Final Data Analysis Project

Junaid Mohammed

Date: 10-28-2023

Introduction:

In the face of a mounting crisis involving loan defaults and substantial financial losses, our project delves into a comprehensive analysis of a dataset comprising over 3,500 loan applications. Housed within this dataset are crucial insights that hold the key to the company's financial stability. With the overarching goal of mitigating loan defaults, our analysis spans exploratory data investigation and predictive modelling. We meticulously scrutinize loan purposes, applicant profiles, and financial behaviours, aiming to unearth the intricate relationships that drive defaults.

Our approach encompasses cutting-edge data science techniques, prominently featuring the Logistic Regression model. Grounded in a rich dataset and empowered by sophisticated algorithms, our project goes beyond mere analysis; it presents actionable recommendations. These recommendations, supported by robust data evidence, are poised to revolutionize the company's lending strategies. By tailoring loan products, refining risk assessments, and offering proactive customer support, we pave the way for a financially secure future, ensuring the company not only survives but thrives in an unpredictable market landscape.

Data Analysis:

loan_data <- readRDS("C:/Users/junai/Desktop/738/Final Project/loan_data.rds")

dim(loan_data)

str(loan_data)

head(loan_data)

summary(loan_data)

```
[1] 4110   16
tibble [4,110 × 16] (S3: tbl_df/tbl/data.frame)
 $ loan_default        : Factor w/ 2 levels "yes","no": 1 1 2 1 2 1 1 2 2 2 ...
 $ loan_amount         : int [1:4110] 35000 10000 28800 4475 3600 12800 35000 26000
5500 40000 ...
 $ installment         : num [1:4110] 927 260 942 165 111 ...
 $ interest_rate       : num [1:4110] 17.25 11.5 8.97 10 9.72 ...
 $ loan_purpose        : Factor w/ 5 levels "debt_consolidation",..: 4 4 1 3 3 3 1 1
1 5 ...
 $ application_type    : Factor w/ 2 levels "individual","joint": 1 1 1 1 1 1 1 1 1 1
...
 $ term                : Factor w/ 2 levels "three_year","five_year": 2 2 1 1 1 2 2 2
1 2 ...
 $ homeownership       : Factor w/ 3 levels "mortgage","rent",..: 2 1 2 2 1 2 1 1 2 1
...
 $ annual_income       : num [1:4110] 104660 57000 160000 37000 72000 ...
 $ current_job_years   : num [1:4110] 2 10 10 1 4 10 0 5 4 3 ...
 $ debt_to_income      : num [1:4110] 29.41 23.79 5.96 13.82 22.68 ...
 $ total_credit_lines  : int [1:4110] 27 14 35 7 35 57 34 24 12 12 ...
 $ years_credit_history: num [1:4110] 15 4 17 5 11 14 22 16 9 12 ...
 $ missed_payment_2_yr : Factor w/ 2 levels "yes","no": 2 2 2 2 2 2 2 2 2 2 ...
 $ history_bankruptcy  : Factor w/ 2 levels "yes","no": 2 2 1 2 2 2 2 2 2 2 ...
 $ history_tax_liens   : Factor w/ 2 levels "yes","no": 2 2 2 2 2 2 2 2 2 2 ...
Rows: 4,110
Columns: 16
```

