

**LOAN STATUS PREDICTION**

Submitted By:

**Romel Jude Melville (STB03003)**

Data Science Trainee: Batch 1

Under Esteemed Guidance of:

**Ms. Urooj Khan**

Submitted to Scifor Technologies as the final project during the training

**TABLE OF CONTENTS**

* Acknowledgement
* Abstract
* Introduction
* Motivation
* Problem Statement
* Objective
* Proposed Method
* Methodology
* Understanding the data
* Feature Scaling
* Train Test Split and K-Fold cross validation
* Hyperparameter Tuning
* Output
* Conclusion
* References

**ACKNOWLEDGEMENT**

It gives me a great sense of pleasure to present the report of the Data science project undertaken by me during the training period. First and foremost, I extend my heartfelt thanks to Ms. Urooj Khan my supervisor, for their unwavering support, invaluable guidance, and constructive feedback throughout the entire process. Their expertise and encouragement have been instrumental in shaping the content and structure of this report.

**Abstract**

This project utilizes machine learning techniques to predict the status of loan applications, streamlining the decision-making process for financial institutions. By analyzing historical loan data and employing various models, including logistic regression and decision trees, the system aims to accurately identify potential risks and improve overall efficiency. The model's performance is assessed through key metrics, highlighting its potential to enhance the loan approval process and minimize defaults.

**INTRODUCTION**

Getting a loan can be a game-changer for individuals and businesses, but traditional ways of deciding who gets approved can be slow and not always accurate. This project is about using smart computer programs, known as machine learning, to predict whether a loan application will be approved or not.

Traditionally, people at banks manually review your application, but it's time-consuming, and sometimes they might miss important details. We're turning to machine learning, a type of smart computer technology, to analyze heaps of data from past loan applications. This helps our computer learn patterns and make predictions about whether a new application should be approved or not.

We're trying different computer models to see which one is the quickest and most accurate. We're using metrics like accuracy, which tells us how often our computer is right, and other fancy metrics to ensure it's making good decisions. If successful, this approach could help banks and lenders make quicker and better decisions about who gets a loan, making the whole process smoother and more efficient.

**Motivation**

Loan prediction is a very common real-life problem that every finance company faces in their lending operations. If the loan approval process is automated, it can save a lot of man hours and improve the speed of service to the customers. The increase in customer satisfaction and savings in operational costs are significant. However, the benefits can only be reaped if the bank has a robust model to accurately predict which customer's loan it should approve and which to reject, in order to minimize the risk of loan default. Our goal is to leverage technology to make loan approvals efficient, accurate, and accessible to a broader spectrum of applicants.

**Problem Statement**

Banks, Housing Finance Companies and some NBFC deal in various types of loans like housing loan, personal loan, business loan etc. in all over the part of countries. These companies have existence in Rural, Semi Urban and Urban areas. After applying loan by customer these companies validate the eligibility of customers to get the loan or not. This project provides a solution to automate this process by employing machine learning algorithm. So, the customer will fill an online loan application form. This form consists details like Sex, Marital Status, Qualification, Details of Dependents, Annual Income, Amount of Loan, Credit History of Applicant and others.

**Objective**

This project aims to streamline the loan approval process by leveraging machine learning. It implements a system that automates loan approval, reducing manual workload and expediting decision-making. Training machine learning models to predict loan statuses with high accuracy, ensuring dependable outcomes. Improves the efficiency of loan approvals, providing timely responses to applicants and reducing processing times. Promotes a fairer system by mitigating discrimination and biases, ensuring equal opportunities for approval across diverse applicant profiles. These objectives collectively aim to modernize and optimize the loan approval process, making it more efficient, accurate, and equitable.

