Sarang Tirmanwar

2023-10-29

R MARKDOWN

INTRODUCTION

In this project, we are analyzing a dataset of personal loans granted by a national bank in 2017. The main objectives are to identify the factors leading to loan defaults, create a predictive model for future defaults, and minimize financial losses. The bank is facing increasing loan defaults, and they want to improve their risk assessment.

Key Questions:

What factors contribute to loan defaults?

Can we accurately predict loan defaults?

How many costly prediction errors may occur?

Are there actions or policies to reduce default risks?

Dataset Overview

The dataset contains information about individuals who applied for personal loans from a national bank in 2017. It includes financial details and applicant behavior, such as income, debt ratios, loan amount, interest rate, and historical payment records. The main focus is on the "loan_default" variable, which indicates whether applicants eventually defaulted on their loans, causing financial losses for the bank. Other variables include loan purpose, application type, homeownership status, income, employment duration, credit history, and more. The goal is to analyze these factors to predict loan defaults and reduce financial losses.

loan_default: Indicates whether the borrower defaulted on their loan (yes/no).

loan_amount: Represents the total loan amount borrowed by an individual.

installment: Denotes the monthly installment amount to be paid.

interest_rate: Specifies the loan's interest rate in percentage.

loan_purpose: Describes the purpose for which the personal loan is taken.

application_type: Indicates whether the loan application is made individually or jointly.

term: Refers to the duration of the loan, which can be three or five years.

homeownership: Provides information about the borrower's current homeownership status.

annual_income: Represents the annual income of the person applying for the loan.

current_job_years: Indicates the number of years the applicant has been in their current job.

debt_to_income: Denotes the individual's debt-to-income ratio at the time of loan application.

total_credit_lines: Represents the total count of open credit lines for the applicant.

years_credit_history: Specifies the length of the applicant's credit history.

missed_payment_2_yr: Indicates whether there have been any missed payments in the last 2 years (yes/no).

history_bankruptcy: Describes the presence or absence of a history of bankruptcy (yes/no).

history_tax_liens: Indicates whether there is a history of tax liens (yes/no).

DATA ANALYSIS

```
#EDA
loan_df <- readRDS("/Users/sarangtirmanwar/Downloads/loan_data.rds")
# Load necessary libraries
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(skimr)
library(caret)

## Loading required package: lattice

library(tidyr)
dim(loan_df)
```

```
## [1] 4110   16
```

```r
str(loan_df)
```

```
## 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 35
000 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 ...
```

```r
head(loan_df)
```

```
## # A tibble: 6 × 16
##   loan_default loan_amount installment interest_rate loan_purpose
##   <fct>              <int>       <dbl>         <dbl> <fct>
## 1 yes                35000        927.         17.2  small_business
## 2 yes                10000        260.         11.5  small_business
## 3 no                 28800        942.          8.97 debt_consolidation
## 4 yes                 4475        165.         10    medical
## 5 no                  3600        111.          9.72 medical
## 6 yes                12800        389.         20    medical
## # ℹ 11 more variables: application_type <fct>, term <fct>, homeownership <
fct>,
## #   annual_income <dbl>, current_job_years <dbl>, debt_to_income <dbl>,
## #   total_credit_lines <int>, years_credit_history <dbl>,
## #   missed_payment_2_yr <fct>, history_bankruptcy <fct>,
## #   history_tax_liens <fct>
```

```r
glimpse(loan_df)
```

```
## Rows: 4,110
## Columns: 16
## $ loan_default        <fct> yes, yes, no, yes, no, yes, yes, no, no, no, 
no, …
## $ loan_amount         <int> 35000, 10000, 28800, 4475, 3600, 12800, 35000
, 26…
## $ installment         <dbl> 927.29, 259.58, 941.65, 164.99, 110.70, 389.1
0, 9…
## $ interest_rate       <dbl> 17.25, 11.50, 8.97, 10.00, 9.72, 20.00, 18.25
, 11…
## $ loan_purpose        <fct> small_business, small_business, debt_consolid
atio…
## $ application_type    <fct> individual, individual, individual, individua
l, i…
## $ term                <fct> five_year, five_year, three_year, three_year, 
thr…
## $ homeownership       <fct> rent, mortgage, rent, rent, mortgage, rent, m
ortg…
## $ annual_income       <dbl> 104660, 57000, 160000, 37000, 72000, 73000, 1
6700…
## $ current_job_years   <dbl> 2, 10, 10, 1, 4, 10, 0, 5, 4, 3, 10, 10, 5, 1
0, 1…
## $ debt_to_income      <dbl> 29.41, 23.79, 5.96, 13.82, 22.68, 30.94, 25.9
1, 7…
## $ total_credit_lines  <int> 27, 14, 35, 7, 35, 57, 34, 24, 12, 12, 16, 9, 
17,…
## $ years_credit_history <dbl> 15, 4, 17, 5, 11, 14, 22, 16, 9, 12, 22, 9, 8
, 17…
## $ missed_payment_2_yr <fct> no, no, no, no, no, no, no, no, no, no, no, n
o, n…
## $ history_bankruptcy  <fct> no, no, yes, no, no, no, no, no, no, no, no, 
no, …
## $ history_tax_liens   <fct> no, no, no, no, no, no, no, no, no, no, no, n
o, n…

skim(loan_df)
```

Data summary

Name	loan_df
Number of rows	4110
Number of columns	16
_______________________	
Column type frequency:	
factor	8
numeric	8
________________________	

Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
loan_default	0	1	FALSE	2	no: 2580, yes: 1530
loan_purpose	0	1	FALSE	5	deb: 1218, cre: 879, sma: 853, med: 635
application_type	0	1	FALSE	2	ind: 3494, joi: 616
term	0	1	FALSE	2	thr: 2588, fiv: 1522
homeownership	0	1	FALSE	3	mor: 1937, ren: 1666, own: 507
missed_payment_2_yr	0	1	FALSE	2	no: 3640, yes: 470
history_bankruptcy	0	1	FALSE	2	no: 3624, yes: 486
history_tax_liens	0	1	FALSE	2	no: 4050, yes: 60

