Lending Club Predict Interest Rates

Project Report

Tzega Abera

7/23/23

Data Science

Contents

1 Introduction

1.1 Problem statement and objective

LendingClub is a leading peer-to-peer lending platform that connects borrowers with investors. Evaluating loan applications requires a rigorous process. Today, LendingClub assigns an interest rate to each loan by considering factors, such as credit score, employment history, loan amount, and loan purpose. This interest rate determination process, however, may involve subjective judgment, leading to potential inconsistencies and suboptimal decisions. By leveraging data and building a regression model, LendingClub can automate and enhance this process, resulting in improved loan pricing and risk management.

The objective of this project is to develop a regression model to predict the interest rates on loans issued by LendingClub. In order to predict the interest rates, we will focus on features related to the borrower's creditworthiness, loan characteristics, loan term, and loan amount. Zip code and other factors can be used to accurately predict the interest rate assigned to a loan. An accurate model will facilitate LendingClub improving its risk assessment processes and making more informed decisions regarding interest rates for future loan applications. This will allow the organization to cut its operational cost and increase its profit.

1.2 Dataset

In the field of peer-to-peer lending, the Lending Club dataset is a widely used dataset. This popular online lending platform contains information about loans issued to borrowers and provides a comprehensive view of loan applications, borrower characteristics, and loan performance. Because of these reasons, the dataset is ideal for conducting credit risk analysis and determining interest rates based on borrower’s credit worthiness.

Below are the broad categories of the variables involved in the dataset:

1. Loan Information:
    - Loan amount: The amount requested by the borrower.
    - Loan term: The duration of the loan (e.g., 36 months, 60 months).
    - Interest rate: The interest rate charged on the loan.
    - Installment: The monthly payment amount.
    - Loan status: The current status of the loan (e.g., fully paid, charged off, default).
    - Grade and subgrade: Lending Club's rating system for assessing loan risk.
2. Borrower Information:
    - Employment length: The length of time the borrower has been employed.
    - Home ownership: Indicates whether the borrower owns a home or rents.
    - Annual income: The annual income reported by the borrower.
    - Verification status: Indicates whether the income source was verified by Lending Club.
    - Purpose of the loan: The intended use of the loan funds (e.g., debt consolidation, home improvement).
3. Credit History:
    - Grade and Subgrade score: The credit score assigned to the borrower by Lending club.

- Revolving balance: The outstanding balance on revolving accounts (e.g., credit cards).
- Revolving utilization: The percentage of available credit utilized by the borrower.
- Number of open credit lines: The total number of open credit lines (e.g., credit cards, loans).
- Open public records: Indicates whether the borrower has any derogatory public records (e.g., bankruptcies, tax liens).

4. Payment History:
   - Delinquencies: The number of late payments on the borrower's credit history.
   - Months since last delinquency: The number of months since the borrower's last delinquency.
   - Months since last record: The number of months since the last derogatory public record.

These variables provide valuable information about the borrowers and their loan applications, allowing for analysis and prediction of credit risk and interest rates.

2 Data Preprocessing and Exploration

The data frame contained 887, 379 rows and 74 features.

2.1 Cleaning the Data

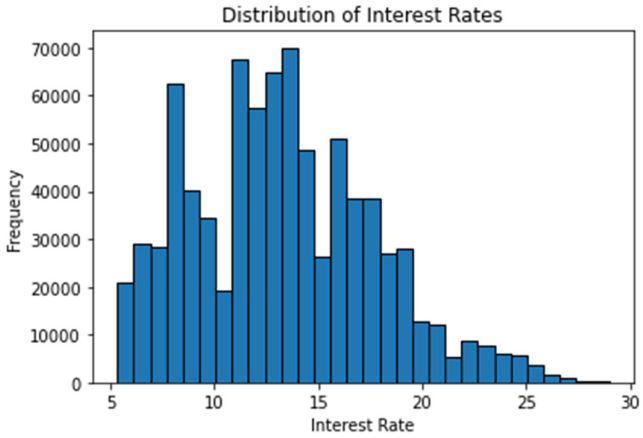
- Removed unnecessary columns, resulting in 55 columns.
- Handled missing values: Dropped columns with more than 80% missing values (19 columns).
- Investigated remaining columns for NaN entries.