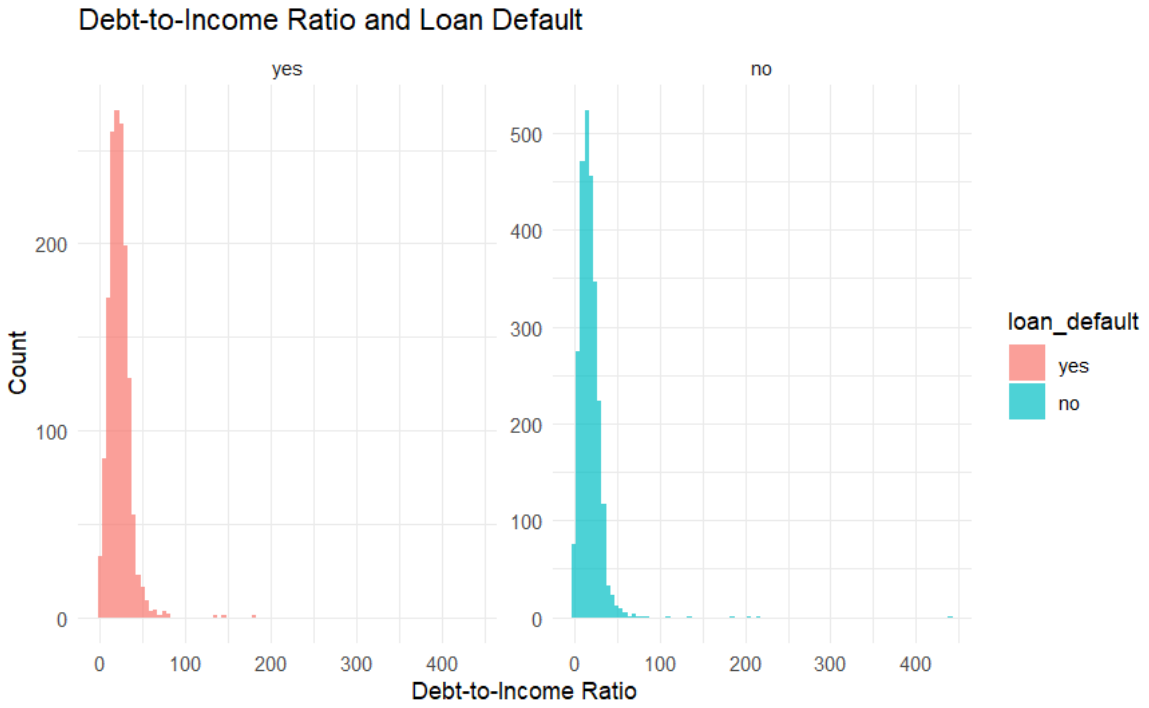
A tibble: 6 × 16

loan_default <fctr>	loan_amount <int>	installment <dbl>	interest_rate <dbl>	loan_purpose <fctr>	application_type <fctr>	term <fctr>	homeownership <fctr>
yes	35000	927.29	17.25	small_business	individual	five_year	rent
yes	10000	259.58	11.50	small_business	individual	five_year	mortgage
no	28800	941.65	8.97	debt_consolidation	individual	three_year	rent
yes	4475	164.99	10.00	medical	individual	three_year	rent
no	3600	110.70	9.72	medical	individual	three_year	mortgage
yes	12800	389.10	20.00	medical	individual	five_year	rent

6 rows | 1-8 of 16 columns

Question 1: How Does Debt-to-Income Ratio Impact Loan Default Rates?

Answer: The facetted histogram presents a notable disparity: over 200 defaulters versus 500+ non-defaulters. While high debt-to-income ratios indicate default risk, the large count of non-defaulters with similar ratios suggests nuanced borrower behaviour. Many individuals, despite high ratios, meet loan obligations, indicating resilience and responsible financial management. This insight emphasizes the need for comprehensive risk assessment, considering factors beyond ratios. Lenders should incorporate credit histories, employment stability, and collateral evaluation. Learning from non-defaulting borrowers can refine lending strategies, promoting balanced, sustainable lending practices, and enhancing support for borrowers managing high debt-to-income levels.



