Credit Risk Modeling Using CRISP-DM: A Capstone Report

Dat 690: Data Analytics capstone

Southern New Hampshire University

Instructor: Anzhelika Lyubenko

Author: Rachel Goldsbury

May 18, 2025

Credit Risk Project Summary and Analytic Plan

Project Summary

For my capstone project, I am developing a predictive model to assess the likelihood of loan default using historical credit applicant data. Loan defaults present a significant financial risk to lenders, and a reliable predictive model can help reduce these risks by supporting smarter, data-informed lending decisions. The core research question is: Can I accurately predict the probability of loan default using applicant and credit history features? This milestone is the initial plan and a part of a larger project that applies the data analysis, modeling, and communication skills I've developed throughout the program and will be carried out using the industry standard CRISP-DM framework. I will primarily use RStudio and Excel for analysis and modeling, applying techniques such as Naïve Bayes, Logistic Regression, PCA, and Decision Trees to build a model that is both accurate and interpretable for business use.

CRISP-DM Phase 1: Business Understanding

In the first phase, I explore the business problem. Loan defaults pose significant financial and reputational risks for lending institutions, making accurate risk assessment a top priority. This project aims to develop a predictive model that estimates the probability of default based on historical credit application data. The core research question guiding this project is whether or not I can accurately predict the probability of loan default using applicant and credit history features? By generating risk scores for each applicant, the model will support better lending decisions, reduce potential losses, and improve overall portfolio performance. The final model will also be designed for scalability and integration into loan application systems to enhance operational efficiency.

CRISP-DM Phase 2: Data Understanding

Understanding the data is incredibly important. This is what I do in the data understanding phase. This dataset includes information on 1,000 past credit applicants, featuring variables such as age, employment status, credit history, and whether the applicant defaulted on a loan. Initial exploration using Excel and R will involve generating descriptive statistics, checking correlations, and creating visual summaries like histograms and boxplots to uncover patterns or outliers. Special attention will be paid to identifying any class imbalances between default and non-default outcomes, as this can impact model accuracy and fairness. This will be the time to evaluate the dataset for potential bias related to gender, age, or other sensitive attributes, given the ethical and legal importance of ensuring equitable access to credit. Furthermore, Data quality will be verified by checking for missing values, duplicate records, and ensuring that data types are correctly assigned, since misclassification of variables can significantly affect model performance. Once I have a clear understanding of the business and the data available, I move on to the data preparation phase.

CRISP-DM Phase 3: Data Preparation

During the data preparation phase, I identify and choose the most relevant features, such as income level, previous defaults, and credit amount, that have the strongest influence on loan default prediction. To reduce dimensionality and simplify the model, I may apply principal component analysis (PCA) or other feature selection techniques. Missing values will be addressed through imputation or removal, and variables will be standardized or normalized as needed to ensure consistency across the dataset. At this point, I can also construct new variables where appropriate, for example a debt-to-income ratio which will add predictive value, and categorical variables will be encoded using one-hot or label encoding to make them suitable for modeling. These steps will help ensure the dataset is clean, concise, and ready for effective model training.

CRISP-DM Phase 4: Modeling

For the modeling phase, the plan is to begin by testing both Naïve Bayes and Logistic Regression, as they are well-suited for binary classification problems like loan default prediction and offer strong interpretability. I plan to also explore Decision Trees to compare performance and gain additional insights into feature importance. The dataset will be split into training and testing sets using an 80/20 split, and cross-validation will be applied to ensure model reliability. Evaluation metrics such as the confusion matrix, ROC-AUC, and precision-recall will be used to assess model performance. The final model will be selected based on its balance of predictive accuracy and ease of interpretation for stakeholders.

CRISP-DM Phase 5: Evaluation

In the evaluation phase, I will present and reflect on key performance metrics such as AUC, accuracy, and F1-score to assess how well the model predicts loan defaults. Attention will be given to analyzing false positives and false negatives, as these outcomes have important business implications such as denying credit to a qualified applicant or approving a risky one. I will also pause here to reflect on the impact of feature selection, modeling choices, and any assumptions made throughout the process. Based on the results, I will edit and finalize the model and begin documenting it for deployment. If necessary, this is where I'll propose further iterations or additional data collection initiatives.

CRISP-DM Phase 6: Deployment

The finalized credit risk model will be integrated into the organization's loan approval system, either through batch scoring or a real-time API, to support fast and consistent data backed decision making. Dashboards or reporting tools will be developed to communicate model insights and risk scores to key stakeholders consistently. To ensure continued performance, the model will undergo regular validation and be retrained periodically as new applicant data becomes available.

Conclusion

This credit risk modeling project highlights the importance of using data analytics to support more accurate, equitable, and efficient credit decision making. By applying the full CRISP-DM framework, I've demonstrated my ability to translate a real-world business problem into a structured, data backed solution. Throughout this process, I will leverage tools like RStudio and Excel, and apply key modeling techniques. Considering both technical performance and ethical responsibility is essential for success. This capstone reflects the knowledge and skills I've developed in this program and reinforces my readiness to contribute to further analytics driven projects. I'm confident that this experience has prepared me to take on real world challenges and build meaningful, production ready models with other team members.

Milestone Two: Data Understanding and Data Preparation

Data Understanding and Data Preparation

Introduction

In the credit risk modeling project use case, the business challenge is to predict the likelihood of loan default in order to reduce financial losses and improve lending decisions. Here in milestone Two, we continue the analytic plan by focusing on the Data Understanding and Data Preparation phases of the CRISP-DM framework. The analysis and transformations for this milestone will be carried out using RStudio and Excel within the uCertify virtual desktop environment.

