Financial Risk Assessment For Loan Approvals

Paul Kim
Department of Computer Science
Georgia State University)
Atlanta, Georgia
paulkim62.2001@gmail.com

***Abstract*—This report explores the financial risk assessment for loan approvals using machine learning algorithms. The research evaluates three models – Random Forest, Logistic Regression, and Decision Tree. The dataset has 20,000 entries with 24 cleaned features. Results show that Logistic Regression has the highest accuracy of 0.96, with a high speed and performance evaluation. Random Forest has an accuracy of 0.93, demonstrating robustness and ability to handle complex and nonlinear relationships while providing feature importance insights. Decision Tree offers an accuracy of 0.88 but has high interpretability and transparency, making it a great choice for stakeholder presentations. The study concludes that Random Forest is the most effective model for loan risk assessment as it balances sensitivity and specificity. However, Decision Tree is recommended for scenarios requiring high interpretability such as a presentation for stakeholders.**

***Keywords*— Financial Risk Assessment, Loan Approvals, Data Mining, Machine Learning Algorithms, Random Forest, Logistic Regression, Decision Tree, Feature Importance**

I. Introduction

As the years pass, it is shown that financial risk assessment is an important part in the loan approval process. This is because having a reliable system to assess financial risks directly influences the sustainability of financial institutions and the economy. Accurately evaluating financial risks help lenders figure out if a borrower has the ability to pay back their loans. Financial risk assessments have become more important as the global financial system expands, and as more regulations are made to have become stricter to prevent another economic collapse. The 2008 economic collapse is a great example as to why we need to build models and methods that can assess the risk of giving out loans so that the economy will not collapse.

The 2008 financial crisis revealed major flaws in the financial risk assessment methods that banks and lenders were using previously. Many of these risk assessment methods were outdated and as a result, led to a lot of people failing to pay back their loans, especially in the housing market. This crisis showed that traditional ways of assessing financial risk, like using credit scores, were not enough to predict when people would default on loans nor to prevent large-scale financial problems. Because of this, there was a need for more advanced, data-driven methods that could use real-time data to better understand the economic changes, and spot financial issues earlier. The purpose of these new methods is to make financial risk assessment more accurate and help avoid another financial crisis.

This project will provide guidelines for making informed decisions for both lenders and borrowers when it comes to giving and receiving loans. By examining various methods, it aims to propose an effective model selection process for loan assessment and identifying potential financial risks. To achieve this, the paper will incorporate data mining and analysis concepts, and algorithms that were covered in class.

We will assess which methods are the most reliable for assessing financial risk which will help to ensure an accurate loan approval process. Therefore, this research will not only enhance our understanding of risk assessment but also contribute to the development of improved, data-driven strategies for financial decision-making.

https://github.com/pkim6201/finriskassessment

II. Materials and Methods

A. Data explanation and characterization

The dataset used for this project was provided by the user LORENZO ZOPPELLETTO, through Kaggle.com. It contains 20,000 entries of personal and financial data, which is used to predict the financial risk assessment for the loan approval process.

According to the website, this dataset serves two purposes: (1) to predict a continuous risk score associated with each individual's likelihood of not being able to pay their loans back, and (2) to determine the outcome of loan approval, whether the applicant is likely to be approved or denied for a loan.

This dataset will be used to show the relationship between various personal and financial features. The dataset includes a variety of features like demographic information, credit history and score, employment status, income, current debt, and other financial information to provide a solid foundation for any data-driven analysis and decision-making.

B. Data preprocessing

Originally, the dataset included 36 features. However, 12 features were deemed redundant or irrelevant to the current analysis due to lack of correlation with the target feature 'Loan Approved.' One of such features was 'Total Assets' due to the redundancy with the outcome of a loan approval. As a result, the final dataset includes 24 features with 20,000 entries.

The random state in all the methods used in this project is set to 42 to create consistency. To make sure effective training and testing was used for the methods, the training size was set to 70% and the testing size to 30%. The goal of this project is to assess various methods by using variables that have the most positive/negative correlation to loan approvals.