2.2 Transforming the Data

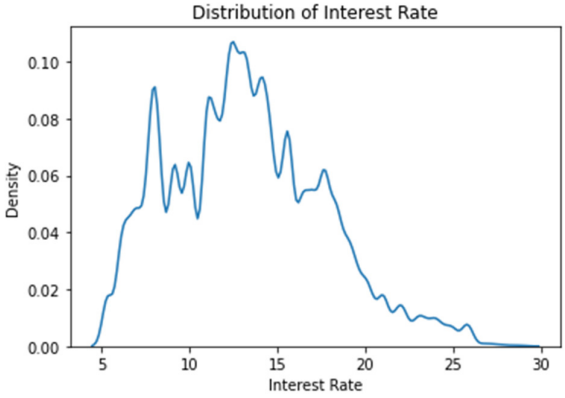
- Converted 'emp_length' column to numbers only.
- Extracted numeric values from the 'term' column and added a new column 'term_in_month'.
- Converted the 'earliest_cr_line' column to the datetime format.
- Calculated the average values of the 'earliest_cr_line' and 'last_credit_pull_d' columns and filled NaN values with their average.
- Investigated three columns ('tot_cur_bal', 'total_rev_hi_lim', 'tot_coll_amt') with the same amount of missing values.
- Checked if these three columns have missing values in the same rows and share a common pattern of missing data.
- Checked the number of missing values in each column and created a heatmap of missing values.
- Dropped rows with missing values in all three columns.
- Calculated the average value of the 'last_pymnt_d' column and filled NaN values with the average.
- Conducted analysis to ensure no remaining missing values in the dataset.
- Replaced NaN values with the median value in certain columns.

2.3 Exploratory data analysis

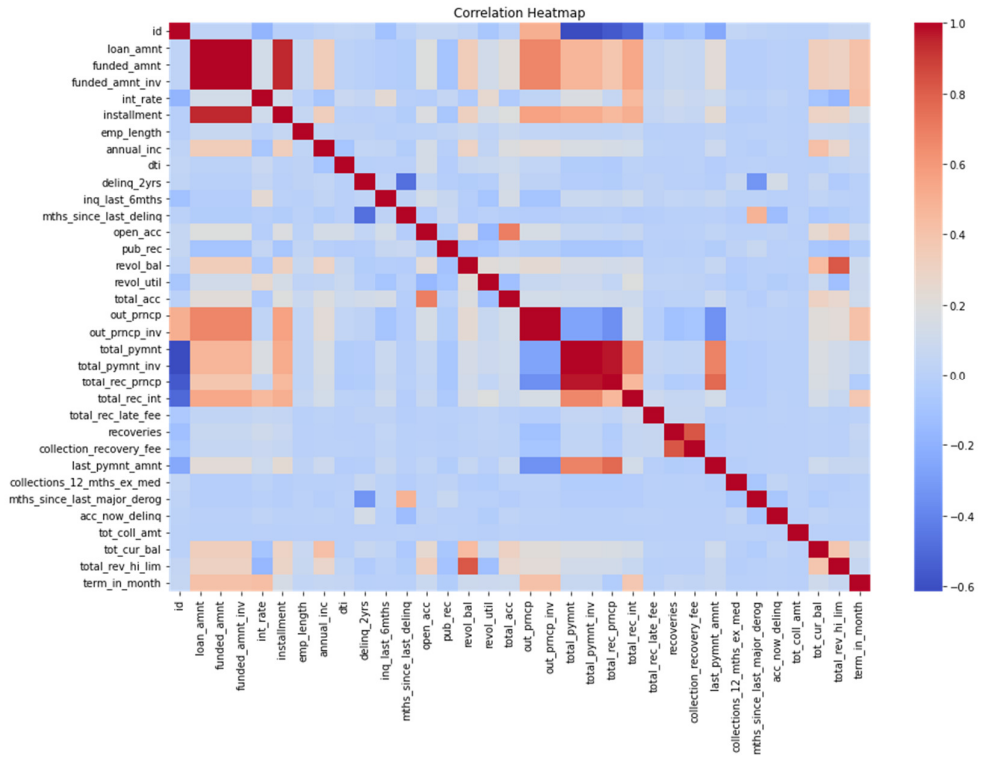
1. Dependent Variable Selection:
    - Chose 'int_rate' as the dependent variable for regression analysis.
2. Univariate Analysis



