Loan default analysis

Data importation

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(tidymodels)
```

```
## ── Attaching packages ─────────────────────────────────────── tidymodels
1.1.1 ──
## ✓ broom        1.0.5     ✓ rsample      1.2.0
## ✓ dials        1.2.0     ✓ tune         1.1.2
## ✓ infer        1.0.5     ✓ workflows    1.1.3
## ✓ modeldata    1.2.0     ✓ workflowsets 1.0.1
## ✓ parsnip      1.1.1     ✓ yardstick    1.2.0
## ✓ recipes      1.0.8
## ── Conflicts ────────────────────────────────────────
tidymodels_conflicts() ──
## ✖ scales::discard() masks purrr::discard()
## ✖ dplyr::filter()   masks stats::filter()
## ✖ recipes::fixed()  masks stringr::fixed()
## ✖ dplyr::lag()      masks stats::lag()
## ✖ yardstick::spec() masks readr::spec()
## ✖ recipes::step()   masks stats::step()
## • Use tidymodels_prefer() to resolve common conflicts.
```

```
loans_df <- read_rds("C:/Users/Administrator/Desktop/loan_data.rds")
loans_df<-data.frame(loans_df)
head(loans_df)
```

```
##   loan_default loan_amount installment interest_rate       loan_purpose
## 1          yes       35000      927.29         17.25     small_business
```

```
## 2          yes       10000      259.58         11.50     small_business
## 3           no       28800      941.65          8.97 debt_consolidation
## 4          yes        4475      164.99         10.00            medical
## 5           no        3600      110.70          9.72            medical
## 6          yes       12800      389.10         20.00            medical
##   application_type       term homeownership annual_income
current_job_years
## 1       individual  five_year          rent        104660
2
## 2       individual  five_year      mortgage         57000
10
## 3       individual three_year          rent        160000
10
## 4       individual three_year          rent         37000
1
## 5       individual three_year      mortgage         72000
4
## 6       individual  five_year          rent         73000
10
##   debt_to_income total_credit_lines years_credit_history
missed_payment_2_yr
## 1          29.41                27                   15
no
## 2          23.79                14                    4
no
## 3           5.96                35                   17
no
## 4          13.82                 7                    5
no
## 5          22.68                35                   11
no
## 6          30.94                57                   14
no
##   history_bankruptcy history_tax_liens
## 1                 no                no
## 2                 no                no
## 3                yes                no
## 4                 no                no
## 5                 no                no
## 6                 no                no
```

Data analysis

Question 1: Is there a significant difference in the loan amount between borrowers who defaulted on their loans and those who did not?

Summary dataframe showing Average loan amount by loan default

```
plotdata <- loans_df %>%
group_by(loan_default) %>%
summarize(mean_loan_amount = mean(loan_amount))
plotdata
```

```
## # A tibble: 2 × 2
##   loan_default mean_loan_amount
##   <fct>                   <dbl>
## 1 yes                    17448.
## 2 no                     16245.
```

Bar plot showing Average loan amount by loan default

