Assignment Stage 2

Student Name: Shaqran Bin Saleh
Student ID: 25010238

Student Name: Hasnaine Ahmed Dihan
Student ID: 25382363

Student Name: Al Jobyer Swajoy
Student ID: 25389483

Student Name: Masuma Tasnim Jerin
Student ID: 25097799

Student Name: Md Eusuf Ali Rinku
Student ID: 25215411

36100 Data Science for Innovation
TD School
University of Technology Sydney

Contents

Table of Contents

Literature Review

The biggest risk to banks is credit risk, which is the possibility of suffering losses if borrowers don't pay back their loans. A bank's ability to successfully manage credit risk depends on its ability to keep liability within reasonable bounds. Institutional constraints, poor practices, fluctuating interest rates and insufficient oversight are some of the elements that contribute to credit risk. A bank's liquidity and solvency may suffer if credit risk is not managed. Credit risk management is an important issue for banks all over the world, as profitability is nearly dependent on credit or asset return (ROA). Credit risk consistently has an impact on bank performance, according to a study that investigated the quantitative impact on the performance of commercial banks in Nigeria during an 11-year period (KOLAPO, 2012). This study was measured by return on assets (ROA) across banks in Nigeria. The higher levels of non-performing loans and loan loss provisions negatively impact a bank's profitability. Good management of credit risks affects banks profitability. A study conducted on financial performance of commercial banks in Nepal shows two important things: how often people couldn't pay back their loans on time and how much it cost the bank to give out loans (Poudel, 2012). When banks manage the credit risks, they do well. The relationship between credit risk and commercial bank performance has been a focus of emerging research in both developed and developing countries. But there is a significant inverse relationship found between commercial banks measured by ROA and credit risk measured by default rate and capital adequacy ratio (Poudel, 2012). An analysis on Kenyan banks suggests that while credit risk management may have led to a decrease in non performing loans, it did not significantly influence bank profitability (Kithinji, 2010). During the study period, profitability grew just slightly despite fluctuations in loan levels. The impact of different risk management strategies on profitability is unknown but the relationship between risk management and profits in Kenyan banks by implying that banks may relax credit standards in order to increase earnings during periods of strong credit growth. Another study of six commercial banks in Ghana indicate that there is a positive and significant relationship between bank performance and credit risk (Boahene SH, 2012). According to the author, Ghanaian banks have significant profitability during periods of elevated credit risk parameters. Unreasonably high loan rates, fees and commissions may have contributed to this. While many practical studies show negative and significant association between credit risk and the performance of commercial banks, it might be challenging to draw a conclusion on this topic because different papers have produced differing findings (Million GIzaw, 2015).