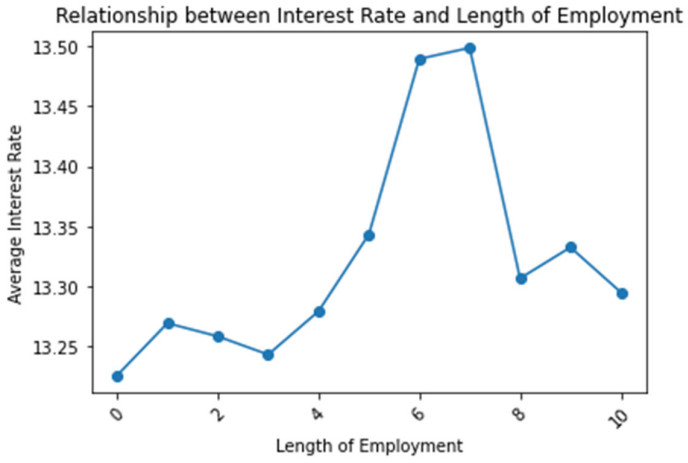
    - Examined the distribution of interest rates using a histogram.
    - Found that the interest rates ranged from approximately 6% to 28%.
    - Calculated descriptive statistics such as mean, median, mode, minimum, maximum, and standard deviation.
    - Observed a slightly positively skewed distribution of interest rates.



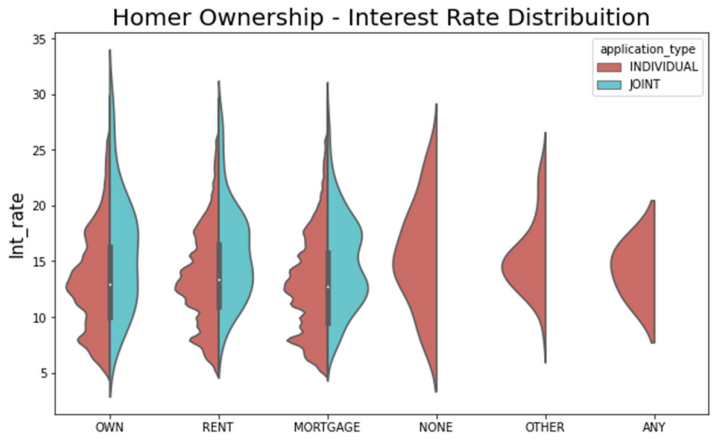
3. Bivariate Analysis:



- Generated a correlation heatmap to visualize the relationships between numerical variables.
- Identified variables such as 'term_in_month', 'total_rev_hi_lim', 'total_rec_int', and 'annual_inc' that have some correlation with interest rate, though not very strong.
- Explored the relationship between interest rate and length of employment using a line plot.

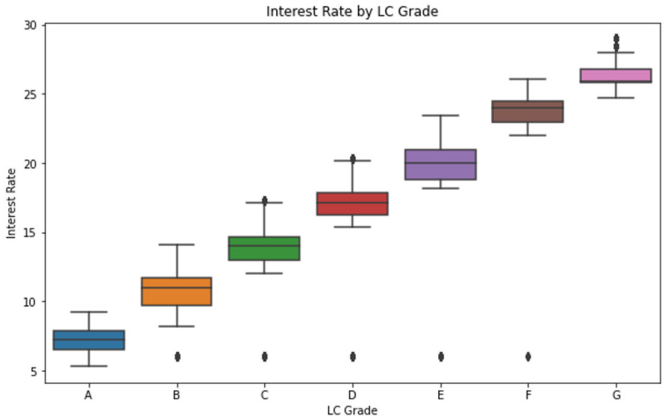


- Found a non-linear relationship, with a drop in average interest rate for individuals with 8 to 10 years of employment.
- Visualized the distribution of interest rates based on homeownership status using a violin plot.



