Section B, Team 56

Boqing (Niko) Zheng, Genna Barge, Kainat Nazir, Yiqiao (Kevin) Wang, Rajat Bajaj

Predictive Modeling for Loan Default Risk: Balance of Return and Risk

I. Business Understanding

The banking industry encounters persistent challenges in accurately assessing the risk level associated with loan applicants. Loan defaults not only result in immediate financial losses, but also drive up the cost of capital, reduce liquidity, and impact a bank’s long-term profitability as a financial institution. The ability to predict the likelihood of loan defaults essentially enables firms to take proactive measures in managing this risk, which is essential for maintaining financial stability, competitiveness, and customer loyalty.

Failure to properly assess this default risk can lead to two problematic and prominent outcomes: excessively conservative lending practices, or overly lenient lending. Regarding the former, conservative practices may turn away high-value, low-risk customers, while lenient lending may inherently approve too many high-risk loans that end in default. Both of these situations adversely affect business operations, and therefore, influence decision-making. A more precise risk prediction enables banks to make more confident decisions regarding upfront loan approvals. Additionally, by identifying which loans are likely to default, the bank can allocate its resources more effectively. For instance, high-risk loans can be combated with safer investments, allowing the firm to create a balanced portfolio that maximizes revenue while controlling for potential losses. Overall, a more comprehensive loan approval decision-making system up front is a prerequisite for the safety and profitability of a bank's operations.

II. Expected Profit Formula

In order to assess the business impact of loan default predictions, we will utilize an expected profit formula that balances the potential risks and rewards based on the predicted default probability:

$$\text{Expected Profit} = \text{Loan Interest} \times (1 - P) - P \times \text{Loan Amount}$$

$$\text{Loan Interest} = \text{Loan Amount} \times \left(\left(1 + \frac{\text{Interest Rate}}{12} \right)^{\text{Loan Term}} - 1 \right)$$

P is the predicted probability of the borrower defaulting. Loan amount is the principal amount lended. Loan interest is calculated based on the customer's loan amount, loan term (in months) and interest rate (annual). We believe this formula to be the best encapsulation of our business problem, as it includes the potential gain from the interest earned on loans that are not defaulted, as well as the expected losses from defaults. Approving based on whether or not the expected return is positive will allow banks to consider the risk and return of each loan in a more holistic manner than an approval that only bases on the default rate, realizing an improvement in the overall return on the loan business.

III. Data Understanding

To develop a robust predictive model, we are utilizing a dataset from Kaggle containing 255,347 records of loan applications. This dataset includes key features that influence the risk of loan default, such as: borrower demographics, financial data, loan information, and our target variable, loan default. The data has been compiled from historical records within the banking sector, incorporating both demographic and financial information for each applicant. While the dataset is comprehensive, it’s crucial to acknowledge the potential biases that may be present. For example, certain demographic groups could be underrepresented, potentially leading to skewed predictions. We are committed to identifying and addressing such biases to ensure that our model remains fair and equitable in the loan approval process. Additionally, any historical bias toward groups with higher default rates must be carefully monitored to prevent disproportionate outcomes in our model's predictions.

IV. Data Preparation

After an extensive review of the dataset, we determined that no values were missing. Therefore, it was not necessary to impute or remove values or variables from the dataset. Based on the findings from our visualization results later, we transformed age into a categorical variable for different age groups to allow the model to better capture the impact of age. Next, we transformed binary variables such as the presence of a mortgage and dependents to a binary format from the given Yes/No imputation. In addition, for those categorical variables that do not have only two values (Yes/No), we processed their values with dummy coding to generate multiple columns of dummy variables to ensure that these categorical variables can be properly understood by the model. We also removed the variable InterestRate from the training data, this is because we believe that banks will have other models that give different interest rates based on each customer characteristic and market data, this is not a relevant customer characteristic and is not considered in this project. Through these various steps, we achieved a clean and uniform dataset that is optimal for model development and testing, ensuring our predictive model's reliability and efficiency.

V. Data Visualisation

We analyzed the relationship between Credit Score and Average Loan Amount, revealing a downward trend—lower credit scores are linked to higher average loan amounts(Exhibit 1). Furthermore, we introduced two new columns: CreditScoreRating and AgeGroup.