Setup

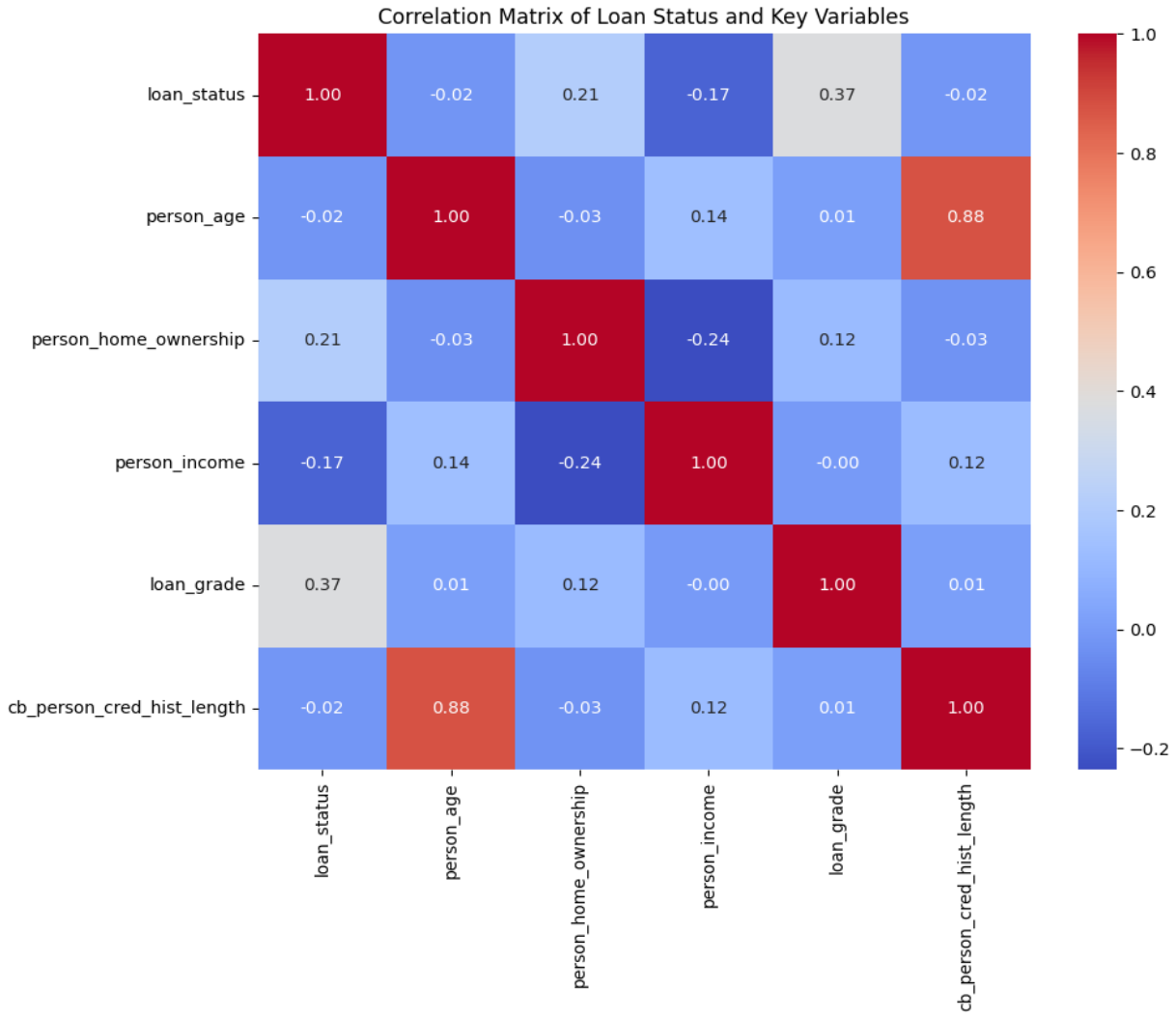

fig: Correlation Matrix

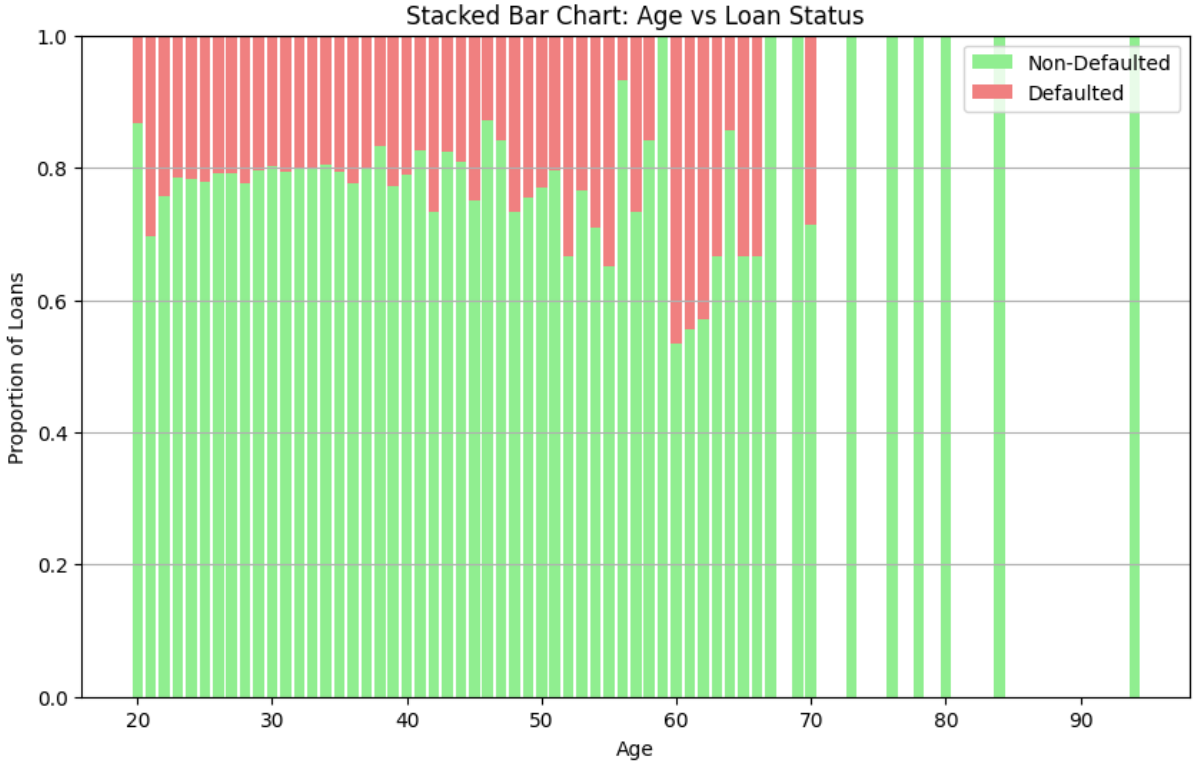

Fig: Relationship between Age and Loan Status

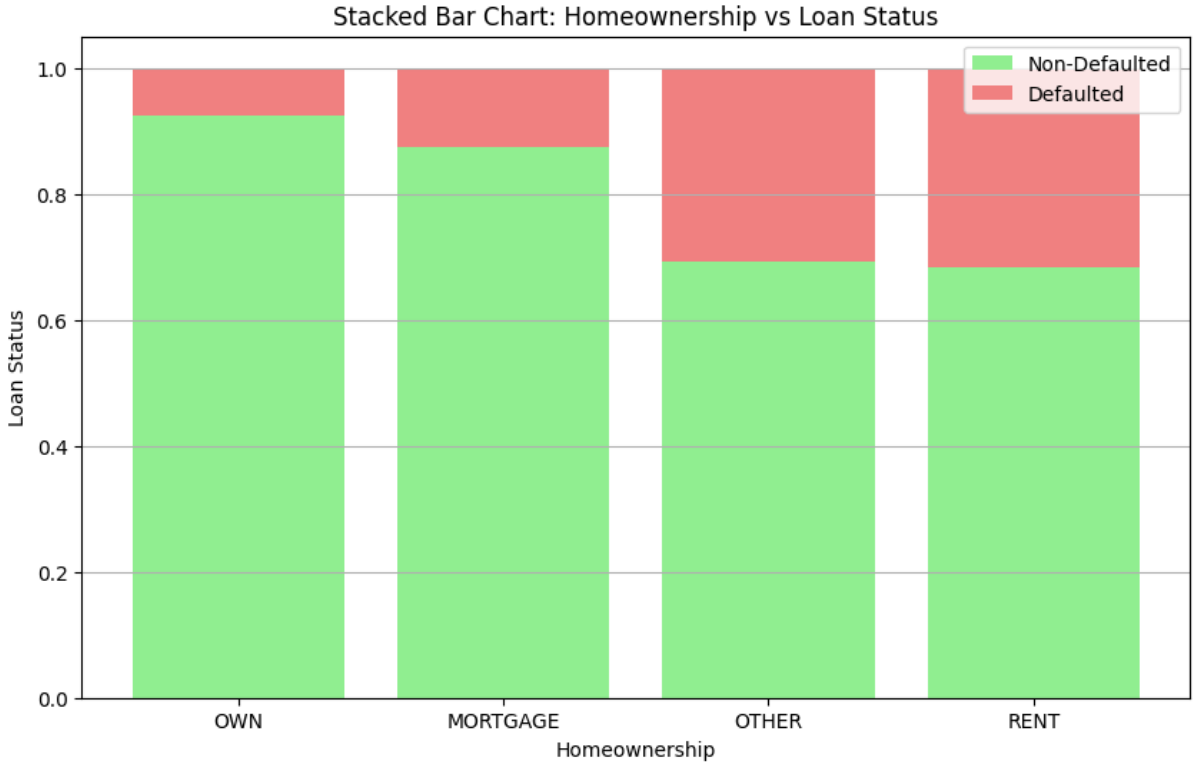

Fig: Relationship between Home Ownership and Loan Status

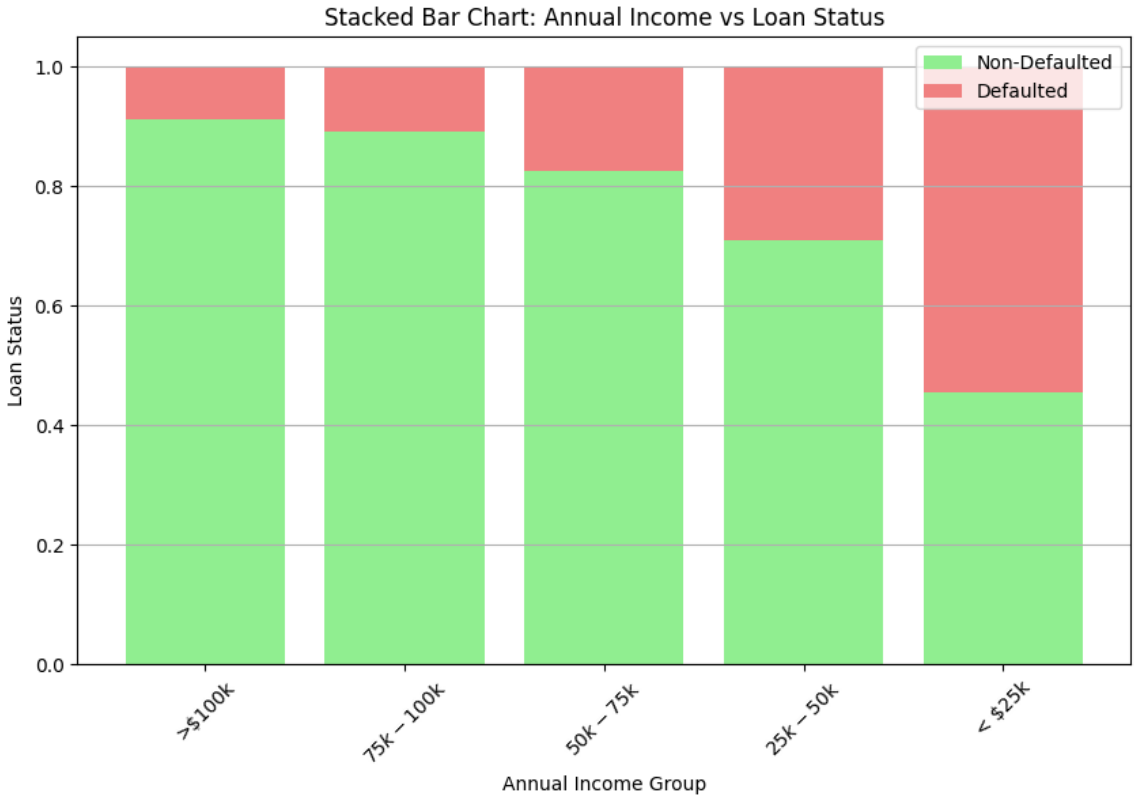

Fig: Relationship between Annual Income and Loan Status

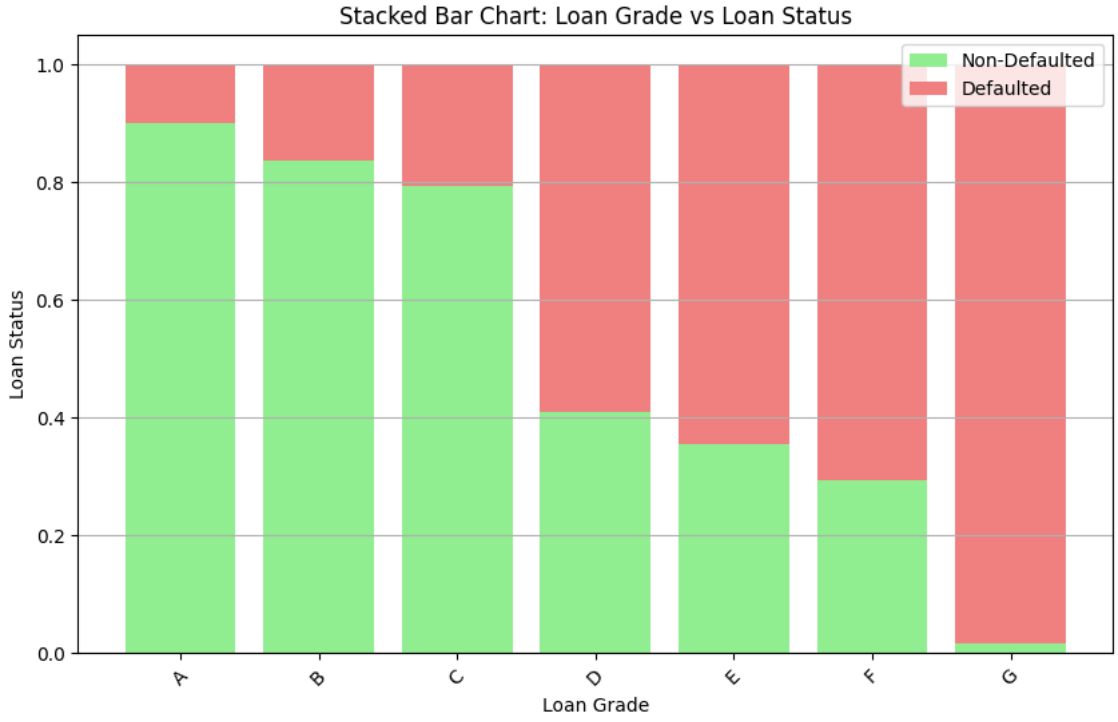

Fig: Relationship between Loan Grade and Loan Status

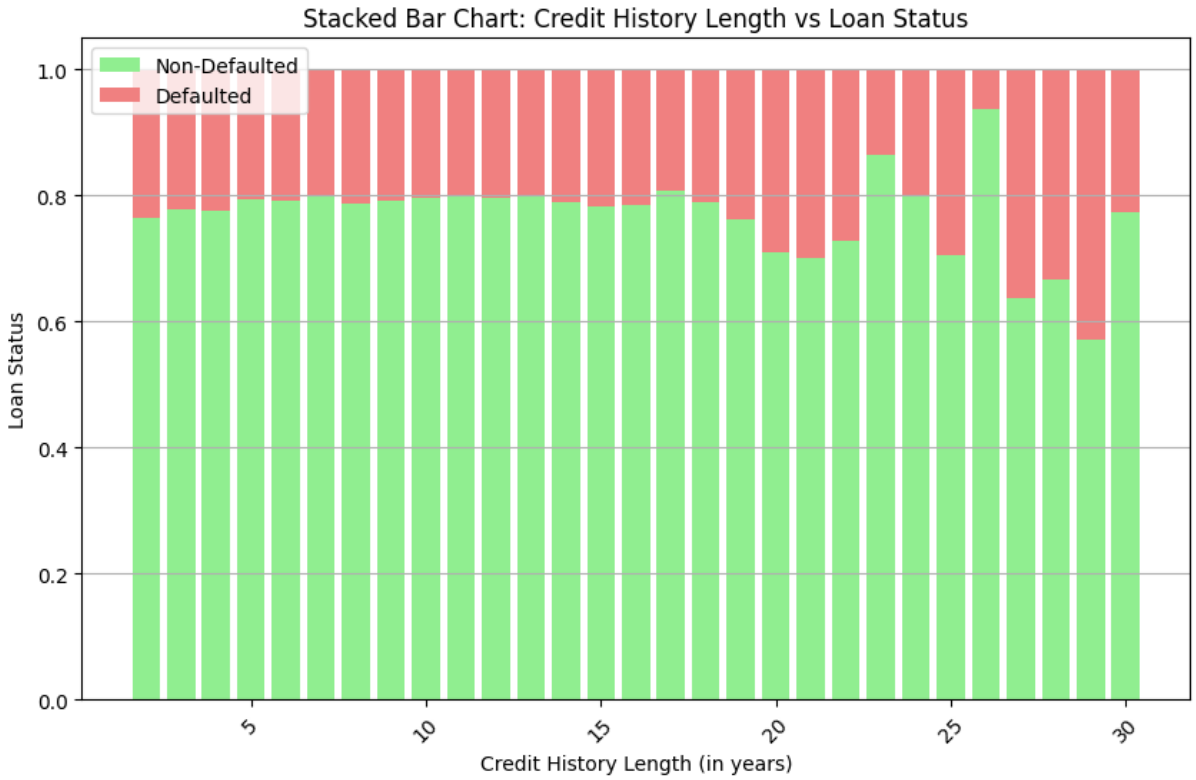

Fig: Relationship between Credit History Length and Loan Status

We've selected the study questions based on the heatmap's strong demonstration of the important correlation between loan defaults and features such as age, homeownership, income, loan grade, and credit history.

Our analysis of the correlation matrix reveals several key insights into the relationship between loan status and borrower characteristics. Person_home_ownership shows a moderate positive correlation with loan status. Loan grade exhibits a similar trend, with a moderate positive correlation to loan status. Interestingly, age and credit history length