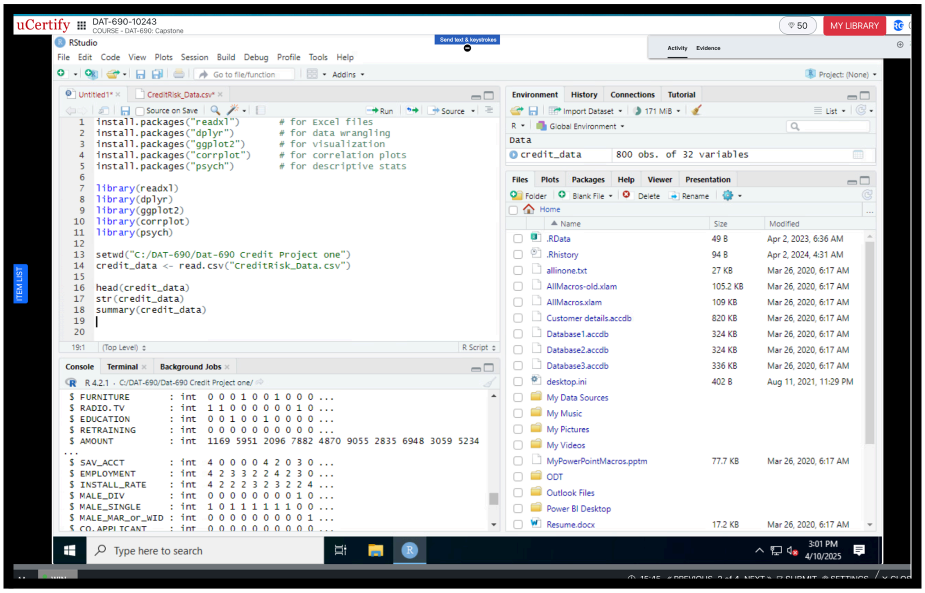
I. Select Data and Discuss Data Types

The dataset used in this project comes from the Credit Risk folder in the DAT-690 directory of the uCertify virtual desktop. It includes three files: the uncleaned raw dataset, a verified version, and an Excel file containing both the data and a code list with variable names, types, and descriptions. The dataset includes 800 observation records and 32 variables, consisting of categorical, numerical, and binary data types. Variables selected for inclusion in the analysis are those most relevant to assessing creditworthiness, such as age, duration of credit, credit amount, savings and checking account status, employment length, and credit history. These features are strong indicators of a borrower's financial behavior and risk. Some purpose related binary fields for example: NEW_CAR, RADIO/TV and redundant demographic flags such as MALE_DIV, MALE_SINGLE were excluded due to overlap, low variability, or limited contribution to predictive power. Any features with consistently missing or unknown values, for example something like PROP_UNKN_NONE, may also be excluded to maintain model accuracy and avoid noise.

II. Descriptive Statistics and Correlation Analysis

To begin exploring the dataset, R packages were installed and loaded, including dplyr for data manipulation, ggplot2 for visualization, corrplot for creating heatmaps, and psych for descriptive statistics. The dataset CreditRisk_Data.csv was loaded from the project directory using the read.csv() function. The functions head(), str(), and summary() were used to inspect the structure and contents of the dataset, revealing the distribution and types of variables and confirming that the data loaded correctly (figure 1). Descriptive statistics, including mean, median, and standard deviation, were generated for the numeric variables to understand central tendency and variation across features (figure 2). And finally, a correlation matrix was then calculated using the cor() function and visualized with corrplot(), producing a heatmap, seen in figure 3 to assess linear relationships between variables. The heatmap showed a moderate positive correlation between DURATION and AMOUNT, suggesting that longer loans tend to have higher amounts. Other variables such as AGE, NUM_CREDITS, and NUM_DEPENDENTS demonstrated little to no correlation with each other, indicating they may contribute unique value to the model. No multicollinearity was detected, meaning all selected features can be retained for now, though the relationship between DURATION and AMOUNT will be monitored in later modeling stages.

Figure 1:The functions head(), str(), and summary() functions



```r
install.packages("readxl")        # for Excel files
install.packages("dplyr")         # for data wrangling
install.packages("ggplot2")       # for visualization
install.packages("corrplot")      # for correlation plots
install.packages("psych")         # for descriptive stats

library(readxl)
library(dplyr)
library(ggplot2)
library(corrplot)
library(psych)

setwd("C:/DAT-690/Dat-690 Credit Project one")
credit_data <- read.csv("CreditRisk_Data.csv")

head(credit_data)
str(credit_data)
summary(credit_data)
```

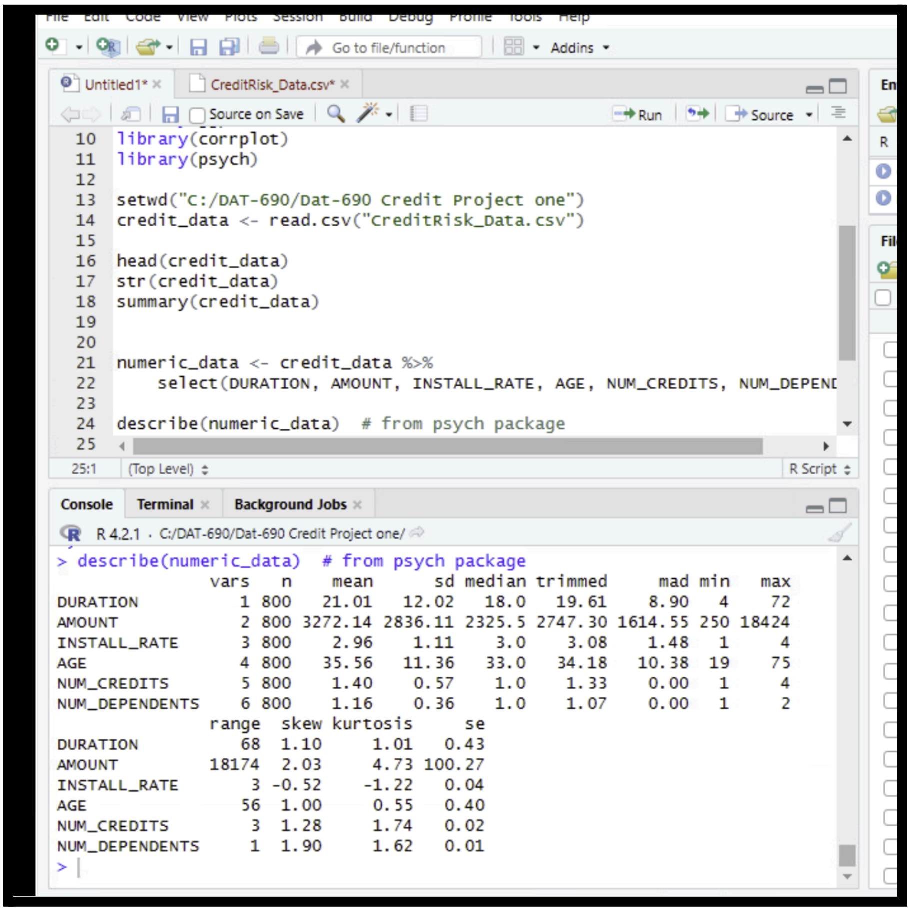
```
R 4.2.1 · C:/DAT-690/Dat-690 Credit Project one/
 $ FURNITURE       : int  0 0 0 1 0 0 1 0 0 0 ...
 $ RADIO.TV        : int  1 1 0 0 0 0 0 0 1 0 ...
 $ EDUCATION       : int  0 0 1 0 0 1 0 0 0 0 ...
 $ RETRAINING      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMOUNT          : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234
...
 $ SAV_ACCT        : int  4 0 0 0 0 4 2 0 3 0 ...
 $ EMPLOYMENT      : int  4 2 3 3 2 2 4 2 3 0 ...
 $ INSTALL_RATE    : int  4 2 2 2 3 2 3 2 2 4 ...
 $ MALE_DIV        : int  0 0 0 0 0 0 0 0 1 0 ...
 $ MALE_SINGLE     : int  1 0 1 1 1 1 1 0 0 ...
 $ MALE_MAR_or_WID : int  0 0 0 0 0 0 0 0 1 ...
 $ CO.APPLICANT    : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
Data
credit_data    800 obs. of 32 variables
```

