Teesside University, MIDDLESBROUGH - TS1 3BA

School of Computing, Engineering & Digital Technologies

Machine Learning, CIS4035-N

Machine Learning Application and Report

Name: Hiteesh Pushpendra Kondepati

MSc. Computer Science

Student ID: S3021074

Email: S3021074@live.tees.ac.uk

Submission Date: 08/05/2024

Word Count: 2058

LOAN APPROVAL PREDICTION WITH MACHINE LEARNING APPROACHES

Abstract:

In the realm of financial services, loan approval stands as a critical decision-making process with significant implications for both lenders and borrowers. Leveraging machine learning approaches, this study delves into the predictive modelling of loan approval. Our dataset encompasses a variety of features including the number of dependents, educational background, self-employment status, annual income, loan amount, loan term, credit score (CIBIL), and various asset values. Through the analysis of these features, we aim to develop robust predictive models to forecast loan approval outcomes. By employing state-of-the-art machine learning techniques, we endeavour to enhance the efficiency and accuracy of loan approval processes, thus facilitating more informed lending decisions and potentially mitigating risks associated with loan default. This research contributes to the ongoing discourse on the application of machine learning in financial domains and underscores its potential to revolutionise traditional lending practices.

Keywords:

Random Forest Classifier, Gradient Boost, Logistic Regression, Gaussian NB.

Introduction:

In the contemporary landscape of financial services, the process of loan approval stands as a pivotal undertaking, embodying the delicate balance between risk assessment and financial inclusion. With the advent of advanced computational techniques and the proliferation of data analytics, the application of machine learning methodologies has emerged as a

a promising avenue for enhancing the efficiency and accuracy of loan approval processes.

The loan business is one of the bank's products which significantly promotes the economy and majorly helps banks drive their revenue growth. With the development of technologies, the loan business became a widespread business with the emergence of various lending institutions other than banks. The popularity of this loan business exposes the problem of loan default which both banks and other lending institutions are the victims. This does not only affect the institutions but also negatively impacts the economy which may lead to an economic crisis.

However, machine learning models are not one size fits all, impliedly some models perform better for loan default prediction than other subjects to the project at hand. In this study, I examined the predictive performance of four prominent machine learning models in anticipating loan default: Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Random Forest Classifier (RF), and Gradient Boost (GB). Beyond assessing the effectiveness of these models, I also did exploratory data analysis and data pre-processing. The goal of this is to evaluate and compare the predictive capabilities of various machine learning techniques for loan default prediction.

The objective of this project is to compare the performance of different machine learning models for loan default prediction and determine the most effective algorithm(s) and important variables. To reach the stated objectives, it is important to answer the following questions:

1. Which machine learning models are most effective for loan default prediction?

2. Which variables are important in the prediction of loan default based on the dataset?

References:

1. A study on Predicting Loan Default Using a Random Forest Algorithm

(https://www.sciencedirect.com/science/article/pii/S1877050919320277)

2.https://www.analyticsvidhya.com/blog/2022/02/loan-approval-prediction-machine-learning/

Methods:

This study evaluates the predictive capabilities of various machine learning models, including Logistic Regression, KNN, Gaussian Naïve Bayes, Random Forest Classifier, and Gradient Boost in forecasting loan defaults. The methodology employed for this comparative analysis is explained below.

Steps Followed:

1. Data Collection
2. Data Evaluation
3. Data cleaning and preprocessing
4. Feature Engineering
5. Train and Test data split
6. Model Training
7. Model optimization
8. Model Evaluation

Data Exploration and Feature Selection:

EDA is something that helps us understand and analyse data. Exploratory Data Analysis (EDA) serves as a crucial step in understanding the inherent characteristics of our dataset, unveiling trends, and identifying patterns. To conduct our analysis effectively, we leveraged essential libraries such as NumPy, Pandas, and Seaborn. These tools empowered us to delve into the data, performing tasks like value counts to gauge the frequency of categorical variables. Through our EDA process, we discerned a notable imbalance within the dataset, as illustrated in Figure 1. Additionally, by scrutinizing the distribution of loan applications by credit score, we observed that high credit score applicants exhibit the lowest incidence of loan approvement. Our exploration extended to visual aids like histograms and scatter plots, which provided insights into the distribution of numerical and categorical features. These visualizations not only shed light on the frequency of distributions but also enabled us to explore relationships with our target variables effectively.