CreditScoreRating

1. **Excellent**: Scores from 720 to 850. Low risk for lenders, receiving the best interest rates
2. **Good**: Scores from 690 to 719. Still favorable, but may not receive most optimal terms

3. **Fair**: Scores from 630 to 689. May find it difficult to get loans, facing higher rates
4. **Poor**: Scores from 300 to 629. May struggle to get loans or face extremely high rates

AgeGroup

- **18-25**: Young adults who may be entering the credit market.
- **26-35**: Individuals establishing their careers and have significant financial obligations.
- **36-45**: Middle-aged adults seeking larger loans for mortgages or other investments.
- **46-55**: Older adults who may be consolidating debt or preparing for retirement.
- **56 and above**: Senior individuals who might not have any different borrowing needs

Our findings show that individuals with Poor credit rating have the highest default rates (Exhibit 2).

Insights by Age Group

Age Group 26-35: This demographic has the poorest credit scores, and the second-highest default count but takes lower avg. loan amounts. We also found high loan count indicating they take multiple loans, which strains their ability to manage payments, leading to more defaults and consequently their poor rating(Exhibit 3- Exhibit 6). We also found educational debt prevalent in this age group which may hinder their ability to repay other loans, increasing their financial strain.

Age Group 18-25: This age group records highest default count alongside the highest average loan amount secured. They have the lowest count of loans compared to other working-age groups, which can be attributed to many individuals pursuing higher education or not yet actively participating in the workforce so they don't have home or auto loans. Despite having the highest defaults, they possess the second lowest count of Poor credit ratings. This anomaly suggests that while some individuals may default on larger loans, their overall credit profiles might still be relatively favorable, possibly due to a lack of diverse credit accounts.

Conclusion

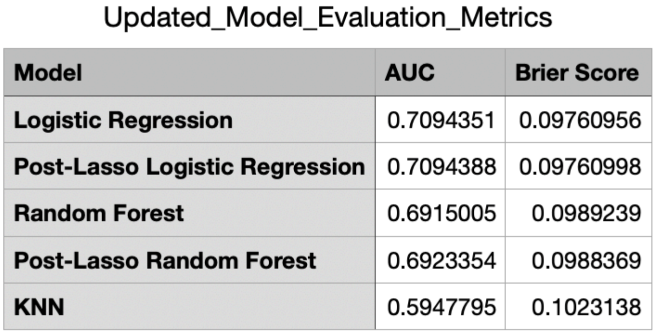
The data highlights significant variances in credit behaviors across age groups. The 26-35 age group, despite having lower loan amounts, faces challenges in managing multiple loans, particularly educational debts. Conversely, the 18-25 age group, while experiencing high defaults, appears to navigate their credit landscape differently, suggesting that age-related factors play a crucial role in shaping credit behavior and default rates. Based on the findings, we process the Age variable in the dataset.

VI. Modeling

In terms of model consideration, we did the following design. First, we chose the logistic regression model as the baseline model. Logistic regression is a linear model for binary classification problems that provides the probability of predicting that a sample belongs to a certain class. Second, we chose the random forest model. Random forest is an integrated learning method that improves the stability and accuracy of the model by constructing multiple decision trees and averaging their results. We trained the random forest model on the dataset and compared its performance with logistic regression. Next, to further improve the performance of the models, we tried the Post Lasso method (L1 regularization.) The Lasso model prevents overfitting by adding regularization terms to select important features and reduce the weights of unimportant features. By selecting the best regularization parameter (Lambda), we take all the features of the dataset and filter out some important features. We use the filtered features to train the logistic regression and random forest. Finally, we also used the KNN model as another nonparametric classification method. By traversing different K values (number of neighbors) during the cross-validation process, we finally selected the best model with a K value of 170. We intend to compare this nonparametric approach with others at its best case.