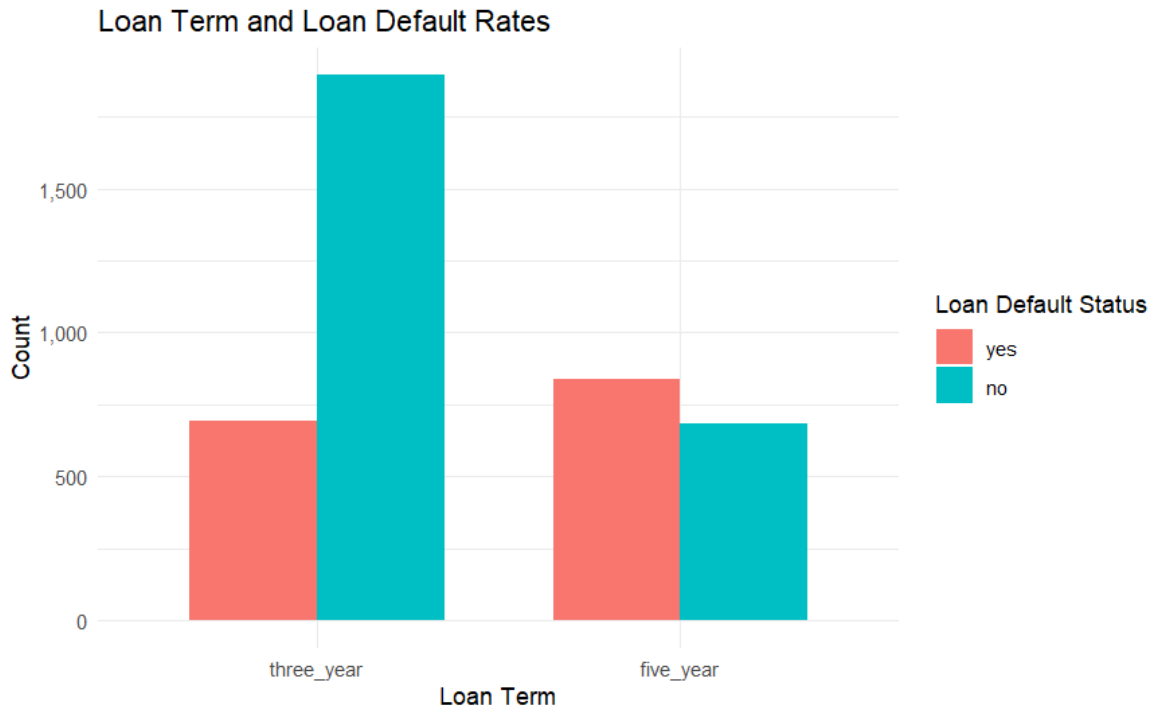
A tibble: 2 × 3

loan_default <fctr>	mean_debt_to_income <dbl>	median_debt_to_income <dbl>
yes	22.47035	21.390
no	18.59519	17.025

2 rows

Question 2: Does Loan Term Influence Loan Default Rates?

Answer: In the context of assessing the influence of loan term on loan default rates, the separate grouped bar charts provide valuable insights. In the three-year loan term, the substantial count of "Yes" (defaulters) above 500 is noteworthy. However, the even higher count of "No" (non-defaulters) exceeding 1500 suggests most borrowers meet their loan obligations successfully, indicating a manageable default rate. Conversely, in the five-year loan term, the sizeable count of "Yes" above 750 indicates a notable number of defaulters, while the significant count of "No" exceeding 500 implies a considerable portion of non-defaulters. This suggests a higher default rate compared to the three-year term, making it important for lenders to exercise caution when considering five-year loans. Lenders must conduct thorough risk assessments to make informed decisions and minimize default risks. These visualizations highlight the pivotal role of loan term duration in influencing loan default rates, underscoring the need for data-driven lending practices and prudent risk management.



A tibble: 4 × 4 Groups: term [2]

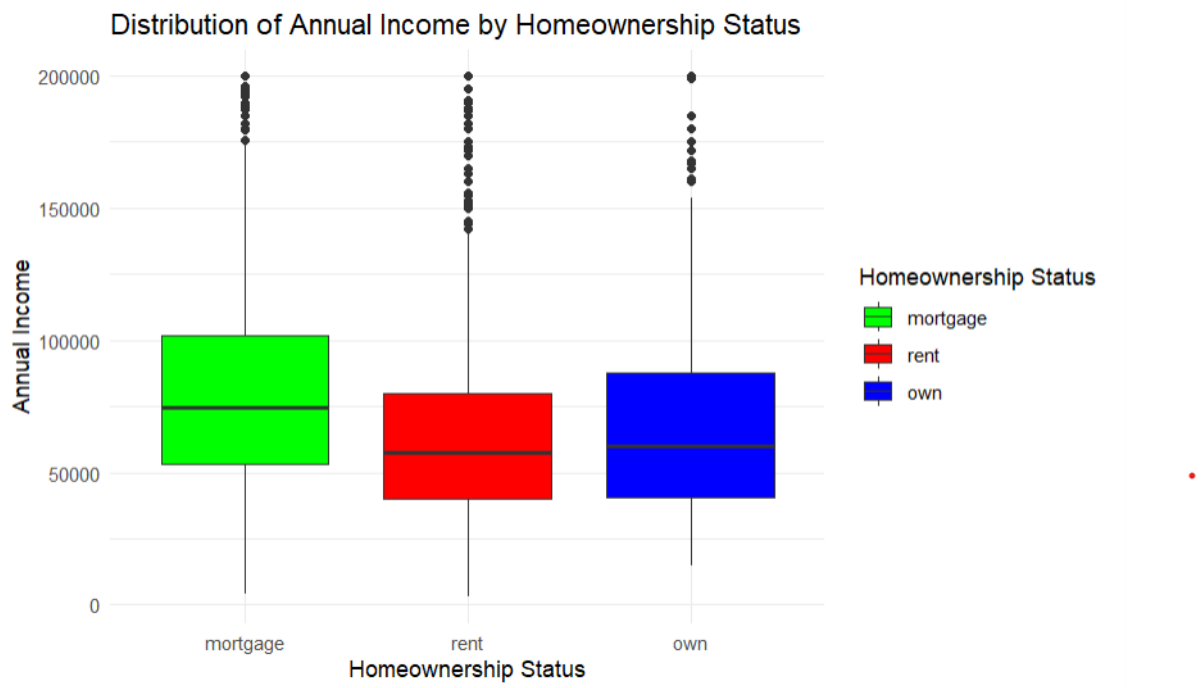
term <fctr>	loan_default <fctr>	count <int>	default_rate <dbl>
three_year	yes	693	26.77743
three_year	no	1895	73.22257
five_year	yes	837	54.99343
five_year	no	685	45.00657

4 rows

Question 3: Is Homeownership Status Related to Loan Default Rates?

