tanejane

Objective: To assess the influence of loan duration and instalment commitment on the credit risk class of bank customers.

Q1: What are the average loan duration and instalment commitment for customers in each credit risk class?

```
# 1. Average Loan Duration and Installment Commitment for Each Credit Risk Class
library(dplyr)
library(ggplot2)
```

1.1 Average statistics

```
avg = df %>%
  group_by(class) %>%
  summarize(avg_duration = mean(duration),
            avg_installment_commitment = mean(installment_commitment),
            sd_duration = sd(duration),
            sd_installment_commitment = sd(installment_commitment))
```

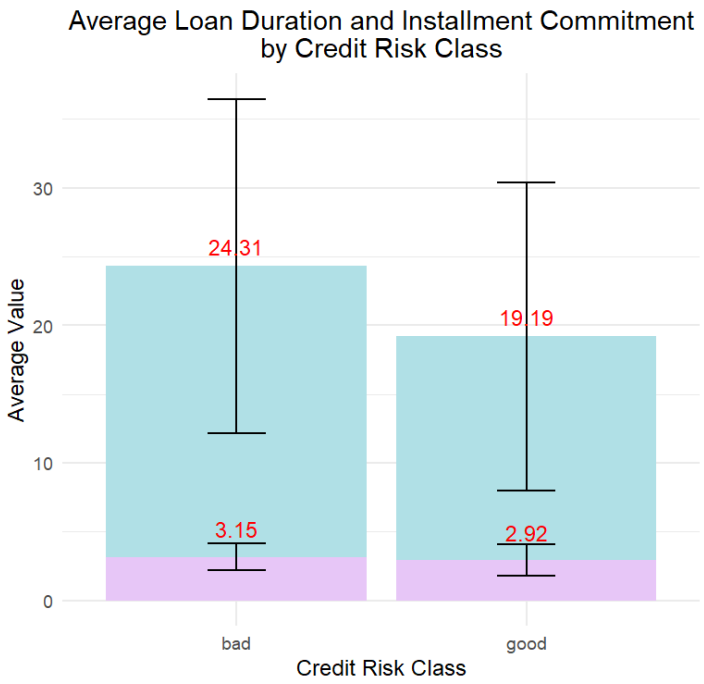
```
> print(avg)
# A tibble: 2 × 5
  class avg_duration avg_installment_commitment sd_duration sd_installment_commitment
  <chr>        <dbl>                       <dbl>       <dbl>                    <dbl>
1 bad           24.3                        3.15        12.2                    0.968
2 good          19.2                        2.92        11.2                    1.14
```

It shows the average statistic by each class.

1.2 Bar Plot Visualization

```
ggplot(avg, aes(x = class)) +
  # Bar and error bars for Loan Duration
  geom_bar(aes(y = avg_duration), fill = "powderblue",
           stat = "identity", position = position_dodge(width = 0.7)) +
  geom_errorbar(aes(ymin = avg_duration - sd_duration,
                    ymax = avg_duration + sd_duration),
                width = 0.2, position = position_dodge(width = 0.7)) +
  # Add text labels for Loan Duration
  geom_text(aes(y = avg_duration, label = round(avg_duration, 2)),
            position = position_dodge(width = 0.7), vjust = -0.7, color = "red") +

  # Bar and error bars for Installment Commitment
  geom_bar(aes(y = avg_installment_commitment), fill = "plum1",
           stat = "identity", position = position_dodge(width = 0.7), alpha = 0.7) +
  geom_errorbar(aes(ymin = avg_installment_commitment - sd_installment_commitment,
                    ymax = avg_installment_commitment + sd_installment_commitment),
                width = 0.2, position = position_dodge(width = 0.7)) +
  # Add text labels for Installment Commitment
  geom_text(aes(y = avg_installment_commitment, label = round(avg_installment_commitment, 2)),
            position = position_dodge(width = 0.7), vjust = -0.7, color = "red") +

  labs(title = "Average Loan Duration and Installment Commitment\nby Credit Risk Class",
       y = "Average Value", x = "Credit Risk Class") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) # Make the title in the middle
```

tanejane



**plum = instalment commitment; blue = loan duration*

This plot compares **loan duration** and **instalment commitment** across **good and bad credit risk classes**. The bad credit class has an average loan duration of 24.31 months with high variability (error bars ranging from 10 to 35), while the good credit class has an average of 19.19 months with less variability. The bad credit class also has a higher average instalment commitment (3.15) compared to the good credit class (2.92). Overall, **bad credit customers have longer loan durations and higher instalment commitments**.

tanejane

Q2: How are the credit risk class distribution across different range of loan duration and instalment commitment?

```
#2. How are the credit risk class distribution across different range of loan duration and installment commitment?
library(dplyr)
library(ggplot2)
library(scales) # For percentage formatting
```

2.1 Categorize loan duration and instalment commitment

```
# Categorize loan duration and installment commitment into ranges
df <- df %>%
  mutate(loan_range = cut(duration, breaks = c(0, 2, 5, 10, Inf),
                          labels = c("0-2", "3-5", "6-10", "10+")),
         installment_commitment_range = cut(installment_commitment,
                                            breaks = c(0, 2, 4, 6, Inf),
                                            labels = c("0-2%", "2-4%", "4-6%", "6%+"))
  )
```

By creating loan ranges (0-2, 3-5, 6-10, 10+) and instalment commitment ranges (0-2%, 2-4%, 4-6%, 6%+), the distribution of credit risk classes becomes clearer.

