Loan Prediction and Risk Assessment Using Regression💰

- Ayush Srivastava(01854763), Chakshu Patel(01975869), Chinmay Brahmabatt (01943857), Sonakshi Gupta(01976552)

Abstract

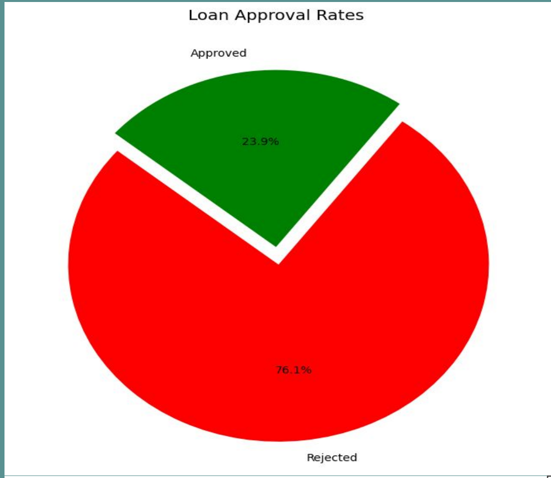
Our project focused on building a predictive model to help assess loan approval and risk using regression techniques. We worked with a dataset of 20,000 rows and 36 columns, applying logistic regression and other methods like decision trees and random forests to find the best-performing model. We evaluated our results using metrics like precision, accuracy, and the F1 score. In the end, we created a model that can help people better understand their chances of getting a loan, making the process a little less confusing and more accessible.

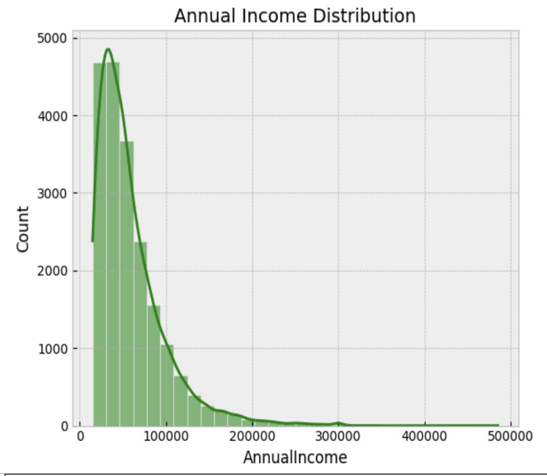
Introduction

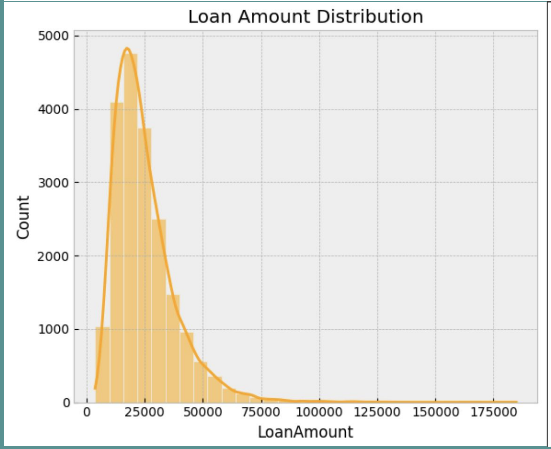
- More than 20% of loan applications are denied every year
- Just having good credit score isn’t enough to get approved
- Banks use complicated criteria to approve loans that is not widely known
- A tool that could predict loan approval would be useful for both banks and people
  - There would be less people applying so banks would not need to review as many loan applications
  - People would not need to go to the bank just to get rejected saving them time
- The loan predicting model we made can not only predict whether or not a person will get a loan but also give them a risk score for the loan

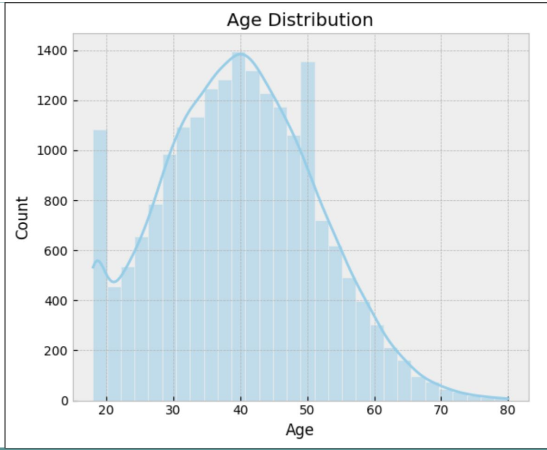
Data

- Data obtained from Kaggle dataset
- 20,000 people's information recorded
- 36 different features per person
- 80/20 split for training/testing
- 70% of the data is people who were rejected from a loan
- The average income of the people is about $60,000
- The average loan amount was $25,000
- Average age of people in the dataset is 40 years