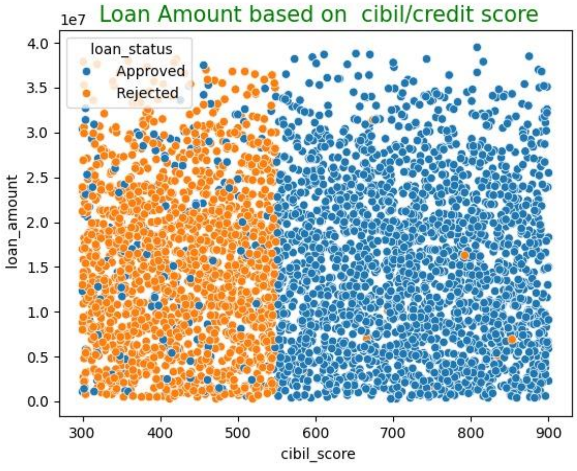

Fig 1

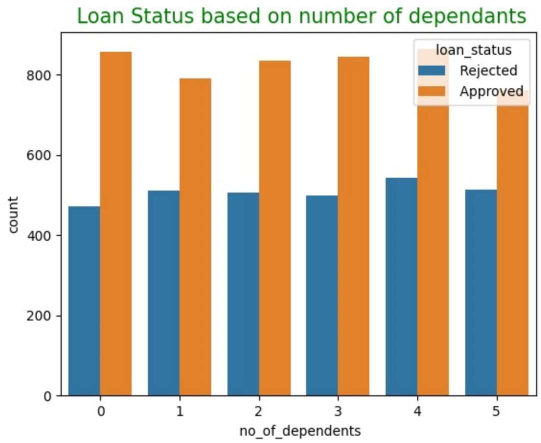

Fig 2

Additionally, we utilized a correlation matrix to discern the relationships between our numerical variables, as depicted in Figure 3. From this matrix, it became evident that there is a strong correlation between the income per annum and the value of luxury assets; these factors are highly correlated with the loan amount. Specifically, we observed that as the loan amount increases, so does the income and assets value.

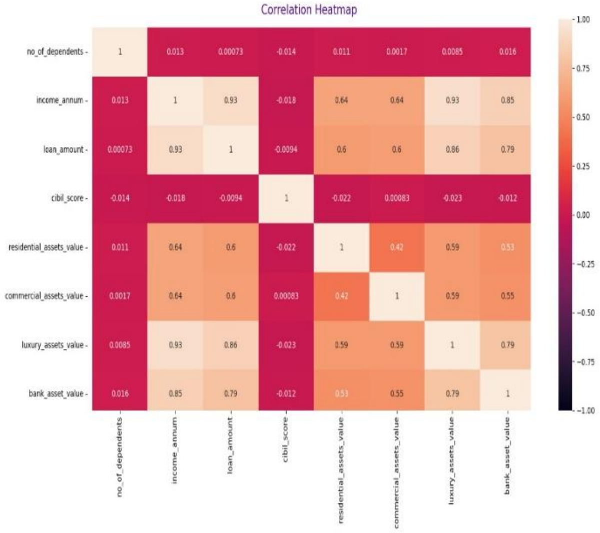

Fig 3

Feature Engineering:

1. Encoding of categorical features:

In our research, we implemented several techniques to prepare our data for machine learning model application. Given that machine learning operates effectively with numerical data, we employed encoding methods.

Specifically, we utilized label encoding to convert our categorical variables into numerical features. Among these features, the column with the highest number had 2 variables. Therefore, we utilized change_datatype during the process and converted the categorical values to numerical values.

2. Handling Missing Values:

To ensure compatibility with our machine learning models and mitigate potential errors, addressing missing values within the dataset is imperative. The presence of missing values can introduce bias into the performance of our machine learning models. Therefore, it's crucial to handle them appropriately. In this regard, we employed imputation techniques where numerical data were replaced with the median, considering their unequal distribution. For text data, we utilized mode replacement, which contributed to enhancing the overall quality of the dataset.

3. Addressing Outliers:

Through our exploratory data analysis (EDA), we gained valuable insights into the loan dataset, revealing that there are no outliers or anomalies present in the numerical columns of this dataset.

Model Implementation:

In this research, we deployed four distinct machine learning algorithms, namely Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Random Forest Classifier (RF), and Gradient Boost. These algorithms were selected due to their long-standing reputation and extensive usage in binary classification tasks.