2.2 Proportion for Each Range Combination

```
# Validate and explore class distribution
class_distribution <- df %>%
  group_by(loan_range, installment_commitment_range, class) %>%
  summarise(count = n(), .groups = "drop") %>%
  mutate(proportion = count / sum(count))
  # Calculate proportions for each combination

print(class_distribution)
```

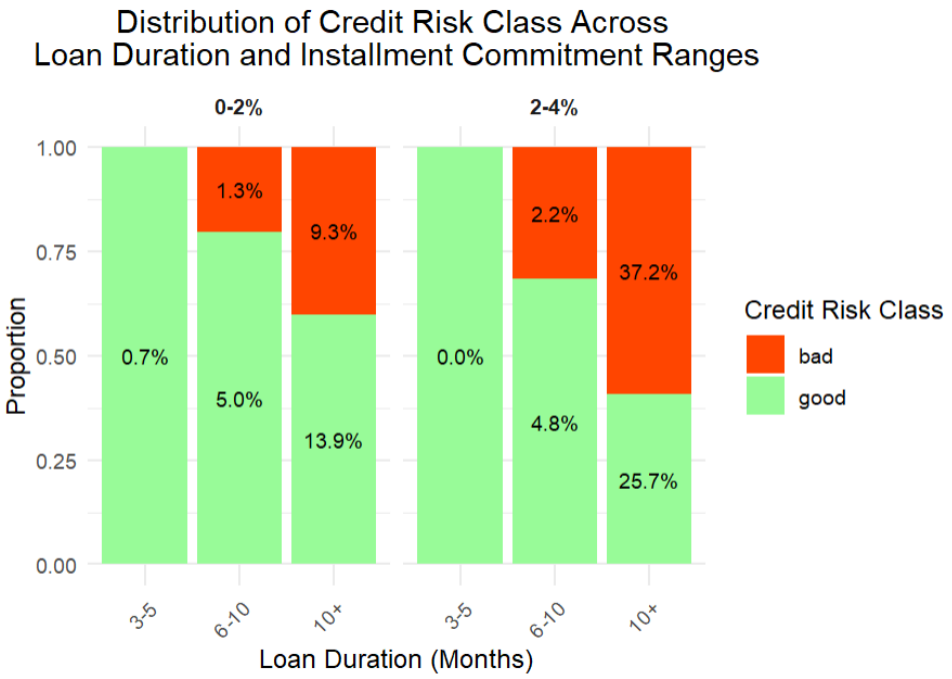
```
> print(class_distribution)
# A tibble: 10 × 5
   loan_range installment_commitment_ra…¹ class count proportion
   <fct>      <fct>                       <chr> <int>      <dbl>
 1 3-5        0-2%                        good     22   0.00677
 2 3-5        2-4%                        good      1   0.000308
 3 6-10       0-2%                        bad      41   0.0126
 4 6-10       0-2%                        good    161   0.0495
 5 6-10       2-4%                        bad      72   0.0222
 6 6-10       2-4%                        good    156   0.048
 7 10+        0-2%                        bad     303   0.0932
 8 10+        0-2%                        good    451   0.139
 9 10+        2-4%                        bad    1208   0.372
10 10+        2-4%                        good    835   0.257
```

This tibble shows the count and proportion of credit risk class across of loan_range and installment_commitment_range. Loans in the 10+ months duration and 2-4% instalment dominate the dataset, comprising 25.7% 'good' and 37.2% 'bad' loans. For 10+ months and 0-2% instalment, the 'good' class has higher count (451) than 'bad' (303). However, loans

between 3-5 months of duration have lower counts and proportions compared to longer durations.

2.3 Stacked Bar Plot Visualization

```
# Visualize proportions with percentage labels
ggplot(class_distribution, aes(x = loan_range, y = proportion, fill = class)) +
  geom_bar(stat = "identity", position = "fill") +
  geom_text(
    aes(label = percent(proportion, accuracy = 0.1)),
    position = position_fill(vjust = 0.5),
    size = 3
  ) +
  facet_wrap(~ installment_commitment_range, ncol = 2) +
  labs(
    title = "Distribution of Credit Risk Class Across \nLoan Duration and Installment Commitment Ranges",
    x = "Loan Duration (Months)",
    y = "Proportion",
    fill = "Credit Risk Class"
  ) +
  scale_fill_manual(values = c("good" = "palegreen", "bad" = "orangered")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1),
    strip.text = element_text(face = "bold")
  )
```



This plot visualizes the distribution of credit risk classes (good vs. bad) across loan duration with two facets based on instalment commitment ranges (0-2% and 2-4%).

For instalment commitment range **(0-2%), 'good' credit class is more prevalent**. However, in 2-4% range, the bad credit class dominates for loan durations over 10 months (37.2% vs. 25.7% for the good class).

tanejane

Overall, **higher instalment commitment with longer loan duration tends to be classified in 'bad' credit risk class**.

tanejane

Q3: What are the underlying patterns and factors for loan durations and instalment commitments in credit class classification?

```
#3. What are the underlying patterns and factors for loan durations and installment commitments
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(caret)
```

3.1 Descriptive Statistics

```
summary_by_class <- df %>%
  group_by(class) %>%
  summarise(duration_mean = mean(duration),
            duration_median = median(duration),
            duration_sd = sd(duration),
            installment_commitment_mean = mean(installment_commitment),
            installment_commitment_median = median(installment_commitment),
            installment_commitment_sd = sd(installment_commitment))

print(summary_by_class)
```

```
> print(summary_by_class)
# A tibble: 2 × 7
  class duration_mean duration_median duration_sd installment_commitment_mean installment_commitment_m…¹ installment_commitme…²
  <chr>         <db1>           <db1>       <db1>                       <db1>                       <db1>                   <db1>
1 bad            24.3            22.2        12.2                        3.15                        3.55                   0.968
2 good           19.2            18          11.2                        2.92                        3                      1.14
# ℹ abbreviated names: ¹installment_commitment_median, ²installment_commitment_sd
```

For **loan duration**, **bad credit customers have a higher average loan duration** (24.3 months) and median (22.2 months) compared to the good class (19.2 months and 18 months, respectively). The standard deviation indicates more variability in loan duration for the bad class (12.2 months) than for the good class (11.2 months).