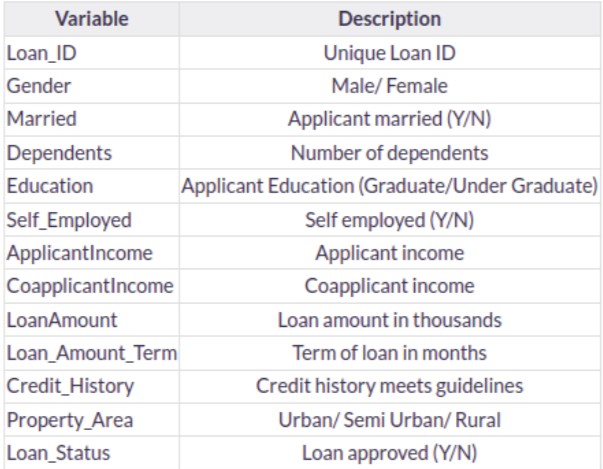
**Proposed Method**

The proposed method will be cleaning and preprocessing the loan dataset, addressing missing values and outliers. Applying feature engineering to extract relevant information from various data points. Utilizing a mix of models, including logistic regression and decision trees. Training the model and optimizing parameters to achieve the best predictive performance. Implementing cross-validation to ensure the model's effectiveness on new data. Evaluating the model using standard metrics like accuracy, precision. This proposed method aims to develop a reliable and accurate model, contributing to an efficient loan approval process.

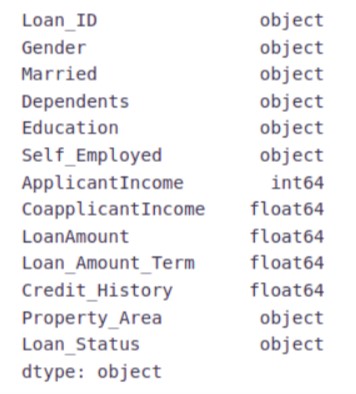
**Methodology**

The methodology involves collecting a comprehensive dataset of historical loan records, followed by rigorous data preprocessing and feature engineering to optimize model performance. An ensemble learning approach, incorporating models like logistic regression and decision trees, is employed to construct a robust predictive model. The training process includes hyperparameter tuning and cross-validation to enhance the model's accuracy.