```
ggplot(plotdata,
aes(x = loan_default,
y = mean_loan_amount)) +
geom_bar(stat = "identity", fill="blue")+
  labs(title = "Average loan amount by loan default")
```

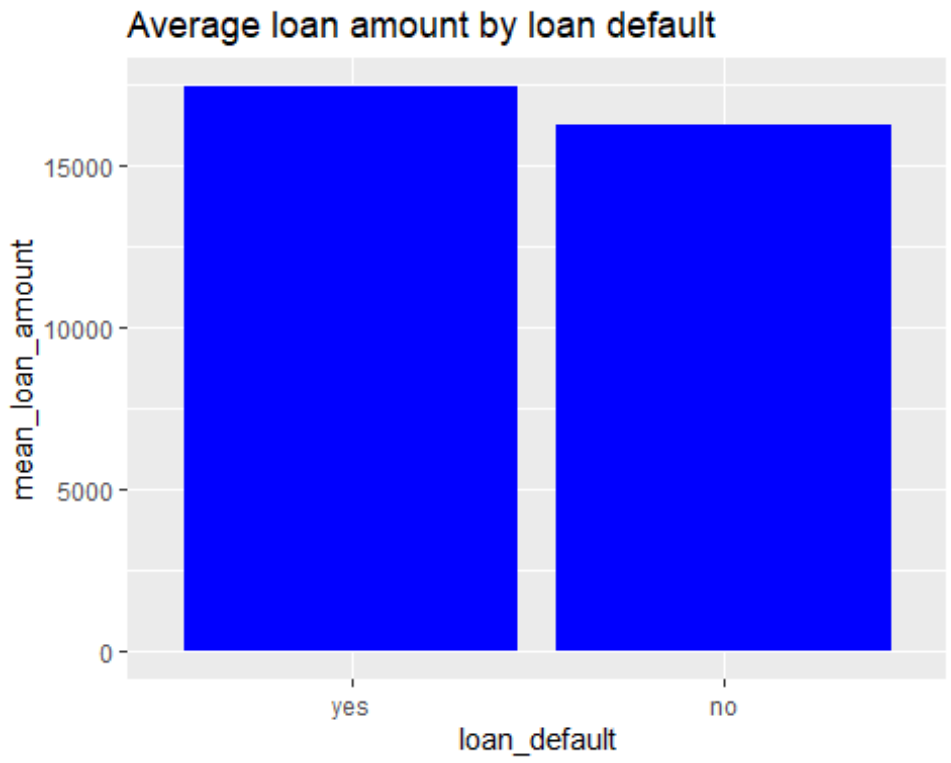

Figure 1

Answer: Based on the results from the bar graph above and the summary data frame above, there is a difference in the loan amount between borrowers who defaulted on their loans and those who did not. The average amount taken by those who defaulted on payment was 17,447.53, while the average amount for those who did not default on loan payment was 16,245.21. Those who took bigger loans were more likely to default.

Question 2: Does the interest rate offered to borrowers have an impact on their likelihood of defaulting on loan payments?

Summary dataframe showing Interest rate distribution by likelihood of loan defaulting

```
plotdata <- loans_df %>%
group_by(loan_default) %>%
summarize(mean_interest_rate = mean(interest_rate))
plotdata
```

```
## # A tibble: 2 × 2
##   loan_default mean_interest_rate
##   <fct>                     <dbl>
## 1 yes                       14.9
## 2 no                         9.30
```

Boxplot showing Interest rate distribution by likelihood of loan defaulting

```
ggplot(loans_df, aes(x = loan_default, y = interest_rate)) +
geom_boxplot(notch = TRUE,
fill = "cornflowerblue",
alpha = .7) +
labs(title = "Interest rate distribution by likelihood of loan defaulting")
```

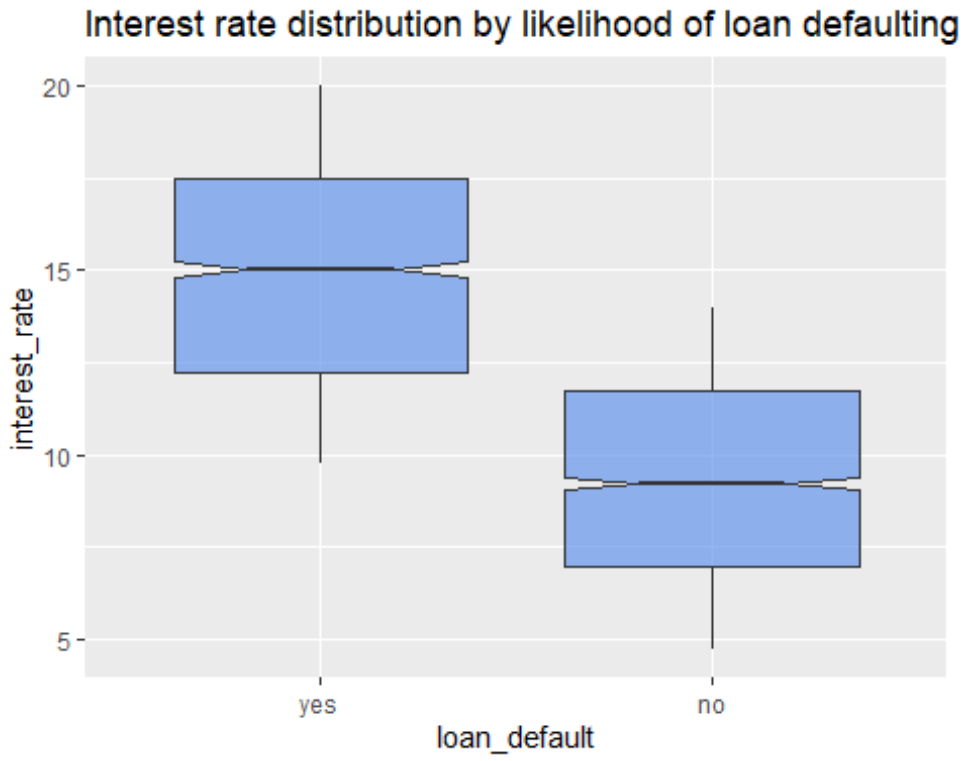

Figure 2

Answer: Yes, the interest rate offered to borrowers have an impact on their likelihood of defaulting on loan payments. The average interest rate among those who defaulted on loan payment was 14.89% while the average interest rate among those who did not default was 9.30%.

Question 3: Is there an association between the loan term and the likelihood of loan default?

Grouped bar chart showing Lona Default By Loan Term