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
loan_amount	0	1	16692.79	10038.89	1000.00	9600.00	15000.00	24000.00	40000.00	▆▇▅▃▂
installment	0	1	489.42	289.50	31.04	274.82	421.97	663.98	1566.59	▇▇▅▂▁
interest_rate	0	1	11.38	3.92	4.72	8.22	11.25	13.75	20.00	▅▅▇▃▃
annual_income	0	1	73015.01	37203.11	3000.00	45000.00	65000.00	92000.00	200000.00	▃▇▃▁▁
current_job_years	0	1	5.80	3.69	0.00	2.00	5.00	10.00	10.00	▇▃▂▂▇
debt_to_income	0	1	20.04	14.23	0.00	11.85	18.59	26.13	437.61	▇▁▁▁▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
total_credit_lines	0	1	22.47	12.03	2.00	14.00	20.00	29.00	87.00	▇▇▂▁▁
years_credit_history	0	1	15.76	7.22	3.00	11.00	14.00	19.00	51.00	▆▇▂▁▁

```
summary(loan_df)
```

```
## loan_default  loan_amount     installment      interest_rate
## yes:1530     Min.   : 1000   Min.   :  31.04   Min.   : 4.72
## no :2580     1st Qu.: 9600   1st Qu.: 274.82   1st Qu.: 8.22
##              Median :15000   Median : 421.97   Median :11.25
##              Mean   :16693   Mean   : 489.42   Mean   :11.38
##              3rd Qu.:24000   3rd Qu.: 663.99   3rd Qu.:13.75
##              Max.   :40000   Max.   :1566.59   Max.   :20.00
##             loan_purpose    application_type        term        homeowner
ship
## debt_consolidation:1218   individual:3494    three_year:2588   mortgage:1
937
## credit_card       : 879   joint     : 616    five_year :1522   rent    :1
666
## medical           : 635                                        own     :
507
## small_business    : 853
## home_improvement  : 525
##
## annual_income    current_job_years debt_to_income   total_credit_lines
## Min.   :  3000   Min.   : 0.000    Min.   :  0.00   Min.   : 2.00
## 1st Qu.: 45000   1st Qu.: 2.000    1st Qu.: 11.85   1st Qu.:14.00
## Median : 65000   Median : 5.000    Median : 18.59   Median :20.00
## Mean   : 73015   Mean   : 5.802    Mean   : 20.04   Mean   :22.47
## 3rd Qu.: 92000   3rd Qu.:10.000    3rd Qu.: 26.13   3rd Qu.:29.00
## Max.   :200000   Max.   :10.000    Max.   :437.61   Max.   :87.00
## years_credit_history missed_payment_2_yr history_bankruptcy history_tax_l
iens
## Min.   : 3.00        yes: 470            yes: 486           yes:  60
## 1st Qu.:11.00        no :3640            no :3624           no :4050
## Median :14.00
## Mean   :15.76
## 3rd Qu.:19.00
## Max.   :51.00
```

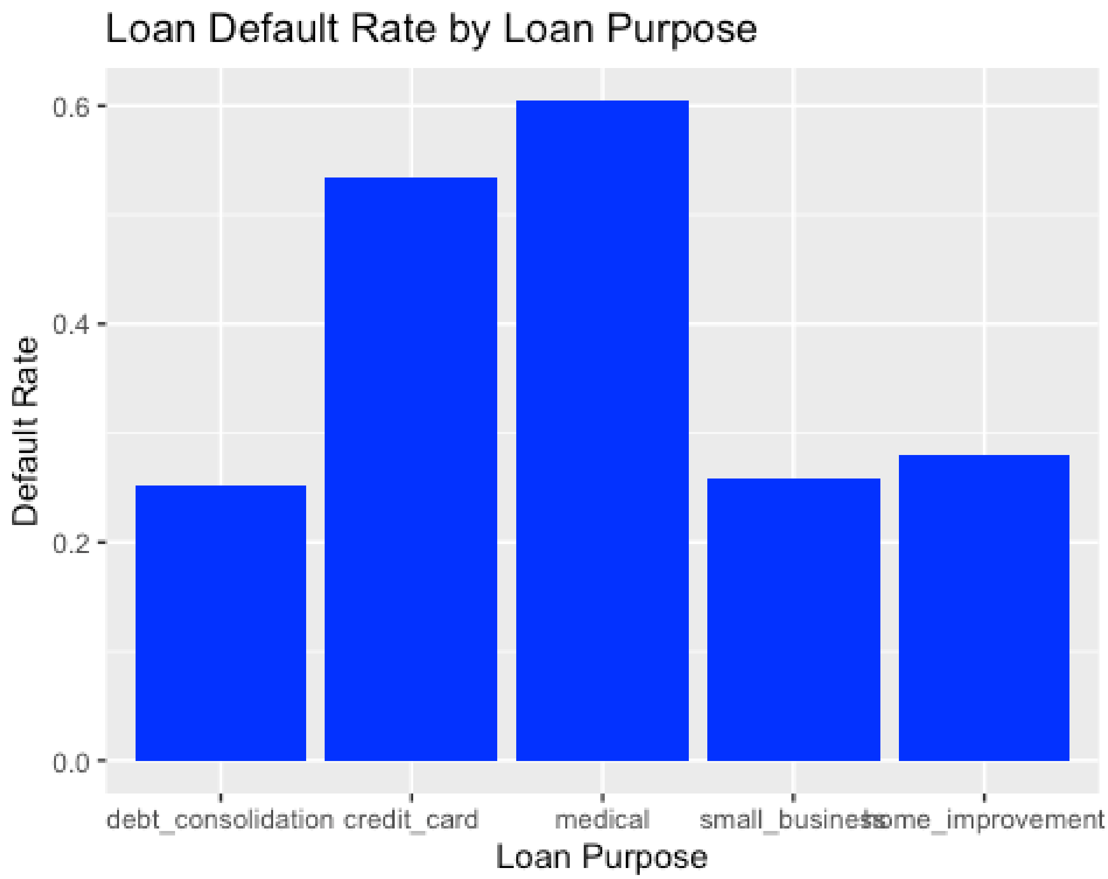
1. Is there a relationship between loan default and the loan purpose?

Answer: There appears to be a significant relationship between the purpose of the loan and the likelihood of default. The default rates vary across different loan purposes. The loan purpose significantly influences the default rate. Loans for medical and credit card purposes have high default rates (around 60.5% and 53.5%, respectively), while debt consolidation and small business loans have lower default rates (25.3% and 25.9%, respectively). This analysis suggests that the loan purpose is an important factor in predicting loan default. Borrowers taking loans for specific purposes, such as medical expenses, appear to have a higher risk of default compared to those using loans for other purposes.

```r
# Calculate the default rate for each loan purpose
default_by_purpose <- loan_df %>%
  group_by(loan_purpose) %>%
  summarise(default_rate = mean(loan_default == 'yes'))
default_by_purpose
```

```
## # A tibble: 5 × 2
##   loan_purpose       default_rate
##   <fct>                     <dbl>
## 1 debt_consolidation        0.253
## 2 credit_card               0.535
## 3 medical                   0.605
## 4 small_business            0.259
## 5 home_improvement          0.28
```

```r
# Create a bar chart
ggplot(default_by_purpose, aes(x = loan_purpose, y = default_rate)) +
  geom_bar(stat = 'identity', fill = 'blue') +
  labs(x = "Loan Purpose", y = "Default Rate") +
  ggtitle("Loan Default Rate by Loan Purpose")
```

