BUSINESS REQUIREMENT DOCUMENT

Loan Approval Prediction using Machine Learning

Work By:

Rimi Mondal

Introduction

In today's world, financial institutions play a crucial role in facilitating economic growth by providing funds to businesses and individuals. Loans are one of the primary financial products offered by banks and other financial organizations. However, with the increase in the number of loan applications, it has become challenging for financial institutions to assess the creditworthiness of borrowers accurately.

To address this issue, we have developed a machine learning model that predicts an individual's creditworthiness based on various demographic and financial factors. The goal of this business case document is to present the results of our data analysis and demonstrate the value of our machine learning model to financial institutions.

Our analysis aims to identify the key factors that contribute to creditworthiness and create a model that can accurately predict whether a loan application will be approved. With our model, financial institutions can make more informed decisions concerning loan approval, reduce the risk of default, and increase profitability.

In this document, we will discuss the data sources, data cleaning, and preprocessing techniques used to prepare the data for analysis. We will also present the findings of our exploratory data analysis and modeling techniques used to create the final machine learning model. Finally, we will discuss the potential impact of our model on financial institutions and the broader community.

Purpose and Use Case

The purpose of this project is to predict an individual's creditworthiness based on various demographic and economic factors. The use case of this project is to provide financial organizations with a methodology to make more informed decisions concerning loan approval, using machine learning models to forecast loan eligibility. The project aims to identify the key factors that contribute to creditworthiness and create a model that can accurately predict whether a loan application will be approved.

Our Audiences

The audiences for this project could be professionals in the finance and banking industry who are interested in using machine learning to improve loan approval processes, as well as data scientists and researchers interested in exploring the application of machine learning algorithms on financial data. Students and educators in the field of data science and finance may also find this project relevant for learning and teaching purposes. Additionally, individuals who are interested in personal finance and loan eligibility may find the insights from this project informative.

Data Source and Description

The dataset used in this project is called "Loan Prediction" and was uploaded to Kaggle by user Vikas U Kani. The dataset includes information on 614 loan applicants, each represented by a row in the dataset.

The 12 independent variables include: Gender, Married, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area, and Total_Income (a newly engineered feature created by combining ApplicantIncome and CoapplicantIncome).

The dependent variable is Loan_Status, which indicates whether a loan application was approved (labeled as 'Y') or not approved (labeled as 'N').

The data cleaning and preprocessing steps involved identifying and handling missing values in the dataset. The following variables had missing values: Gender, Married, Dependents, Self-Employed, LoanAmount, Loan_Amount_Term, and Credit_History. Missing categorical values were imputed using the mode, and missing numerical values were imputed using the median.

To make the data suitable for analysis, categorical variables were converted to numerical data using one-hot encoding, and numerical data was standardized by normalizing to have a mean of 0 and standard deviation of 1.

Workflow

Data collection: Gather the dataset from a reliable source, such as Kaggle.

Data preprocessing: Preprocess the data by removing duplicates and missing values, converting categorical data into numerical data, and standardizing numerical data.

Data analysis: Conduct exploratory data analysis to identify relationships between the variables and the loan status. Visualize the data using plots and charts to gain insights into the data.

Feature engineering: Create new features from existing data that may be more predictive of the loan status.

Model selection: Select an appropriate model for the task of predicting loan eligibility. Consider using classification models like logistic regression, decision trees, or random forests.

Model training: Train the selected model on the preprocessed data using appropriate training techniques.

Model evaluation: Evaluate the performance of the model using evaluation metrics such as accuracy, precision, recall, and F1 score.

Hyperparameter tuning: Tune the hyperparameters of the model using techniques like GridSearchCV to improve its performance.

Model deployment: Deploy the trained and optimized model in a production environment where it can make accurate loan eligibility predictions.

Monitoring and maintenance: Monitor the model's performance in the production environment, retrain it periodically, and update it as necessary to maintain its accuracy and relevance.

