. Instead of filling all missing values with the mode, a more in-depth analysis was performed, ensuring a more accurate imputation.
- **Better Feature Engineering**: Applied robust scaling techniques to minimize the influence of extreme values, helping models make fairer predictions.
- **Addressing Class Imbalance**: By using SMOTE (Synthetic Minority Over-sampling Technique), we balanced the dataset, allowing the models to learn from both approved and rejected loans more effectively.
- **Optimized Model Performance**: We fine-tuned hyperparameters using GridSearch, ensuring each model operates at its best.
- **Expanded Model Selection**: Different machine learning models were tested to evaluate whether they outperformed those used in the paper. The models applied included Logistic Regression, Support Vector Machines (SVM), and XGBoost. After comparing the results, it was observed that all models demonstrated good performance; however, there was still room for improvement. To address this, SMOTE was applied to handle class imbalance, followed by a second iteration to assess any performance enhancements.

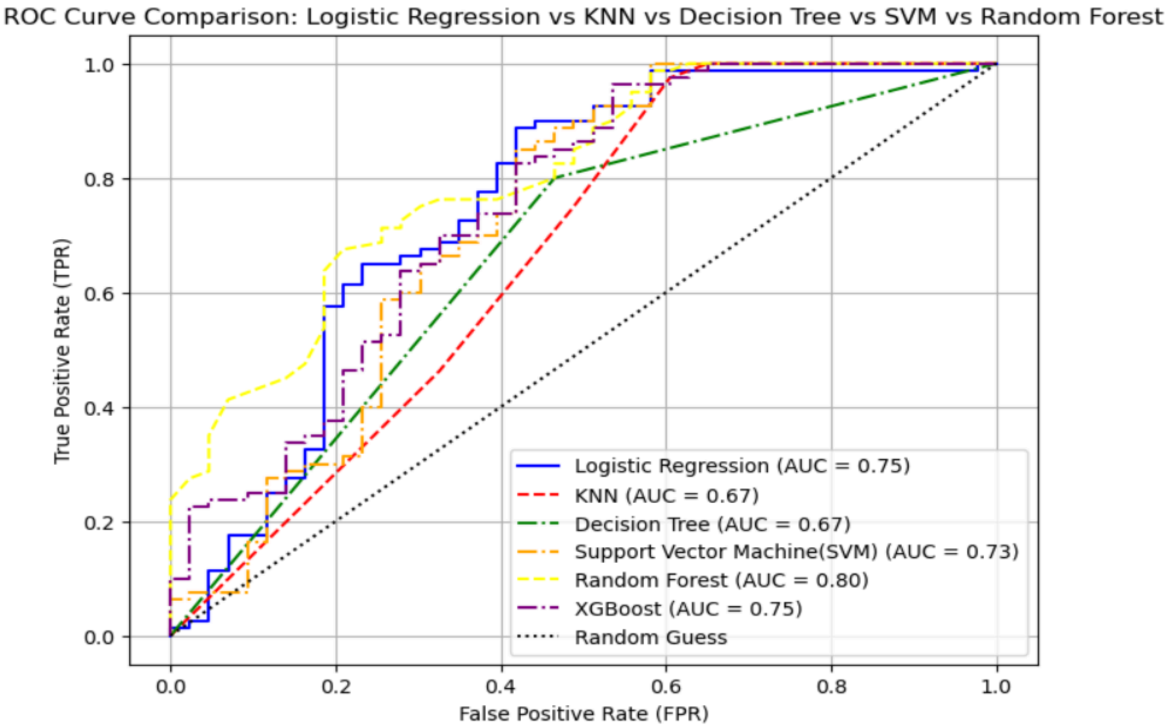
These improvements not only enhanced prediction accuracy but also made the system more practical and insightful for real-world banking decisions.