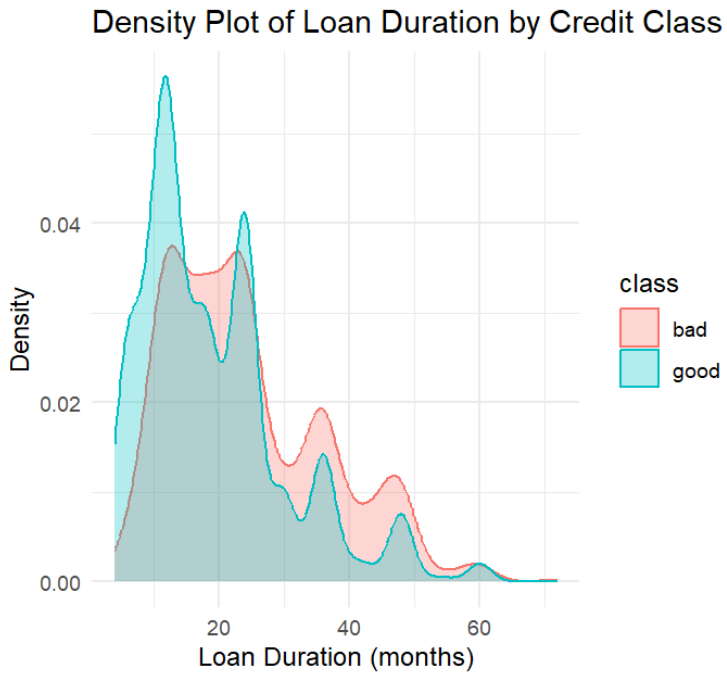
Conversely, **bad credit customers have a slightly higher average instalment commitment** (3.15) than the good one (2.92) and more variability. However, the difference in median values is not very large.

In summary, **bad credit customers tend to have longer loan durations and slightly higher instalment commitments**.

tanejane

3.2 Density Plot Visualization

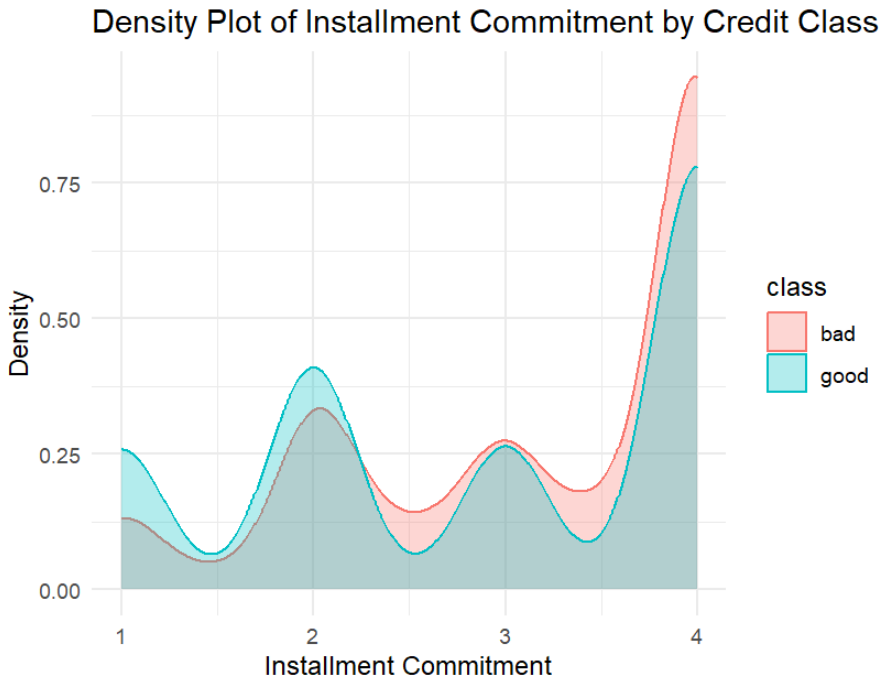
```
# Density Plot of Loan Duration by Credit Class
ggplot(df, aes(x = duration, fill = class, color = class)) +
  geom_density(alpha = 0.3) +
  labs(title = "Density Plot of Loan Duration by Credit Class",
       x = "Loan Duration (months)",
       y = "Density") +
  theme_minimal()
```



For the good class, the peak is at around 10 months, indicating that most of customers in the good credit class have loan durations close to 10 months. At around 30 months, bad and good class curves start to overlap, with the bad class curve being higher. This means that as loan durations increase, the likelihood of a customer belonging to the bad credit class becomes higher. This suggests that **longer loan durations are associated with a higher probability of having a bad credit risk**.

tanejane

```
# Density Plot of Installment Commitment by Credit Class
ggplot(df, aes(x = installment_commitment, fill = class, color = class)) +
  geom_density(alpha = 0.3) +
  labs(title = "Density Plot of Installment Commitment by Credit Class",
       x = "Installment Commitment",
       y = "Density") +
  theme_minimal()
```