display a very weak negative correlation with loan status, indicating minimal influence. However, a very strong positive correlation exists between age and credit history length, implying that older individuals tend to have established credit histories. Finally, a moderately negative correlation between homeownership and income is observed. This suggests that individuals with higher incomes may be less likely to own homes, potentially reflecting lifestyle preferences or demographic trends, such as a higher prevalence of renting in urban areas. While these findings provide a foundation for understanding how borrower characteristics are linked to loan status, further investigation is necessary.

By hypothesis testing, we can prove that the relationships mentioned are statistically significant. This allows us to better understand which characteristics contribute a lot to loan default. And it's very important for optimizing loan distribution.

Quantifying reliability(Significance testing and confidence intervals)

We'll use significance testing and confidence intervals to quantify the reliability of our results.

In significance testing, we'll use Chi-Squared Statistics and Kruskal-Wallis-Test which will help us confirm the correlation between loan defaults and features like age, home ownership, income, loan grade, and credit history length. On the other hand, by confidence intervals, we'll predict the range where the true results can fall in.

Stating null and alternative hypothesis

We ran hypothesis testing(Chi-Squared-Test and Kruskal-Wallis-Test) on age, home ownership, income, loan grade, and credit history length to observe their effects on loan defaults. For every case, we'll calculate the P value. If the P value is less than the threshold(significance = 0.05), then we'll reject the null hypothesis.

Chi-Squared-Test:

1. **Age vs loan status:** The Null Hypothesis (H0) is "There is no association between age groups and loan defaults" and the Alternative Hypothesis (H1) is "There is an association between age groups and loan defaults". Here, the P-value is 0.000508. So we'll reject the null hypothesis.
2. **Homeownership vs loan status:** The Null Hypothesis (H0) is " There is no association between homeownership status and loan defaults" and the Alternative Hypothesis (H1) is " There is an association between homeownership status and loan defaults". As the P value is 0.0 we'll reject the null hypothesis.
3. **Annual income vs loan status:** The Null Hypothesis (H0) is " There is no association between income groups and loan defaults" and The Alternative Hypothesis (H1) is " There is an association between income groups and loan defaults". The P value is 0.0 so we'll reject the null hypothesis.
4. **Loan grade vs loan status:** The Null Hypothesis (H0) is " There is no association between loan grade groups and loan defaults" and the Alternative Hypothesis (H1) is " There is an association between loan grade groups and loan defaults". The P value is 0.0. So we'll reject the null hypothesis.
5. **Credit history length vs loan status:** The Null Hypothesis (H0) is " There is no association between credit history length and loan defaults" and the Alternative Hypothesis (H1) is " There is an association between credit history length and loan defaults". As the P value is 1.386994281147343e-05, we'll reject the null hypothesis.