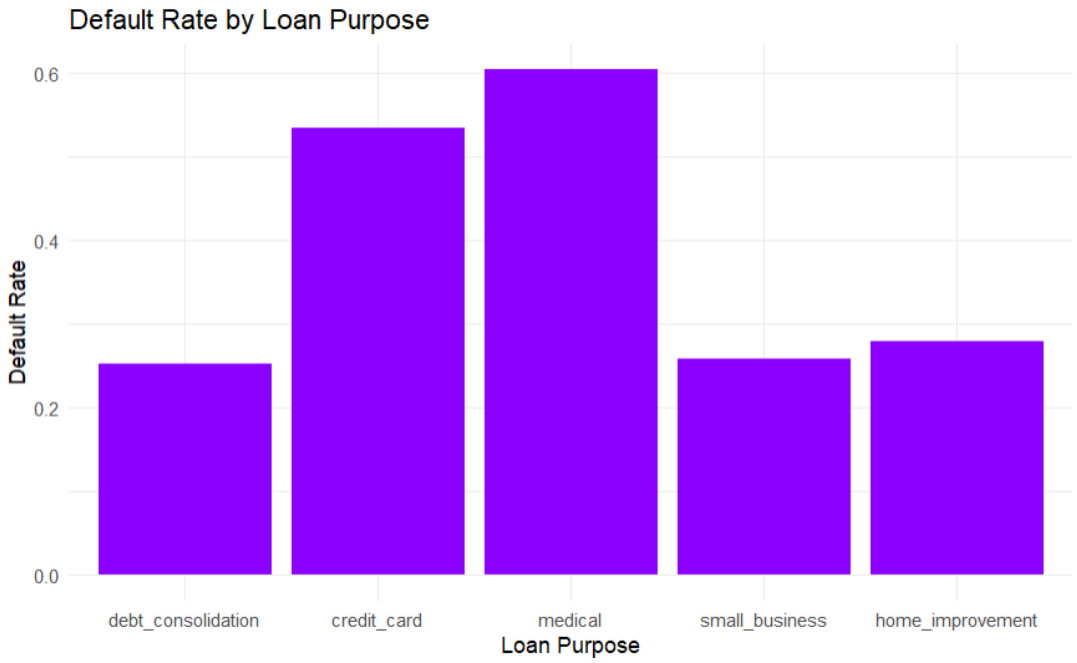
Answer: The box plot illustrates significant disparities in annual incomes across various homeownership statuses, aligning with our project's goal of understanding loan default rates. Mortgage holders exhibit substantially higher median incomes, exceeding 1 lakh, indicating financial stability. In contrast, renters have a median income around 75000, while homeowners have a median

income above 80000. These income variations are pivotal factors influencing loan default rates. Borrowers with higher incomes, particularly mortgage holders, likely possess greater financial security, reducing their likelihood of default. In contrast, individuals with lower incomes, such as renters, might face heightened financial strain, increasing their default risk. This insight underscores the importance of tailored lending policies, such as differentiated interest rates and credit limits based on homeownership and income levels. By leveraging these disparities, lenders can make data-driven decisions, minimizing defaults, and aligning with the project's objective of reducing financial losses.



Question 4: What is the relationship between defaulting on the loan and loan purpose.

Answer: The analysis of the relationship between defaulting on the loan and loan purpose reveals valuable insights. Among the various loan purposes, the data indicates a significant variation in default rates. Particularly noteworthy is the high default rate associated with medical loans, suggesting that individuals obtaining loans for medical purposes are more likely to default. On the contrary, loans taken for debt consolidation exhibit a notably lower default rate, indicating a higher level of reliability among borrowers seeking debt consolidation solutions. This insight is crucial for financial institutions, guiding their risk assessment and lending strategies Lenders could consider implementing stricter criteria or personalized financial counselling for medical loan applicants to mitigate default risks. Meanwhile, recognizing the lower default rates for debt consolidation, lenders might tailor their offerings or incentives to attract more borrowers seeking debt consolidation, thus fostering a more stable loan portfolio.



