Loan Prediction and Risk Assessment Using Regression

Ayush Srivastava 01854763, Chinmay Brahmbhatt 01943857, Sonakshi Gupta 01976552, Chakshu Patel 01975869

Abstract

This project aimed to predict loan approval (classification) and risk score (regression) using a dataset of financial and personal information. Data preprocessing involved handling missing values, encoding categorical features, and scaling numerical features. The dataset was split into training, validation, and test sets, and models were trained using Logistic Regression, Random Forest, and XGBoost for classification, and Linear Regression, Random Forest Regressor, and Ridge Regression for regression. Random Forest emerged as the most accurate model for both tasks, achieving high validation accuracy for loan approval and low mean squared error for risk score prediction. The models provide actionable insights to improve loan approval chances and reduce risk, showcasing the potential of machine learning in financial decision-making. Future work could explore additional features and advanced optimization techniques to further enhance performance.

1) Introduction

In recent years, loan rejection rates in the United States have been rising at an alarming level, with over 20% of loan applications rejected in the last year—hitting a five-year high! Loans are an essential financial resource in today's world. As students who are soon to graduate, we will soon be in a situation where getting a loan will be a necessity, such as for things like buying a car and maybe even a house in the far future.

The challenge of getting approved for a loan itself becomes even more convoluting as banks do not disclose any minimum requirements or clear metrics and rather approve or deny loans at their discretion based on different combinations of factors.

Additionally, as students, many of us with student loans, face a bigger challenge. The fact is it is much harder to get a loan with student debt, and more debt makes getting rejected all the more likely. While a single loan rejection does not majorly impact a person's credit score, repeated hard credit checks (due to multiple rejections) can negatively impact it. This limits borrowing ability in the future.

1.1) Problem and Significance

As getting a loan continues to become a more difficult endeavor, the demand for a tool that can help you predict whether or not you will get approved for a loan has never been higher. Many people think loans come down to credit score and annual income, however, that is not always the case. More often than not lenders take into account a variety of factors when deciding on whether or not to approve a loan. Having a tool that people can use that replicates the banks' process would be immensely helpful and time-saving. We created a model that can not only predict whether or not a person will be approved for a loan but also give their risk factor for the loan.

2) Method

This project used machine learning techniques to predict loan approval (classification) and risk score (regression). The dataset, containing both numerical and categorical features, underwent extensive preprocessing to ensure readiness for modeling. Missing values were checked, and categorical variables (MaritalStatus, EmploymentStatus, etc.) were encoded using one-hot encoding. Numerical features like Age, AnnualIncome, and DebtToIncomeRatio were normalized to the [0,1] range using Min-Max Scaling to standardize their influence on the models. The data was split into training (60%), validation (20%), and testing (20%) sets to enable robust training, hyperparameter tuning, and performance evaluation.

For the regression task (RiskScore), three models were trained: Linear Regression, Ridge Regression, and XGBoost Regressor. Linear Regression provided a baseline for comparison, Ridge Regression addressed overfitting through L2 regularization, and XGBoost Regressor captured non-linear relationships and feature interactions. For classification (LoanApproved), three models were trained: Logistic Regression, Random Forest Classifier, and XGBoost Classifier. Logistic Regression served as an interpretable baseline, while the tree-based models captured complex decision boundaries and provided feature-importance insights. Metrics like MSE, R^2, accuracy, and F1-score were used to evaluate and compare model performance across validation and test datasets. This systematic approach ensured reliable predictions and actionable insights into loan approvals and risk assessment.

3) Data

We trained the model on a dataset found on Kaggle. The data has 20,000 people's personal and financial information. There were 36 columns of information recorded for each person including whether or not they were approved for a loan. The information is all filled in, and there is a 70/30 split between people who have been denied a loan versus people who have been accepted. The data is split into 60% for training, 20% for validation, and another 20% for tests.