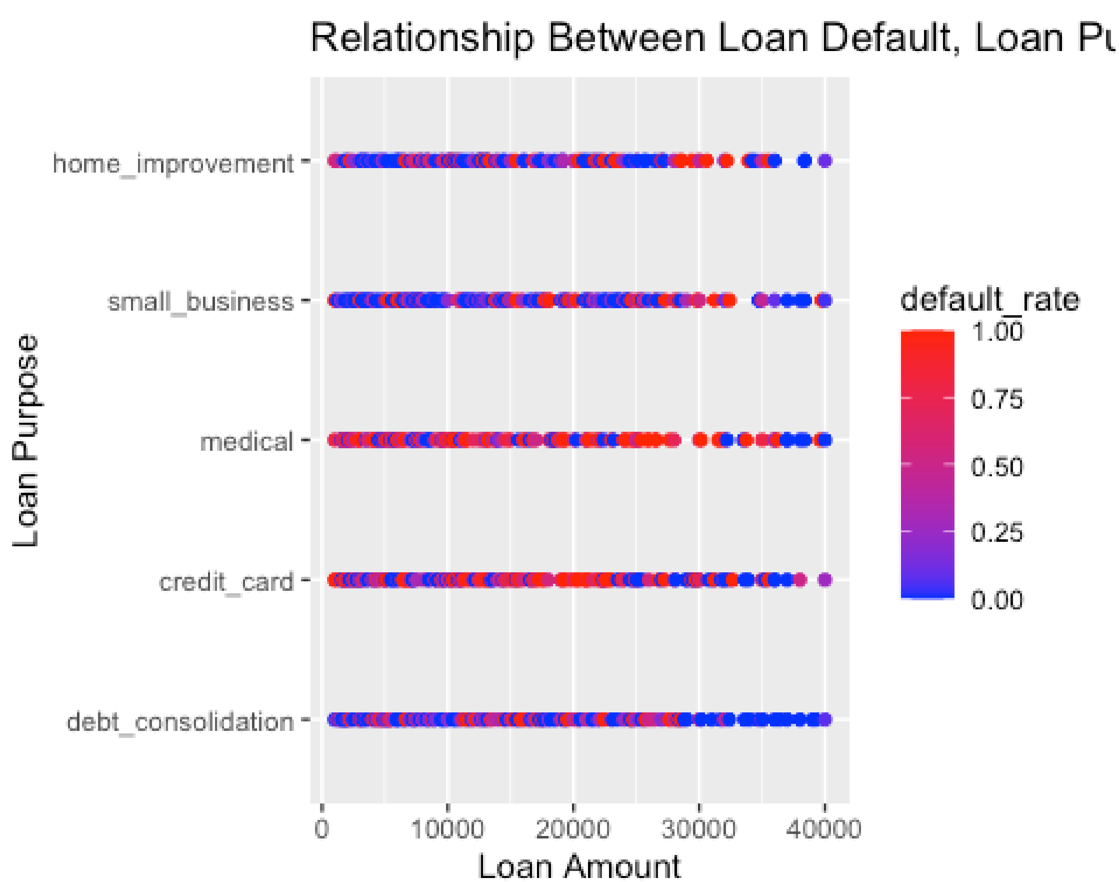


2.Is there a relationship between default rates for different combinations of loan purpose and loan amount ?

Answer: The scatter plot shows the relationship between loan default, loan purpose, and loan amount. It appears that loan amount and loan purpose are both factors influencing the default rate. Some loan amounts within the “debt_consolidation” category have higher default rates, while others have no defaults. This suggests that loan amount plays a role in loan default, with certain amounts being riskier than others, especially within the “debt_consolidation” purpose.

```
# Calculate default rates for different combinations of loan purpose and loan
amount
default_rates <- loan_df %>%
  group_by(loan_purpose, loan_amount) %>%
  summarise(default_rate = mean(loan_default == "yes"))

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

default_rates
```

```
## # A tibble: 743 × 3
## # Groups:   loan_purpose [5]
##    loan_purpose       loan_amount default_rate
##    <fct>                    <int>        <dbl>
##  1 debt_consolidation        1000        0
##  2 debt_consolidation        1200        0.333
##  3 debt_consolidation        1375        0
##  4 debt_consolidation        1450        1
##  5 debt_consolidation        1500        0.25
##  6 debt_consolidation        1600        0
##  7 debt_consolidation        1700        0
##  8 debt_consolidation        1800        0
##  9 debt_consolidation        1825        0
## 10 debt_consolidation        1925        0
## # ℹ 733 more rows

# Create a scatter plot to visualize the relationship
ggplot(default_rates, aes(x = loan_amount, y = loan_purpose, color = default_
rate)) +
  geom_point() +
  labs(x = "Loan Amount", y = "Loan Purpose") +
  scale_color_gradient(low = "blue", high = "red") +
  ggtitle("Relationship Between Loan Default, Loan Purpose, and Loan Amount")
```



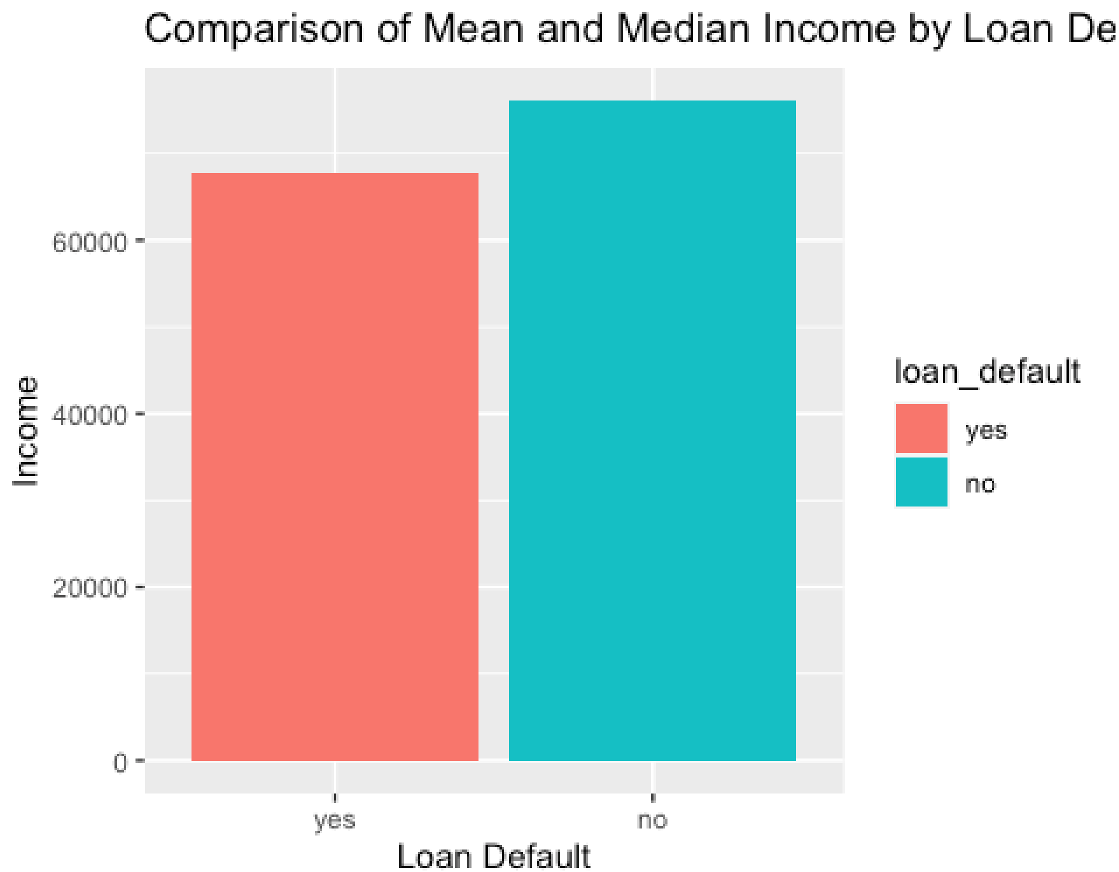
3. How does the applicant’s annual income impact loan defaults?