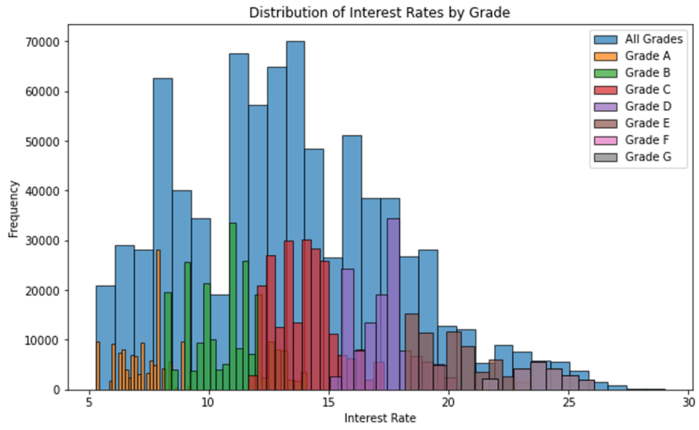
- Calculated descriptive statistics for interest rates based on homeownership categories.

4. Multivariate Analysis:
   - Investigated the relationship between interest rate and LC grade using a box plot.



   - Found that higher LC grades (lower credit risk) were associated with lower interest rates.

5. Regression Model:



- Identified the relationship between various features and the dependent variable, interest rate.
- Suggested fitting a regression model with 'int_rate' as the dependent variable and other variables as independent variables.

Based on the analysis, it is evident that interest rate is influenced by factors such as employment length, homeownership status, LC grade, and other numerical variables. The findings provide insights into the lending market dynamics and borrower characteristics. Recommendations for the client could include implementing risk-based pricing strategies, considering employment history as a factor in loan evaluation, and leveraging credit grades to assess interest rates for borrowers. Further research could focus on analyzing the impact of other categorical variables and exploring non-linear relationships between features and interest rates.

3 Model Building

In the analysis, we used three different models: Ridge regression, Random Forest Regression, and XGBoost.

1. Ridge Regression:
   - Ridge regression is a linear regression technique that incorporates a regularization term to prevent overfitting.
   - It works by adding a penalty term to the loss function, which encourages the model to have smaller coefficients.
   - The regularization term helps reduce the impact of irrelevant features and can improve the model's generalization performance.
   - Ridge regression is suitable when dealing with multicollinearity, where independent variables are highly correlated.
2. Random Forest Regression:
   - Random Forest Regression is an ensemble learning method based on decision trees.
   - It combines multiple decision trees and makes predictions by averaging the predictions of individual trees.
   - Each tree is trained on a random subset of the data and uses a random subset of features, reducing overfitting and improving robustness.
   - Random Forest Regression can handle both numerical and categorical features, and it can capture non-linear relationships between variables.
3. XGBoost:
   - XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting framework.
   - It is designed to improve the performance of gradient boosting algorithms by employing regularization techniques and parallel computing.
   - XGBoost builds an ensemble of weak prediction models (typically decision trees) in a sequential manner, where each new model corrects the errors made by the previous models.
   - It uses a combination of gradient descent optimization and tree-based learning to minimize the loss function and make predictions.
   - XGBoost is known for its high predictive accuracy and ability to handle complex data patterns.

These models were chosen for the analysis because they offer different advantages and capabilities. Ridge regression provides a simple and interpretable linear model, while Random Forest Regression and XGBoost can handle non-linear relationships and capture complex interactions between variables. The combination of these models allows for a comprehensive analysis and comparison of their performance in predicting the interest rate in the lending dataset.