VII. Model Evaluation and Selection

We used two main metrics for evaluating the performance of each model AUC and Brier Score. AUC (Area Under the ROC Curve) measures the ability of the model to distinguish between classes (i.e., default and non-default). A higher AUC value indicates better model performance in distinguishing between both default and non-default events. And the Brier Score measures the accuracy of probabilistic predictions. It represents the mean squared difference between the predicted probability and the actual outcome (0 for non-default and 1 for default). Lower Brier scores indicate better model performance. We selected AUC because it gives a clear indication of how well the model can distinguish between defaults and non-defaults without being sensitive to the threshold chosen for classification. Brier Score was chosen because it directly measures the accuracy of the predicted probabilities, which is important for expectation calculation and decision making. The performance of our models is shown below:



Updated_Model_Evaluation_Metrics

Model	AUC	Brier Score
Logistic Regression	0.7094351	0.09760956
Post-Lasso Logistic Regression	0.7094388	0.09760998
Random Forest	0.6915005	0.0989239
Post-Lasso Random Forest	0.6923354	0.0988369
KNN	0.5947795	0.1023138

Based on our evaluation, the logistic regression model demonstrated more stable performance across iterations, with an average AUC of 0.7094 and a Brier score of 0.0976, making it more reliable compared to the other models. We found that the Post-Lasso method only excluded one feature, suggesting that most features are important in the model. Thus, the original logistic regression model slightly outperforms the Post-Lasso version overall. KNN performed the worst, with an AUC of only 0.5948 and a Brier score of 0.1023, indicating that it is unsuitable for our task. While the Random Forest model can capture complex non-linear relationships in certain cases, it achieved an average AUC of only 0.6915 and a relatively high Brier score of 0.0989, making its overall performance inferior to the logistic regression model. We placed greater emphasis on the Brier score in our model evaluation because it not only considers the accuracy of the model's predictions but also assesses the calibration of the predicted probabilities, which will later be directly used in subsequent calculations.

VIII. Simulation

In our project, the purpose of the simulation is to evaluate the performance of different models in real applications. We compared the total return and prediction accuracy of two approaches through multiple sampling and model training: the expected return approach and the traditional threshold approach. First, for each simulation, we randomly selected 5,000 samples from the complete loan dataset to ensure that the data were different for each simulation and to assess the robustness of the model under different sample scenarios. Then, for each sampled dataset, we first train a logistic regression model that is used to predict the probability that a loan is in default. Next, we calculate the expected return on each loan based on the formula defined earlier and make loan approval decisions. We use two different decision-making approaches to predict default and compare:

The expected return approach: if the expected return on a loan is negative, we predict that the loan will default (i.e., not lend the loan); if the expected return is not negative, we predict that it will not default (i.e., lend the loan).

The traditional threshold approach: when the probability of default is greater than or equal to 0.5, it is predicted to default; when the probability of default is less than 0.5, it is predicted not to default. This is a simple thrA compliance assessment procedure must be established to guarantee that legal specialists consistently evaluate the model's conformity with local rules.

eshold-based classification method.

Finally, we calculated the total return and prediction accuracy for each of the two methods:

Prediction accuracy: we compared the consistency between the predictions of each method and the actual defaults. By calculating the accuracy, we can assess the performance of the different methods on default prediction.

Total return: If it is approved and the customer does not default, it receives a profit on this loan; if it is approved and the customer defaults, it loses the entire amount of the loan; and if it is not approved, it receives nothing. We sum the total returns from each simulation to compare the overall performance of the two approaches.

After 10 cycles of simulations, the expected profit method achieved an accuracy of 76.688%, while the traditional 0.5 threshold method (logistic regression model) had a higher accuracy of 88.256%. Despite this, the expected profit method predicted a higher average total return without defaults: 228,808,084 compared to 220,061,129 for the logistic regression model. These results demonstrate that our new method improves the overall return at the cost of some accuracy loss. This offers banks a new strategy: accepting a certain level of risk in exchange for higher potential returns.

IX. Deployment

The prediction model for loan default risk will be used by the bank to make financing choices. To achieve this, the logistic regression model will calculate the risk of default for each loan application, which is subsequently utilized to compute the anticipated profit for each loan. Upon submission of a new loan application, borrower information will be instantly included into the model. If the anticipated profit is positive, the loan may be sanctioned; if not, it will be declined. This model lets them make more consistent, data-driven lending decisions based on profitability and risk management. Deployment starts with integration with the loan management system for automated forecasts, ongoing performance monitoring to enhance the model in response to evolving borrower characteristics, and regular retraining of the model with updated data to uphold accuracy.

1. Deployment Standards

Data privacy: Data use must be in compliance with the General Data Protection Regulation.

Model performance: While the model worked well during testing, real-world situations may be different. Future monitoring and modifications may be needed.

System scalability: The bank's system should be able to handle a substantial number of applications without affecting performance.