```
# grouped bar plot
ggplot(loans_df,
aes(x = term,
fill = loan_default)) +
geom_bar(position = "dodge")+
  labs(title = "Lona Default By Loan Term", x="Loan Term", y="Frequency")
```

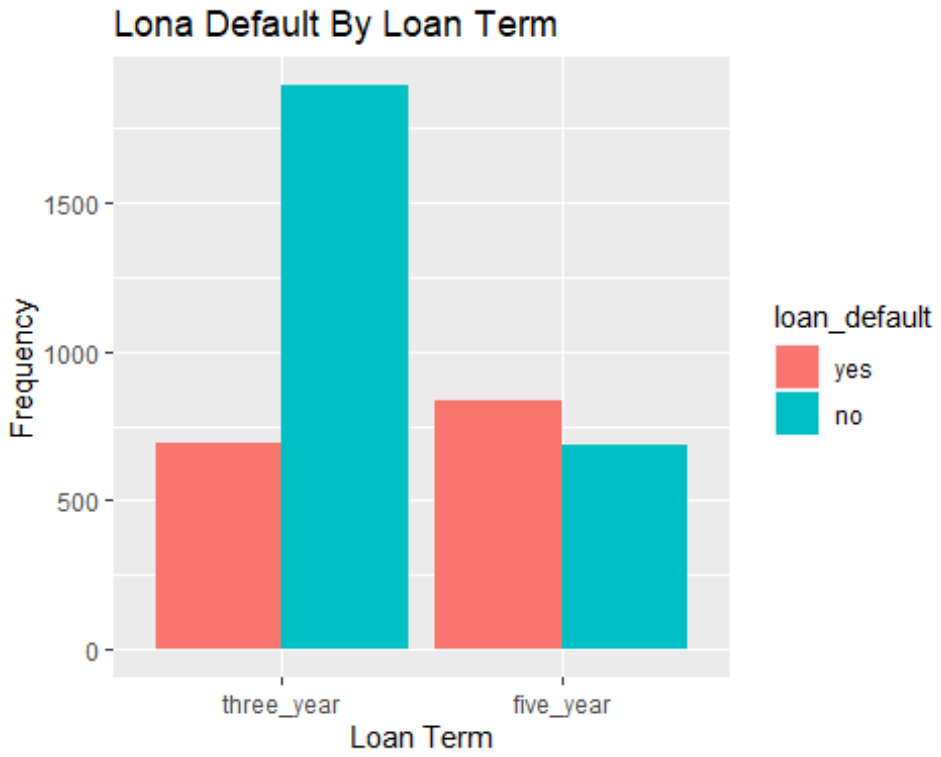

Figure 3

Answer: Yes, there is an association between the loan term and the likelihood of loan default. The results from Figure 3 above show that the likelihood of loan default was higher among five-year loan terms compared to three-year loan terms. The rate of default in the five-year loan term was higher than the rate of payment within the five-year loan term.

Question 4: Is there a relationship between defaulting on the loan and loan purpose?

Grouped bar chart showing the relationship between defaulting on the loan and loan purpose