Name	Size	Modified
.RData	49 B	Apr 2, 2023, 6:36 AM
.Rhistory	94 B	Apr 2, 2024, 4:31 AM
allinone.txt	27 KB	Mar 26, 2020, 6:17 AM
AllMacros-old.xlam	105.2 KB	Mar 26, 2020, 6:17 AM
AllMacros.xlam	109 KB	Mar 26, 2020, 6:17 AM
Customer details.accdb	820 KB	Mar 26, 2020, 6:17 AM
Database1.accdb	324 KB	Mar 26, 2020, 6:17 AM
Database2.accdb	324 KB	Mar 26, 2020, 6:17 AM
Database3.accdb	336 KB	Mar 26, 2020, 6:17 AM
desktop.ini	402 B	Aug 11, 2021, 11:29 PM
My Data Sources		
My Music		
My Pictures		
My Videos		
MyPowerPointMacros.pptm	77.7 KB	Mar 26, 2020, 6:17 AM
ODT		
Outlook Files		
Power BI Desktop		
Resume.docx	17.2 KB	Mar 26, 2020, 6:17 AM

uCertify
DAT-690-10243
COURSE - DAT-690: Capstone
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled1*
CreditRisk_Data.csv*
Source on Save
Run
Source
Console
Terminal
Background Jobs
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Type here to search
3:01 PM
4/10/2025

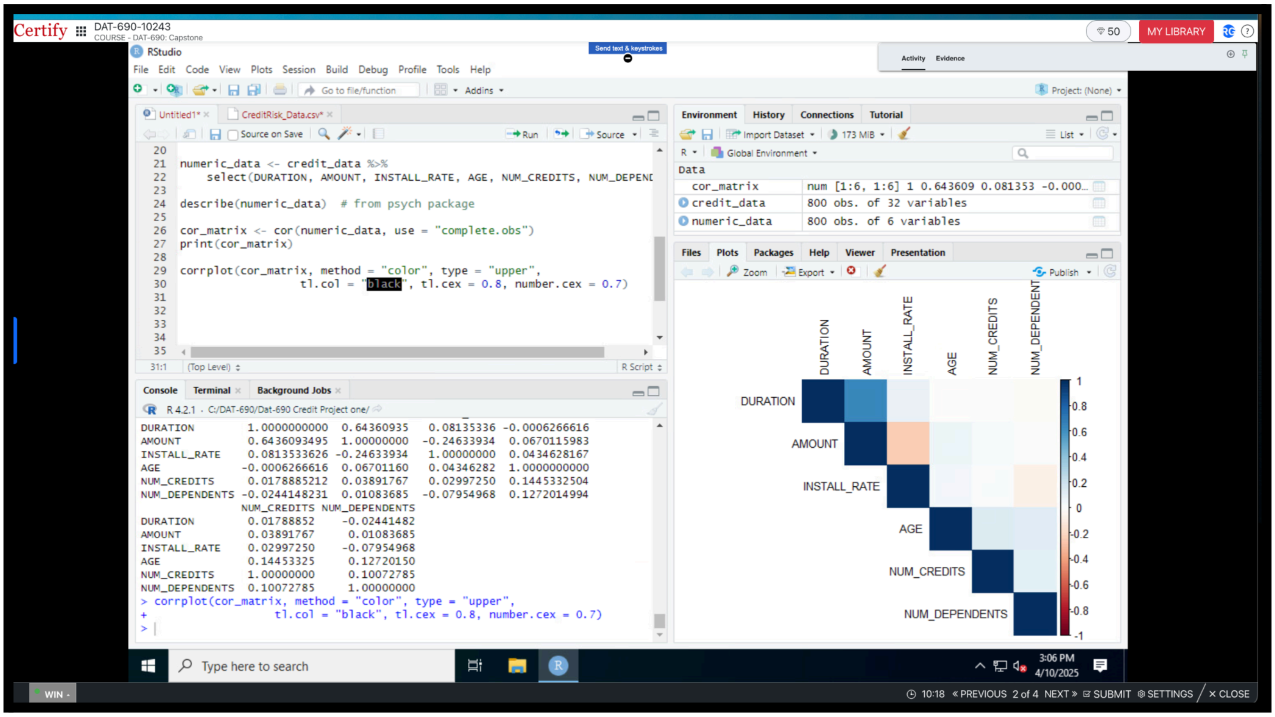
Figure 2: Descriptive statistics



```r
library(corrplot)
library(psych)

setwd("C:/DAT-690/Dat-690 Credit Project one")
credit_data <- read.csv("CreditRisk_Data.csv")

head(credit_data)
str(credit_data)
summary(credit_data)


numeric_data <- credit_data %>%
    select(DURATION, AMOUNT, INSTALL_RATE, AGE, NUM_CREDITS, NUM_DEPEND

describe(numeric_data)  # from psych package
```

```
R 4.2.1 · C:/DAT-690/Dat-690 Credit Project one/
> describe(numeric_data)  # from psych package
               vars   n    mean      sd  median trimmed     mad min   max
DURATION          1 800   21.01   12.02    18.0   19.61    8.90   4    72
AMOUNT            2 800 3272.14 2836.11  2325.5 2747.30 1614.55 250 18424
INSTALL_RATE      3 800    2.96    1.11     3.0    3.08    1.48   1     4
AGE               4 800   35.56   11.36    33.0   34.18   10.38  19    75
NUM_CREDITS       5 800    1.40    0.57     1.0    1.33    0.00   1     4
NUM_DEPENDENTS    6 800    1.16    0.36     1.0    1.07    0.00   1     2
               range  skew kurtosis     se
DURATION          68  1.10     1.01   0.43
AMOUNT         18174  2.03     4.73 100.27
INSTALL_RATE       3 -0.52    -1.22   0.04
AGE               56  1.00     0.55   0.40
NUM_CREDITS        3  1.28     1.74   0.02
NUM_DEPENDENTS     1  1.90     1.62   0.01
> |
```