Kruskal-Wallis-Test:

1. **Age vs loan status:** The Null Hypothesis (H0) is " There is no difference in default rates across different age groups" and the Alternative Hypothesis (H1) is " There is a difference in default rates across different age groups". The P value is 9.06372671051147e-05. We'll reject the null hypothesis.
2. **Homeownership vs loan status:** The Null Hypothesis (H0) is " There is no difference in default rates across different homeownership categories" and the Alternative Hypothesis (H1) is " There is a difference in default rates across different homeownership categories". The P value is 0.0. We'll reject the null hypothesis.
3. **Annual income vs loan status:** The Null Hypothesis (H0) is " There is no difference in default rates across different income groups" and the Alternative Hypothesis (H1) is "There is a difference in default rates across different income groups". The P value is 0.0. We'll reject the null hypothesis.
4. **Loan grade vs loan status:** The Null Hypothesis (H0) is " There is no difference in default rates across different loan grades" and the Alternative Hypothesis (H1) is "There is a difference in default rates across different loan grades". The P value is 0.0. We'll reject the null hypothesis.
5. **Credit history length vs loan status:** The Null Hypothesis (H0) is " There is no difference in default rates across different credit history length groups" and the Alternative Hypothesis (H1) is " There is a difference in default rates across different credit history length groups". The P value is 0.00014531599804522142. We'll reject the null hypothesis.

Our results indicate that all have a significant influence on loan defaults. In particular, homeownership and loan grade are shown to be closely associated with greater default rates. Through these experiments, we can confirm that in predictive analysis, the loan default is heavily dependent on these characteristics, which can help lenders enrich their loan risk assessments.

Measuring model effectiveness

We'll measure the effectiveness of our model by these specific metrics: R-squared for regression analysis which represents how accurate the model's predictions are compared to the actual data; V-measure, which we'll use for analyzing how closely the clusters match true group; and lastly the F1-score for classification, which we'll utilize to evaluate the balance between precision and recall in predicting loan defaults.

Approach

Proposed Model and Benchmark Models

In addressing the credit risk assessment with the provided dataset, the approach employed involves using two distinct machine learning models tailored to capture different aspects of the data characteristics. These models were chosen due to their suitability in handling binary classification tasks such as predicting loan defaults.:

Logistic Regression: This model is implemented as the benchmark due to its widespread use and effectiveness in binary classification scenarios, such as predicting whether a loan will default. Logistic Regression is particularly beneficial for its interpretability and the ease with which it handles binary outcomes. The best hyperparameters, determined through grid search, were a regularization strength (`C`) of 1 and a penalty type of `l2`. This configuration helps balance model complexity and generalization, preventing overfitting while maintaining a decent predictive capability.

Decision Tree Classifier: Chosen for its ability to handle complex, non-linear relationships between features, the Decision Tree serves as a complementary model to the Logistic Regression. It can uncover interactions and hierarchies between variables that logistic regression might miss. The model was optimized with a `gini` criterion for split quality measurement, a maximum depth of 6 to control for overfitting, and a `best` strategy for choosing splits, which are the optimal settings found via grid search.

Both models were rigorously tuned to achieve optimal performance, with grid search playing a crucial role in identifying and setting the most effective hyperparameters.

Method of Analysis

The analytical framework designed to answer the research questions on the impact of various factors on loan defaults includes the following steps:

1. **Data Preparation**: Before any analysis, the data undergoes pre-processing which includes dealing with missing values, encoding categorical variables (like home ownership and loan grade), and normalizing numerical variables to ensure each has an equal opportunity to influence the model outcome.

2. **Feature Engineering**: Insights from the correlation matrix guide the engineering of features that could potentially increase model predictive power. For instance, the stronger correlation of loan grade with loan status might suggest combining this with other features like income or home ownership to create interaction terms or new composite indicators.

3. **Correlation Analysis:** As initially observed, some variables show more promise than others in predicting defaults. For instance, loan grade shows a significant correlation with loan status, suggesting its potential as a strong predictor. This analysis helps refine the features to include in the predictive models.
4. **Hypothesis Testing:** We performed the hypothesis testing using Chi_square Statistics and Kruskal-Wallis-Test.
5. **Predictive Modeling**: Logistic Regression and Decision Tree classifiers are deployed, with performance metrics such as precision, recall, F1-score, and accuracy evaluated against a confusion matrix. This helps ascertain the effectiveness of each model in predicting defaults and allows for the comparison of their performance.
6. **Model Evaluation and Selection**: Each model is assessed based on its ability to answer the research questions effectively. The evaluation considers not only the accuracy but also the interpretability of the model, ensuring that the results can be understood and utilized by stakeholders to make informed decisions.

In this way, we can get a complete picture of the data. We'll use statistical methods and machine learning tools together to clearly identify the factors that affect loan defaults. By combining Logistic Regression and Decision Tree models, we can capture both simple and more complex relationships within the data. This will help us understand exactly what's going on behind the scenes.

Results

Our analysis looked at two different methods (logistic regression and decision trees) to predict whether someone would default on a loan (Class 1) or not (Class 0). The first method, logistic regression, worked well for identifying people who wouldn't default, but it wasn't as good at finding those who would. The second method, decision trees, was improved by fine-tuning its settings, which made it perform much better overall.