**Understanding the Data**

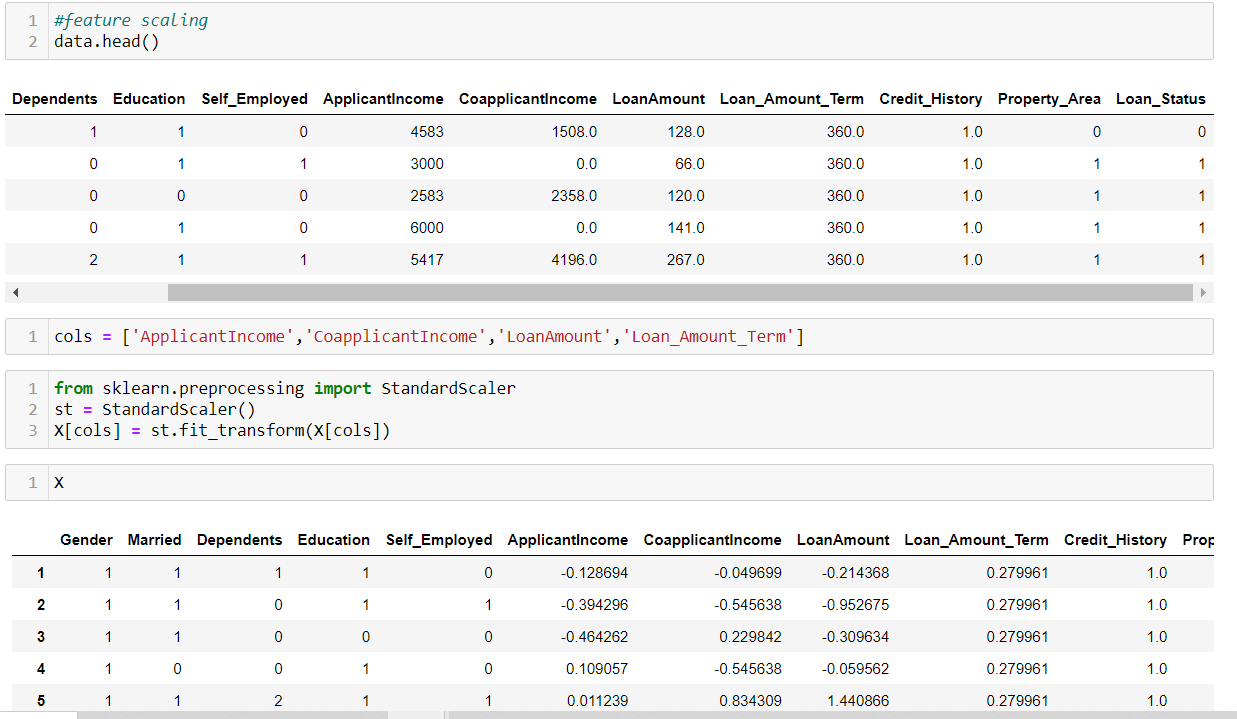
****

**Loan\_Status** is the target variable.



**Feature Scaling**

Feature scaling is essential for ensuring that various input features, such as income and loan amount, are on a comparable scale. This is crucial because certain machine learning algorithms, like support vector machines or k-nearest neighbors, calculate distances between data points or coefficients for features, and these calculations can be influenced by the scale of the input features. Standardizing or normalizing the features through scaling helps prevent disparities in magnitudes from skewing the predictive model. For instance, if one feature like income is on the order of thousands and another feature like age is on the order of tens, without feature scaling, the income feature could dominate the prediction process. By applying feature scaling to ensure uniformity in scale, the loan status prediction model can capture better relationships between features, resulting in a more accurate and reliable outcome.

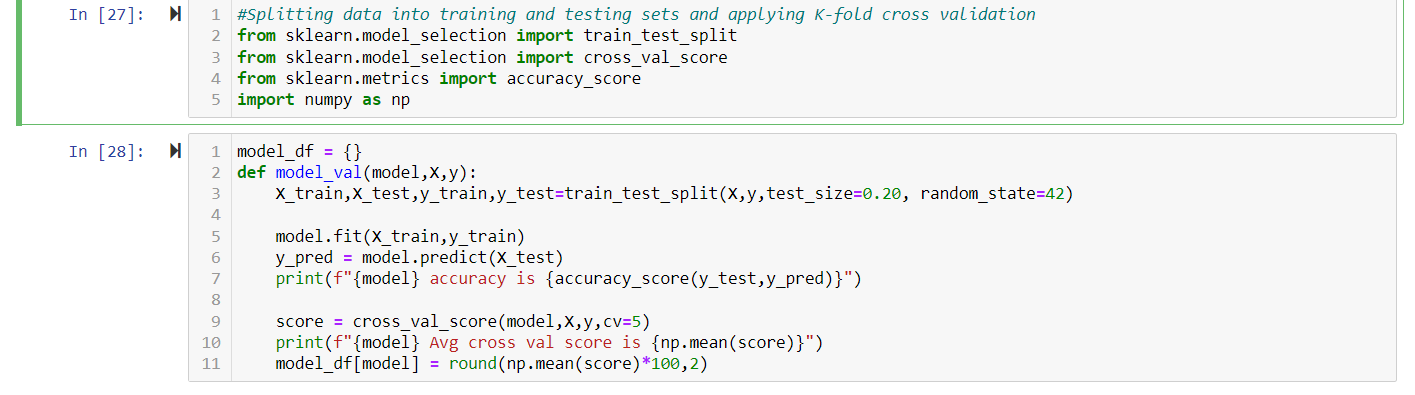


**Train Test Split**

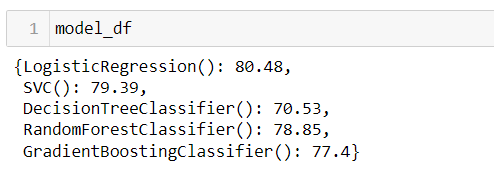
Divide the dataset into two subsets – a larger training set used to train the machine learning model and a smaller testing set used to assess its performance on new, unseen loan applications. For instance, allocate 80% of the data for training and reserve 20% for testing. This ensures that the model learns from a substantial portion of historical loan data and is evaluated on a representative sample.

**K-Fold Cross Validation**

Mitigate the impact of variability in model performance due to a single train-test split. It involves dividing the dataset into ‘k’ folds, training the model ‘k’ times, and evaluating its performance on different subsets of the data. If choosing, for example, 5-fold cross-validation, the dataset is split into five parts, and the model is trained and tested five times, with each part serving as the testing set exactly once.