Integration with Current Systems: Seamless integration into existing workflows is key to avoid disrupting current operations.

Data Quality: The model depends on uniform and precise input data for the model's efficacy.

Model Drift: The model must be retrained often to retain credibility as borrower behavior or economic condition changes.

2. Ethical Considerations

Bias, Fairness and Transparency: Since the model uses demographic data, it may propagate social prejudices. Frequent audits and comprehensive review are essential to guarantee equity in lending practices. Financial institutions must also be transparent with borrowers about the decision-making process. Thus, reasons for loan application approvals and rejections must be explained clearly.

Impact on vulnerable groups: Ethical considerations must be made when handling high-risk individuals who may rely on loans for financial well-being. The bank should carefully decide how much risk it is willing to take and how to support these customers.

3. Risks and Mitigation Strategies

Over Reliance on the Model: Loan officers may over-rely on model output, resulting in the neglection of a few qualitative elements. Staff should be trained to use the model as a tool rather than an exclusive criterion. To mitigate this, human judgment must be used and manual overrides must be allowed when necessary.

Model Risk: Incorrect predictions may result in inefficient lending decisions. Continuous performance testing with updated data and feedback loop can help mitigate this by maintaining accuracy.

Reputational Risk: If the model is perceived as biased, the bank's reputation may suffer and they may face legal repercussions. Hence, an ethical review board must be formed for frequent evaluation and collaborations with stakeholders must be encouraged to uphold trust.

Operational Risk: Technical malfunctions may jeopardize or delay accurate loan decisions. For this, comprehensive testing before deployment is crucial, and IT assistance must be readily accessible.

Regulatory and Compliance Risk: Since lending laws may change, the model must be updated to accommodate any new constraints. A compliance assessment procedure must be established to guarantee that legal specialists consistently evaluate the model's conformity with local rules.

Bibliography

Opa, Valentine Ojong, and Wendy Tabe-Ebob. *The Effects of Loan Default on Commercial Banks Profitability: Case Study BICEC Limbe*. 21 Dec. 2020.

Provost, Foster, and Tom Fawcett. *Data Science for Business*. O'Reilly Media, Inc., 2013.

"What Happens if You Default on a Business Loan?" *Bankrate*, 21 Sept. 2023, www.bankrate.com/loans/small-business/what-if-you-default-on-business-loan/?tpt=a. Accessed 13 Oct. 2024.

Data

Nikhil. "Loan Default Prediction Dataset." *Kaggle*, 1 Oct. 2022, www.kaggle.com/datasets/nikhil1e9/loan-default.