I. SELECTED VARIABLES AFTER DATA CLEANING.

Variable	Explanation/Parameters
CreditScore	300 ~ 850
EducationLevel	0: High School, 1: Associate, 2: Bachelor, 3: Master, 4: Doctorate
Experience	Years of work experience
LoanAmount	Requested loan amount
LoanDuration	Loan repayment period in months 12,24,36,48,60,72,84,96,108,120
MonthlyDebtPayments	Monthly debt obligations
CreditCardUtilizationRate	Credit card usage in %
DebtToIncomeRatio	Debt to income proportion
BankruptcyHistory	0: None, 1: One or more
PreviousLoanDefaults	Past loan defaults 0: None, 1: One or more
PaymentHistory	Past payment behavior
LengthOfCreditHistory	Credit history duration
TotalLiabilities	Total owed debts
TotalDebtToIncomeRatio	Total debt against income
EmploymentStatus	Self-Employed, Unemployed
HomeOwnershipStatus	Rent, Other
LoanPurpose	Education, Other

C. Data Analysis/Mining

Decision tree uses a "tree"-like structure, where it has nodes and zero to two branches. It predicts the value of a target node by learning decision rules inferred from the data features it gathers. Decision trees benefit from having depth which will result in complex decision rules and a fitter model. This method is simple to understand and interpret as it has little data preparation, uses a white box model where the explanation for each condition is easily explained by Boolean logic, and performs well even if its assumptions are violated by the true model the data was generated from.

Random forest is an ensemble method where each of the classifiers in the ensemble is a decision tree classifier so that the collection of classifiers is grouped together like a "forest." The decision trees are randomly generated using a selection of attributes at each node to determine the split. The accuracy of a random forest depends on the classifiers and the dependency between each one. They are robust to errors and outliers. It is ideal to maintain the strength of the classifiers without increasing their correlation. The number of attributes selected for consideration at each split do not matter for random forests. This method is efficient for very large databases because they consider fewer attributes for each split.

Logistic regression is a similar method to linear regression; however, it solves classification issues by predicting categorical results, whereas linear regression would predict a continuous result. As it is a model for binary classification, it is likely that the results will be between 0 and 1.

D. Evaluation and interpretations

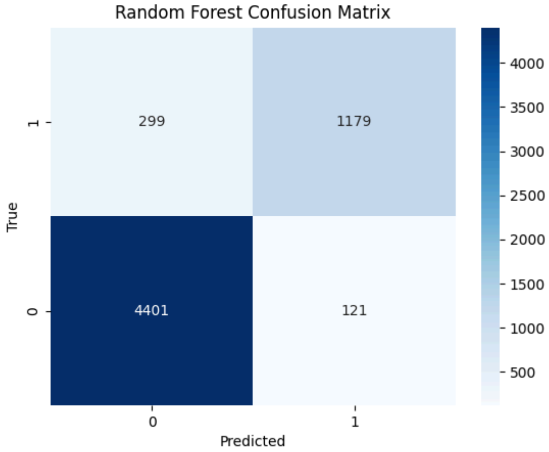
Confusion matrices have been used in this project to define the performance of a classification algorithm. They visualize and summarize their performance on a matrix. It compares the actual labels to the predicted labels and provides insight into the performance of the method. This matrix consists of four metrics: 1) True Positives, which are correctly predicted positive results, 2) True Negatives, which are correctly predicted negative results, 3) False Positives, which are incorrectly predicted positive results, 4) False Negatives, which are incorrectly predicted negative results.

Metrics used based on the confusion matrix in this project were "accuracy" to measure the performance of the model as it compares the total number of approvals given by the model to the total number of approvals given by the dataset.

