1. Dataset Sourcing

The dataset, sourced from Kaggle, Zoppelletto (2023), contains 20,000 rows and 35 variables related to factors influencing loan approval decisions, along with a binary response variable (Loan Approved). It spans from January 2018 to October 2072 but the ApplicationDate column will be disregarded, as the focus is on identifying the key variables that impact loan approval. The target variable, Loan Approved, is represented as 0 (not approved) and 1 (approved). All 35 variables will be initially considered for analysis.

2. Application Description

Financial institutions face significant risks from loan defaults, making accurate risk prediction models essential for their business strategy. The goal of this analysis is to identify the key factors that influence loan approval or denial, allowing institutions to better allocate resources and minimize risk. By refining these models, institutions can avoid rejecting creditworthy applicants while minimizing lending to potential defaulters, ultimately improving profitability.

3. Exploratory Data Analysis

The dataset consists of both continuous and categorical variables. There are 6 categorical variables, such as MaritalStatus and EmploymentStatus, and 30 numerical variables, 9 of which are floats and 21 are integers, including CreditScore, LoanAmount, and AnnualIncome. The response variable is LoanApproved, which is a numerical type containing binary outcomes.

I visualized the distributions of categorical and numerical variables and calculated the skewness for key variables like TotalAssets (5.311), LoanAmount (1.834), and AnnualIncome (2.089), which showed high right-skewness. Log transformation was applied to predictors to reduce the impact of large values and normalize their distributions. This also helped mitigate the effect of outliers (detailed outliers analysis in supplementary).

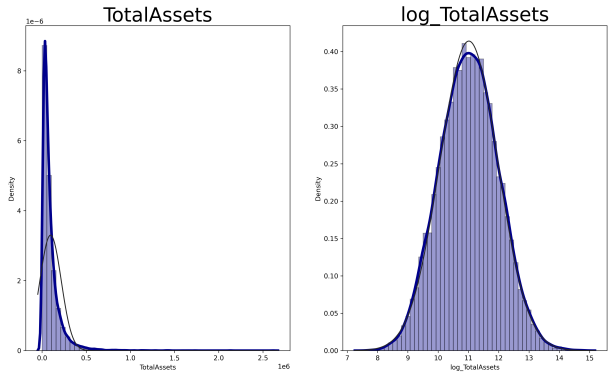

Figure 1: Log Transformation

No missing values were detected in the dataset. For example, in the variable HomeOwnershipStatus, the Other category is used to handle where the ownership information may not have been clearly specified.

For the correlation analysis, I first identified 14 variables with over 10% correlation to loan approval. I then checked for multicollinearity to ensure none were too similar, which could affect model accuracy. From a business view, TotalAssets was included to account for cases where assets could be sold if the loan defaults. In the end, 6 key variables were selected based on their importance to risk management: CreditScore, TotalAssets, LoanAmount, InterestRate, AnnualIncome, and RiskScore.

Figure 2: Correlation Analysis

4. Training / Testing Data Split

Studies show that the best results come from using 20-30% of data for testing and 70-80% for training. I chose an 80/20 split to keep more data for training, which should improve model performance. According to Gholamy et al. (2018), an 80% training split is the most effective.

5. Logistic Regression and K-Nearest Neighbor Models

1. The six predictors were selected based on their correlation and alignment with the analysis assumptions, and categorical variables were converted to numerical to generate a complete correlation map. Log transformation was applied to positively skewed predictors to stabilize variance and achieve a more normal distribution. All predictors were scaled before fitting the models, allowing for comparisons on the same scale.

2. For K-Nearest Neighbors (KNN) classification, used 10-fold cross-validation to select the optimal value of k, Jeganathan (2024), and k = 11 provides the best rusult.

3. Since KNN is a non-parametric model, it doesn't produce coefficients for predictors, so I cannot directly identify the most important variable. While there might be alternative methods to estimate feature importance, but I did not apply any of alternative methods in this assignment.

4. The results confirm the initial expectations, with Risk Score having the highest coefficient (6.70) and the most significant impact on loan approval. For each unit increase in risk score, the odds of approval rise significantly, assuming other factors remain constant. Financial institutions often prioritize risk scores in their loan decisions, making it a critical factor in the process.

6. Model evaluation

The evaluation on the test set shows that the KNN model has a slightly lower misclassification error (0.01325) compared to the logistic regression classifier (0.01875). However, the logistic regression model is better at identifying positive cases of loan approvals, as it has a higher sensitivity than KNN.

7. Logistic Regression with Shrinkage

1. I selected the Lasso shrinkage method, and the best shrinkage value (alpha) identified by LassoCV is 0.0001.

2. Both logistic regression models indicate the variable RiskScore is the most significant predictor. In the model without Lasso, RiskScore had a strong positive impact on loan approval, with a large coefficient 6.6997. After applying Lasso, the coefficient shrank to -0.2587, indicating a smaller and negative effect, where higher RiskScores slightly reduce the chance of approval.

3. For the logistic regression models, the one without Lasso performed with significantly higher accuracy (0.98125) compared to the model with Lasso (0.7165) on the testing set.

8. Conclusions

The goal of this project was to build a model to predict loan approvals and help banks manage risk. I focused on six key factors: CreditScore, TotalAssets, LoanAmount, InterestRate, AnnualIncome, and RiskScore.

I tested logistic regression (with and without Lasso) and K-Nearest Neighbors (KNN). The logistic regression without Lasso performed best with 98.9% accuracy, followed closely by KNN with 98.67%. Overall, the logistic regression model without Lasso was the most effective for predicting loan approvals.