```
# grouped bar plot
ggplot(loans_df,
aes(x = loan_purpose,
fill = loan_default)) +
geom_bar(position = "dodge")+
  labs(title = "Lona Default By Loan Purpose", x="Loan Purpose",
y="Frequency")
```

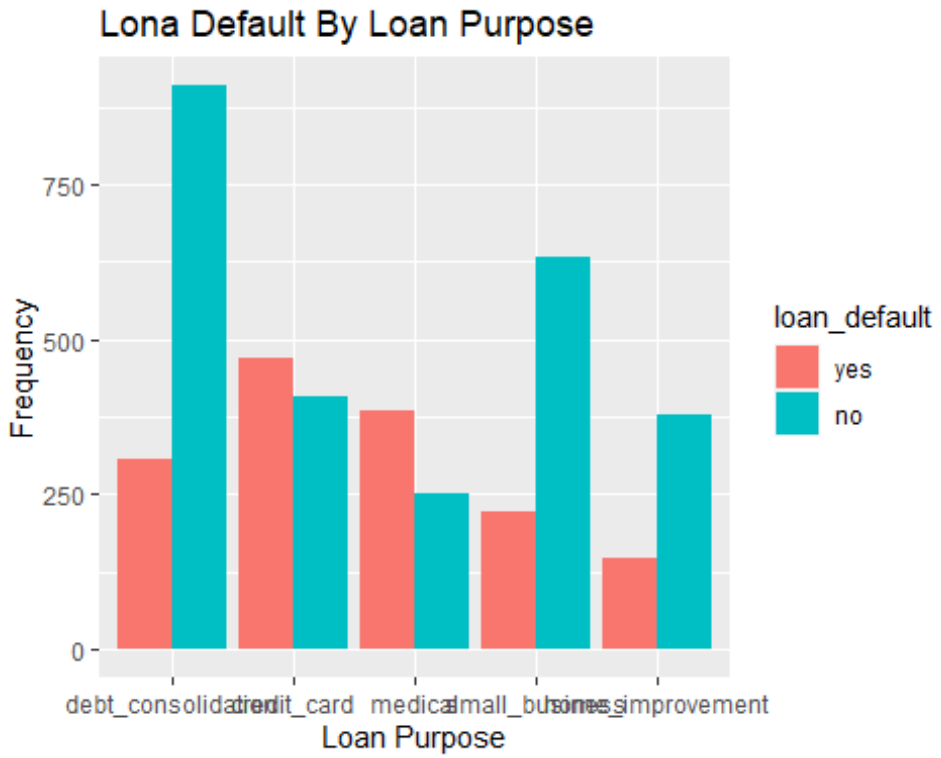

Figure 4

The results in figure 4 above shows that the loan default rate was associated with the purpose of the loan. The default rate was higher among those people who took the loan for credit card purposes, and medical expenses. The default rate was low among those who took the loan for debt consolidation, small business and home improvement.

Question 5: What is the relationship between loan defaulting and loan purpose as well as loan amount?

Summary dataframe showing Loan amount by Loan purpose and Loan default

```
plotdata <- loans_df %>%
group_by(loan_default, loan_purpose) %>%
summarize(mean_loan_amount = mean(loan_amount))

## `summarise()` has grouped output by 'loan_default'. You can override using
the
## `.groups` argument.

plotdata

## # A tibble: 10 × 3
## # Groups:   loan_default [2]
##    loan_default loan_purpose       mean_loan_amount
##    <fct>        <fct>                         <dbl>
##  1 yes          debt_consolidation           17704.
```

```
##  2 yes          credit_card                17076.
##  3 yes          medical                    17058.
##  4 yes          small_business             18351.
##  5 yes          home_improvement           17755.
##  6 no           debt_consolidation         16224.
##  7 no           credit_card                16173.
##  8 no           medical                    16635.
##  9 no           small_business             16116.
## 10 no           home_improvement           16330.
```

Bar graph showing Loan default based on Loan amount and Loan purpose