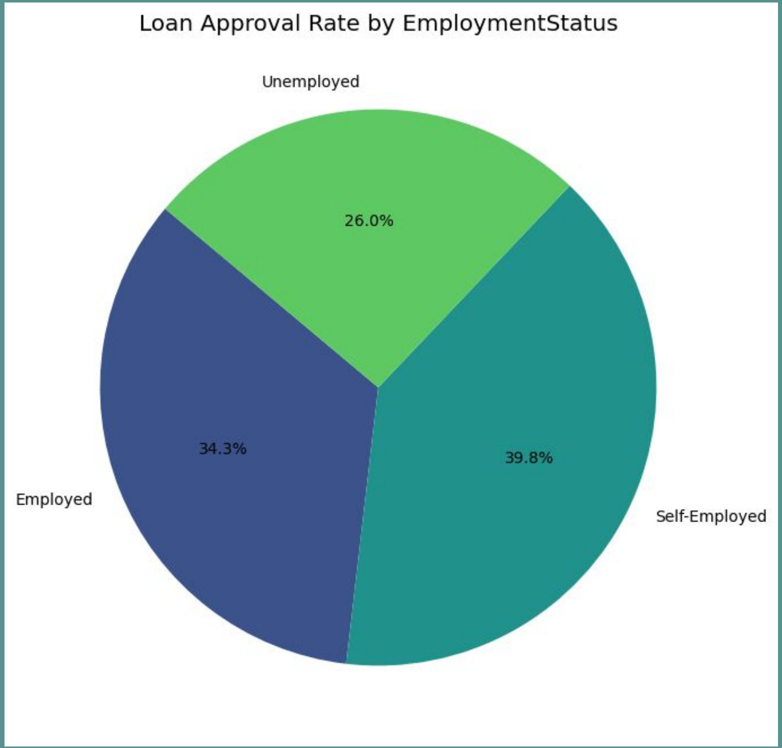


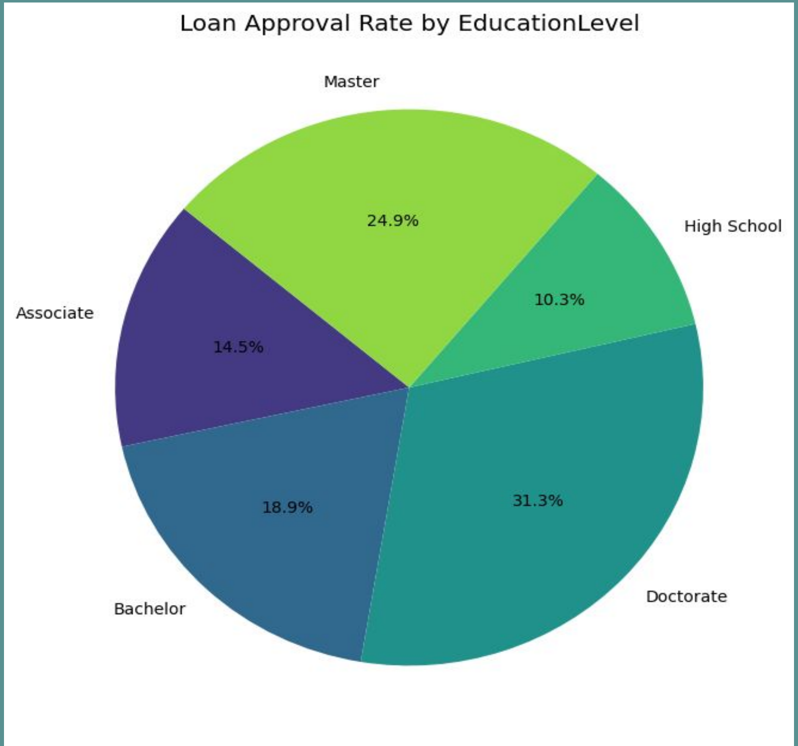




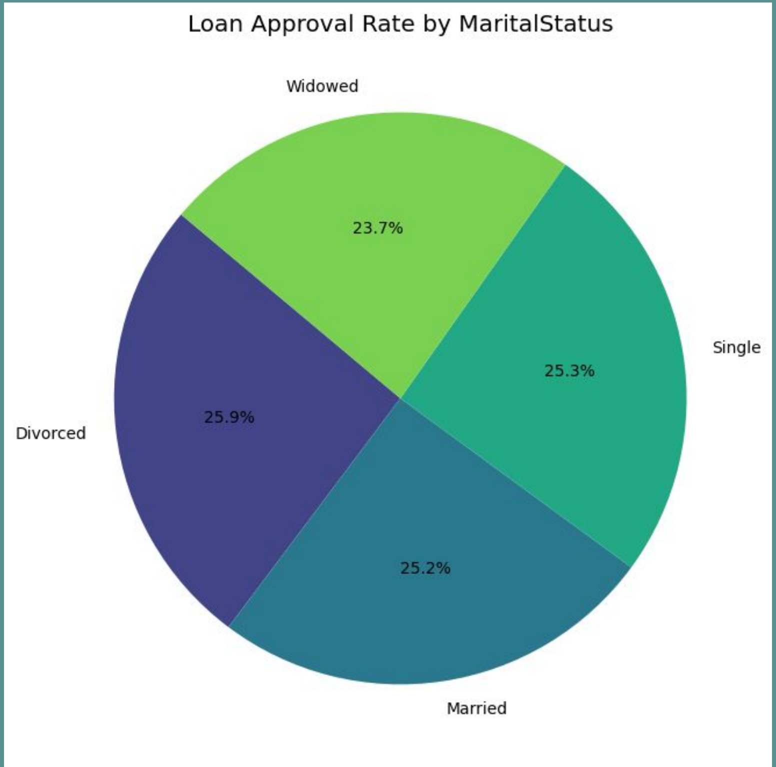


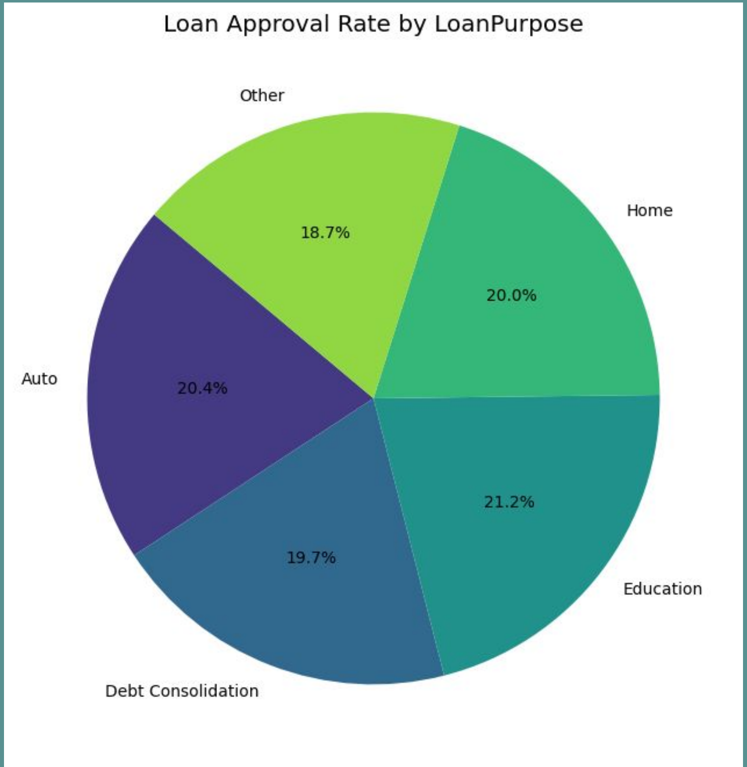
Data Insights





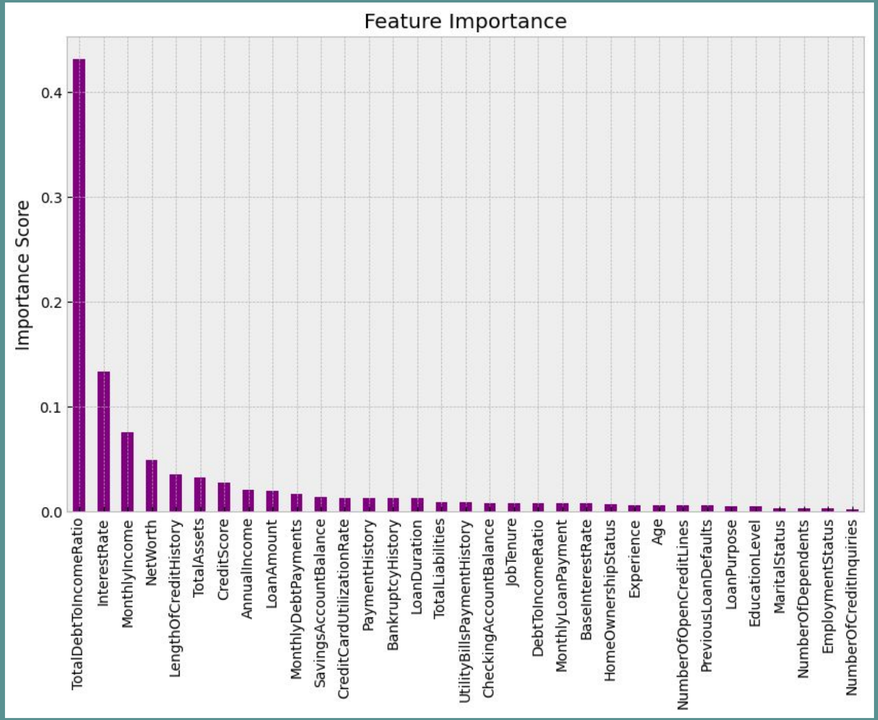
Data Insights





Data Insights

- Most important factors to determine whether loan got approved: Debt to Income Ratio
  - This is the main factor when it comes to determining one's risk score
- Risk Score itself would've been the highest
- Smaller factors include
  - Interest Rate
  - Monthly Income
  - Networth
  - Length of Credit History



Method

🔍 **Preprocessing**

- 🔢 **Numerical Features**:
  - 📏 Scaled Age, Annual Income, and Debt-to-Income Ratio using **Min-Max Scaling** ([0,1]) to standardize their influence.
- 📝 **Categorical Features**:
  - 🔠 Encoded variables like **Marital Status** and **Employment Status** using **One-Hot Encoding**.
- ⚠️ **Checked for Missing Values** and handled them appropriately to ensure data integrity.

🔀 **Data Splitting**

- 📊 **Training Set**: 60% – for model learning.
- 📈 **Validation Set**: 20% – for tuning.