Significant improvements

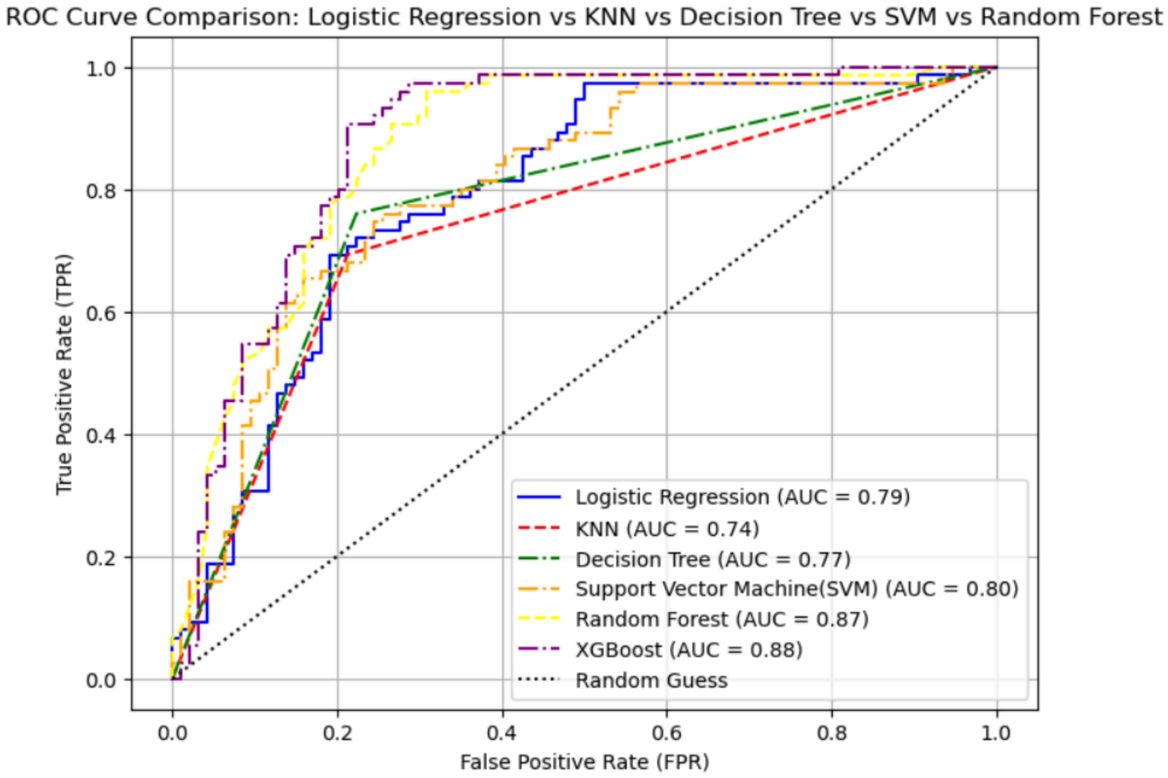
1. Additional Models Version:

The paper originally utilized Random Forest, Decision Tree, and K-Nearest Neighbors (KNN). To compare performance, additional models—Logistic Regression, SVM, and XGBoost—were implemented. After the first iteration, SMOTE was applied to address class imbalance, and a second iteration was conducted to evaluate any improvements in performance.

AUC-ROC Curve before SMOTE



AUC-ROC Curve after SMOTE



Results:

	Research Paper		Research Paper + SMOTE	
Model	**Accuracy score**	**AUC score**	**Accuracy score**	**AUC score**
Random Forest	0.77	0.77	0.79	0.87
Decision Tree	0.71	0.66	0.76	0.76
KNN	0.77	0.67	0.76	0.85
Logistic Regression	0.79	0.75	0.69	0.79
SVM	0.79	0.73	0.70	0.73
XGBoost	0.78	0.75	0.83	0.88

Insights:

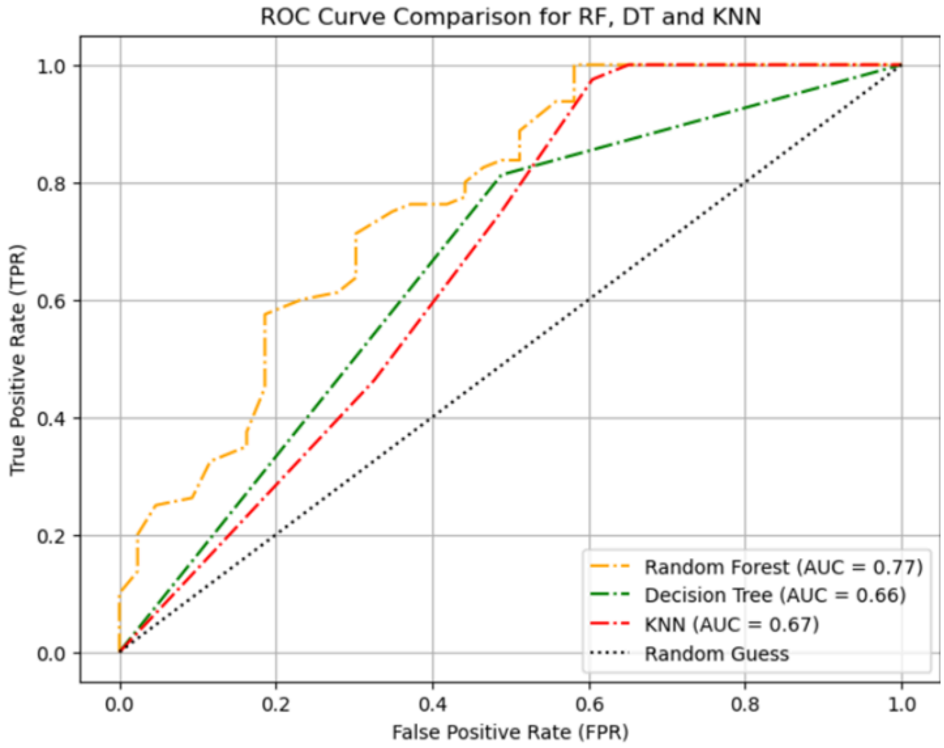
After applying SMOTE, most models showed an improvement, with XGBoost achieving the best performance among them.