Answer : The mean annual income for borrowers who defaulted on their loans (yes) is $67,819, while for non-defaulters (no), it’s higher at $76,096. Similarly, the median annual income for borrowers who defaulted on their loans is $60,000, while for non-defaulters, it’s higher at $69,000. This suggests that, on average, borrowers who default on their loans tend to have lower annual incomes compared to those who do not default.

```
# Calculate the mean and median annual income for default and non-default cas
es using dplyr
income_summary <- loan_df %>%
  group_by(loan_default) %>%
  summarise(mean_income = mean(annual_income), median_income = median(annual_
income))
income_summary
```

```
## # A tibble: 2 × 3
##   loan_default mean_income median_income
##   <fct>              <dbl>         <dbl>
## 1 yes               67819.         60000
## 2 no                76096.         69000
```

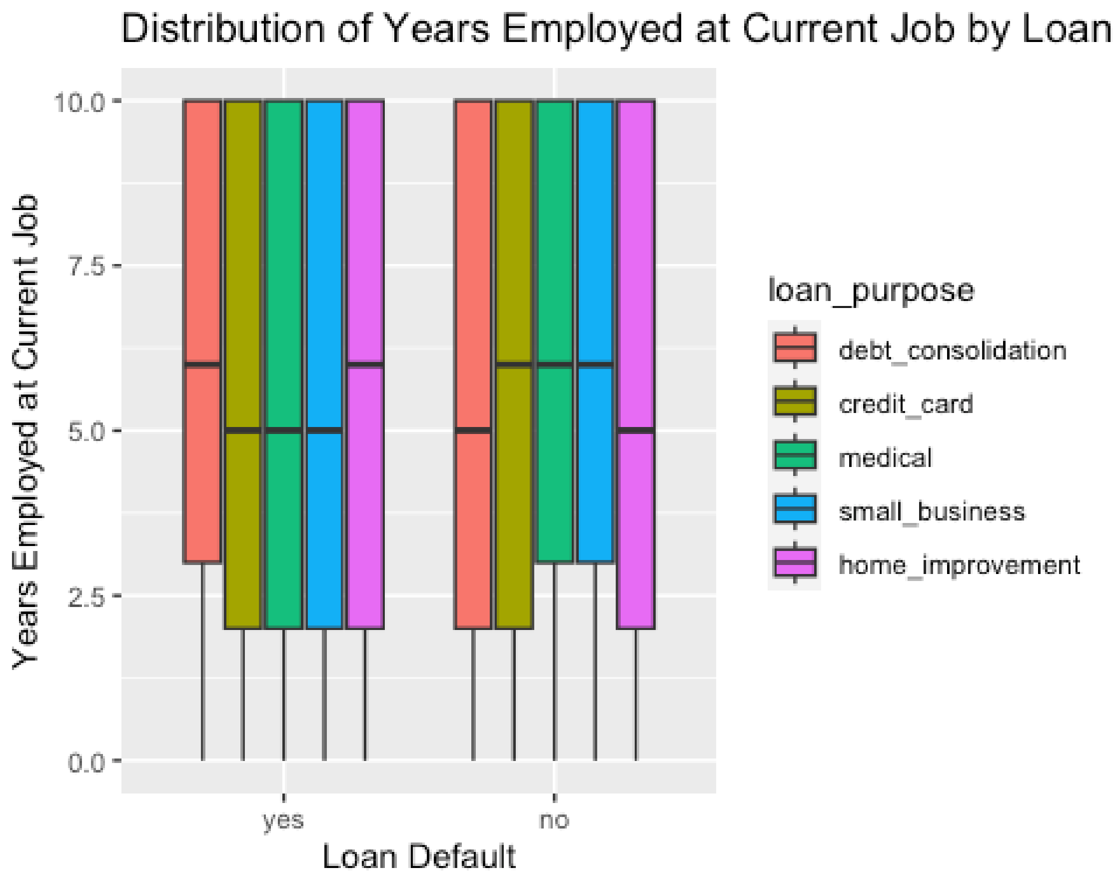
```
# Create a grouped bar chart to compare mean and median income by loan defaul
t
ggplot(income_summary, aes(x = loan_default, y = mean_income, fill = loan_def
ault)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  geom_bar(aes(y = median_income), stat = 'identity', position = 'dodge', alp
ha = 0.6, width = 0.4) +
  labs(x = "Loan Default", y = "Income") +
  ggtitle("Comparison of Mean and Median Income by Loan Default")
```



4. Is there a correlation between the applicant's job stability (current_job_years) and loan defaults, and does this correlation differ based on loan purpose?

Answer: The box plot analysis indicates that borrowers who defaulted on their loans tend to have a lower median value for years employed at their current job compared to those who did not default. This suggests that less job stability is associated with a higher likelihood of loan defaults. Additionally, the box plot differentiates loan purposes, and it appears that the impact of job stability on loan defaults is consistent across different loan purposes. In other words, job stability is a significant factor in loan defaults, and this relationship is not significantly influenced by the specific purpose of the loan.