Figure 3: Heatmap



```
numeric_data <- credit_data %>%
    select(DURATION, AMOUNT, INSTALL_RATE, AGE, NUM_CREDITS, NUM_DEPEND

describe(numeric_data)  # from psych package

cor_matrix <- cor(numeric_data, use = "complete.obs")
print(cor_matrix)

corrplot(cor_matrix, method = "color", type = "upper",
                  tl.col = "black", tl.cex = 0.8, number.cex = 0.7)
```

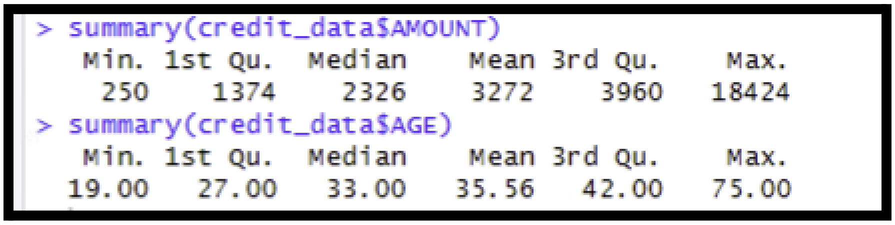
```
Data
  cor_matrix       num [1:6, 1:6] 1 0.643609 0.081353 -0.000...
  credit_data      800 obs. of 32 variables
  numeric_data     800 obs. of 6 variables
```

```
DURATION         1.0000000000  0.64360935   0.08135336 -0.0006266616
AMOUNT           0.6436093495  1.00000000  -0.24633934  0.0670115983
INSTALL_RATE     0.0813533626 -0.24633934   1.00000000  0.0434628167
AGE             -0.0006266616  0.06701160   0.04346282  1.0000000000
NUM_CREDITS      0.0178885212  0.03891767   0.02997250  0.1445332504
NUM_DEPENDENTS  -0.0244148231  0.01083685  -0.07954968  0.1272014994
                NUM_CREDITS NUM_DEPENDENTS
DURATION         0.01788852    -0.02441482
AMOUNT           0.03891767     0.01083685
INSTALL_RATE     0.02997250    -0.07954968
AGE              0.14453325     0.12720150
NUM_CREDITS      1.00000000     0.10072785
NUM_DEPENDENTS   0.10072785     1.00000000
> corrplot(cor_matrix, method = "color", type = "upper",
+                  tl.col = "black", tl.cex = 0.8, number.cex = 0.7)
> |
```

III. Descriptive Statistics for Selected Variables

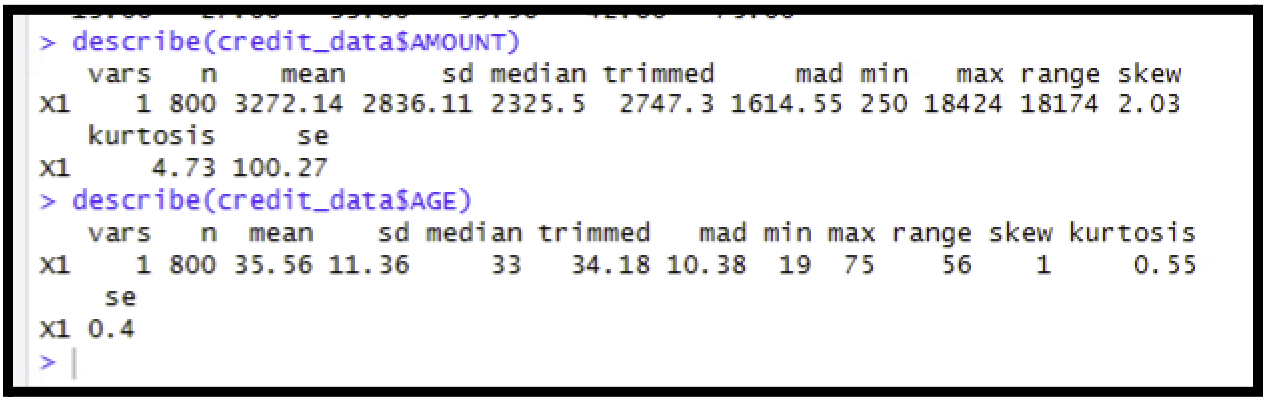
For this section, the variables AMOUNT and AGE were selected for detailed analysis due to their strong relevance to loan default risk. AMOUNT reflects the financial exposure of the lender, while AGE may reveal important demographic trends and is often a factor in assessing risk and fairness in credit decisions. Figure 4 and figure 5 display the summary statistics and for each chosen variable which includes the mean, median, standard deviation, skewness, kurtosis, and range.