To determine the relevance of columns in predicting interest rates, a domain knowledge-based analysis was conducted. The following columns were assessed for their potential impact on interest rates:

1. 'int_rate': This column represents the target variable, the interest rate, and is directly relevant for the prediction task.
2. 'loan_amnt': Higher loan amounts may be associated with higher interest rates due to increased risk.
3. 'annual_inc': Borrower's annual income is an important factor considered by lenders, where higher incomes may result in lower interest rates.
4. 'mths_since_last_delinq': The number of months since the borrower's last delinquency can impact creditworthiness and, in turn, interest rates.
5. 'revol_bal' and 'revol_util': Higher revolving credit balances and utilization may indicate a higher risk for lenders, potentially leading to higher interest rates.
6. 'tot_coll_amt' and 'tot_cur_bal': These columns reflect the borrower's credit history and financial stability, which can influence interest rates.
7. 'total_rev_hi_lim': The total revolving credit limit provides information about the borrower's available credit and debt capacity, influencing interest rates.
8. 'term_in_month': Loan term in months can affect interest rates, with longer terms potentially carrying higher risks and rates.
9. 'empl_length': The length of employment can be an indicator of job stability, where longer employment histories may result in lower interest rates.

Considering these factors, it is expected that these columns have varying degrees of relevance in predicting interest rates. By incorporating these variables into the predictive models, we aim to capture their potential impact on the target variable and improve the accuracy of our predictions.

To reduce the size of the dataset and save computational time, a random sampling technique was applied. Specifically, only 6% of the original dataframe was sampled for further analysis. This sampling approach helps to create a smaller dataset while retaining the overall representation of the data. By working with a reduced dataset, computational resources and processing time can be optimized, allowing for faster model training and analysis. This sampling strategy enables efficient exploration of the data while still providing meaningful insights and accurate predictions for the interest rate.

4 Model Evaluation

In this project, a dummy baseline model was implemented by predicting the mean of the target variable (interest rate). The performance of more complex regression models was then compared against this baseline to assess their effectiveness.

Metric	Value
Mean Squared Error (Dummy Baseline)	19.5506
Mean Absolute Error (Dummy Baseline)	3.5397
R^2 Score (Dummy Baseline)	0

We also created three separate regression models: ridge_model, rf_model, and xgbr_model and analyzed the performance metrics of each model.

Model	MSE	MAE	R^2
Ridge	14.2707	3.03048	0.26332
Random Forest	12.6739	2.8279	0.345752
XGBoost	12.2303	2.77868	0.368651

Although the regression models (Ridge regression, Random Forest Regression, and XGBoost) outperformed the dummy baseline model, there is still room for improvement in their performance.

Based on the results obtained, the XGBoost model outperforms the other two models (Ridge regression and Random Forest Regression) in terms of mean squared error (MSE) and mean absolute error (MAE). The XGBoost model achieved an MSE of 12.23 and an MAE of 2.7, indicating that, on average, its predictions have a smaller deviation from the actual interest rates compared to the other models.

Furthermore, the R^2 score, which represents the goodness of fit of the model, indicates that the XGBoost model provides a better fit to the data compared to the other models. A higher R^2 score suggests that a larger proportion of the variance in the target variable (interest rate) is explained by the model. Therefore, the XGBoost model demonstrates a stronger ability to capture the underlying patterns in the data and make accurate predictions.

Based on these results, it can be concluded that the XGBoost model is the most suitable choice among the three models for predicting interest rates in this dataset. However, we will conduct further analysis and hyperparameter tuning to optimize the model's performance and ensure robustness.