At the beginning, it shows that **fewer individuals in the bad class have low instalment commitments** compared to the good class. The bad class curve becomes higher than the good class curve around 2.5. Toward the higher end of instalment commitment, the bad class consistently shows a higher density than the good class. In summary, the plot highlights that **individuals with bad credit are more likely to have higher instalment commitments**, while those with good credit tend to have lower instalment commitments.

tanejane

3.3 Welch's T-test

```
# T-test for loan duration by credit class
t_test_duration <- t.test(duration ~ class, data = df)
print(t_test_duration)

# T-test for installment commitment by credit class
t_test_installment <- t.test(installment_commitment ~ class, data = df)
print(t_test_installment)
```

```
        Welch Two Sample t-test

data:  duration by class
t = 12.479, df = 3225.3, p-value < 2.2e-16
alternative hypothesis: true difference in means between group bad and group good is not equal to 0
95 percent confidence interval:
 4.312063 5.919693
sample estimates:
 mean in group bad mean in group good
          24.30874           19.19287
```

From the sample estimates, the mean **loan duration** for the bad class is 24.31, while for the good class, it is 19.20. This confirms that customers in the **bad credit class tend to have significantly longer loan durations** compared to good credit class.

There is an estimated difference in means is approximately **5.12 months** (midpoint of the confidence interval). Specifically, loans in the **bad class have an average duration approximately 5.6 months longer than those in the good class**. This difference is highly statistically significant, as indicated by the very low p-value (<2.2e-16).

```
        Welch Two Sample t-test

data:  installment_commitment by class
t = 6.4377, df = 3169.4, p-value = 1.397e-10
alternative hypothesis: true difference in means between group bad and group good is not equal to 0
95 percent confidence interval:
 0.1657942 0.3110162
sample estimates:
 mean in group bad mean in group good
          3.153430           2.915025
```

The t-statistic **(6.4377)** indicates a noticeable difference between the means of bad and good classes. On average, **'bad' credit customers allocate 3.15%** of their income to loan repayment whereas **good credit customers allocate 2.92%**. The small p-value (1.397e-10) and the confidence interval confirm the reliability of this result, confirming that customers in the **bad credit class tend to have slightly higher instalment commitments**.

tanejane

3.4 Correlation Analysis

```
# Correlation between loan duration and installment commitment by credit class
cor_good <- cor(df %>% filter(class == "good")
                %>% select(duration, installment_commitment))
cor_bad <- cor(df %>% filter(class == "bad")
               %>% select(duration, installment_commitment))

print(paste("Correlation for good class: ", cor_good))
print(paste("Correlation for bad class: ", cor_bad))
```

```
> print(paste("Correlation for good class: ", cor_good))
[1] "Correlation for good class:  1"
[2] "Correlation for good class:  0.0967057779698294"
[3] "Correlation for good class:  0.0967057779698294"
[4] "Correlation for good class:  1"
> print(paste("Correlation for bad class: ", cor_bad))
[1] "Correlation for bad class:  1"
[2] "Correlation for bad class:  -0.0355623030784395"
[3] "Correlation for bad class:  -0.0355623030784395"
[4] "Correlation for bad class:  1"
```

The correlation between duration and instalment commitment in the **good class** is **0.0967**, indicating a very **weak positive relationship** between the two variables.

In **bad class**, the correlation is **-0.0356**, which is also **close to 0**, showing a very **weak negative relationship** between the two variables.

In practical terms, for both class, there is almost no linear relationship between duration and instalment commitment.

3.5 Logistic Regression Model

```
df$class <- factor(df$class, levels = c("good", "bad"), labels = c(0, 1))

model <- glm(class ~ duration + installment_commitment, family = binomial, data = df)
summary(model)
```

```
Call:
glm(formula = class ~ duration + installment_commitment, family = binomial,
    data = df)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.41025    0.12929 -10.907  < 2e-16 ***
duration               0.03712    0.00319  11.635  < 2e-16 ***
installment_commitment  0.19977    0.03432   5.821 5.87e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4505.5  on 3249  degrees of freedom
Residual deviance: 4318.4  on 3247  degrees of freedom
AIC: 4324.4

Number of Fisher Scoring iterations: 4
```

tanejane

The coefficient for duration indicates that **1-month increase in loan duration increases the log-odds of class being 'bad' by 0.037**. Similarly, **1-unit raises in installment commitment increases the log-odds of class being 'bad' by 0.199**. Both **duration (<2e-16)** and **installment commitment (5.87e-09)** are significant predictors of the credit class (p-values < 0.001), meaning they have a **meaningful impact** on the outcome.