Figure 4: Summary Statistics



```
> summary(credit_data$AMOUNT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    250    1374    2326    3272    3960   18424 
> summary(credit_data$AGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  19.00   27.00   33.00   35.56   42.00   75.00 
```

Figure 5: Mean, Median, Standard deviation, Skewness, Kurtosis, and Range

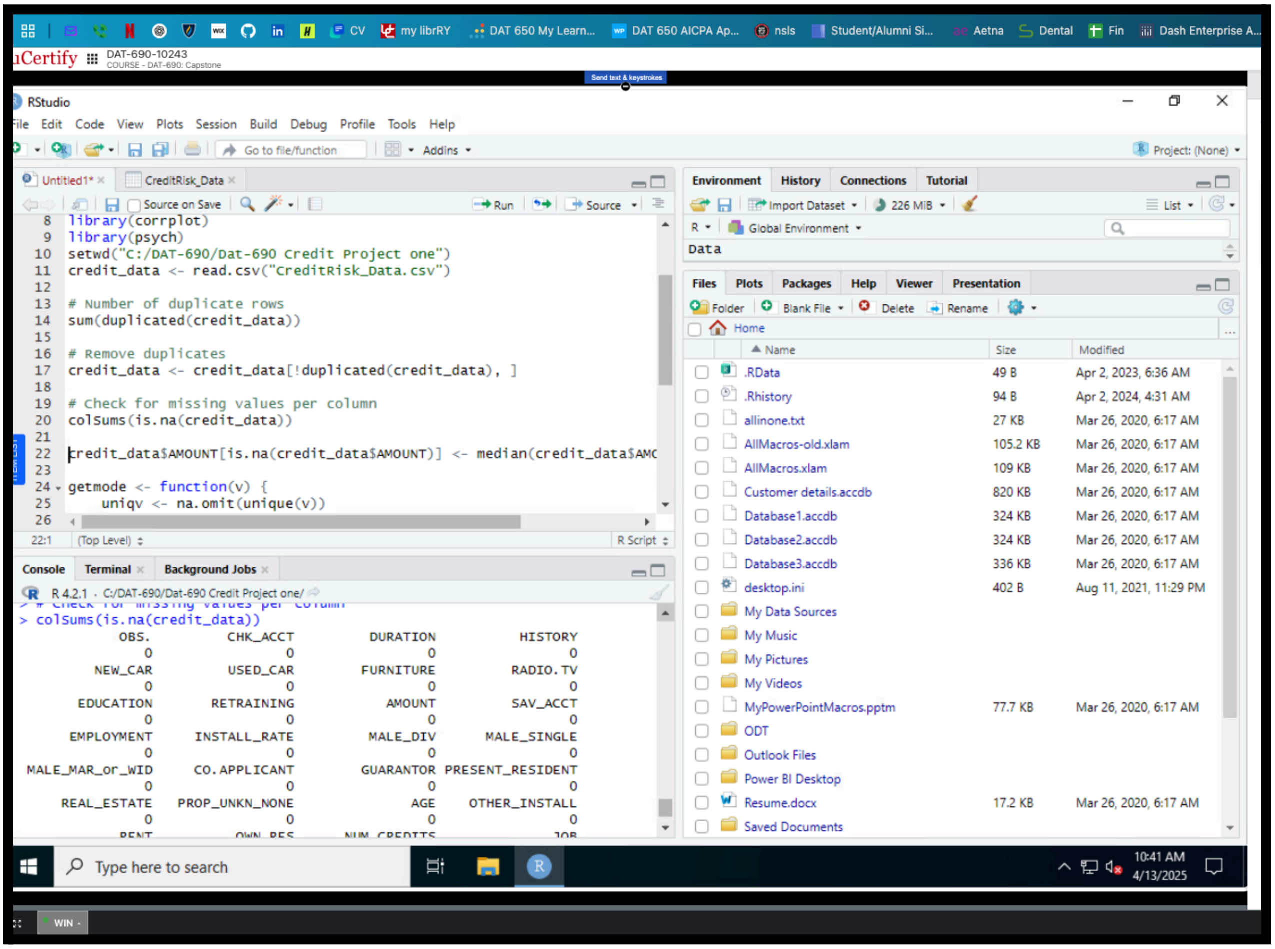


```
> describe(credit_data$AMOUNT)
   vars   n    mean      sd median trimmed     mad min   max range skew
X1    1 800 3272.14 2836.11 2325.5  2747.3 1614.55 250 18424 18174 2.03
   kurtosis     se
X1     4.73 100.27
> describe(credit_data$AGE)
   vars   n  mean    sd median trimmed   mad min max range skew kurtosis
X1    1 800 35.56 11.36     33   34.18 10.38  19  75    56    1     0.55
    se
X1 0.4
> 
```

IV. Data Preparation and Cleaning Process

In the Data Preparation phase, several important quality checks were performed to ensure the dataset was ready for modeling. First, the dataset was examined for duplicate records using the duplicated() function in R, and it was confirmed that there were zero duplicate rows, meaning no records needed to be removed. Next, missing values were assessed across all variables using colSums(), and the output showed that the dataset contained no missing data in either numeric or categorical fields. Although imputation was not necessary in this case, planned strategies such as median imputation for numeric variables and mode imputation for categorical variables were in place in case issues were detected. Additionally, the structure of the dataset was reviewed using the str() function to verify that each variable was assigned the correct data type. Categorical variables such as SAV_ACCT and CHK_ACCT were explicitly converted to factor types to ensure compatibility with modeling functions, while numeric variables were confirmed as either integers or numerics. These steps verified that the dataset was complete, consistent, and properly formatted, making it suitable for feature engineering and future modeling in the next CRISP-DM phase.