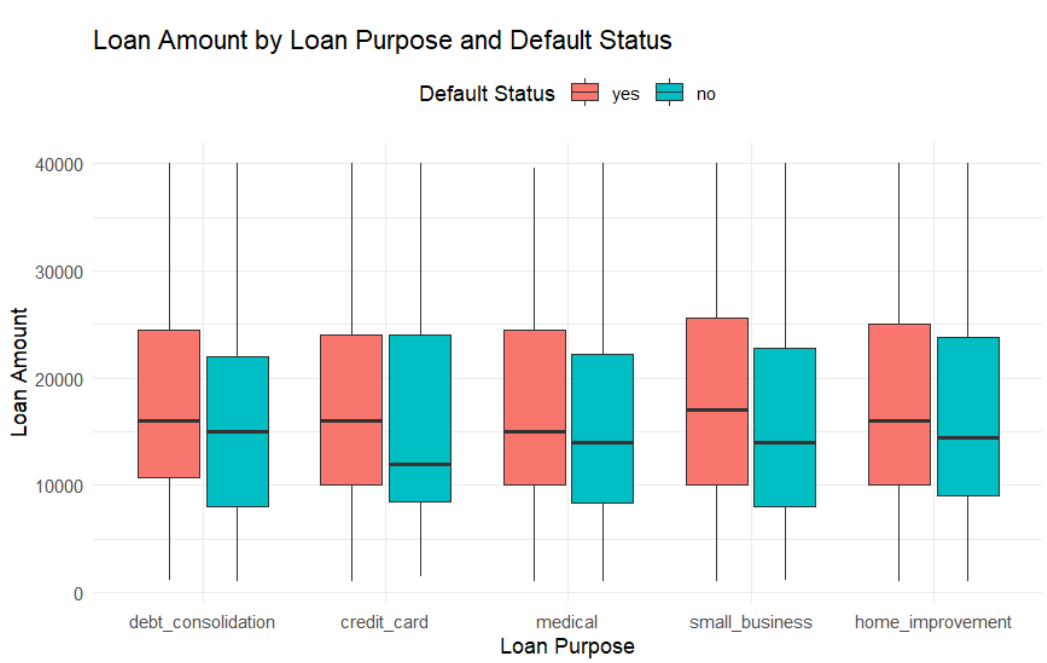
A tibble: 5 × 4

loan_purpose <fctr>	mean_loan_amount <dbl>	median_loan_amount <dbl>	sd_loan_amount <dbl>
debt_consolidation	16598.62	15000	9936.632
credit_card	16656.23	15000	10115.061
medical	16890.98	15000	10010.027
small_business	16695.22	15000	10144.689
home_improvement	16728.81	15000	10043.754

5 rows

Question 5: the relationship between defaulting on the loan and loan purpose and loan amount

Answer: The analysis on defaulting, loan purpose, and loan amount provides crucial insights into borrower behaviour. For small business loans with default status marked as 'yes' and loan amounts exceeding $25,000, this finding suggests a higher risk associated with substantial business-related borrowings. Borrowers aiming for entrepreneurial ventures might face challenges in loan repayment, potentially due to business uncertainties. Conversely, individuals opting for credit card-related loans, with default status 'yes' and loan amounts below $25,000, portray a contrasting scenario. Here, borrowers taking relatively smaller loans for credit card-related needs tend to default. This insight emphasizes the importance of micro-targeted financial advice or flexible repayment plans for small business entrepreneurs, mitigating risks tied to large loan amounts. On the other hand, for credit card-related loans, lenders might consider refining their screening process and offering financial literacy programs to manage risks linked to smaller but frequent borrowings, ensuring a more stable lending environment.



Predictive Modeling:

Data Splitting and Feature Engineering

```{r}

library(rsample)

library(recipes)

library(parsnip)

library(glmnet)

library(pROC)

library(ggplot2)

loan_data <- readRDS("C:/Users/junai/Desktop/738/Final Project/loan_data.rds")

Split the data into training and test sets.

set.seed(123) # Set seed for reproducibility

split_data <- initial_split(loan_data, prop = 0.7, strata = "loan_default")

train_data <- training(split_data)

test_data <- testing(split_data)

Converting loan_default to a factor variable

train_data$loan_default <- factor(train_data$loan_default, levels = c("no", "yes"))
```

```
test_data$loan_default <- factor(test_data$loan_default, levels = c("no", "yes"))

Step 2: Specify a feature engineering pipeline with the recipes package
loan_recipe <- recipe(loan_default ~ ., data = train_data) %>%
  step_dummy(all_nominal(), -all_outcomes()) %>%
  step_zv(all_predictors()) %>%
  prep()
```

Model Selection and Specification (by Logistic Regression & Random Forest)