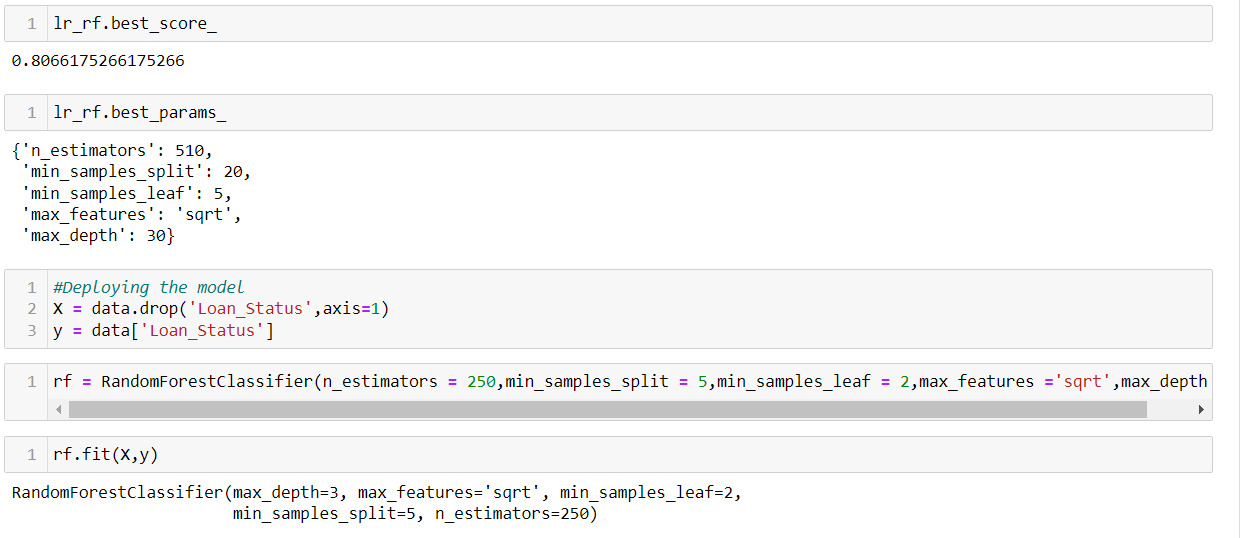
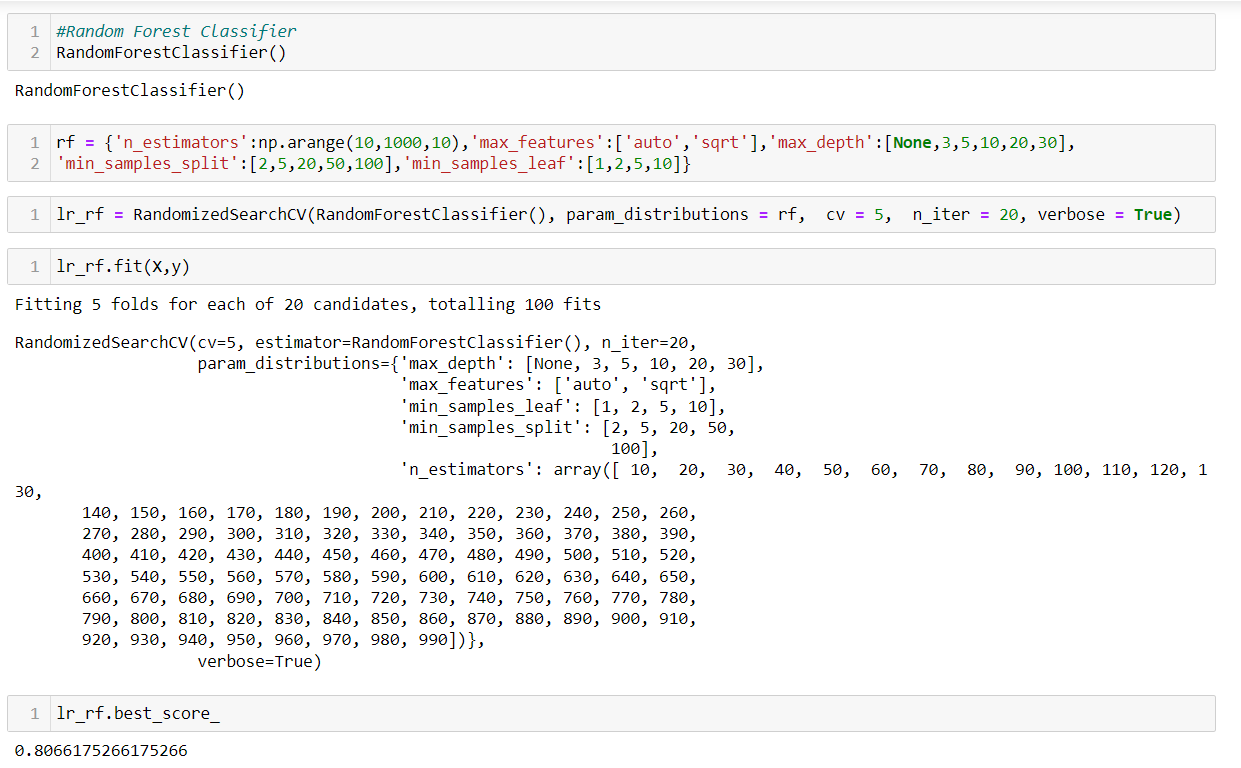