Figure 6: Preparation and cleaning of data.



```r
library(corrplot)
library(psych)
setwd("C:/DAT-690/Dat-690 Credit Project one")
credit_data <- read.csv("CreditRisk_Data.csv")

# Number of duplicate rows
sum(duplicated(credit_data))

# Remove duplicates
credit_data <- credit_data[!duplicated(credit_data), ]

# Check for missing values per column
colSums(is.na(credit_data))

credit_data$AMOUNT[is.na(credit_data$AMOUNT)] <- median(credit_data$AMC

getmode <- function(v) {
    uniqv <- na.omit(unique(v))
```

```
> # Check for missing values per column
> colSums(is.na(credit_data))
           OBS.        CHK_ACCT        DURATION         HISTORY
              0               0               0               0
        NEW_CAR        USED_CAR       FURNITURE        RADIO.TV
              0               0               0               0
      EDUCATION      RETRAINING          AMOUNT        SAV_ACCT
              0               0               0               0
     EMPLOYMENT    INSTALL_RATE        MALE_DIV     MALE_SINGLE
              0               0               0               0
MALE_MAR_or_WID    CO.APPLICANT       GUARANTOR PRESENT_RESIDENT
              0               0               0               0
    REAL_ESTATE  PROP_UNKN_NONE             AGE   OTHER_INSTALL
              0               0               0               0
           RENT         OWN_RES     NUM_CREDITS             JOB
```

Name	Size	Modified
.RData	49 B	Apr 2, 2023, 6:36 AM
.Rhistory	94 B	Apr 2, 2024, 4:31 AM
allinone.txt	27 KB	Mar 26, 2020, 6:17 AM
AllMacros-old.xlam	105.2 KB	Mar 26, 2020, 6:17 AM
AllMacros.xlam	109 KB	Mar 26, 2020, 6:17 AM
Customer details.accdb	820 KB	Mar 26, 2020, 6:17 AM
Database1.accdb	324 KB	Mar 26, 2020, 6:17 AM
Database2.accdb	324 KB	Mar 26, 2020, 6:17 AM
Database3.accdb	336 KB	Mar 26, 2020, 6:17 AM
desktop.ini	402 B	Aug 11, 2021, 11:29 PM
My Data Sources		
My Music		
My Pictures		
My Videos		
MyPowerPointMacros.pptm	77.7 KB	Mar 26, 2020, 6:17 AM
ODT		
Outlook Files		
Power BI Desktop		
Resume.docx	17.2 KB	Mar 26, 2020, 6:17 AM
Saved Documents		

V. Feature Construction, Integration, and Formatting

In this phase, I focused on preparing the dataset for modeling by refining feature formatting and structure. No derived attributes were added and key numeric features such as DURATION, AMOUNT, and AGE were retained due to their importance in evaluating credit risk. Categorical variables, including SAV_ACCT, were transformed using one-hot encoding to convert each category into a binary column, ensuring compatibility with most machine learning algorithms. Additionally, the selected numeric variables were standardized using the scale() function in R to ensure consistent scaling across features, which helps prevent bias in algorithms sensitive to feature magnitude. No external data sources were merged, as the dataset provided was already complete for the scope of this pilot. A final structure check confirmed the correct number of rows and columns, and that all variables were appropriately typed and formatted, confirming the dataset is fully ready for the modeling phase.

Conclusion

This milestone provides a critical foundation for the predictive modeling process by focusing on data understanding and preparation. Through exploration, cleaning, and formatting of the dataset, we ensured that all variables are relevant, consistent, and compatible with machine learning algorithms. These steps not only improve the reliability and interpretability of the model but also minimize bias and ensure ethical data use. With a clean dataset in place, the project is now ready to move into the modeling phase of the CRISP-DM framework.

Milestone Three: Modeling and Evaluation

This milestone focuses on the modeling and evaluation phases of the CRISP-DM process. Key steps include selecting and fitting the model, assessing data quality, analyzing data structure, and evaluating the model fit and results. Based on the business problem and the data characteristics, logistic regression is what I had selected as the primary modeling technique. Logistic regression was chosen for its strong interpretability, making it easy for stakeholders to understand how different variables impact the probability of loan default. This model is highly suitable for binary classification problems like predicting default versus non-default outcomes, offering a balance between simplicity, transparency, and predictive power.

CRISP-DM Modeling Phase: Assessing Data Quality

The data quality checks are performed here during the preparation phase which helps ensure the integrity and reliability of the modeling process. Just to recap; duplicate records were not present, missing values were confirmed to be zero, and categorical variables were properly factored for modeling compatibility. This step minimized the risk of bias or technical errors during model training and evaluation.