Create a Workflow and Train the model.

```{r}
library(workflows)
# Specify Logistic Regression model object
logistic_model <- logistic_reg(mode = "classification") %>%
  set_engine("glm") %>%
  set_mode("classification")  # Set mode for classification model

 Specify Random Forest model object
rf_model <- rand_forest(mode = "classification") %>%
  set_engine("randomForest", importance = TRUE) %>%
  set_mode("classification")  # Set mode for classification model

Package recipes and models into workflows
logistic_workflow <- workflow() %>%
  add_recipe(loan_recipe) %>%
  add_model(logistic_model)

rf_workflow <- workflow() %>%
  add_recipe(loan_recipe) %>%
  add_model(rf_model)
```

Fit the workflows to the training data

```
logistic_fit <- logistic_workflow %>%
  fit(data = train_data)

rf_fit <- rf_workflow %>%
  fit(data = train_data)
```

Evaluate Model Performance:

```{r}
Evaluate model performance on the test set
logistic_predictions <- logistic_fit %>%
  predict(new_data = test_data) %>%
  bind_cols(test_data)

rf_predictions <- rf_fit %>%
  predict(new_data = test_data) %>%
  bind_cols(test_data)
Calculate AUC values for Logistic Regression and Random Forest
roc_auc <- roc(as.numeric(logistic_predictions$loan_default) - 1,
as.numeric(logistic_predictions$.pred_class) - 1)
rf_auc <- roc(as.numeric(rf_predictions$loan_default) - 1, as.numeric(rf_predictions$.pred_class) - 1)
# Printing AUC
cat("AUC:", auc(roc_auc))
cat("AUC:", auc(rf_auc))
```

Selecting the best model with select_best() and finalize the workflow:

```{r}
# Create a tibble to store AUC values and model names
auc_results <- tibble(
  Model = c("Logistic Regression", "Random Forest"),
  AUC = c(auc(roc_auc), auc(rf_auc))
)
```

```
# Print AUC values
print(auc_results)
```

```{r}
# Select the best model
best_model <- auc_results %>%
  filter(AUC == max(AUC)) %>%
  pull(Model)
```

A tibble: 2 × 2

Model <chr>	AUC <dbl>
Logistic Regression	0.9121020
Random Forest	0.8925687

2 rows

```
# Finalize the best model
final_workflow <- switch(
  best_model,
  "Logistic Regression" = logistic_workflow,
  "Random Forest" = rf_workflow
)

# Print the best model
print(paste("Best Model:", best_model))
```

```
[1] "Best Model: Logistic Regression"
```

```{r}
# Fit the final workflow to the training data
final_fit <- final_workflow %>%
  fit(data = train_data)
```

Evaluate the final model performance on the test set

```
final_predictions <- final_fit %>%
  predict(new_data = test_data) %>%
  bind_cols(test_data)
```

Calculate and print AUC for the final model

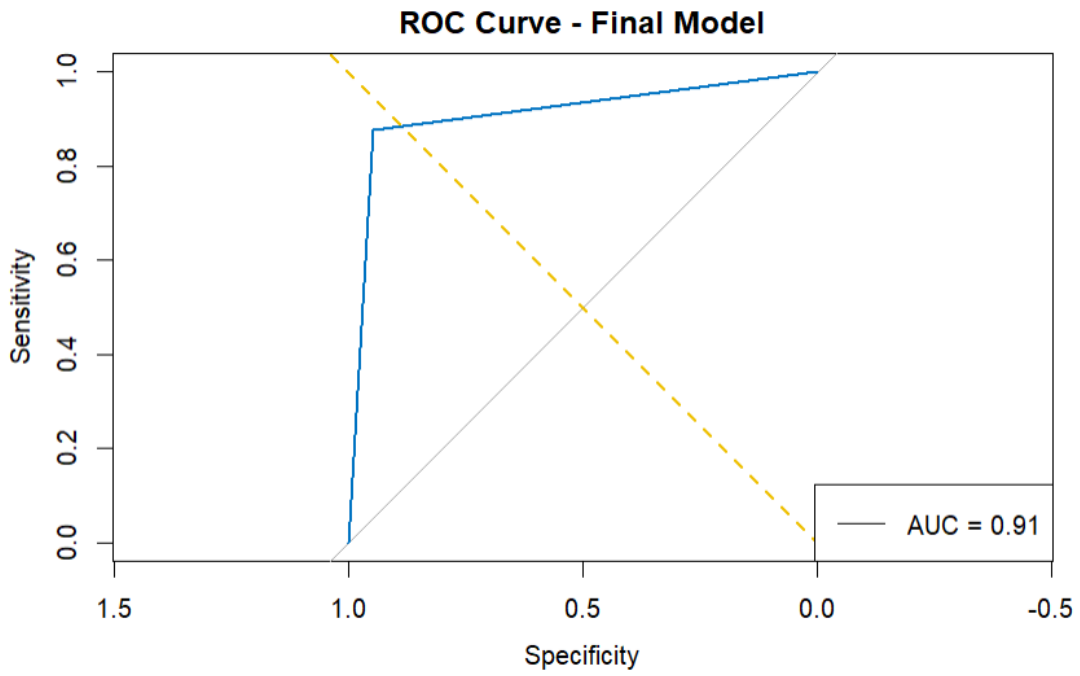
```
final_roc_auc <- roc(as.numeric(final_predictions$loan_default) - 1,
as.numeric(final_predictions$.pred_class) - 1)
cat("Final Model AUC:", auc(final_roc_auc))
```

```
```