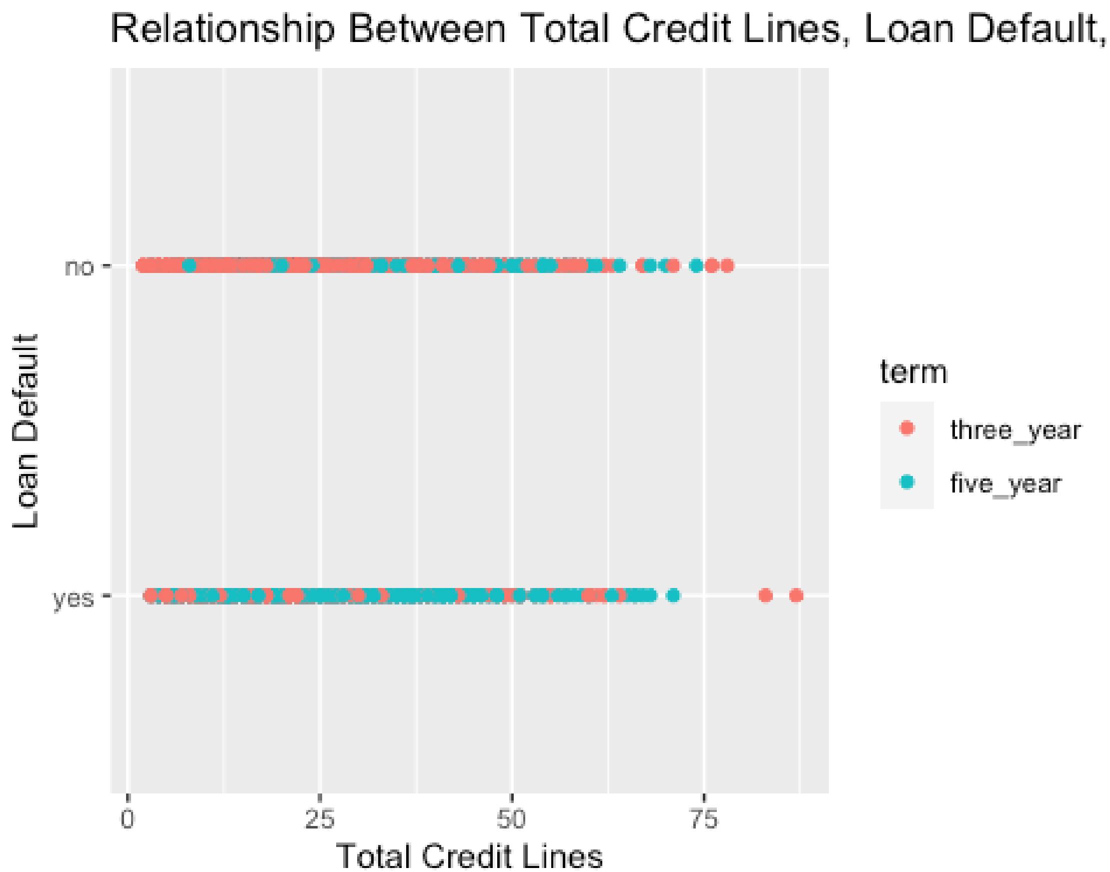
```
# Create a box plot to compare years employed at the current job for default and non-default cases, grouped by loan purpose
ggplot(loan_df, aes(x = loan_default, y = current_job_years, fill = loan_purpose)) +
  geom_boxplot() +
  labs(x = "Loan Default", y = "Years Employed at Current Job") +
  ggtitle("Distribution of Years Employed at Current Job by Loan Default and Loan Purpose")
```



5.Is there a relationship between total credit lines, loan default, and loan term ?

Answer: The scatter plot illustrates the relationship between the total number of credit lines and loan default. Each point represents an individual borrower. We can observe that there is no clear linear pattern or trend in the relationship between the total credit lines and loan default. Both default (yes) and non-default (no) cases are scattered across different values of total credit lines. The color of the points represents the loan term, with one color indicating three-year loans and another indicating five-year loans. Within both loan term categories, we see a mix of default and non-default cases across various total credit lines. In summary, the scatter plot suggests that the relationship between total credit lines, loan default, and loan term is not easily characterized by a simple linear trend.

```
#Create a scatter plot to explore the relationship between total credit lines
, loan default, and loan term
ggplot(loan_df, aes(x = total_credit_lines, y = loan_default, color = term))
+
  geom_point() +
  labs(x = "Total Credit Lines", y = "Loan Default") +
  ggtitle("Relationship Between Total Credit Lines, Loan Default, and Loan Te
rm")
```

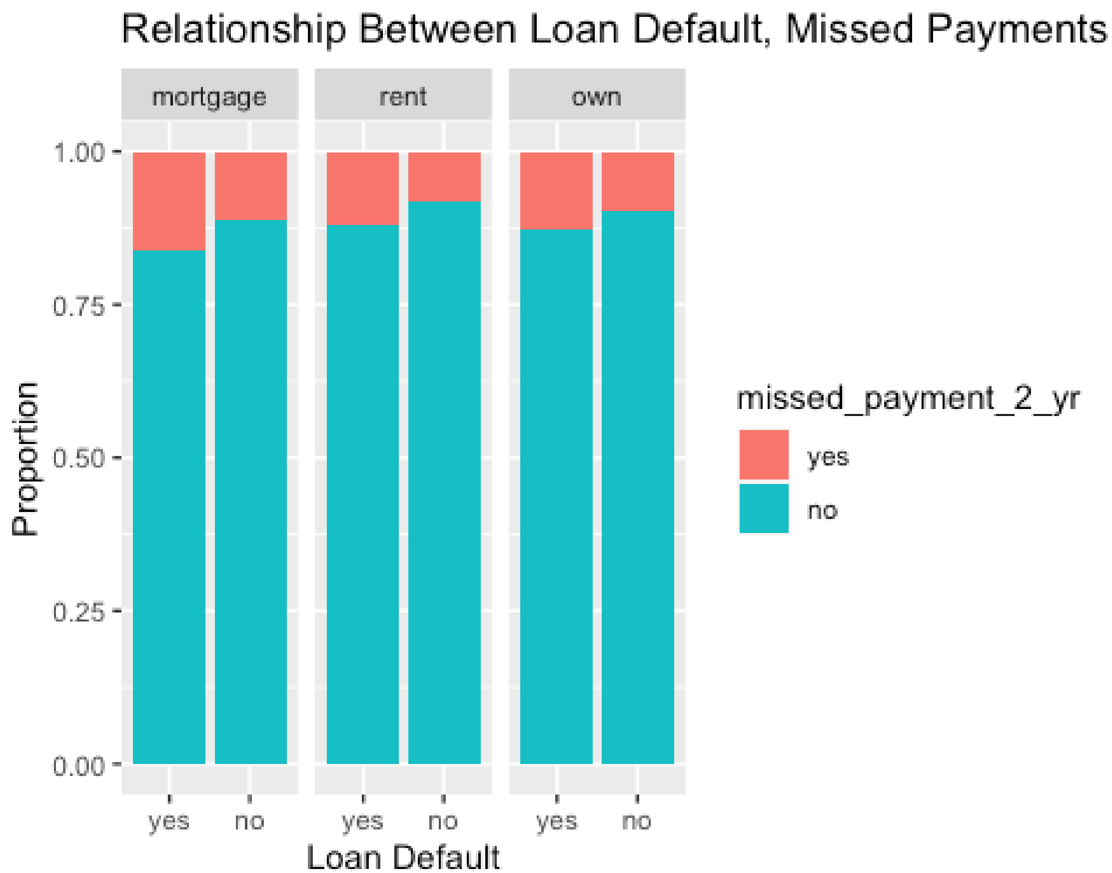


6.Is there a relationship between loan default, missed payments, and homeownership?