This analysis assume that **longer loan durations and higher installment commitments increase the likelihood of being classified as 'bad' credit risk**.

tanejane

3.6 Log-loss Calculation

```
# Predict probabilities from the logistic regression model
df$predicted_prob <- predict(model, type = "response")

# Convert 'class' to numeric for log-loss calculation (0 for 'good' and 1 for 'bad')
df$class_numeric <- as.numeric(df$class)  # 'good' = 0, 'bad' = 1

# Calculate log-loss
log_loss <- -mean(df$class_numeric * log(df$predicted_prob) + (1 - df$class_numeric) * log(1 - df$predicted_prob))

# Print the log-loss value
print(paste("Log-Loss: ", log_loss))
```

```
# Calculate the proportion of 'good' and 'bad' classes
p_good <- mean(df$class_numeric == 0)  # Proportion of good cases
p_bad <- mean(df$class_numeric == 1)   # Proportion of bad cases

# Baseline probabilities
prob_good <- p_good
prob_bad <- p_bad

# Baseline log-loss calculation
baseline_log_loss <- -mean(df$class_numeric * log(prob_bad) + (1 - df$class_numeric) * log(prob_good))

# Print the baseline log-loss
print(paste("Baseline Log-Loss: ", baseline_log_loss))
```

```
> print(paste("Log-Loss: ", log_loss))
[1] "Log-Loss:  0.661250157540023"
```

```
> print(paste("Baseline Log-Loss: ", baseline_log_loss))
[1] "Baseline Log-Loss:  0.693146991210821"
```

The **log-loss of 0.6613**, lower than the **baseline log-loss of 0.6931** demonstrate good model performance. The baseline assumes naive predictions based on class proportions (~50% for each class). A log-loss of **0.6613** shows the model provides more accurate predictions than simply guessing based on the class distribution. In general, **the lower the log-loss, the better the predictions align with the true outcomes**.

tanejane

Q4: Can loan duration and instalment commitment be used to predict a customer's credit risk class?

```
#4. Can loan duration and installment commitment be used to predict a customer's credit risk class?
library(ggplot2)
library(randomForest)
library(caret)
library(pdp)
library(pROC)
```

4.1 Data Preprocessing

```
# Ensure the 'class' variable is a factor
df$class = as.factor(df$class)
df$class = factor(df$class, levels = c("good", "bad"), labels = c(0, 1))
```

This step is to ensure that the 'class' is a factor to be used for the model analysis.

4.2 Cross Validation

```
# Set up cross-validation (10-fold)
train_control <- trainControl(method = "cv", number = 10)

# Perform training with cross-validation
set.seed(123)  # For reproducibility
rf_model_cv <- train(class ~ duration + installment_commitment,
                     data = df,
                     method = "rf",
                     trControl = train_control,
                     ntree = 500)

# Print cross-validation results
print(rf_model_cv)

# Extract cross-validation results
print(rf_model_cv$results)
```

```
> print(rf_model_cv)
Random Forest

3250 samples
   2 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2924, 2925, 2926, 2925, 2925, 2926, ...
Resampling results:

  Accuracy   Kappa
  0.7627598  0.5254283
```

tanejane

This output summarizes a 10-fold cross-validation of Random Forest model with 3250 rows, two predictor variables (duration, installment commitment) and two levels of target variable (class).

The model trains on 90% of the data and tested on the remaining 10%, repeating this process 10 times. The average accuracy is **76.28%**, showing the proportion of correct prediction. Moreover, Kappa statistic **(0.5254)** indicates moderate agreement between predicted and actual classes.

```
> print(rf_model_cv$results)
  mtry  Accuracy     Kappa AccuracySD    KappaSD
1    2 0.7627598 0.5254283 0.02442143 0.04872641
```

This result displays the tuning parameter (mtry) with two predictors. The model's accuracy is 76.28% and standard deviation is **±2.44%** across the folds. Alternatively, the average Kappa statistic is 0.5254 where the Kappa values may vary by about **±0.0487**. **Low standard deviations** indicate **stable performance** across the 10 folds.

tanejane

4.3 Confusion Matrix

```
# Confusion Matrix and Accuracy
rf_predictions_cv <- predict(rf_model_cv, df)
confusionMatrix(rf_predictions_cv, df$class)
```

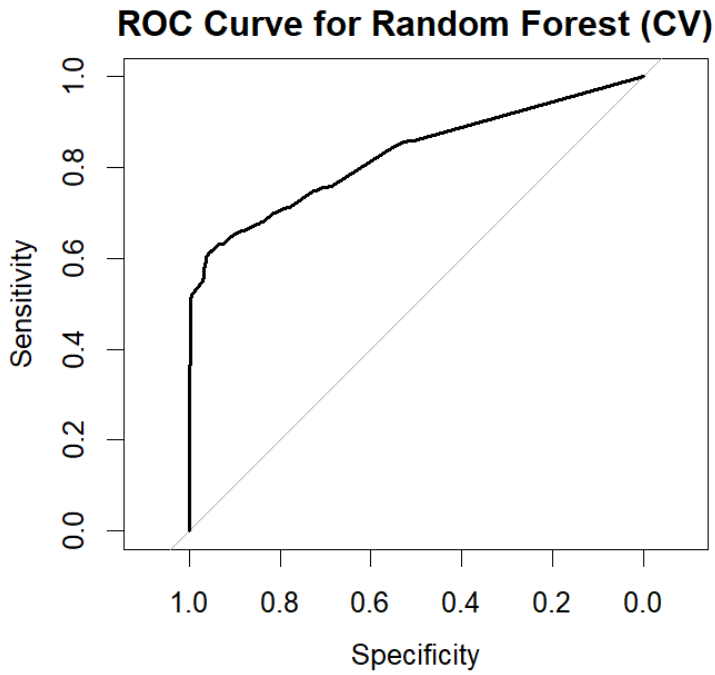
```
> confusionMatrix(rf_predictions_cv, df$class)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1546  625
         1   80  999

               Accuracy : 0.7831
                 95% CI : (0.7685, 0.7971)
    No Information Rate : 0.5003
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5661

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9508
            Specificity : 0.6151
         Pos Pred Value : 0.7121
         Neg Pred Value : 0.9259
             Prevalence : 0.5003
         Detection Rate : 0.4757
   Detection Prevalence : 0.6680
      Balanced Accuracy : 0.7830

       'Positive' Class : 0
```