```
ggplot(plotdata,
aes(x = loan_purpose,
y = mean_loan_amount)) +
geom_bar(stat = "identity", fill="blue")+
  labs(title = "Loan default based on Loan amount and Loan purpose")+
  facet_wrap(~loan_default) +
  coord_flip()
```

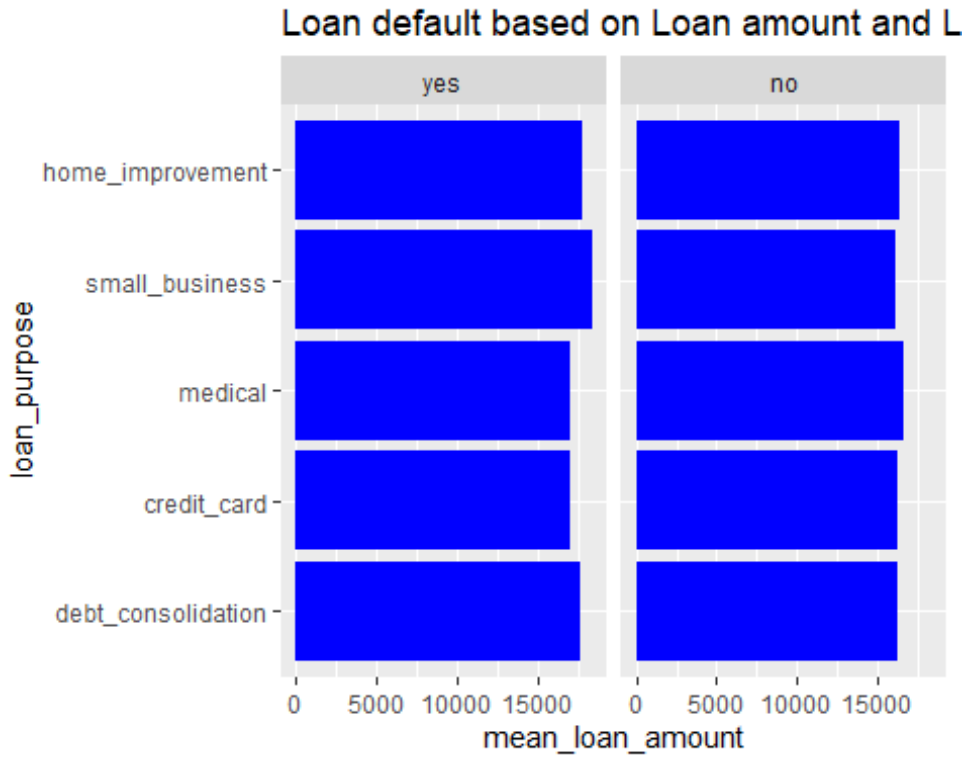

Figure 5

The results from Figure 5 above indicate that the rate of default was lower among individuals who took a loan of a larger amount (M =16224) for debt consolidation compared to those who took a loan of a lower amount (M = 17704) for debt consolidation. The loan amount had a bigger influence on whether a person defaulted on loan payments for individuals who took the loan for small business purposes. Individuals who took a

bigger loan for small businesses were more likely to default on payment. The amount of loan taken did not have a bigger impact on whether a person defaulted on loan payments or not for loans taken for medical purposes.

Predictive analysis

In this section, we fitted two classification algorithms (logit model and Linear Discriminant Analysis (LDA) model) to predict the response variable, loan_default. This study used all of the other variables in the loans_df data as predictor variables for each model.

Loading necessary libraries

```
# Load necessary libraries
library(caret)

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity

## The following object is masked from 'package:purrr':
##
##     lift

library(recipes)
library(parsnip)
library(yardstick)
library(ROCR)
library(caret)
library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

library(tidymodels)
library(tidyverse)
library(MASS)

##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

Data splitting and feature engineering steps were only conducted once so that the models will be using the same data and feature engineering steps for training.

Feature engineering

This study used normalization, scaling and one-hot encoding as prefered feature engineering techniques.

```
# Set a seed for reproducibility
set.seed(123)

# Step 1: Feature engineering on the original data
loan_recipe <- recipe(loan_default ~ ., data = loans_df) %>%
  # Scale numeric variables
  step_scale(all_numeric()) %>%
  # One-hot encode categorical variables
  step_dummy(all_nominal())
```

Splitting the dataset

```
# Split the data into a training and test set using caret
set.seed(123)
splitIndex <- createDataPartition(loans_df$loan_default, p = 0.7,
                                  list = FALSE,
                                  times = 1)
train_data <- loans_df[splitIndex, ]
test_data <- loans_df[-splitIndex, ]

# Split the training data into 5 folds for 5-fold cross-validation
cv_folds <- createFolds(train_data$loan_default, k = 5)
```

Specifying parsnip model object

```
# Specify a parsnip model object (Logistic Regression)
logit_model <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")
```

Fitting the workflow and packaging the recipe

```
# Step 4: Package your recipe and model into a workflow for Logistic
Regression
workflow_logit <- workflow() %>%
  add_recipe(loan_recipe) %>%
  add_model(logit_model)

# Step 5: Fit your workflow to the training data for Logistic Regression
trained_workflow_logit <- train(
  loan_default ~ .,
```

```
  data = train_data,
  method = "glm",
  trControl = trainControl(method = "none", classProbs = TRUE),  # Enable
class probabilities
  metric = "ROC"
)
```

Fitting a logit model

```
# Step 6: Get predicted probabilities for Logistic Regression
test_predictions_logit <- predict(trained_workflow_logit, newdata =
test_data, type = "prob")

# Calculate ROC curve and AUC for Logistic Regression
roc_logit <- roc(test_data$loan_default, test_predictions_logit$yes)

## Setting levels: control = yes, case = no

## Setting direction: controls > cases

roc_auc_logit <- auc(roc_logit)

# Plot ROC curve
plot(roc_logit, print.auc = TRUE, main="ROC of the logit model")
```