Logistic Regression performed well in identifying borrowers who would likely repay their loans (non-defaulters). It achieved a high success rate (precision) of 89% for this group and a good ability to find most non-defaulters (recall) at 95%. However, the model struggled to accurately predict borrowers at high risk of default. The success rate (precision) for defaulters was lower at 77%, and it missed a significant portion of defaulters (recall) at only 55%.

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.89	0.95	0.92	7661
1	0.77	0.55	0.64	2112
accuracy			0.87	9773
macro avg	0.83	0.75	0.78	9773
weighted avg	0.86	0.87	0.86	9773

Confusion Matrix:

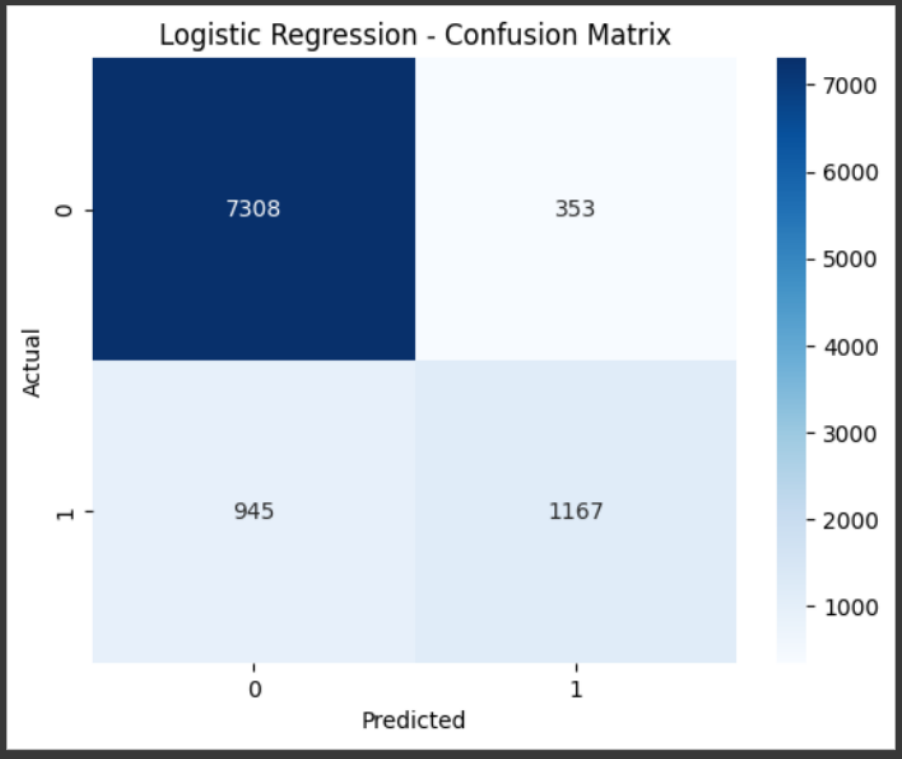

Fig: Performance evaluation of Logistic Regression after Hyperparameter Optimization

Decision Tree Classifier showed promise, particularly after fine-tuning its settings (hyperparameter optimization). Overall accuracy improved to 92%, indicating a more robust model. The ability to identify non-defaulters remained strong with a success rate (precision) of 92% and a high likelihood of finding most non-defaulters (recall) at 99%. Importantly, the optimized model significantly improved its ability to pinpoint defaulters, achieving a success rate (precision) of 94%. There was a slight trade-off, with a small decrease in finding all defaulters (recall) from 77% to 67%.