A hyperparameter search for the RandomForestRegressor model was performed using RandomizedSearchCV and here are the results:

Model	Training MAE	Test MAE	Training MSE	Test MSE	R2 Score
Ridge	3.04803	3.03048	14.5239	14.2707	0.26332
Random Forest Regression	2.58382	2.81333	10.465	12.4642	0.356577
XGBoost	0.582531	2.96478	0.78251	14.0727	0.273543

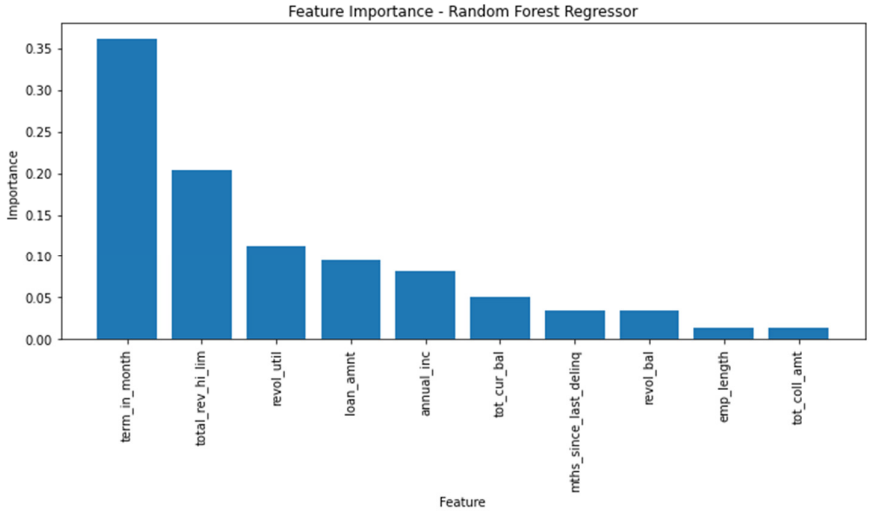
Based on the results, the Random Forest Regression model demonstrates better performance compared to the XGBoost model in terms of capturing the error in both the training and test datasets. The XGBoost model shows lower training MAE and MSE, suggesting that it fits the training data well. However, it exhibits higher test MAE and MSE, indicating potential overfitting.

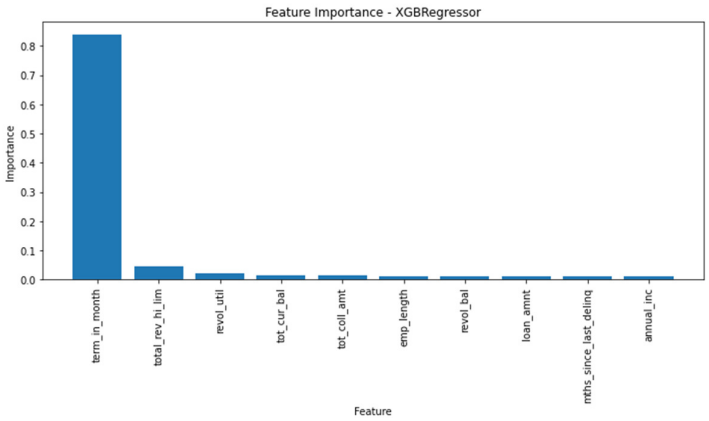
In contrast, the Random Forest Regression model achieves lower test MAE and MSE, implying better generalization to unseen data. Additionally, it obtains a higher R2 score, indicating a better fit and ability to explain the variance in the target variable.

Considering these findings, the Random Forest Regression model is recommended as the preferred choice among the three models. However, further analysis and hyperparameter tuning may be necessary to optimize its performance and ensure its suitability for the specific task at hand.

5 Feature Importance

As shown in the graph below, the term variable exhibits a significant impact on predicting the interest rate when using the Random Forest Regressor. The Random Forest algorithm considers the importance of different features when making predictions. In this case, the term feature has been identified as having a substantial influence on the predicted interest followed by total_rev_hi_lim, and revol_util.





Based on the analysis above, it is evident that the term feature carries substantial predictive power for determining the interest rate in the XGBosst Regressor model. It implies that the term variable is an essential factor in estimating or understanding the interest rates in the given context.