2. Upgraded Version:

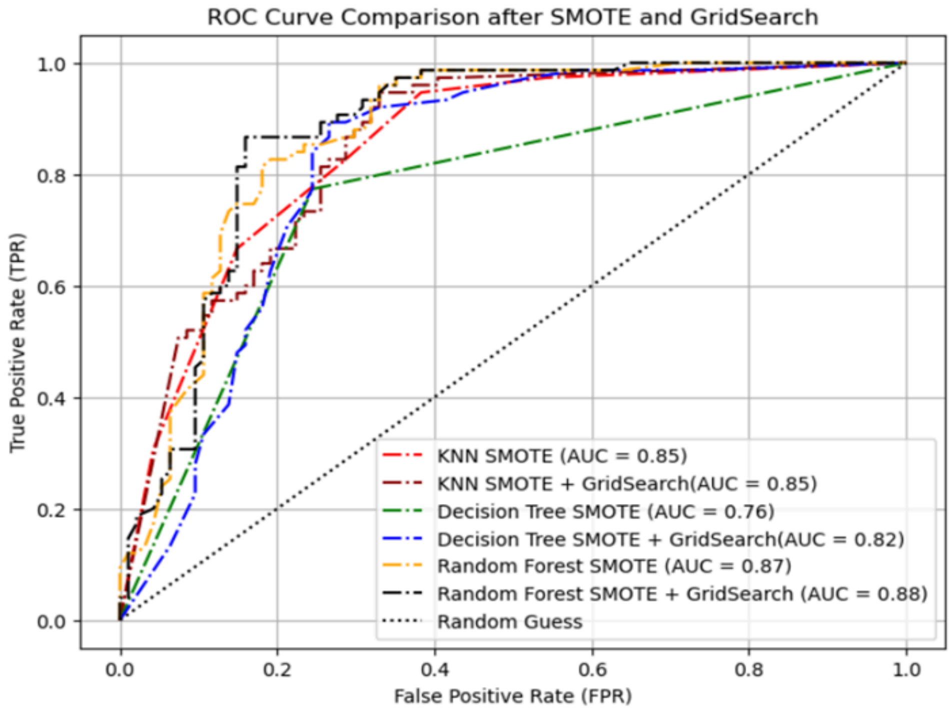
We improved loan approval predictions with better data preprocessing, feature scaling, class balancing (SMOTE), and model optimization (GridSearch) for more accurate and fair results. These upgrades resulted in significant improvements, as detailed in the following insights.

AUC ROC Curves:

AUC-ROC Curve Research Paper



AUC-ROC Curve Upgraded version with SMOTE and GridSearch



Results:

	Research Paper		Upgraded version + SMOTE		Upgraded version + SMOTE + Gridsearch	
Model	**Accuracy score**	**AUC score**	**Accuracy score**	**AUC score**	**Accuracy score**	**AUC score**
Random Forest	0.77	0.77	0.79	0.87	0.80	0.88
Decision Tree	0.71	0.66	0.76	0.76	0.79	0.82
KNN	0.77	0.67	0.76	0.85	0.77	0.85

Insights

Based on the provided model performance data, here are some key insights:

- The most robust model for loan approval prediction was **Random Forest**, which clearly outperforms the other models in both accuracy and AUC score, especially when SMOTE and GridSearch are applied.
- This makes **SMOTE** proven effective in handling class imbalance, boosting performance for all models, particularly in terms of **AUC score**.
- **GridSearch** increments the improvement, specially on models like Random Forest, which is a more complex model and benefits from hyperparameter optimization.
- **Decision Tree** and **KNN** show improvements with SMOTE and GridSearch.

Challenges

- Choosing an appropriate completion method for missing values was a challenge. They filled all missing values with `mode()` but we would say there should be a better way. We took a look into each record and handled missing values properly. For example, if the "Married" column has no value while the "CoapplicantIncome" value exists, that person should be considered as married. So we filled the "Married" with "Yes."
- In the Upgraded Version, we applied a Grid Search to optimize hyper parameters but it was challenging to find better ones. We can try exploring more combinations.

Conclusion and Future Scope

This project successfully implemented and improved upon machine learning techniques for loan approval prediction. By replicating the research paper's methodology and introducing enhancements, we achieved improved accuracy and efficiency. However, there remains room for further advancements:

- **Expanding the dataset:** Incorporating larger and more diverse datasets could enhance model robustness.
- **Implementation of Real-time Processing for Real-world assessment:** Real-time processing is crucial for practical loan approval systems. Introducing models optimized for real-time inference and deploying them on cloud platforms should be considered.

Through these improvements, we aim to further refine machine learning-based loan approval systems and enhance their real-world applicability.