This result indicates **accuracy of 78.31%** which the overall proportion of correctly classified cases. For **sensitivity (0.9508)**, high sensitivity (95%) means the model rarely misses 'good' credit risk customers. A **specificity of 61.5%** indicates the model is moderately good at detecting 'bad' credit risks.

tanejane

4.4 Area Under the Curve for the Receiver Operating Characteristic

```
# AUC from the cross-validated model
rf_probabilities_cv <- predict(rf_model_cv, df, type = "prob")
# Use probabilities for the positive class
roc_curve_cv <- roc(df$class, rf_probabilities_cv[, 2])
plot(roc_curve_cv, main = "ROC Curve for Random Forest (CV)")
auc(roc_curve_cv)
```




```
> auc(roc_curve_cv)
Area under the curve: 0.8276
```

The **AUC** for ROC curve is **0.8276** indicating the Random Forest model has an **82.76% chance of correctly distinguishing between good and bad credit risks** across all classification thresholds.

tanejane

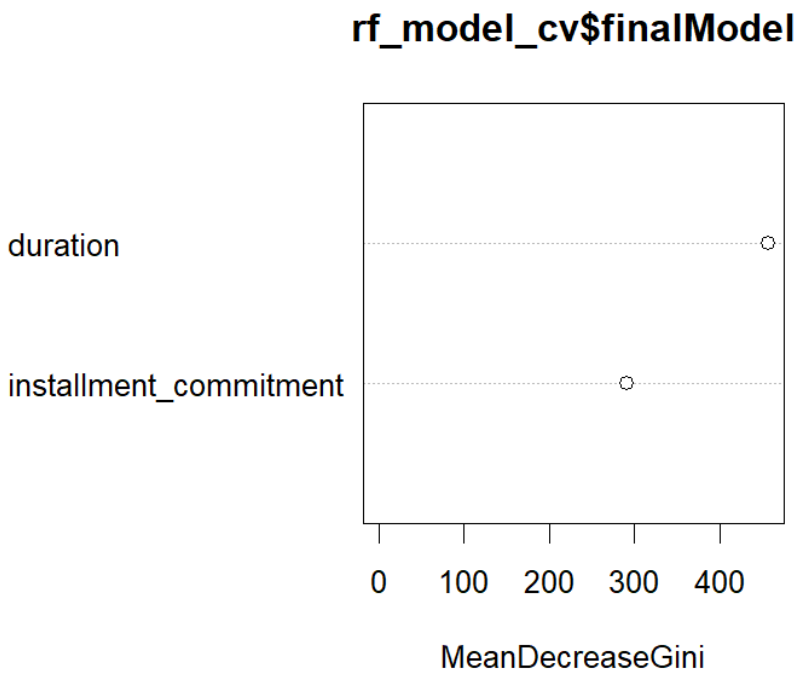
4.5 Variable Importance Plot

```
# Variable importance plot from the cross-validated model
importance(rf_model_cv$finalModel)
varImpPlot(rf_model_cv$finalModel)
```

```
> importance(rf_model_cv$finalModel)
                       MeanDecreaseGini
duration                       457.2248
installment_commitment         290.7741
```

Mean Decrease in Gini Index measures each predictor's contribution to improving the classification accuracy of the model.

Duration (457.2248) is a more **important predictor** in the model, as it results in the largest decrease in the Gini Index, showing **a strong relationship** with customers classification. I**nstallment commitment (290.7741)** also contributes but is **less important than duration**.



Duration is more influential than installment commitment in predicting credit risk, with the model relying more on duration-based splits to classify good and bad credit risks.

tanejane

Q5: What are the loan duration and installment commitment thresholds can minimize the number of bad credit customers?

```
#5. What loan duration and installment commitment values are associated with a higher probability
library(ggplot2)
library(caret)
library(pROC)
library(PRROC)
library(xgboost)
```

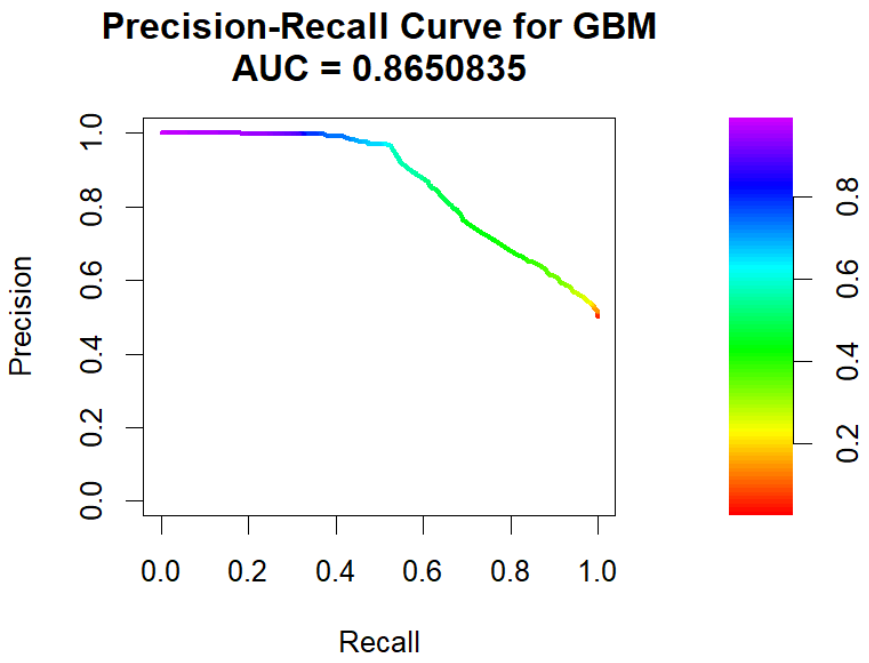
5.1 Data Preprocessing

```
# Ensure the 'class' variable is a factor with 0 for "good" and 1 for "bad"
df$class = as.factor(df$class)
df$class = factor(df$class, levels = c("good", "bad"), labels = c(0, 1))
```

This step is to ensure that the 'class' is a factor to be used for the model analysis.