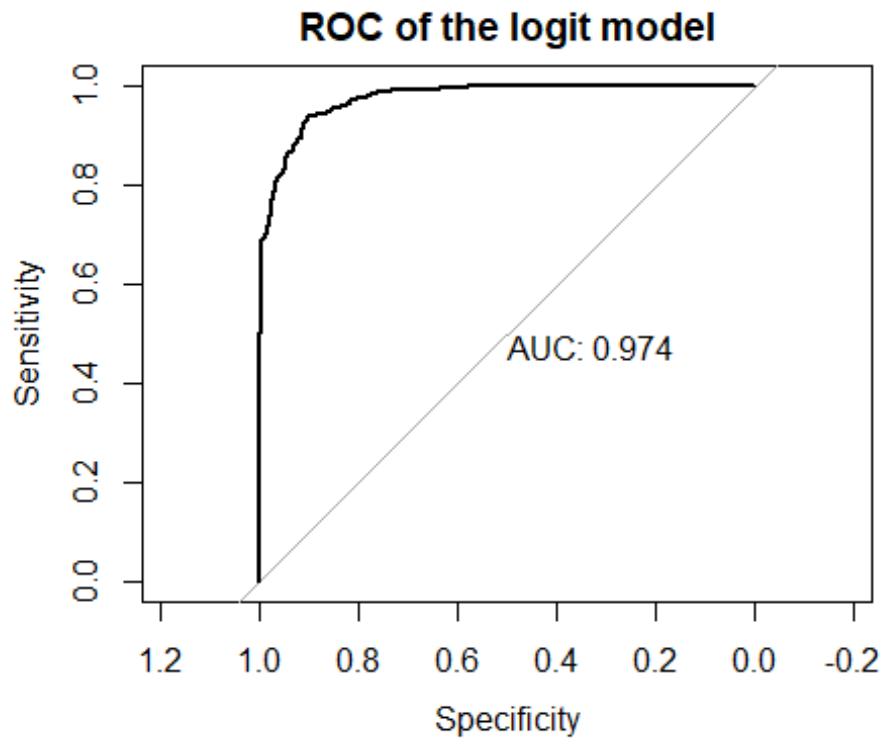

Figure 6.

The results from figure 6 above shows that the AUC for the logistic regression model was 0.974. This indicates that the logistic regression model has a stronger ability to differentiate between individuals who defaulted on loan payments and those who paid in full. The area under the curve 1 is closer to the maximum value of 1. This indicates that the model's predictive ability is more accurate and can be relied upon.

Fitting a LDA

```
# Create the LDA model
lda_model <- lda(loan_default ~ ., data = train_data)

# Fit workflow to the training data
# No hyperparameter tuning needed for LDA
trained_lda_model <- lda_model

# Predict on the test data
test_predictions_lda <- predict(trained_lda_model, newdata = test_data)

# Create an ROC curve and calculate the AUC
roc_lda <- roc(response = ifelse(test_data$loan_default == "yes", 1, 0),
predictor = test_predictions_lda$x)

## Setting levels: control = 0, case = 1

## Warning in roc.default(response = ifelse(test_data$loan_default == "yes",
:
## Deprecated use a matrix as predictor. Unexpected results may be produced,
## please pass a numeric vector.

## Setting direction: controls > cases

# Calculate the AUC (Area Under the Curve)
roc_auc_lda <- auc(roc_lda)

# Plot the ROC curve with the AUC value
plot(roc_lda, print.auc=TRUE ,main = "ROC of the LDA model")
```