6 Findings and Recommendations:

Key Findings

1. Relationship between predictor variables and interest rate: The analysis revealed several predictor variables that are relevant in predicting interest rates. These include loan amount, borrower's annual income, months since last delinquency, revolving credit balance, revolving credit utilization, total collection amount, total current balance, total revolving credit limit, loan term, and length of employment. These variables provide insights into the borrower's creditworthiness, financial stability, and risk profile, which influence the interest rates assigned by lenders.
2. Model performance: The XGBoost and Random Forest Regression models were evaluated for predicting interest rates. While the XGBoost model showed lower training errors, it exhibited higher test errors, indicating potential overfitting. On the other hand, the Random Forest Regression model demonstrated a better balance between training and test errors and achieved a higher R2 score, suggesting better overall performance and generalization.

Further Research and Analysis:

1. Feature engineering: Exploring additional transformations or combinations of the existing predictor variables, such as interaction terms or polynomial features, could enhance the model's predictive power and capture more complex relationships between the predictors and the interest rate.
2. External data sources: Incorporating external data sources, such as macroeconomic indicators or industry-specific data, may provide additional insights into interest rate determinants and improve the model's accuracy and robustness.
3. Time series analysis: Conducting a time series analysis on historical interest rate data can help identify trends, seasonality, or cyclical patterns that influence interest rates. This analysis can be used to build models that consider the temporal dynamics and improve predictions.

Recommendations:

1. Refine the Random Forest Regression model: Further fine-tuning of hyperparameters, such as the number of trees, maximum depth, or minimum sample split, can be performed to optimize the Random Forest Regression model's performance and improve its accuracy in predicting interest rates.
2. Regular model monitoring and updating: Given the dynamic nature of lending markets and changing borrower profiles, it is essential to regularly monitor the model's performance and update it with new data. This ensures that the model remains relevant and effective in predicting interest rates accurately.
3. Use the model for scenario analysis: The developed model can be leveraged to conduct scenario analysis and simulate the impact of different variables on interest rates. This information can help the client make informed decisions and evaluate the potential consequences of various lending strategies or market conditions.

Overall, by leveraging the insights gained from the analysis and implementing the recommendations, the client can enhance their decision-making processes related to interest rates, improve risk assessment, and optimize their lending strategies for better outcomes.

7 Conclusion

- Analysis of Predictor Variables: The report analyzed various predictor variables, such as loan amount, borrower's annual income, credit history, employment length, and others, to understand their relevance in predicting interest rates.
- Baseline Model: A dummy baseline model was used to establish a benchmark for performance comparison with more complex models. This baseline model predicted the mean of the target variable.
- Regression Models: Three regression models, namely Ridge regression, Random Forest Regression, and XGBoost, were evaluated for predicting interest rates. Model performance was assessed using metrics such as mean squared error (MSE), mean absolute error (MAE), and R2 score.
- Model Comparison: The Random Forest Regression model outperformed the other models, with lower test MAE, test MSE, and a higher R2 score. It demonstrated a better balance between training and test errors and exhibited better overall performance.
- Insights and Recommendations: The analysis provided insights into the relationship between predictor variables and interest rates. Further research and analysis, such as feature engineering, incorporating external data sources, and time series analysis, were recommended to enhance the model. Concrete recommendations included refining the Random Forest Regression model, regular model monitoring and updating, and using the model for scenario analysis.

Reflection on Project Success and Potential Impact:

The project can be considered successful in achieving its objectives of analyzing the relationship between predictor variables and interest rates and developing regression models to predict interest rates. The findings provide valuable insights for the client, highlighting the

factors that influence interest rates and suggesting ways to improve decision-making processes related to interest rates.

The potential impact of the findings is significant. By leveraging the developed models and recommendations, the client can make more informed decisions regarding interest rates. This can lead to improved risk assessment, better lending strategies, and enhanced profitability. Additionally, the insights gained from the analysis can help the client better understand their borrowers, optimize loan terms, and adapt to market dynamics. Ultimately, the findings have the potential to positively impact the client's competitiveness, customer satisfaction, and overall financial performance.

8 References

https://www.kaggle.com/datasets/husainsb/lendingclub-issued-loans

https://www.kaggle.com/code/dhananjayashok/lending-club-interest-rate-prediction-and-eda