5.2 Gradient Boosting Machines (GBM)

```
# Prepare data for GBM
dtrain <- xgb.DMatrix(data = as.matrix(df[, c("duration", "installment_commitment")]),
                      label = as.numeric(as.character(df$class)))

# Fit the GBM Model
gbm_model <- xgboost(data = dtrain, max_depth = 3, eta = 0.1, nrounds = 100,
                     objective = "binary:logistic", verbose = 0)

# Predict probabilities using GBM
gbm_prob <- predict(gbm_model, as.matrix(df[, c("duration", "installment_commitment")]))

# Plot the Precision-Recall Curve for GBM
pr_curve_gbm <- pr.curve(scores.class0 = gbm_prob,
                         weights.class0 = as.numeric(as.character(df$class)),
                         curve = TRUE)

plot(pr_curve_gbm, main = "Precision-Recall Curve for GBM",
     xlab = "Recall", ylab = "Precision")
```



Recall measures the proportion of actual bad credit risks correctly identified, with higher recall reducing false negatives. **Precision** measures the proportion of predicted bad credit risks that are accurate, with higher precision reducing false positives.

```
# Calculate and print AUC for the Precision-Recall Curve
pr_auc_gbm <- pr_curve_gbm$auc.integral  # Use the integral version of AUC
cat("Precision-Recall AUC for GBM: ", pr_auc_gbm, "\n")
```

```
> cat("Precision-Recall AUC for GBM: ", pr_auc_gbm, "\n")
Precision-Recall AUC for GBM:  0.8650835
```

The high AUC (0.865) indicates that the GBM model is **effective at predicting bad credit risk**. It can maintain both good precision and recall across a range of thresholds.

tanejane

5.3 Threshold Analysis

```
# Threshold Analysis
threshold_range <- seq(0, 1, by = 0.02)
bad_risk_counts <- sapply(threshold_range, function(thresh) {
  predicted_class <- ifelse(gbm_prob > thresh, 1, 0)  # Classify based on the threshold
  sum(predicted_class == 1)  # Count the number of "bad" predictions
})

# Plot the Threshold Analysis
plot(threshold_range, bad_risk_counts, type = "b", xlab = "Threshold",
     ylab = "Number of Bad Class Customers", main = "Threshold vs Bad Class Customers (GBM)")
```

The count of 'bad risk' customers declines steadily as the threshold increases, indicating a balance between 'good' and 'bad' classifications.

```
thresholds <- seq(0, 1, by = 0.02)
f1_scores <- sapply(thresholds, function(thresh) {
  predicted_class <- ifelse(gbm_prob > thresh, 1, 0)
  cm <- confusionMatrix(factor(predicted_class), df$class, positive = "1")
  precision <- cm$byClass["Precision"]
  recall <- cm$byClass["Recall"]
  f1 <- 2 * ((precision * recall) / (precision + recall))
  return(f1)
})

optimal_threshold <- thresholds[which.max(f1_scores)]
cat("Optimal Threshold (Max F1): ", optimal_threshold, "\n")

> cat("Optimal Threshold (Max F1): ", optimal_threshold, "\n")
Optimal Threshold (Max F1):  0.38
```

The optimal threshold of 0.38 indicates that customers with probabilities above 0.38 are classified as 'bad' credit risk.

tanejane

5.4 Confusion Matrix

```
# Confusion Matrix for GBM
conf_matrix <- confusionMatrix(as.factor(df$predicted_class_gbm),
                               df$class, positive = "1")
print("Confusion Matrix for GBM:")
print(conf_matrix)
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  869  232
         1  757 1392

               Accuracy : 0.6957
                 95% CI : (0.6795, 0.7115)
    No Information Rate : 0.5003
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3915

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8571
            Specificity : 0.5344
         Pos Pred Value : 0.6477
         Neg Pred Value : 0.7893
             Prevalence : 0.4997
         Detection Rate : 0.4283
   Detection Prevalence : 0.6612
      Balanced Accuracy : 0.6958

       'Positive' Class : 1
```