Answer: The chart explores the relationship between loan default (yes or no) and whether borrowers have missed payments in the last 2 years (yes or no). In general, it appears that borrowers who have missed payments are more likely to default on their loans compared to those who have not missed payments. This is indicated by the taller bar for "yes" in the "Missed Payments" category within both "Default" and "No Default" groups. While missed payments seem to be associated with a higher likelihood of default, the impact of homeownership on loan default is not as evident from this chart alone. Further analysis or statistical testing may be required to determine if homeownership significantly influences loan default rates. In summary, the stacked bar chart demonstrates that missed payments in the last 2 years are associated with a higher likelihood of loan default across different homeownership categories.

```
#Create a stacked bar chart to explore the relationship between loan default, missed payments, and homeownership
ggplot(loan_df, aes(x = loan_default, fill = missed_payment_2_yr)) +
  geom_bar(position = "fill") +
  facet_grid(. ~ homeownership) +
```

```
  labs(x = "Loan Default", y = "Proportion") +
  ggtitle("Relationship Between Loan Default, Missed Payments, and Homeowners
hip")
```



Predictive modelling (lOGISTIC REGRESSION AND RANDOM FOREST)

```
#logistic

# Importing necessary libraries
library(tidyverse)  # Comprehensive data manipulation and visualization tools
.

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
## ✓ forcats   1.0.0     ✓ readr     2.1.4
## ✓ lubridate 1.9.3     ✓ stringr   1.5.0
## ✓ purrr     1.0.2     ✓ tibble    3.2.1
## ── Conflicts ──────────────────────────────────────────── tidyverse_conflict
s() ──
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()
## ✗ purrr::lift()   masks caret::lift()
```

```
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```r
library(tidymodels)  # Framework for modeling and machine learning.
```

```
## ── 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_conflict
s() ──
## ✖ scales::discard()        masks purrr::discard()
## ✖ dplyr::filter()          masks stats::filter()
## ✖ recipes::fixed()         masks stringr::fixed()
## ✖ dplyr::lag()             masks stats::lag()
## ✖ purrr::lift()            masks caret::lift()
## ✖ yardstick::precision()   masks caret::precision()
## ✖ yardstick::recall()      masks caret::recall()
## ✖ yardstick::sensitivity() masks caret::sensitivity()
## ✖ yardstick::spec()        masks readr::spec()
## ✖ yardstick::specificity() masks caret::specificity()
## ✖ recipes::step()          masks stats::step()
## • Learn how to get started at https://www.tidymodels.org/start/
```

```r
library(vip)  # Variable Importance Plots.
```

```
##
## Attaching package: 'vip'
##
## The following object is masked from 'package:utils':
##
##     vi
```

```r
# Set Seed
set.seed(123)

# Split the data into training and testing sets
loan_split <- initial_split(loan_df, prop = 0.7, strata = loan_default)
loan_train <- training(loan_split)
loan_test <- testing(loan_split)

# Display the number of rows in the training and testing sets
nrow(loan_train)
```

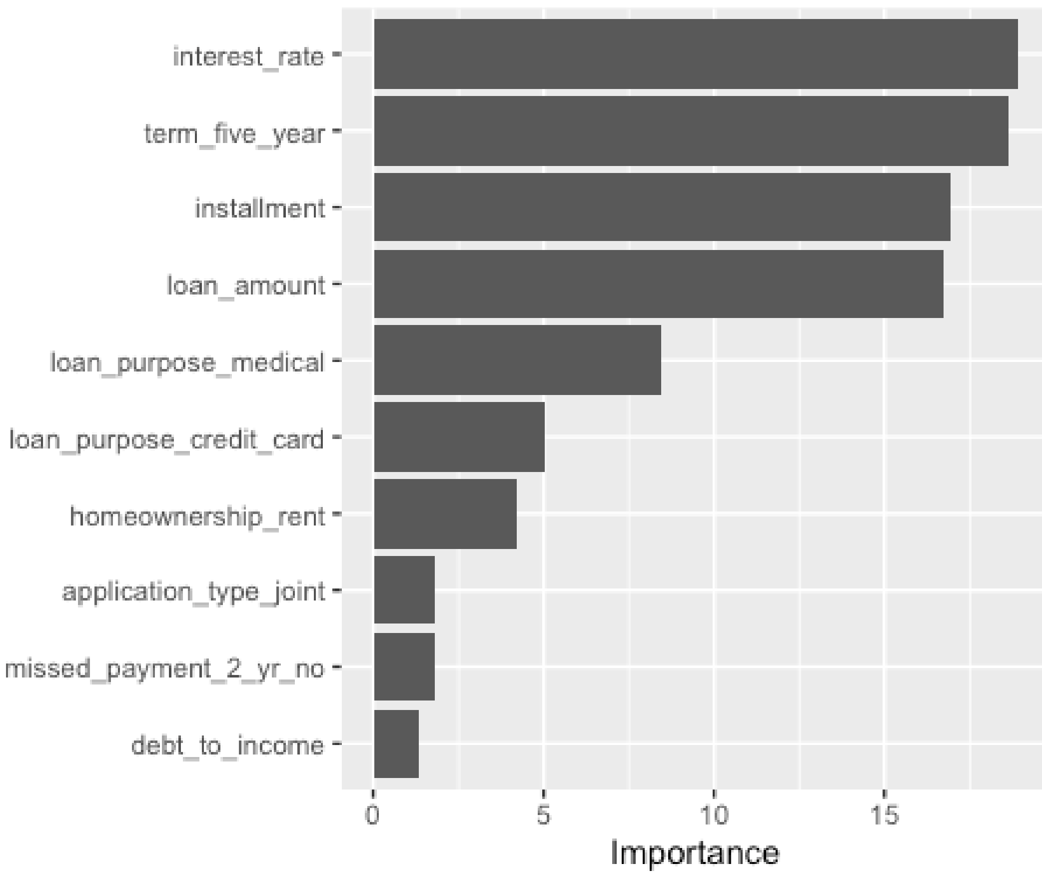
```
## [1] 2876
```

```
nrow(loan_test)

## [1] 1234

# Create a recipe to preprocess the data
loan_recipe <- recipe(loan_default ~ ., data = loan_train) %>%
  step_normalize(all_numeric(), -all_outcomes()) %>%
  step_dummy(all_nominal(), -all_outcomes())

# Prepare the recipe
loan_recipe %>%
  prep(training = loan_train) %>%
  bake(new_data = NULL)

## # A tibble: 2,876 × 20
##    loan_amount installment interest_rate annual_income current_job_years
##          <dbl>       <dbl>         <dbl>         <dbl>             <dbl>
##  1       1.19        1.55         -0.620         2.32              1.13
##  2      -1.31       -1.31         -0.429        -0.0318           -0.504
##  3      -1.12       -1.09         -0.874        -0.0852           -0.504
##  4       2.30        1.58         -0.110        -0.0852           -0.777
##  5      -0.183      -0.140        -0.0464       -0.833             1.13
##  6       1.90        0.934        -1.07          2.99              1.13
##  7      -1.08       -1.02         -0.0464       -1.13              1.13
##  8      -0.679      -0.918         0.463        -1.29             -1.32
##  9      -0.530      -0.864        -0.938        -1.04             -0.232
## 10      -0.679      -0.658        -1.07         -0.352             1.13
## # ℹ 2,866 more rows
## # ℹ 15 more variables: debt_to_income <dbl>, total_credit_lines <dbl>,
## #   years_credit_history <dbl>, loan_default <fct>,
## #   loan_purpose_credit_card <dbl>, loan_purpose_medical <dbl>,
## #   loan_purpose_small_business <dbl>, loan_purpose_home_improvement <dbl>
,
## #   application_type_joint <dbl>, term_five_year <dbl>,
## #   homeownership_rent <dbl>, homeownership_own <dbl>, …

# Create a logistic regression model
lmodel <- logistic_reg() %>%
  set_engine('glm') %>%
  set_mode('classification')

# Create a workflow that includes the model and recipe
loan_workflow <- workflow() %>%
  add_model(lmodel) %>%
  add_recipe(loan_recipe)

# Fit the logistic regression model
logistic_fit <- loan_workflow %>%
  fit(data = loan_train)
```

```r
# Extract the trained model
loan_train_model <- logistic_fit %>%
  pull_workflow_fit()
```