Appendix

Exhibit 1

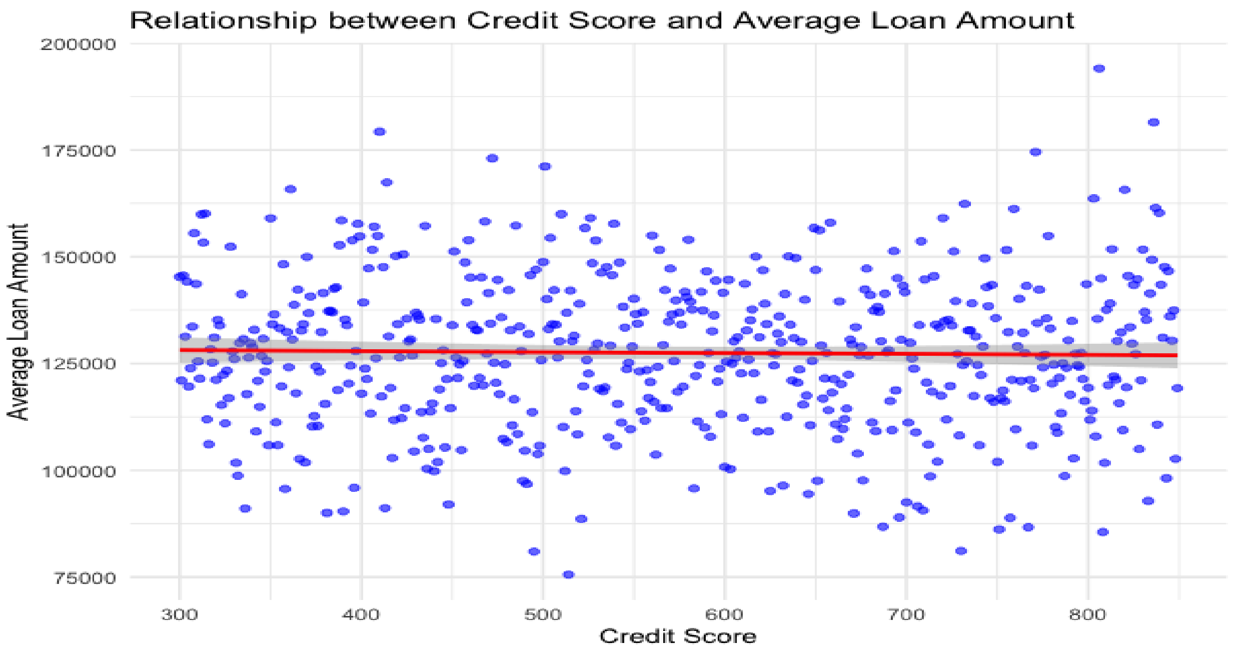


Exhibit 2

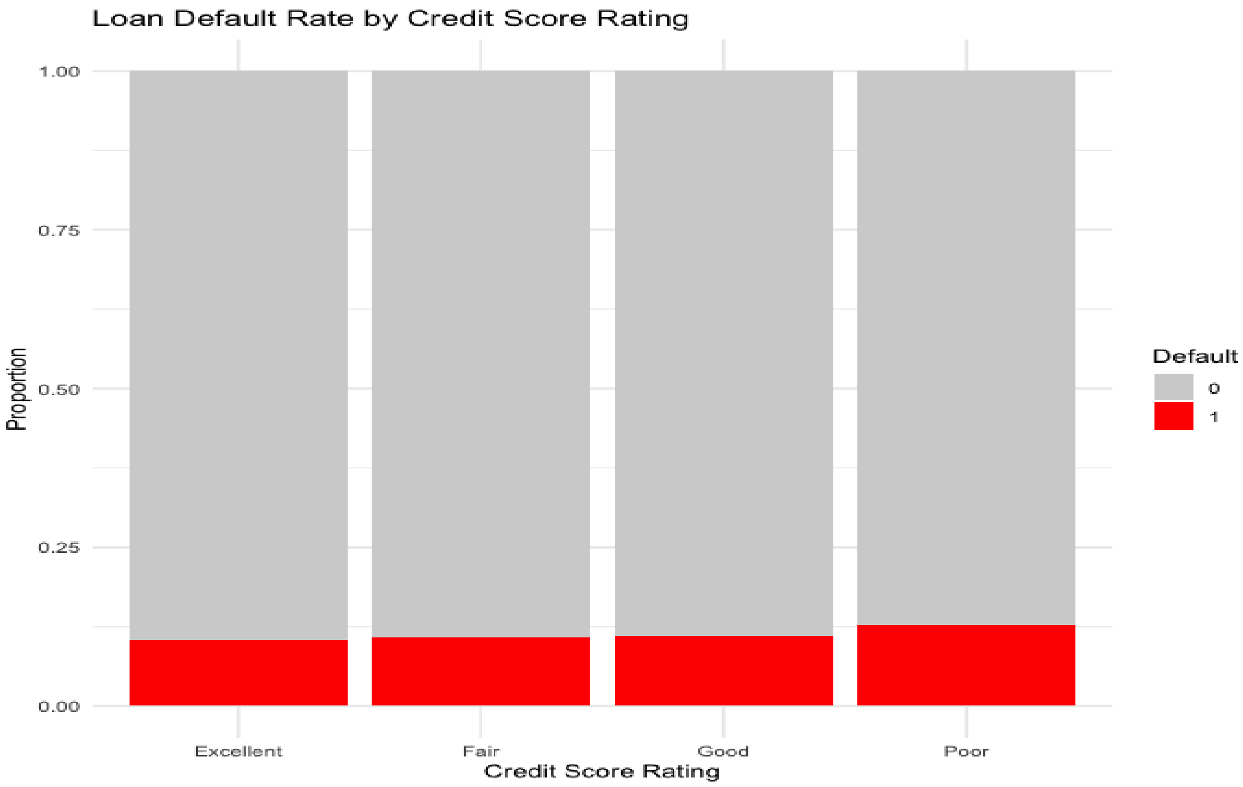


Exhibit 3

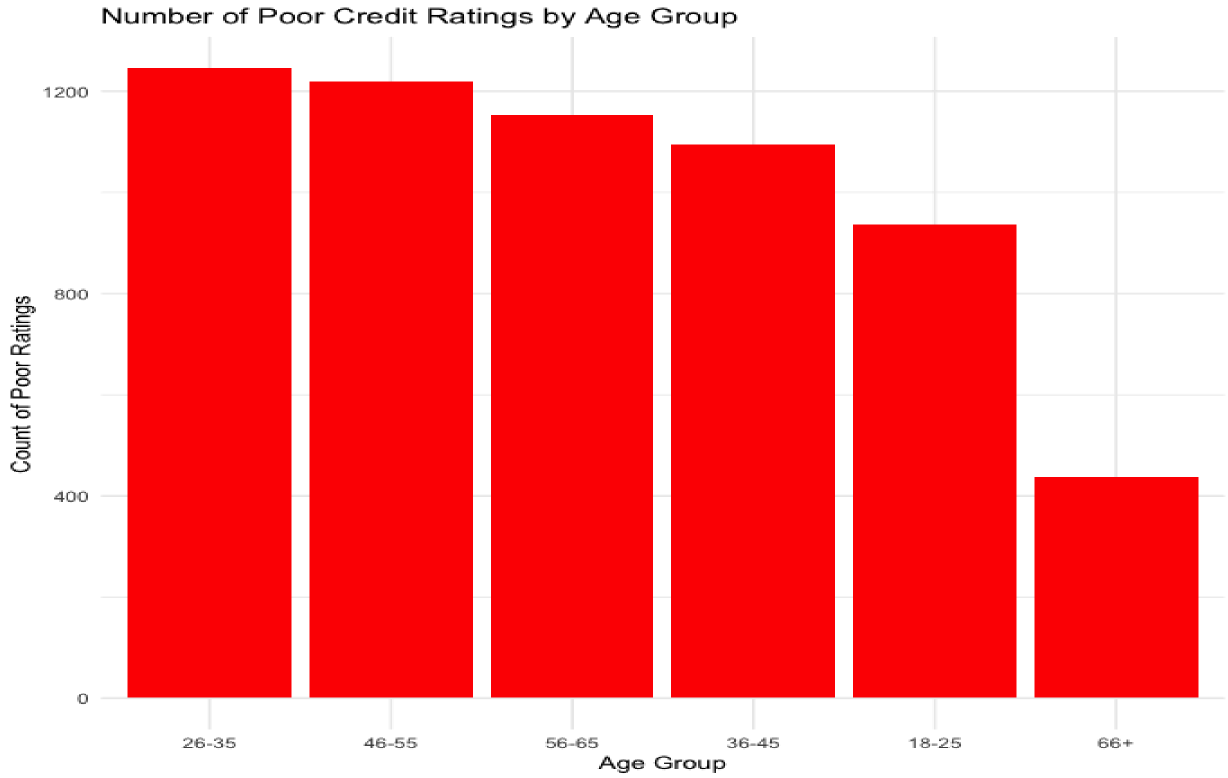