Logistic Regression:

LR is particularly well-suited for scenarios where the probabilities for the dependent variable belong to one of two categories. Given that our project involves binary classification tasks, we opted for logistic regression because it offers probability outputs that can be easily thresholded to derive binary predictions. Furthermore, it stands out as a straightforward and comprehensible algorithm.

Gaussian Naive Bayes (GNB):

GNB stands out as a popular algorithm for binary classification projects. It operates on the assumption that each feature is independent. Despite its occasional oversimplification, GNB performs admirably on our datasets, primarily because the features within our dataset exhibit some degree of independence.

Random Forest Classifier:

This algorithm harnesses ensemble learning techniques to generate predictions based on multiple decision trees. We opted for the Random Forest Classifier in our project due to its resilience against overfitting and its adeptness at handling both categorical and numerical data. Moreover, it excels in regression analysis scenarios.

Gradient Boost:

Gradient Boosting is a highly effective algorithm commonly employed in binary classification and regression tasks. It's prized for its ability to produce accurate predictions and handle complex datasets. Gradient Boosting is a powerful ensemble learning technique that combines the predictive power of multiple weak learners, typically decision trees, to create a strong predictive model.

Model Evaluation:

After choosing our machine learning algorithms, we assessed their performance on the split dataset using various evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics provided insights into how effectively the algorithms could predict loan defaults. To facilitate comparison, we visualised the overall accuracy of the algorithms on both the train and test datasets using bar charts.

We further evaluated the performance of our models using confusion matrices, which included elements such as true positives, false positives, true negatives, and false negatives for each class in the predicted data. This analysis allowed us to assess how effectively the models identified positive and negative instances of loan status. Among these models, Random Forest stood out as the best performer, exhibiting high true positives (826) and true negatives (638), along with low false positives (11) and low false negatives (12).

A summary of the performance metrics was presented in a table, encompassing all four metrics. Additionally, we utilized Receiving Operating Characteristic (ROC) curves to evaluate the models' performance, visualizing the trade-off between True Positive Rate and False Positive Rate at different classification thresholds. Random Forest achieved an AUC score of 1.00, making it the top-performing algorithm in our dataset.

Model optimization:

Under-sampling technique:

Due to the imbalanced nature of our dataset, where the "Approved" class significantly outweighs the "Rejected" class, there's a risk of bias affecting our models' performance. To address this, we employed the random under-sampling technique. This involved randomly selecting a subset of the majority class to match the size of the minority class, thus creating a balanced dataset.

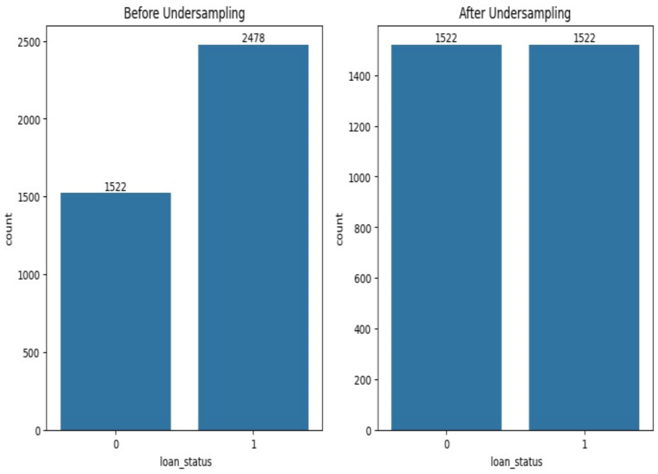

Fig 4

Additionally, we fine-tuned our model parameters through grid search, exploring a range of values for each hyperparameter to optimize performance. We complemented this with K-fold cross-validation, splitting our training dataset into five folds and iteratively training and testing the model on different fold combinations. This approach yielded more robust estimates of model performance and reduced evaluation metric variance.

Following these optimization processes, we reapplied the algorithms and observed improved performance. This was visualized using bar charts, confusion matrices, and ROC curves.

In conclusion, Random Forest and Gradient Boost emerged as the top-performing models for loan approval classification, achieving 100% and 98% accuracy on both training and test datasets, respectively. Gaussian Naive Bayes and logistic regression followed closely behind.