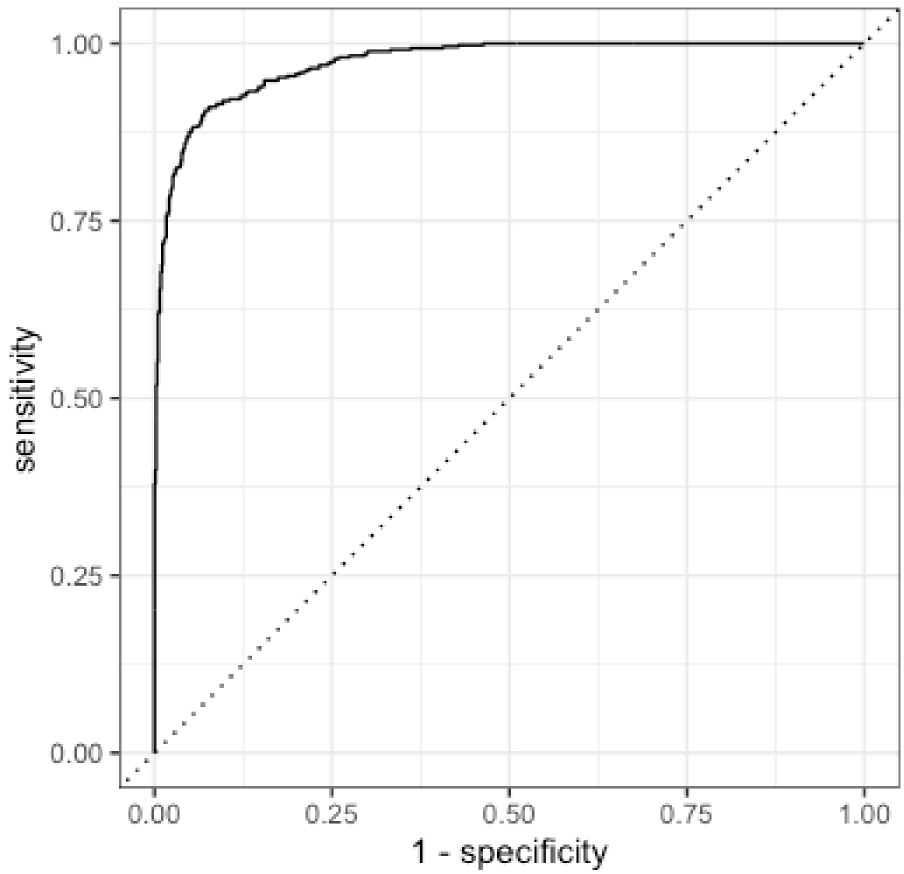
```
## Warning: `pull_workflow_fit()` was deprecated in workflows 0.2.3.
## ℹ Please use `extract_fit_parsnip()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
# Visualize variable importance using VIP package
vip(loan_train_model)
```