Detailed Steps

1. The project started by importing the necessary libraries for data analysis and machine learning, including pandas, numpy, scikit-learn, and matplotlib.
2. Null values, duplicates, and NAN values were dropped from the dataset to ensure that the data was clean and accurate.
3. The data was transposed to make it easier to work with and analyze.
4. NAN values were checked for and replaced if necessary using methods like mean or median.
5. Inconsistent values were checked by looking at unique values in each column to ensure they were within the expected range.
6. Outliers were cleaned from the test data using methods like z-score to identify and remove data points that were significantly different from the rest of the data.
7. The z-score method was used to detect outliers in the test data, which were then removed to improve the accuracy of the model.
8. The IQR method was used to remove outliers to ensure that the data was accurate and representative of the population.
9. Outliers were cleaned from the train data using the same methods as for the test data.
10. Once the data had been cleaned, it was exported to a CSV file for further analysis.
11. A random forest algorithm was used to build a machine learning model that could predict outcomes based on the data.
12. The confusion matrix and error rate were calculated to evaluate the accuracy of the model and identify areas for improvement.

Data Cleaning and Preprocessing

In this project, data cleaning and preprocessing are critical steps to ensure the accuracy and effectiveness of the machine learning model. The following steps were taken to clean and preprocess the data:

First, the necessary libraries for data analysis and machine learning were imported, such as pandas, numpy, scikit-learn, and matplotlib.

Next, any null values, duplicates, or NAN values were dropped from the dataset to ensure that the data is clean and accurate. Transposing the data was done to make it easier to work with and analyze.

NAN values were checked and replaced if necessary using methods like mean or median. Inconsistent values were checked by looking at unique values in each column and ensuring they were within the expected range.

Outliers were cleaned from the test data using methods like z-score to identify and remove data points that were significantly different from the rest of the data. The z-score method was used to detect outliers in the test data and remove them to improve the accuracy of the model. Outliers were also removed using the IQR method to ensure that the data was accurate and representative of the population. The same cleaning methods were used for the train data.

Once the data had been cleaned, it was exported to a CSV file for further analysis. A random forest algorithm was used to build a machine learning model that could predict outcomes based on the data.

Finally, the confusion matrix and error rate were calculated to evaluate the accuracy of the model and identify areas for improvement. These steps ensured that the data was accurate and representative of the population, and the machine learning model was effective in making accurate predictions.

Exploring data analysis

During the exploration of the data analysis phase, the project team examined the various features and variables of the dataset to gain a deeper understanding of the relationships and patterns within the data. This involved identifying the distribution of the variables, examining the correlation between variables, and identifying any outliers or missing values.

The team used various tools and techniques, including histograms, scatterplots, and correlation matrices, to visualize and analyze the data. Through this process, the team was able to gain insights into the key drivers of the target variable and identify any potential biases or limitations in the data. This information was then used to inform the selection of features for the machine learning model and to improve the accuracy and reliability of the predictions.

Modeling

The modelling phase of the project involved the use of various classification techniques such as logistic regression, decision trees, and random forests. Both the training and test sets used the same dataset, and GridSearchCV was employed to adjust the hyperparameters to improve the performance of the models. Metrics such as accuracy, precision, recall, and F1 score were used to evaluate the performance of the models.

The random forest algorithm was found to have the highest F1 score, accuracy, and precision compared to the other classification techniques, with an efficiency of 74% and accuracy of 74%. The recall rate was found to be 89%, indicating the model's ability to correctly identify positive cases. The test size used in the random forest was only 20%. The confusion matrix for the model showed that there were no false positives or false negatives, with an error rate of 0.0. Hence, random forest was chosen as the final model for this project.

Conclusion

Based on our analysis, we can conclude that the loan data performed well in this project, even though it started as a small dataset of just 614 rows. After cleaning, removing duplicates, null values, characters, and outliers, it reduced to around 367 rows. The data showed great progression with an accuracy rate of 74%, which is a significant improvement from the initial dataset. However, the error percentage is still at around 26%, which means there is still room for improvement.

One of the limitations of this project is that the data is very illogical when it comes to giving loans to individuals, as the amount is very low, and loans were given to everyone regardless of their income. This could potentially affect the accuracy of the model and limit its effectiveness in real-world scenarios. The recall rate or true positive rate is also only at 89%, which is not ideal for a loan prediction model. Therefore, further improvements and optimizations are necessary to enhance the performance of the model and make it more reliable and accurate.

Overall, the project demonstrated the effectiveness of using machine learning algorithms, such as random forest, logistic regression, and decision trees, for loan prediction. The grid search CV helped to optimize the hyperparameters, and metrics such as F1 score, recall, accuracy, and precision were used to evaluate the model's performance. In the end, the random forest algorithm was selected as the final model, as it showed the highest F1 score, accuracy, and precision.

References

https://www.kaggle.com/code/vikasukani/loan-eligibility-prediction-machine-lea rning/input

https://www.sample.net/reports/data-analysis-report/