Feature Importance:

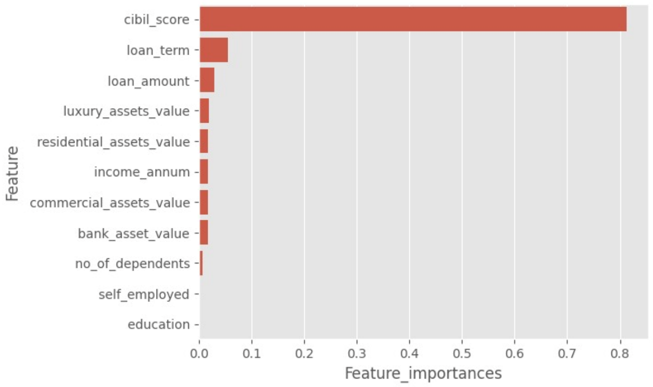

Fig 5

Having established the Random Forest classifier as our top-performing algorithm, we proceeded to conduct a detailed analysis of the importance of the features selected for the project and their contributions to the model's performance. Utilizing a scale ranging from 0 to 1, we visualized these features in descending order of importance. This provided valuable insights into the extent to which each feature contributed to the overall performance of the model.

RESULTS AND DISCUSSION:

To confirm our claim that the Random Forest classifier outperforms the other models utilized in this research, we employed accuracy, recall, F1 score, and ROC curve as evaluation metrics. These metrics were applied to assess the performance of Logistic Regression, Random Forest, Gradient Boost, and Gaussian Naive Bayes models.

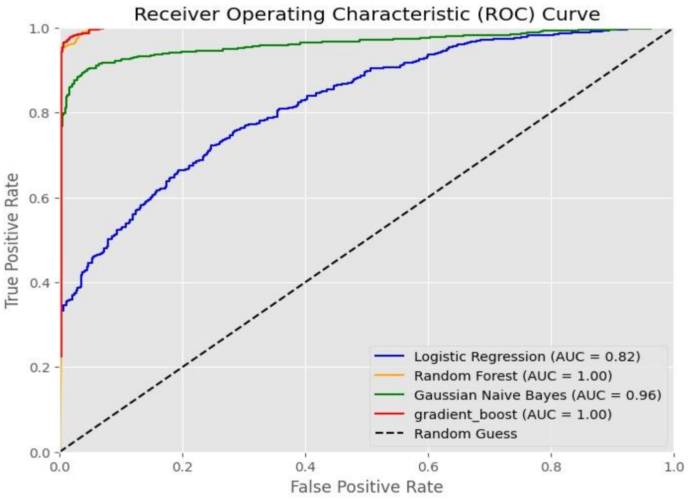

Fig 6

Examining the ROC curves of the four algorithms depicted in Figure 6 above, it is evident that Random Forest is positioned closest to the left side of the curve and exhibits a larger area under the curve, indicating superior performance.

The table below presents an analysis of the other metrics for the four algorithms. Random Forest exhibits the highest accuracy concerning the test data compared to Gradient Boost. Despite Logistic Regression showing lower accuracy on the training dataset compared to Gaussian Naive Bayes, it demonstrates a higher probability of yielding favourable results when applied to new data in comparison to Gaussian Naive Bayes.

In summary, both the Random Forest classifier and Gradient Boost models emerge as the top-performing models. However, Random Forest's performance slightly surpasses that of Gradient Boost.

	Train accuracy	Precision	Recall	F1 score
LR	0.61	0.57	0.61	0.47
RF	1.00	1.0	1.0	1.0
GNB	0.76	0.80	0.76	0.73
GB	0.98	0.98	0.98	0.98

FUTURE WORKS & CONCLUSION:

The model developed in this project has shown promising results. However, there are still areas for improvement in future iterations. These improvements could involve acquiring more data with additional features or exploring new features that could be relevant to the task at hand. Additionally, experimenting with different ensemble learning techniques may help enhance the performance of our current model.

In conclusion, the primary goal of this research was to analyze the performance of selected algorithms for loan approval prediction. It was found that machine learning models offer a reliable means for banks and other lending institutions to manage their risks and make informed decisions.

REFERENCE:

[1] https://www.analyticsvidhya.com/blog/2022/02/loan-approval-prediction-machine-learning/

[2] https://towardsdatascience.com/predict-loan-eligibility-using-machine-learning-models-7a14ef904057

[3] https://www.geeksforgeeks.org/loan-eligibility-prediction-using-machine-learning-models-in-python/

[4] https://www.researchgate.net/publication/372909313_Prediction_of_Loan_Approval_in_Banks_using_Machine_Learning_Approach