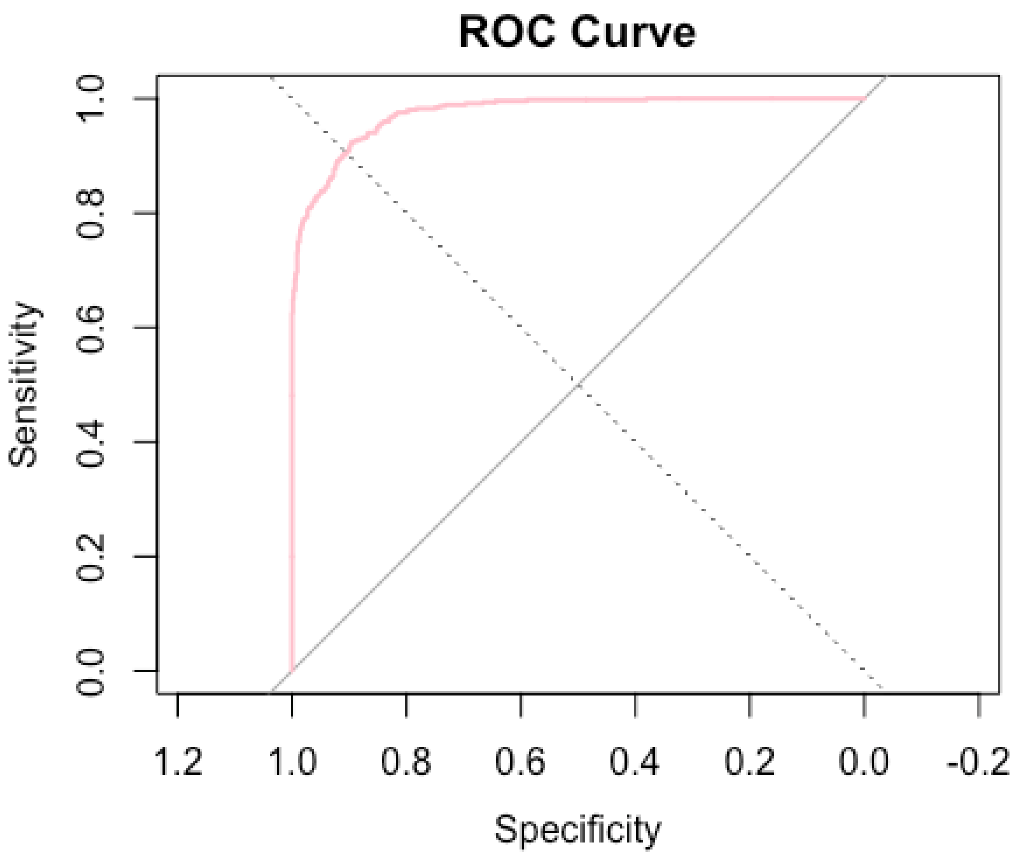

```r
# Make predictions with the logistic regression model
class_preds <- predict(logistic_fit, new_data = loan_test, type = 'class')
prob_preds <- predict(logistic_fit, new_data = loan_test, type = 'prob')

# Combine predictions with actual loan_default values
loan_result <- loan_test %>%
  select(loan_default) %>%
  bind_cols(class_preds, prob_preds)

# Create cross-validation folds
```

```r
loan_folds <- vfold_cv(loan_train, v = 5)

# Calculate confusion matrix
conf_mat(loan_result, truth = loan_default, estimate = .pred_class)

##           Truth
## Prediction yes  no
##        yes 402  40
##        no   57 735

# Calculate F1-measure
f_meas(loan_result, truth = loan_default, estimate = .pred_class)

## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 f_meas  binary         0.892

# Define a set of metrics including accuracy and sensitivity
loan_metric <- metric_set(accuracy, sens)

# Evaluate the model using the defined metrics
loan_metric(loan_result, truth = loan_default, estimate = .pred_class)

## # A tibble: 2 × 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.921
## 2 sens     binary         0.876

# Plot ROC curve
loan_result %>%
  roc_curve(truth = loan_default, .pred_yes) %>%
  autoplot()
```




```
# Random Forest Model
set.seed(223)

# Split the data into training and testing sets
loan_split <- initial_split(loan_df, prop = 0.7)
training_data <- training(loan_split)
testing_data <- testing(loan_split)

# Create a random forest model
randomf_model <- rand_forest() %>%
  set_engine("ranger", importance = "permutation", num.threads = 1) %>%
  set_mode("classification")

# Create a workflow that includes the random forest model and recipe
randomf_wf <- workflow() %>%
  add_model(randomf_model) %>%
  add_recipe(loan_recipe)

# Fit the random forest model
randomf_fit <- randomf_wf %>%
  last_fit(split = loan_split)

# Collect predictions
```

```r
randomf_results <- randomf_fit %>%
  collect_predictions()

# Fit the random forest model for training
train_randomf_workflow <- randomf_wf %>%
  fit(data = training(loan_split))

# Define a set of metrics for evaluation
randomf_metrics <- metric_set(accuracy, f_meas, roc_auc)

# Make predictions with the random forest model
randomf_predictions <- predict(train_randomf_workflow, testing_data) %>%
  bind_cols(testing_data)

# Calculate metrics for the random forest model
randomf_metrics <- metrics(randomf_predictions, truth = loan_default, estimat
e = .pred_class)

# Plot ROC curve for the random forest model
library(pROC)

## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

predictions <- randomf_results$.pred_yes
labels <- ifelse(randomf_results$loan_default == "yes", 1, 0)
roc_curve <- roc(labels, predictions)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

plot(roc_curve, main = "ROC Curve", col = "pink")
abline(a = 0, b = 1, lty = 3, col = "black")
```



SUMMARY OF RESULTS

In our exploratory data analysis, several noteworthy findings emerged regarding the relationship between various factors and loan defaults. One crucial insight is the strong association between the purpose of a loan and the likelihood of default. Loans intended for medical expenses and credit card use exhibit notably higher default rates, whereas debt consolidation and small business loans are linked to lower default rates. This underlines the importance of considering the loan's purpose when assessing credit risk.

Additionally, loan amount appears to be a factor influencing loan default, particularly within the "debt_consolidation" category. Applicants with lower annual incomes were found to be more prone to loan defaults, suggesting the importance of rigorous income verification in the lending process. Moreover, job stability, as measured by the number of years employed at the current job, significantly impacts loan defaults, with those having less job stability being at a higher risk.

While no clear linear relationship was established between the total number of credit lines and loan defaults, it was observed that borrowers who had missed payments in the last two years were more likely to default. However, the analysis did not yield a conclusive result regarding the impact of homeownership on loan defaults. These findings provide valuable insights that can inform lending practices and help reduce loan defaults, ultimately benefitting lending institutions.

Best Classification Model

Logistic Regression excelled over Random Forest, achieving a superior accuracy rate of 0.921 compared to 0.909 and a higher sensitivity level of 0.876. Furthermore, Logistic Regression provides a more transparent and interpretable understanding of how input features relate to the target variable. This favorable combination of high accuracy, interpretability, and resource efficiency collectively establishes Logistic Regression as the preferred choice for effectively predicting loan defaults.

Recommendations:

The analysis yields several important insights that can inform business decisions:

Risk Assessment: Lenders should consider the purpose of the loan when assessing credit risk. Loans for medical and credit card purposes pose higher risks, while debt consolidation and small business loans are associated with lower default rates.

Income Verification: Lenders should pay close attention to the income levels of applicants. Borrowers with lower annual incomes are more likely to default, so income verification and assessment should be a crucial part of the lending process.

Job Stability: Lenders should consider the stability of an applicant's current job. Those with shorter job tenures are at a higher risk of default. This factor should be included in credit risk models.

Missed Payments: Lenders should have stringent policies for borrowers who have missed payments in the last two years. They represent a higher default risk, and additional scrutiny or risk mitigation measures may be necessary.

Further Analysis: While the relationship between homeownership and loan default was not clear in this analysis, a more in-depth investigation or statistical analysis might provide a clearer picture of its impact. Lenders may want to explore this aspect further.

Conclusion:

In conclusion, these findings and recommendations offer lending institutions valuable tools for improving their lending practices, reducing loan defaults, and ultimately fostering a healthier financial environment for both borrowers and lenders. By implementing these insights, lenders can make more informed decisions, mitigate risks, and support responsible lending practices