CRISP-DM Modeling Phase: Analyzing Data Structure

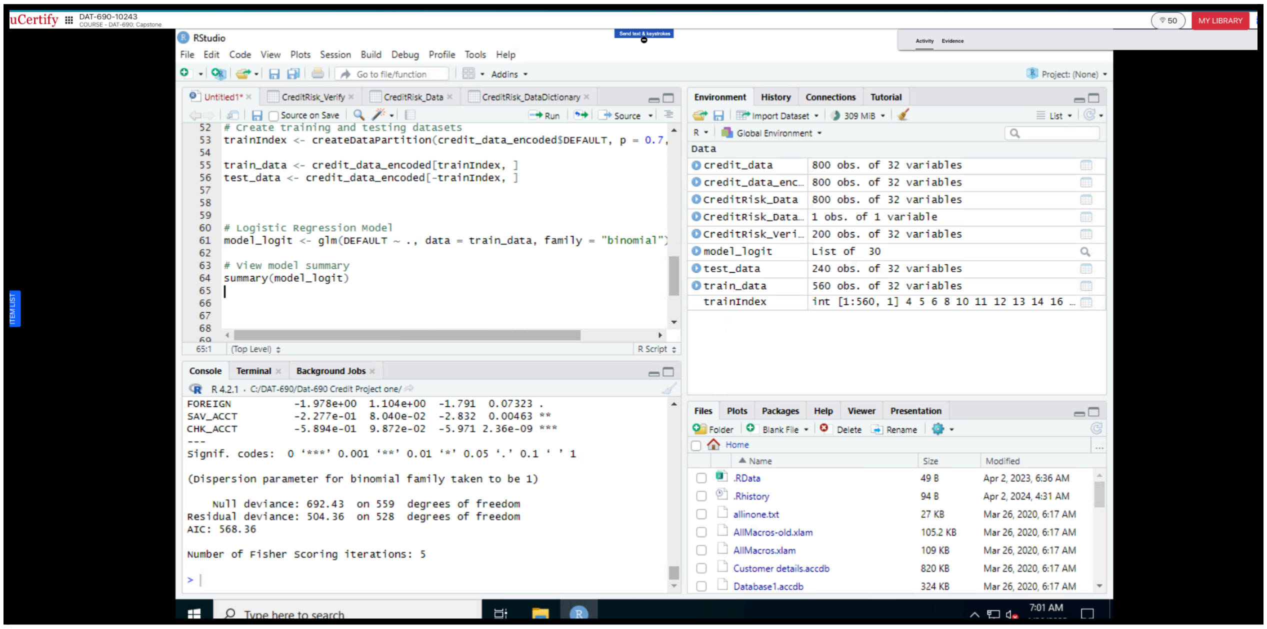
Following the application of one-hot encoding to categorical variables, we found that the dataset contained 32 predictor variables, consisting of a mix of standardized numeric features and binary categorical indicators. Numeric variables such as AMOUNT and AGE were standardized using the scale() function in R to ensure consistent feature scaling across predictors and support model convergence. The target variable, DEFAULT, exhibited moderate class imbalance, with approximately 70% of the observations classified as non-defaults and 30% as defaults. During initial data exploration, no variables were identified for removal based on low variance or redundancy, as all predictors contributed meaningful variation necessary for effective model training and evaluation.

CRISP-DM Modeling Phase: Preliminary Model Artifacts

To evaluate the performance of the logistic regression model developed for credit risk prediction, several key figures were generated.

Model Summary: The logistic regression model was fit to the training dataset using the glm() function with a binomial family. The model summary provides the estimated coefficients and associated p-values for each predictor variable. This output helps determine which variables are statistically significant in predicting loan default. In *Figure 1*, we see the model summary which displays coefficients for predictors such as SAV_ACCT, and CHK_ACCT, along with their standard errors, z-values, p-values, and model fit statistics including null deviance, residual deviance, AIC, and number of iterations, this provides us with a comprehensive view of model strength and predictor influence.

Figure 1: Model summary coefficients and p-values.


uCertify
DAT-690-10243
COURSE - DAT-690: Capstone
MY LIBRARY
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1*
CreditRisk_Verify
CreditRisk_Data
CreditRisk_DataDictionary
Source on Save
Run
Source
Environment History Connections Tutorial
Import Dataset
309 MiB
List
Global Environment
Project: (None)
Activity Evidence
Console Terminal Background Jobs
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Home
Type here to search
7:01 AM

```
# Create training and testing datasets
trainIndex <- createDataPartition(credit_data_encoded$DEFAULT, p = 0.7,

train_data <- credit_data_encoded[trainIndex, ]
test_data <- credit_data_encoded[-trainIndex, ]



# Logistic Regression Model
model_logit <- glm(DEFAULT ~ ., data = train_data, family = "binomial")

# View model summary
summary(model_logit)
```

65:1 (Top Level) R Script

R 4.2.1 · C:/DAT-690/Dat-690 Credit Project one/

```
FOREIGN          -1.978e+00  1.104e+00  -1.791  0.07323 .
SAV_ACCT         -2.277e-01  8.040e-02  -2.832  0.00463 **
CHK_ACCT         -5.894e-01  9.872e-02  -5.971 2.36e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 692.43  on 559  degrees of freedom
Residual deviance: 504.36  on 528  degrees of freedom
AIC: 568.36

Number of Fisher Scoring iterations: 5

> 
```

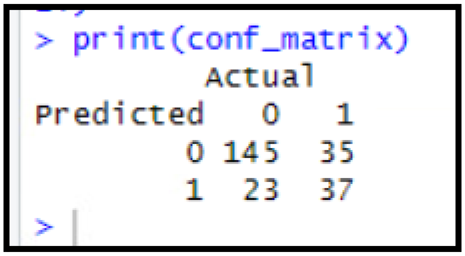
Data

credit_data	800 obs. of 32 variables
credit_data_enc…	800 obs. of 32 variables
CreditRisk_Data	800 obs. of 32 variables
CreditRisk_Data…	1 obs. of 1 variable
CreditRisk_Veri…	200 obs. of 32 variables
model_logit	List of 30
test_data	240 obs. of 32 variables
train_data	560 obs. of 32 variables
trainIndex	int [1:560, 1] 4 5 6 8 10 11 12 13 14 16 …

Name	Size	Modified
.RData	49 B	Apr 2, 2023, 6:36 AM
.Rhistory	94 B	Apr 2, 2024, 4:31 AM
allinone.txt	27 KB	Mar 26, 2020, 6:17 AM
AllMacros-old.xlam	105.2 KB	Mar 26, 2020, 6:17 AM
AllMacros.xlam	109 KB	Mar 26, 2020, 6:17 AM
Customer details.accdb	820 KB	Mar 26, 2020, 6:17 AM
Database1.accdb	324 KB	Mar 26, 2020, 6:17 AM

Confusion Matrix: A confusion matrix was created to evaluate the model's classification performance on the test dataset. It shows the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). *Figure 2:* The confusion matrix output displays 145 true negatives, 37 true positives, 23 false positives, and 35 false negatives, providing insight into the model's ability to correctly classify default and non-default cases and highlighting the balance between sensitivity and specificity.