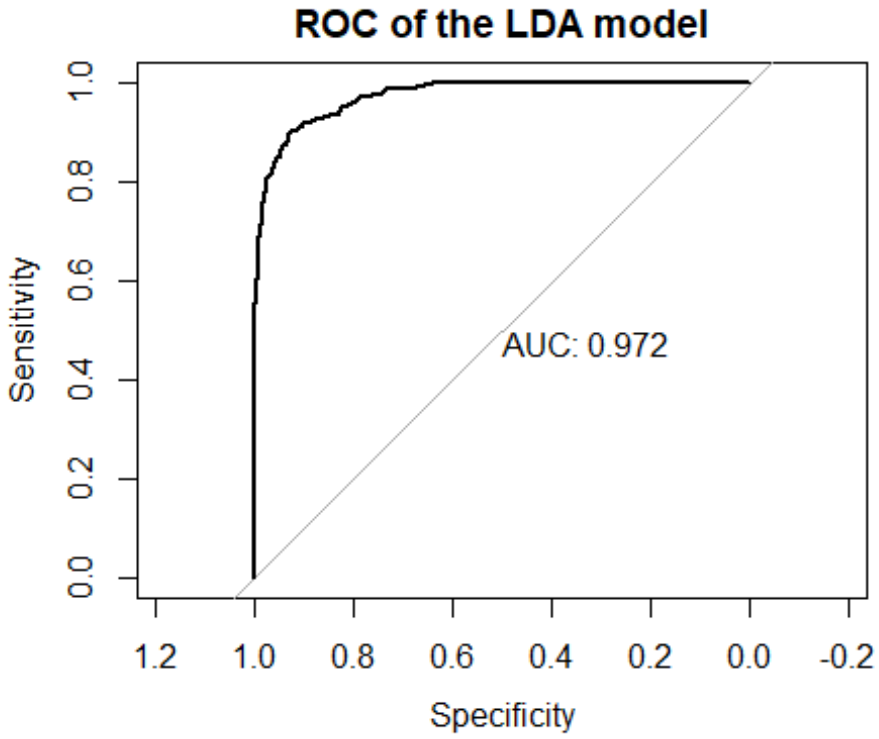

Figure 7.

The results from figure 7 above shows that the AUC for the Linear discriminant analysis (LDA) model was 0.972. This indicates that the Linear discriminant analysis regression model has a stronger ability to differentiate between individuals who defaulted on loan payments and non-defaulter. The area under the curve of 0.972 is closer to the maximum value of 1. This indicates that the model's predictive ability is more accurate and can be relied upon but not better when compared to the logistic regression model.

Model performance metric

```
# Predictions for Logistic Regression
test_predictions_logit <- predict(trained_workflow_logit, test_data)

# Confusion matrix for Logistic Regression
confusion_logit <- confusionMatrix(data = test_predictions_logit, reference =
test_data$loan_default)

# Calculate confusion matrix for LDA model
conf_matrix_lda <- table(Predicted = test_predictions_lda$class, Actual =
test_data$loan_default)

# Calculate accuracy for Logistic Regression
accuracy_logit <- confusion_logit$overall['Accuracy']

# Calculate accuracy for LDA
accuracy_lda <- sum(diag(conf_matrix_lda)) / sum(conf_matrix_lda)
```

```r
# Calculate sensitivity (True Positive Rate) for Logistic Regression
sensitivity_logit <- confusion_logit$byClass['Sensitivity']

# Calculate sensitivity (True Positive Rate) for LDA
actual <- test_data$loan_default
predicted <- predict(trained_lda_model, newdata = test_data)$class

TP <- sum(predicted == "yes" & actual == "yes")
FN <- sum(predicted == "no" & actual == "yes")

sensitivity_lda <- TP / (TP + FN)

cat("Logistic Regression Accuracy:", accuracy_logit, "\n")
```

```
## Logistic Regression Accuracy: 0.918897
```

```r
cat("LDA Accuracy:", accuracy_lda, "\n")
```

```
## LDA Accuracy: 0.9148418
```

```r
cat("Logistic Regression Sensitivity:", sensitivity_logit, "\n")
```

```
## Logistic Regression Sensitivity: 0.8845316
```

```r
cat("LDA Sensitivity:", sensitivity_lda, "\n")
```

```
## LDA Sensitivity: 0.8649237
```

The accuracy of the logistic regression model was 0.9189 with a sensitivity of 0.8845 while the accuracy of the Linear Discriminant model was 0.9148 with a sensitivity of 0.8649.

Selecting the best model

```r
# Selecting the best model
if (roc_auc_lda > roc_auc_logit) {
  best_model <- "LDA"
} else {
  best_model <- "Logistic Regression"
}

cat("The best model is:", best_model)
```

```
## The best model is: Logistic Regression
```

Best of the results above, the best model in the prediction of whether a person would default or not default on loan payment is a logit regression model since it had a better accuracy value compared to the LDA model.

Summary of Results

Introduction

The company faces a critical challenge in reducing loan defaults while still sustaining steady growth and profitability. To address this problem the company aims to leverage data visualization and predictive analytics to determine the factors that are associated with defaulting on loan payments and whether it is possible to accurately predict whether a customer will eventually default on their loan. The findings and insights drawn from this project will be key in decision-making on lending, optimization of risks, and ensuring that there is long-term success in the lending sector. Business problem The primary challenge facing the company (bank) is the record number of customer defaulters experienced by the company over a couple of years and its leading financial losses. The bank is finding it difficult to effectively identify and mitigate the risk of loan defaulters. Solving this problem is essential in maintaining the financial stability of the bank and facilitating business growth. By accurately predicting the likelihood of a person defaulting on the loan payment, the bank will tailor its lending policies and practices to minimize financial loses and maximize profit.