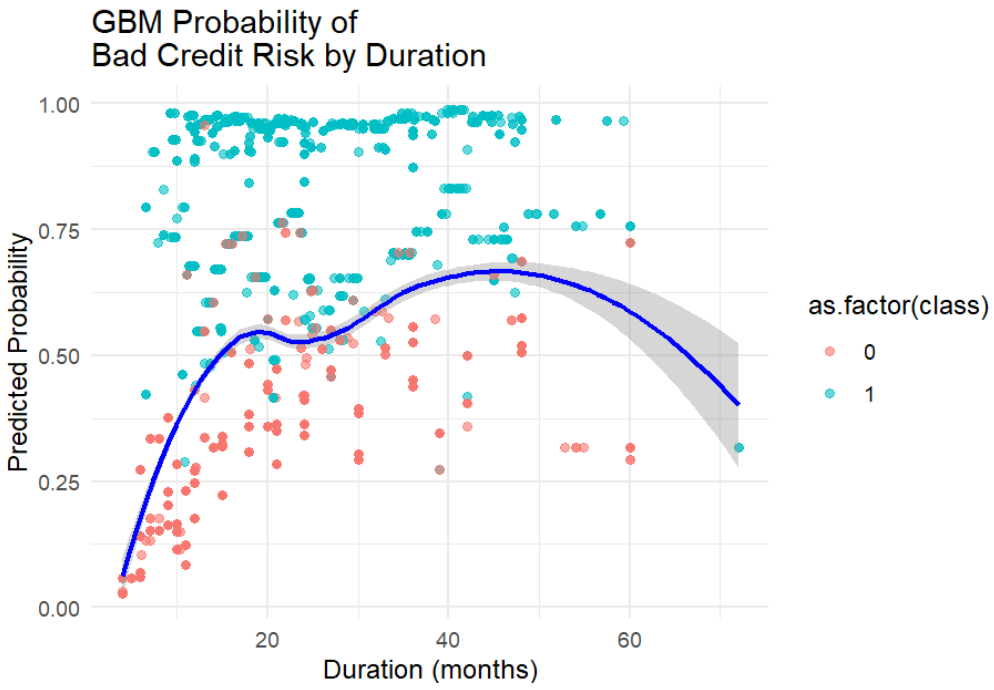
At **69.57%**, the model performs better than random guessing (the No Information Rate of 50.03%). The model correctly identifying 85.71% of the bad customers. At **53.44%**, specificity (correctly identify 'good') is lower than sensitivity.

The model is **better at identifying 'bad' customers (high Sensitivity)** than identifying 'good' customers **(low Specificity)**. This bias could be acceptable in credit risk since it's **safer to classify borderline customers as 'bad'**.

5.5 Scatter Plot Visualization

```
# Duration vs Probability
ggplot(data = df, aes(x = duration, y = gbm_prob)) +
  geom_point(aes(color = as.factor(class)), alpha = 0.6) +
  geom_smooth(method = "loess", color = "blue") +
  labs(title = "GBM Probability of Bad Credit Risk by Duration",
       x = "Duration (months)", y = "Predicted Probability") +
  theme_minimal()
```

tanejane

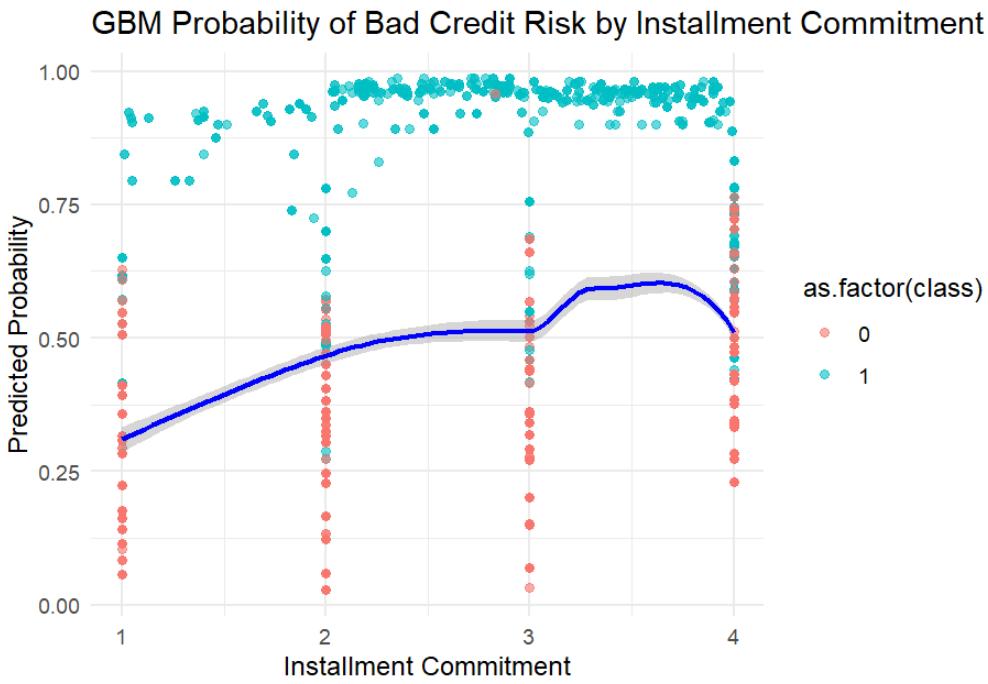


*** 0 = good; 1 = bad*

The **blue line** is a **LOESS** (Locally Weighted Scatterplot Smoothing) curve, showing overall trend of **predicted probabilities across different loan durations**. Concurrently, the **shaded region** around the blue line represents the **confidence interval** of the trendline, indicating the uncertainty of the estimate. From this plot, the **highest predicted probabilities** of bad credit risk occur between **40 and 50 months**.

```
# Installment Commitment vs Probability
ggplot(data = df, aes(x = installment_commitment, y = gbm_prob)) +
  geom_point(aes(color = as.factor(class)), alpha = 0.6) +
  geom_smooth(method = "loess", color = "blue") +
  labs(title = "GBM Probability of Bad Credit Risk by Installment Commitment",
       x = "Installment Commitment", y = "Predicted Probability") +
  theme_minimal()
```

tanejane



*** 0 = good; 1 = bad*

From this plot, customers with installment commitments of **3-4%** are more likely to be classified as bad credit risks.