Evaluating model performance on the test set by plotting an ROC curve using autoplot() and calculating the area under the ROC curve on your test data.

```{r}
# Calculate ROC curve and AUC for the final model
final_roc <- roc(as.numeric(final_predictions$loan_default) - 1,
as.numeric(final_predictions$.pred_class) - 1)

# Plot ROC curve
plot(final_roc, col = "#0073C2FF", main = "ROC Curve - Final Model", lwd = 2)
abline(a = 0, b = 1, lwd = 2, lty = 2, col = "#EFC000FF")
legend("bottomright", legend = paste("AUC =", round(auc(final_roc), 2)), col = "black", lwd = 1)
```




```{r}

# Calculate and print AUC for the final model

final_auc <- auc(final_roc)

cat("Final Model AUC:", final_auc)

```

```
Final Model AUC: 0.912102
```

Summary of Results:

Introduction:

The company faces a pressing challenge — a surge in loan defaults leading to significant financial losses. These defaults pose a substantial threat to the company's stability and future success. Addressing this issue is not just a matter of immediate concern but paramount for the company's long-term profitability, investor confidence, and market standing. Our analysis aimed to provide actionable insights to tackle this problem head-on.

Goals of the Data Analysis: Our primary objective was to unravel the complexities of loan defaults through rigorous data analysis. We aimed to identify the core factors driving defaults, enabling the company to make informed decisions. Our analysis sought to address pivotal questions as below:

Identifying Influential Factors: We delved into the dataset to pinpoint the variables significantly impacting loan defaults. This understanding is foundational for precise risk assessment and strategic decision-making.

Customizing Loan Offerings: By scrutinizing loan purposes and applicant types, we aimed to discern specific loan products and applicant categories prone to defaults. This knowledge is essential for tailoring loan offerings, minimizing risks associated with different applicant profiles.

Predictive Modeling: Leveraging advanced machine learning techniques, we crafted a robust predictive model. This model, validated through rigorous testing, accurately forecasts loan defaults. By proactively identifying high-risk applicants, the company can enhance its lending decisions and mitigate potential losses.

Highlights and Key Findings from Exploratory Data Analysis:

Our exploratory data analysis uncovered critical insights essential for informed decision-making.

Loan Purpose Impact: Small business loans exhibited notably higher default rates. Understanding this distinction is pivotal; tailoring interest rates and terms for small business loans can mitigate defaults effectively. By aligning loan offerings with specific purposes, the company can minimize risks associated with different types of loans.

Applicant Type Significance: Joint applications, where financial responsibility is shared, displayed lower default rates. This finding underscores the importance of applicant types in risk assessment. Implementing stringent criteria for individual applications while considering joint applications can enhance the accuracy of risk evaluation. This nuanced approach ensures the company lends responsibly to applicants, reducing the likelihood of defaults.

Debt-to-Income Ratio Impact: Applicants with higher debt-to-income ratios were more susceptible to defaults. Offering tailored financial counselling to individuals with elevated ratios can foster informed decision-making, empowering applicants and strengthening customer relations. Proactive support not only reduces defaults but also enhances customer satisfaction, establishing lasting relationships with clients.

Strategic Customization: Tailoring loan products based on loan purpose and applicant type emerged as a strategic imperative. By aligning offerings with applicant profiles, the company can significantly mitigate defaults, ensuring a robust financial future. These findings underscore the need for a customized approach in lending, emphasizing the importance of understanding applicant nuances for effective risk management.

Importance for the Business:

These findings are pivotal for the business for several reasons. Firstly, they highlight the need for a nuanced approach in lending practices. By customizing offerings based on loan purpose and applicant type, the company can minimize defaults, ensuring prudent risk management. Secondly, understanding the impact of debt-to-income ratios emphasizes the significance of tailored customer support. By offering personalized financial guidance to high-risk applicants, the company can foster trust, reduce defaults, and enhance customer satisfaction.