Goal of the study

Logit and Linear discriminant analysis (LDA) approach

The development of the predictive model in this project involved the use of logistic regression and linear discriminant analysis (LDA) techniques. These were the most appropriate method as the dependent variable 'loan default' was recorded as binary (yes/no). This project also incorporated elements of the following research questions to determine the factors associated with loan defaulting: Is there a significant difference in the loan amount between borrowers who defaulted on their loans and those who did not?

Does the interest rate offered to borrowers have an impact on their likelihood of defaulting on loan payments?

Is there an association between the loan term and the likelihood of loan default?

Is there a significant relationship between defaulting on the loan and loan purpose?

What is the relationship between loan defaulting and loan purpose as well as loan amount?

These questions will highlight the factors that affect loan repayment which will inform the policies that can be implemented to minimize of defaults arising from those factors. The models will help in determining if it is possible to predict whether a customer will default on their loan and costly errors are the model expected to produce.

Findings

Explanatory Data Analysis

The results from the grouped bar chart on the relationship between loan amount and load defaulting showed that there was a difference in the loan amount between borrowers who defaulted on their loans and those who did not. The difference in the amount between those who defaulted and those who did not default was 1202.32 which is a huge difference. This result indicates that people who received more amount were more likely to default and thus the bank should either request huge collateral, develop better recovery measures or get more details from the customer which will make it hard for them to default.

The results from the boxplot showing the impact of interest rate on defaulting on loan payments showed that loans with a higher interest rate were defaulted the most compared to those with a lower interest rate. On average, loans with an interest rate of 14 attracted the most defaults. This result is important as it displays how the interest rates of the bank are affecting its profitability. Higher default rates from loans with higher interest rates are costing the bank big.

The results from the grouped bar chart showed that loan term and loan default were related. The data showed that loans with a higher term limit were at a higher risk of being defaulted. This finding is important as it helps the company set up more recovery measures and other policy measures on loan term loans that cost the bank more losses.

It was also evident that the loan purpose affected the rate at which people defaulted on the loan. Loans that were taken for medical and credit card purposes were defaulted the most. The amount of money taken did not affect the rate of defaulting on medical loans. The rate of default was higher among small businesses and for home improvement that took larger amounts of loans compared to those that took smaller amounts of loans.

Model results

The best classification model based on the results above was a logistic regression model. The logistic regression model had a higher AUC value of 0.974 compared to 0.972 for the linear discriminant analysis (LDA) model. The accuracy and sensitivity of logit model was also better than the linear discriminant analysis (LDA) model with the accuracy of the logistic regression model being 0.9189 with a sensitivity of 0.8845 while the accuracy of the linear discriminant analysis (LDA) model being 0.9148 with a sensitivity of 0.8649. The logistic regression model correctly predicted the outcome (loan default or non-default) in approximately 91.89% of cases. The Logit model correctly classified 88.45% of defaulters to have defaulted on loan payment which means that only 11.55% of the defaulters were wrongly classified to not have defaulted. The model performance threshold was good.

Recommendation

The results showed that interest rates affected the loan payment. I would recommend that the bank should lower their interest rate so that individuals can find it easier and cheaper to repay their loans. The bank should also use other methods other than higher interest rates on high-risk loans as the interest rate was even worsening the loan default situation. Lowering interest will most probably attract more customers, and attract fewer defaults

thus improving the profitability. The bank should also consider giving out huge loans to individuals who display better repayment features like higher credit scores, more tangible collateral, steady job/business/income flow, and loan term to minimize default rates as the data showed that long-term loans were defaulted the most. I would recommend that more individual information be recorded for long-term loans which might help the bank trace them in case they fail to pay and the company should also attach more collateral options or guarantors who might make it easier for the bank to recover the money given out as a loan. I would also recommend that the bank provides more education and resources for borrowers taking long-term loans which will help them understand the financial commitment and the risk associated with the loans that they have taken.

Conclusion

In conclusion, the analysis of loan default in the bank has yielded important insights. The logistic regression tool selected from this study for future use in the banks for prediction has a higher accuracy of predicting whether a person will default or not on a loan payment and a higher sensitivity of classifying a customer as having defaulted when indeed they have defaulted. This prediction tool and the other findings should be used by the bank to guide their decision-making process to help them optimize the lending policies, minimize potential losses, and enhance risk management.