Most people had an income below $100,000 and very few had an income above $200,000. The mean annual income was about $60,000. Also Employment status and Education level both seemed to have a strong correlation to loan approval. A larger percentage of employed people were approved for a loan compared to unemployed and people with a higher level of education had larger percentages of people being approved as well. Loan amounts were also primarily $50,000 or lower with the average loan amount being about $25,000. The loan approval for the "average person" in the dataset seemed to be just below 25%. In the end, we found that the most impactful factor was the debt-to-income ratio which was the highest in terms of correlation with loan approval.

4) Results

The models were evaluated using MSE and R² for regression, and Accuracy, Precision, Recall, and F1-Score for classification. XGBoost Regressor and XGBoost Classifier outperformed others, achieving the best metrics on test data. Key features like CreditScore, DebtToIncomeRatio, and AnnualIncome significantly influenced predictions, highlighting the effectiveness of tree-based models in capturing complex patterns.

Linear Regression :

Linear Regression Test MSE: 14.150967647433852

Linear Regression Test R² Score: 0.7720800995174577

Ridge Regression:

Ridge Regression Test MSE: 14.160835703642249

Ridge Regression Test R² Score: 0.7719211615250171

XGRegressor:

XGBoost Test MSE: 5.90979577044919

XGBoost Test R² Score: 0.9048149852764883

Logistic Regression Validation Accuracy: 0.8705

Classification Report:

```
         precision   recall  f1-score   support

   0       0.89       0.94      0.92      3044

   1       0.78       0.64      0.70       956
   accuracy                          0.87      4000
  macro avg       0.84      0.79      0.81      4000
weighted avg       0.87      0.87      0.87      4000
```

Random Forest Validation Accuracy: 0.9245

```
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.97      0.95      3044
           1       0.88      0.79      0.83       956

    accuracy                           0.92      4000
   macro avg       0.91      0.88      0.89      4000
weighted avg       0.92      0.92      0.92      4000
```

XGBoost Validation Accuracy: 0.95175

```
Classification Report:
                precision    recall  f1-score   support

           0       0.97          0.97      0.97      3044
           1       0.91          0.89      0.90       956

    accuracy                           0.95      4000
   macro avg       0.94      0.93      0.93      4000
weighted avg       0.95      0.95      0.95      4000
```

5) Conclusion

The goal of the project was to create a model that could help people know whether or not they would be approved for a loan should they go to a bank. The data used to train the model consisted of 20,000 people's information. The logistic regression model performed very well in all the metrics that it was measured by and could give a good indication of whether a person may or may not be ready to go to the bank to get a loan. The model is limited by the data that it is trained on. As more time passes the model may be less accurate as the economic environment may change. These changes would mean that the model would need to be trained again on new more up-to-date data so that it can be accurate.

6) Contribution

Task	Student ID	Description
Model Training and Evaluation	01943857	Developed and evaluated machine learning models to predict loan approval and risk score by preprocessing the dataset, training multiple regression and classification models, and identifying key features influencing predictions.
Analysis of Models and Preparing Presentation	01975869	Worked on the report primarily Introduction, method, results and conclusion. Analysis of data and explanation of how model worked
Put together the final presentation	01976552	Worked with the group to create the final presentation and add all the findings from the project research
EDA	01854763	Analyzed the dataset and identified key insights to help us better understand the data.

References

[1] Will Skipworth. Over 20% of U.S. loans rejected in last year—hitting 5-year high. Forbes. Available: https://www.forbes.com/sites/willskipworth/2023/07/17/over-20-of-us-loans-rejected-in-last-year-hitting-5-year-high/#:~:text=Over%2020%25%20Of%20U.S.%20Loans,Year%E2%80%94Hitting%205%2DYear%20High, Jul. 17, 2023.

[2] Alicia Wallace. "Getting approved for a loan is getting harder." CNN Business. Available: https://www.cnn.com/2023/09/22/economy/getting-approved-for-loan-us/index.html, Sep. 22, 2023.

[3] "What is Logistic Regression?" Statistics Solutions. Available: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/, Accessed: Nov. 4, 2024.

[4] "XGBoost Documentation." Available: https://xgboost.readthedocs.io/en/latest/index.html, Accessed: Nov. 9 2024.