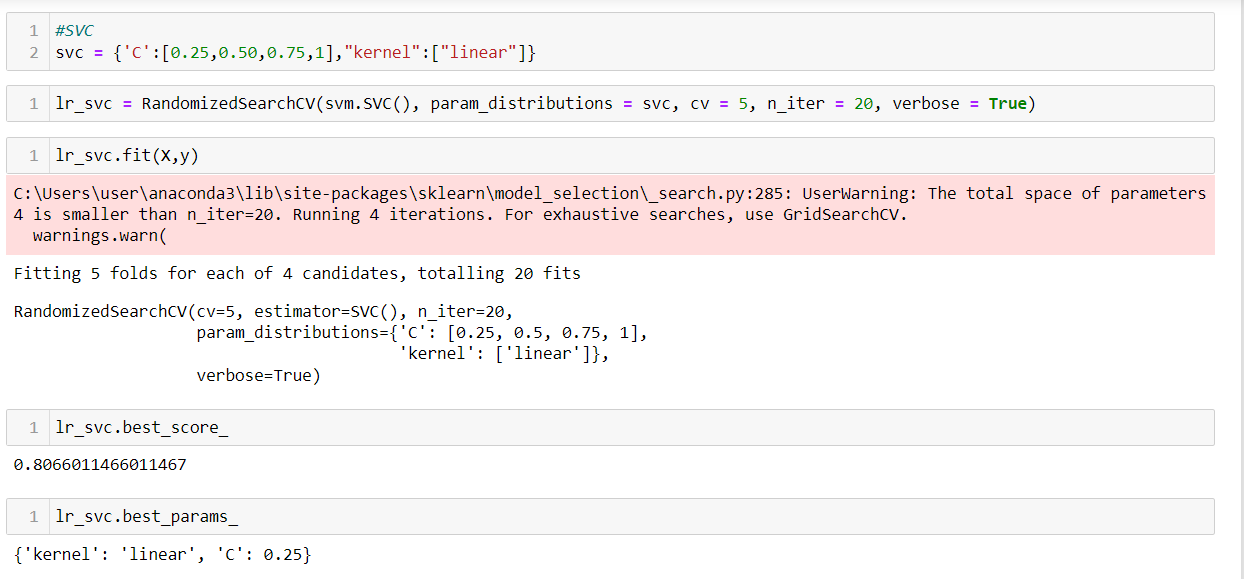
**Accuracy Scores before hyperparameter tuning**