Exhibit 4

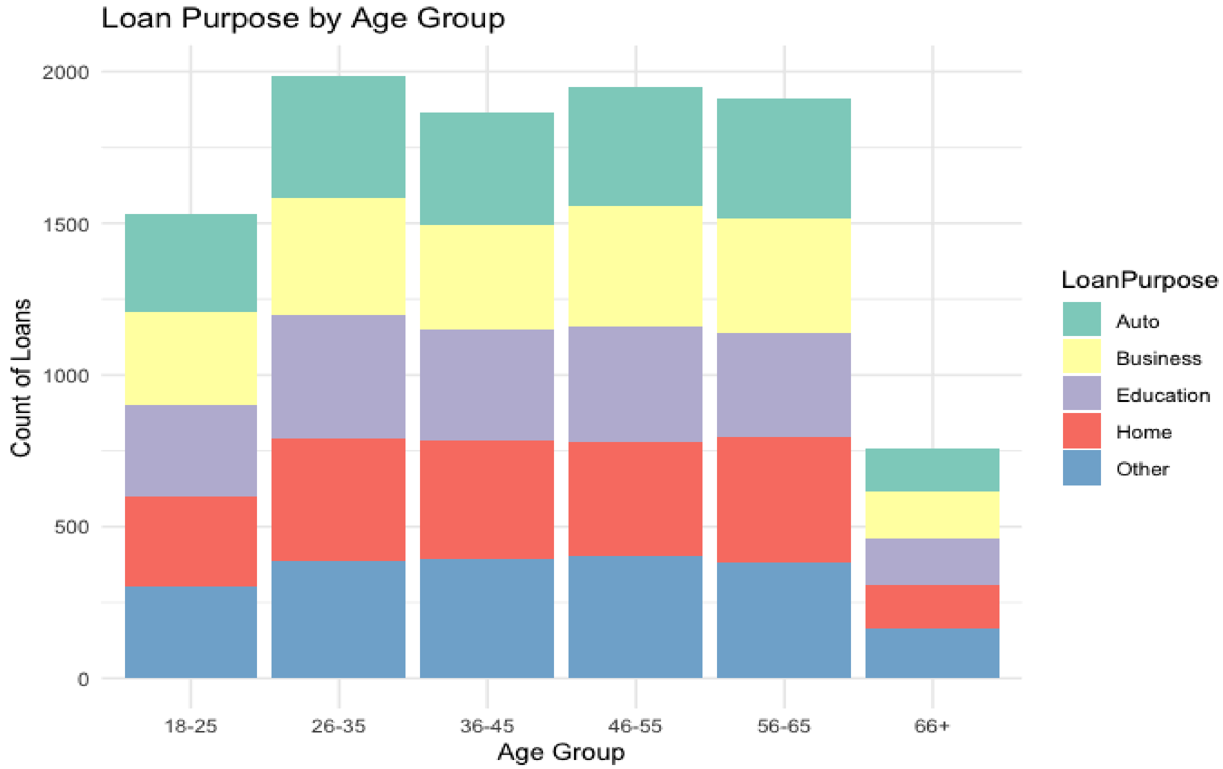


Exhibit 5

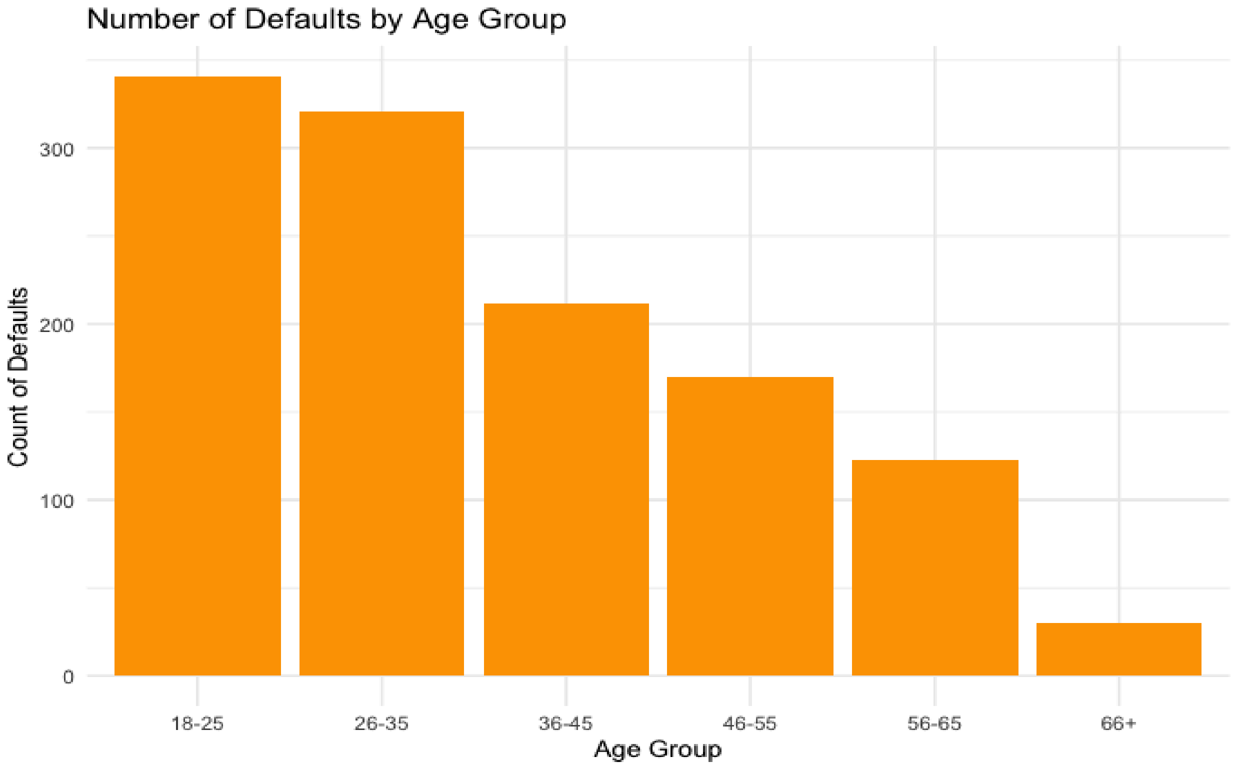
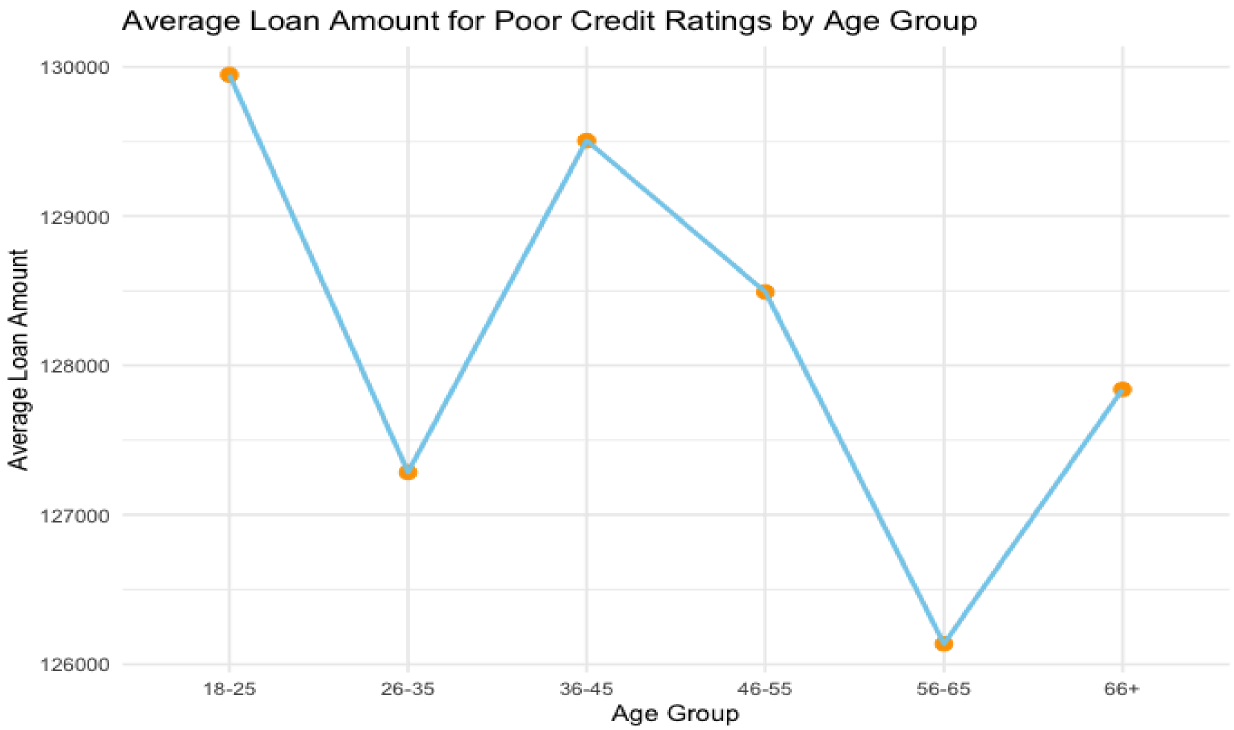


Exhibit 6



Team Member Contributions:

Niko: Modeling, evaluation, simulation, R coding and write-up

Genna: Data cleaning and preparation, business/data understanding refinement and write-up

Kainat: Deployment ideation and write-up, document editing

Kevin: Modeling, evaluation, simulation, R coding and write-up

Rajat: Data visualizations, correlation study and write-up