III. Results

A. Random Forest

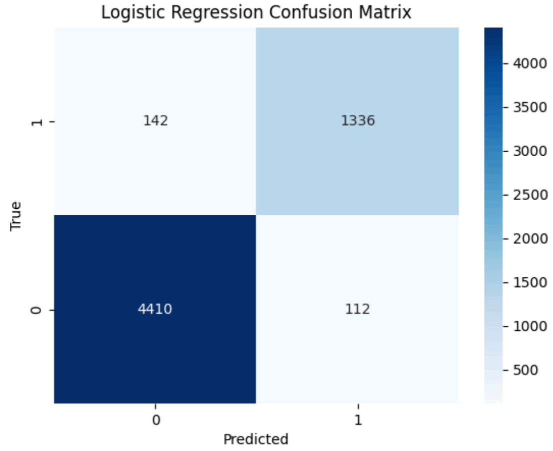
The result of the random forest method is shown in Figure 1. The outcome is biased towards the approval of loan applications, which gives a high true negative rate. The features which are the most important are “Total Debt to Income Ratio”, “Monthly Income”, “Interest Rate”, and “Net Worth.” Changing any of these variables could lead to different outcomes in loan approvals.



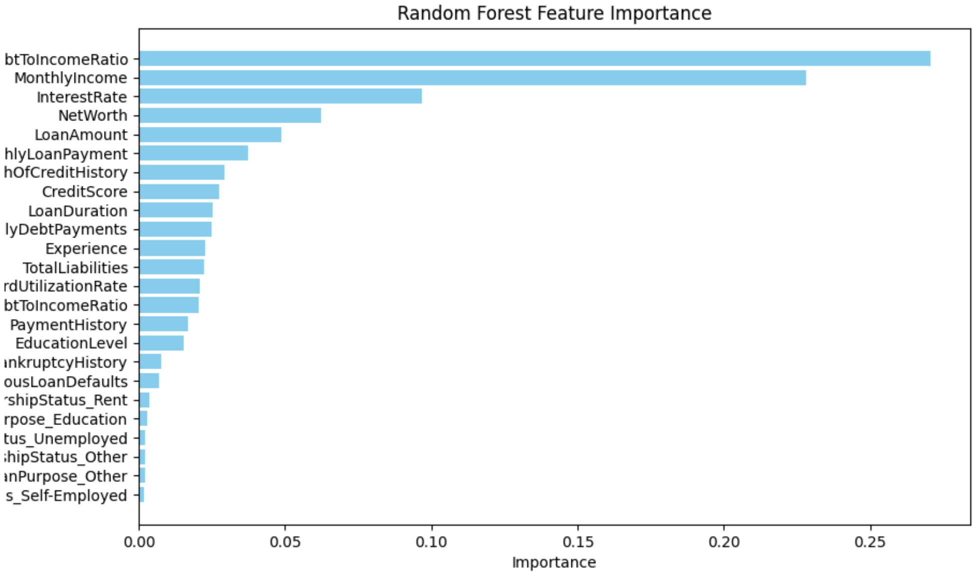
1. Fig. 1: Confusion matrix for random forest classifier

B. Logistic Regression

The result of the logistic regression is shown in Figure 3. This method shows a similar outcome as the random forest method, however, there are less cases of false positives and false negatives proving that this model has a higher accuracy.



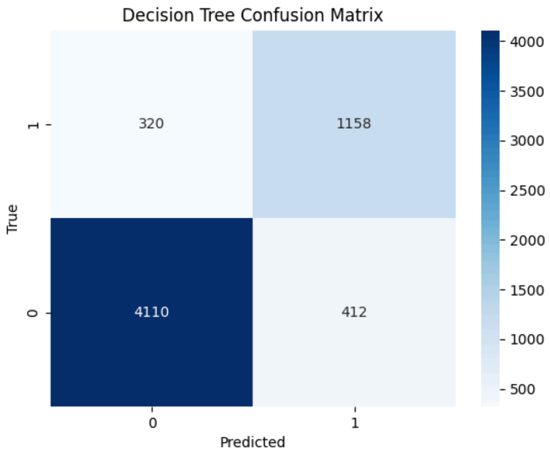
3. Fig. 3: Confusion Matrix for Logistic Regression