- 🧪 **Testing Set**: 20% – for final performance evaluation.

🤖 **Models Used**

- **Regression (Risk Score)**:
  - 📏 **Linear Regression**: Baseline model for comparison.
  - **Ridge Regression**: Added L2 regularization to reduce overfitting.
  - 🌳 **XGBoost Regressor**: Captured non-linear relationships and feature interactions.
- **Classification (Loan Approval)**:
  - 📊 **Logistic Regression**: Simple, interpretable baseline.
  - 🌲 **Random Forest Classifier**: Captured complex decision boundaries.
  - 🚀 **XGBoost Classifier**: Provided robust performance and feature importance insights.

Results

Model	Test MSE	R² Score
📏 Linear Regression	14.15	0.772
🛡️ Ridge Regression	14.16	0.772
🚀 XGBoost Regressor	5.91	0.905

Model	Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (0/1)
📊 Logistic Regression	87.05%	0.89 / 0.78	0.94 / 0.64	0.92 / 0.70
🌲 Random Forest	92.45%	0.94 / 0.88	0.97 / 0.79	0.95 / 0.83
🚀 XGBoost Classifier	95.18%	0.97 / 0.91	0.97 / 0.89	0.97 / 0.90

📉 Regression Results

✅ **Best Regression Model**: XGBoost Regressor

🔍 Classification Results

✅ **Best Classification Model**: XGBoost Classifier

Conclusion

Developed a predictive model for loan approval using diverse numerical and categorical data.

Identified XBG Boost as the most effective model with high accuracy and F1-score.

Logistic Regression offered strong interpretability despite simpler design.

Insights:

- Higher risk scores and high debt-to-income ratios correlated with increased rejection rates.
- Identified areas where applicants might focus to improve their loan approval chances, for example, credit score or their annual income.

Limitations:

- Dependency on historical data limits adaptability in changing economic climates.
- Lack of real-time financial indicators affects predictive accuracy.

Contribution Chart

Task	Student ID	Commentary on Contribution
EDA	01854763	Analyzed the dataset and identified key insights to help us better understand the data.
Research and Presentation	01976552	Worked with the group to research and built the presentation.
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 Report	01975869	Worked on the report, analysis of data and explanation of how model worked.

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.