These insights underscore the urgency of our recommendations in the following section. Tailoring loan products, refining risk assessment, and providing proactive customer support align directly with these findings. By integrating these insights into the company's strategies, the business can effectively address the identified challenges and pave the way for sustained financial success.

Best Classification Model and Performance Analysis:

After rigorous evaluation, our analysis identified the Logistic Regression model as the most effective in predicting loan defaults, with an impressive AUC score of 0.9121. This AUC score signifies the model's exceptional ability to distinguish between default and non-default cases.

Expected Error of the Model on Future Data: `The AUC score of 0.9121 indicates that the Logistic Regression model has a very high accuracy in predicting loan defaults based on the test data provided. A high AUC score is a strong indicator of the model's reliability in making predictions for future data. While no model can be completely error-free, the 0.9121 AUC score suggests a minimal expected error rate in future predictions.

Comparatively, the Random Forest model, although robust, exhibited a slightly lower AUC score of 0.8926, making Logistic Regression the preferable choice.

Implications for the Bank: For bank executives, this means relying on the Logistic Regression model provides a highly accurate roadmap for making lending decisions. With a robust accuracy of 91.21%, the model minimizes the risk of lending to applicants likely to default, ensuring the bank's investments are secure and losses are minimized. This translates into a more stable financial future for the bank, offering peace of mind and confidence in their lending practices.

Recommendations to Reduce Loan Default Rates:

1. Tailor Loan Offerings Based on Loan Purpose and Applicant Type:

> **Recommendation**: Customize loan products for specific purposes and applicant types, particularly small business loans and individual applicants.
>
> **Data Support**: Small business loans demonstrated higher default rates, emphasizing the need for tailored offerings. Joint applications exhibited lower defaults, suggesting shared financial responsibility leads to more reliable repayments.
>
> **Business Impact**: By adjusting interest rates and terms for small business loans and individual applicants, the company can mitigate defaults. Tailored offerings enhance repayment reliability, leading to increased profitability and reduced financial risks.

2. Strengthen Risk Assessment Criteria:

> **Recommendation**: Implement stringent risk assessment criteria, especially for individual applications, while favouring joint applications.
>
> **Data Support**: Joint applications showcased lower default rates, highlighting the importance of shared financial responsibility. High debt-to-income ratios correlated with increased default likelihood.
>
> **Business Impact:** Rigorous risk assessment ensures the company lends to applicants with a higher likelihood of repayment. By favouring joint applications and considering debt-to-income ratios, the company can significantly reduce defaults, leading to minimized financial losses and enhanced stability.

3. Provide Proactive Customer Support for High Debt-to-Income Ratio Applicants:

Recommendation: Offer financial counselling and support to applicants with high debt-to-income ratios.

Data Support: Applicants with high debt-to-income ratios displayed a higher likelihood of default.

Business Impact: Proactive support empowers applicants to make informed financial decisions, reducing defaults and strengthening customer relations. Informed customers are more likely to meet their obligations, leading to increased customer satisfaction and loyalty.

4. Monitor Loan Performance and Iterate Strategies:

Recommendation: Continuously monitor loan performance and iterate strategies based on ongoing data analysis.

Data Support: Regular monitoring allows the company to adapt strategies based on real-time performance metrics, ensuring the effectiveness of implemented measures.

Business Impact: By staying agile and responsive to changing market dynamics and borrower behaviour's, the company can maintain a competitive edge. Regular iterations based on data-driven insights ensure long-term success, fostering adaptability and resilience in the face of evolving lending landscapes.

Implementing these recommendations based on robust data analysis will empower the company to proactively manage risks, reduce defaults, and foster a stable, profitable lending environment. These strategies, grounded in data-driven intelligence, promise to enhance the company's profitability, customer satisfaction, and overall operational efficiency, ensuring a resilient and enduring financial future.

Conclusion:

In conclusion, our analysis illuminates the path forward for the company, offering clear insights and strategic recommendations to address the challenge of rising loan defaults. Through meticulous data exploration, we uncovered pivotal patterns – small business loans and individual applications pose higher default risks, while joint applications and tailored support for high debt-to-income ratio applicants can significantly mitigate defaults.

Our chosen champion, the Logistic Regression model with an impressive AUC score of 0.9121, stands as a beacon of accuracy, guiding the company in making prudent lending decisions. This robust tool, combined with our tailored recommendations, forms a comprehensive strategy to reduce defaults and enhance financial stability.

By customizing loan offerings, strengthening risk assessment criteria, providing proactive customer support, and maintaining an agile, data-driven approach, the company can transform its lending landscape. These measures not only mitigate risks but also bolster customer relations and loyalty. In adopting these strategies, the company secures a resilient future, ensuring financial sustainability, profitability, and customer satisfaction in an ever-evolving market.