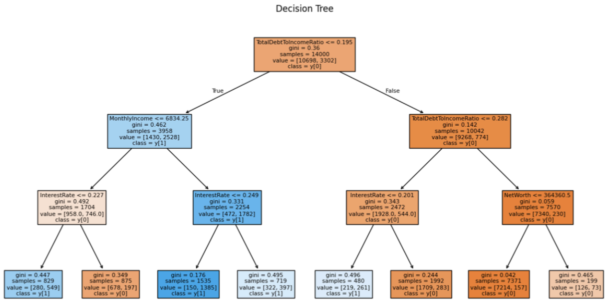
2. Fig. 2: Feature importance for each variable in Random Forest

C. *Decision Tree*

As shown in Figure 4, the decision tree is biased towards the approval of loan applications, which leads to a high true positive rate. The relationship between the selected variables and the approval of the loans is determined using the decision tree method as shown in Figure 5. In Figure 5, each node has a class attached to them, where class = y[0] is rejection and class = y[1] is approval. For example, if the Total Debt to Income Ratio is less than or equal to 0.195, the loan application is more likely to get rejected.

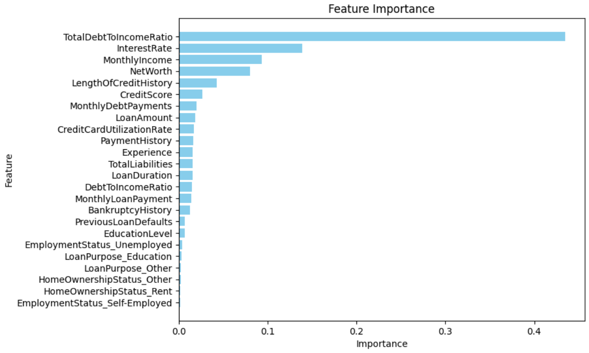


5. Fig. 5: Confusion Matrix for Decision Tree



6. Fig. 6: Decision Tree Visualization

Figure 6 shows that the most important features are "Total Debt to Income Ratio", "Monthly Income", "Interest Rate", and "Net Worth."



7. Fig. 7: Feature importance for each variable in Decision Tree

IV. Discussion and Conclusion

As shown in Table 2, the Logistic Regression method has the highest accuracy score of 0.96, ensuring that most eligible loans are approved. It is simple, fast and moderately interpretable while performing well on linearly separable data. The most optimal use case scenario for this method would be when interpretability and speed are important.

The Random Forest method has an accuracy of 0.93 and is proven that it is robust to overfitting. It provides feature importance insights which make it easier to identify which features influence loan approvals the most. However, it is less interpretable compared to the other two methods and is computationally expensive. The optimal use case scenario would be when the dataset is complex with many interactions with features and classes.

The final method, Decision Trees, has an accuracy of 0.88 which is still reliable but is the lowest out of the three methods. The biggest strength of this method is that it is highly interpretable and works well with small datasets and fewer features. The best use case scenario would be when interpretability and transparency in decision-making is needed in front of stakeholders and investors.

The best method to be used in assessing loan approvals would be the Random Forest method as it balances sensitivity and specificity, handles complex and nonlinear relationships, and provides feature importance insights to refine the loaning process. However, if the purpose is to show stakeholders a visual explanation of the decision process, the Decision Tree method would be the best to be used given its high interpretability.

II. Accuracy of Each Method

Method	Accuracy Score
Logistic Regression	0.96
Random Forest	0.93
Decision Tree	0.88

References

[1] Lorenzo Zoppelletto, “Financial Risk for Loan Approval,” Kaggle , [Online]. Available: https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval [Accessed: Nov. 21, 2024]

[2] Scikit-learn Developers, “RandomForestClassifier,” Scikit-learn Documentation, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed: Nov. 21, 2024]

[3] Scikit-learn Developers, “LogisticRegression,” Scikit-learn Documentation, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. [Accessed: Nov. 21, 2024].

[4] Scikit-learn Developers, “Decision Trees,” Scikit-learn Documentation, [Online]. Available: https://scikit-learn.org/1.5/modules/tree.html. [Accessed: Nov. 21, 2024].