****

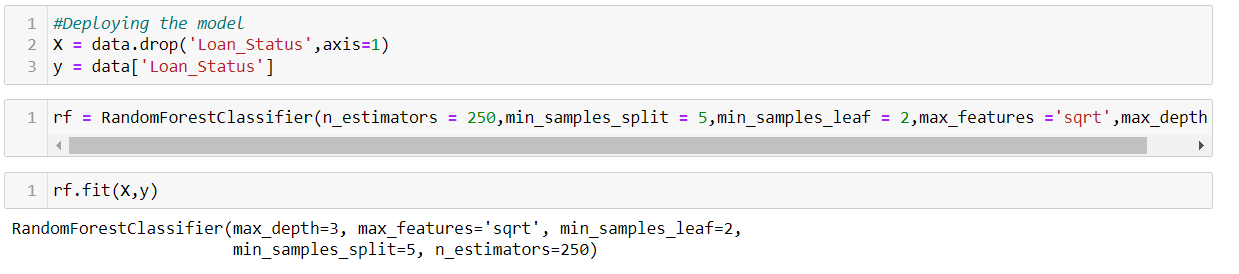
**Hyperparameter Tuning**

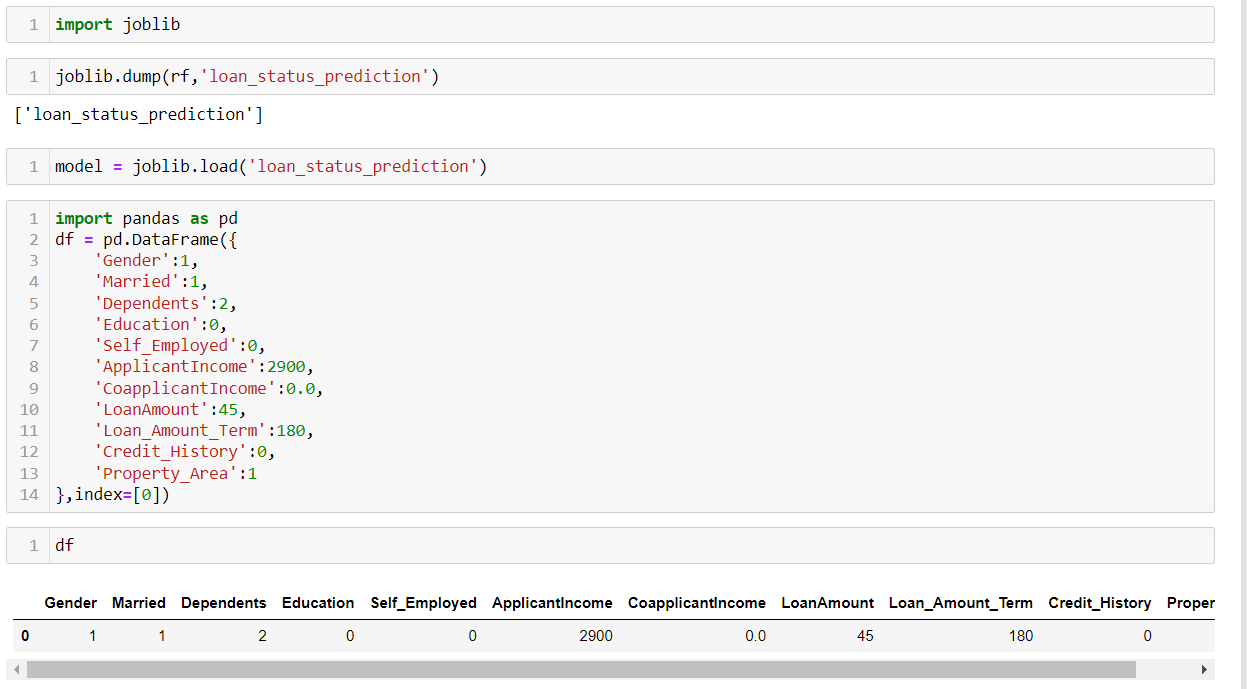
Hyperparameter tuning involves optimizing configuration settings for machine learning models before training. These settings, known as hyperparameters, significantly influence a model's performance. Techniques such as Grid Search or Random Search are employed to systematically explore various combinations of hyperparameter values. Cross-validation ensures that the model's performance is robust across different subsets of the data. The selection of appropriate evaluation metrics, such as accuracy or precision, is crucial in guiding the tuning process. Hyperparameter tuning is an iterative process, refining the search based on initial results.

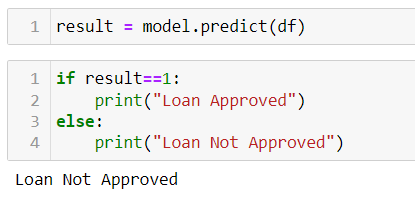
**Accuracy scores after hyperparameter tuning**



**Output**

****

****

****

**Conclusion**

In conclusion, loan status prediction through advanced machine learning models has emerged as a transformative tool with diverse applications across the financial landscape and beyond. The integration of predictive analytics not only enhances the efficiency of loan approval processes but also plays a pivotal role in risk management, fraud detection, and compliance adherence for financial institutions. So here, it can be concluded with confidence that the Random Forest Classifier model is extremely efficient and gives a better result when compared to other models. It works correctly and fulfills all requirements of bankers. This system properly and accurately calculates the result. It predicts the approval or rejection of loan to loan applicant or customer very accurately.

**References**

* <https://www.geeksforgeeks.org/loan-approval-prediction-using-machine-learning/>
* <https://www.irjet.net/archives/V8/i4/IRJET-V8I4785.pdf>
* <https://www.analyticsvidhya.com/blog/2022/02/loan-approval-prediction-machine-learning/>

**THANK YOU**