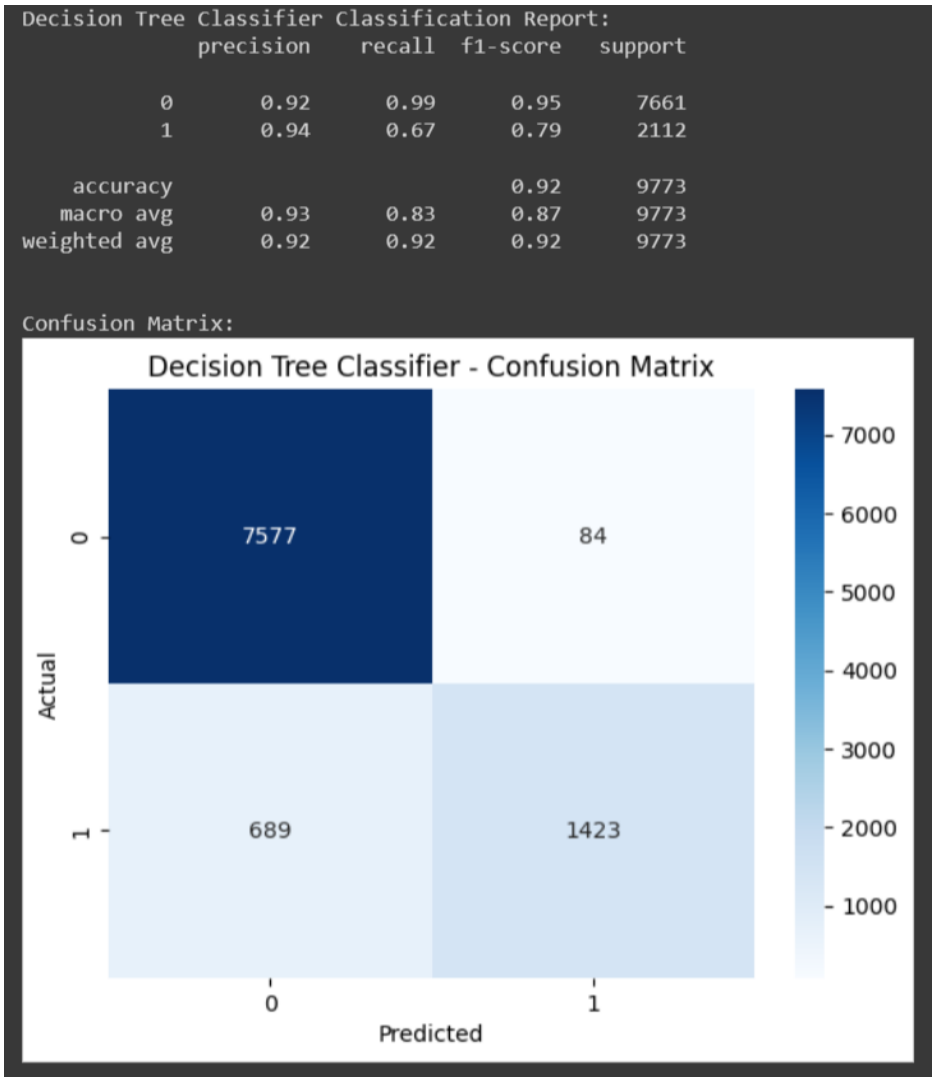

Fig: Performance evaluation of Decision Tree Classifier after Hyperparameter Optimization

In conclusion, while Logistic Regression excelled at identifying non-defaulters, the Decision Tree Classifier offered a more balanced performance across both default categories after optimization.

This suggests the Decision Tree Classifier may be a better choice for our loan default prediction needs.

Limitations:

- Data Challenge**:**
  Both models struggled with the uneven distribution of classes, specifically the underrepresentation of loan defaulters. This imbalance made it difficult to accurately predict defaults. Decision trees, in particular, showed a decrease in their ability to identify defaulters (recall) even after adjustments.
- Model Constraints**:**
  Logistic regression, due to its basic structure, may not be able to capture complex relationships between factors that influence loan default.
  Decision trees, while flexible, are prone to overfitting the training data. This was addressed to some extent by adjusting the model parameters (hyperparameter tuning), but it remains a potential limitation.
- Error Pattern**:**
  Logistic regression tended to incorrectly classify defaulters as non-defaulters (false negatives). This suggests a bias towards predicting the more frequent outcome (majority class).
  Decision trees, despite improvements, also made a significant number of false negative errors, hindering their ability to accurately identify defaulters.

Possible Improvement:

- Balancing the Data: We can improve the model's performance by making sure there are similar numbers of examples for both defaulters and non-defaulters. This can be done using techniques like SMOTE.
- Getting More Information from the Data: We can explore creating new features from the existing data, or including combinations of existing features. This might help the model identify more complex patterns.
- Building Stronger Models: By combining multiple models (like Random Forests or Boosted Trees), we can potentially create a more accurate model overall. This is because averaging the predictions from several models can make them more reliable, especially for both non-defaulters and defaulters.

Conclusion

Our study has provided significant insights into the factors influencing loan defaults, integrating traditional statistical methods with advanced machine learning techniques to enhance the prediction accuracy of credit risk. Through rigorous data exploration, hypothesis testing, and the application of models like Logistic Regression and Decision Tree Classifier, we have successfully identified key predictors of default, such as income level, homeownership, and credit history.

Despite encountering challenges such as class imbalance and model-specific limitations, our approach has led to a deeper understanding of credit risk dynamics. This research not only aids financial institutions in refining their loan approval processes but also contributes to the broader financial market by promoting responsible lending practices.

Future work will focus on addressing the limitations noted, perhaps by exploring more sophisticated ensemble techniques and acquiring more balanced datasets. Our ongoing efforts will continue to push the boundaries of predictive accuracy and risk management in the financial sector.

References

- Boahene SH, D. J. (2012). Credit risk and profitability of selected banks in Ghana. Research journal of finance and accounting.
- Kithinji, A. M. (2010). Credit risk management and profitability of commercial banks in Kenya.
- Kolapo, T. F. (2012). Credit risk and commercial banks' performance in Nigeria: A Panel Model . Australian Journal of Business and Management Research, 31-38.
- Million GIzaw, M. K. (2015). The impact of credit risk on profitability performance of commercial banks in Ethiopia. African Journal of Business Management, 59-66.
- Poudel, R. P. (2012). The impact of credit risk management on financial performance of commercial banks in Nepal. International Journal of Arts and Commerce.