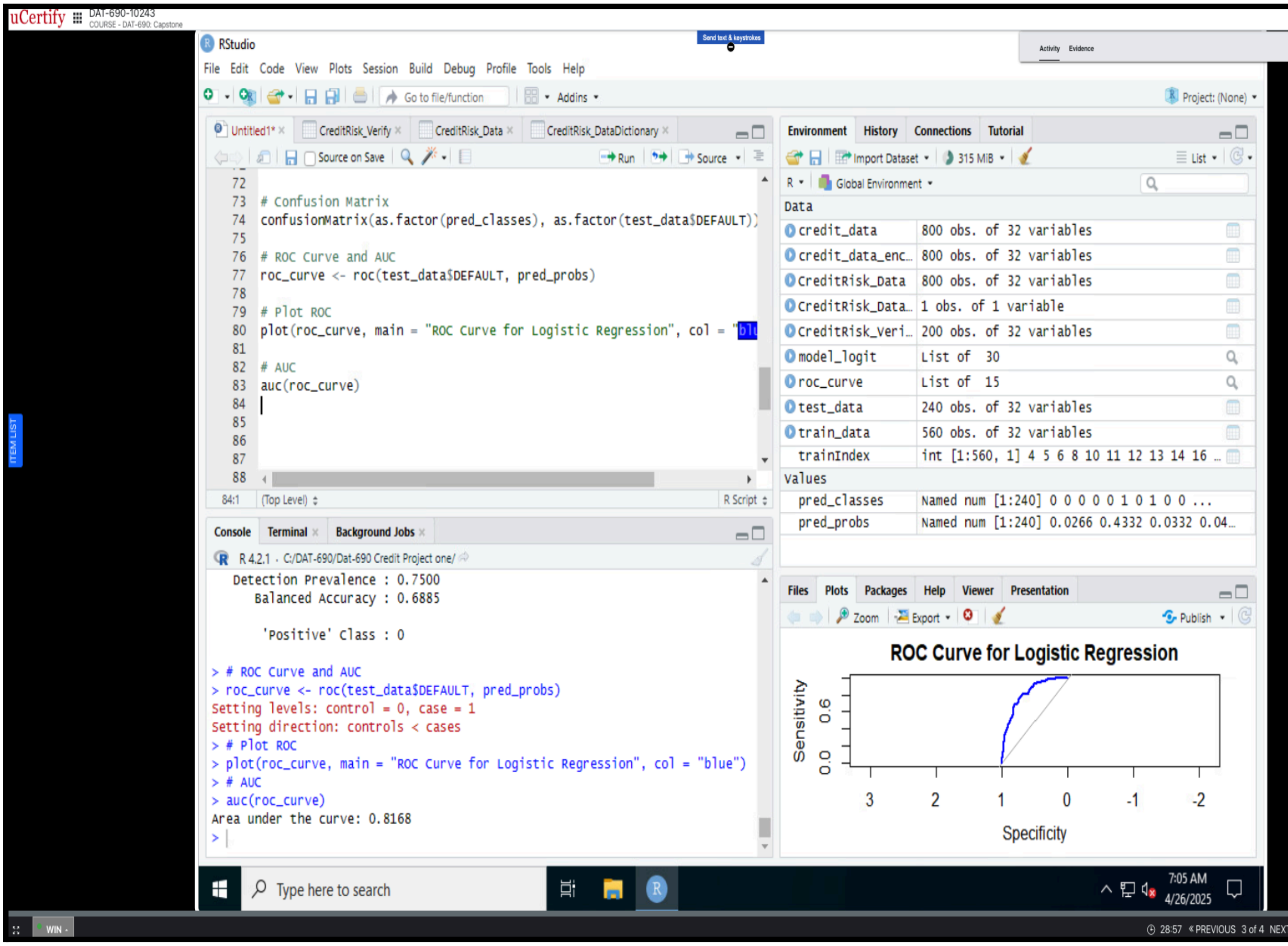
Figure 2: confusion matrix output



```
> print(conf_matrix)
         Actual
Predicted   0   1
        0 145  35
        1  23  37
> 
```

ROC Curve and AUC Score: The Receiver Operating Characteristic (ROC) curve was plotted to visualize the trade-off between sensitivity and specificity. The Area Under the Curve (AUC) score was also calculated to quantify the model's overall performance. A higher AUC indicates better model performance. In *Figure 3*, The ROC curve shows the model's ability to discriminate between defaults and non-defaults, with the curve bending well above the diagonal random guess line. The AUC value of 0.8168 indicates strong model performance, reflecting a high probability that the model ranks a randomly chosen default applicant higher than a non-default.

Figure 3: ROC curve plot and AUC score result



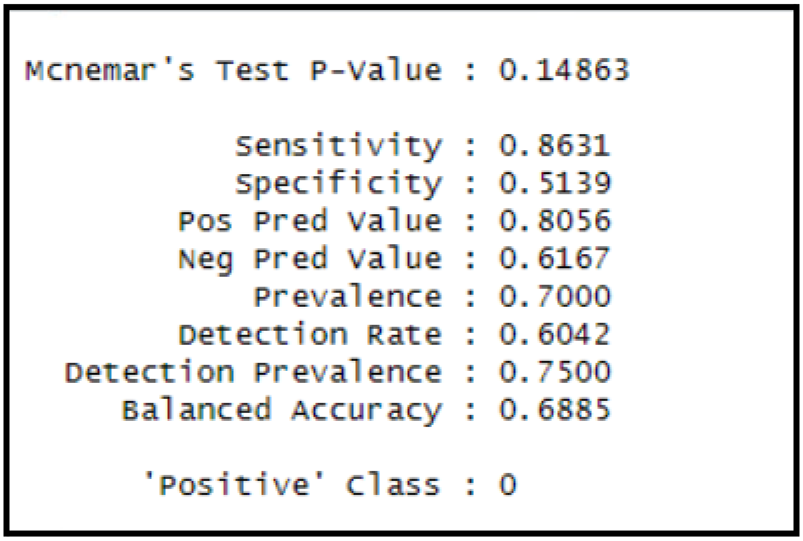
```
# Confusion Matrix
confusionMatrix(as.factor(pred_classes), as.factor(test_data$DEFAULT))

# ROC Curve and AUC
roc_curve <- roc(test_data$DEFAULT, pred_probs)

# Plot ROC
plot(roc_curve, main = "ROC Curve for Logistic Regression", col = "blu

# AUC
auc(roc_curve)
```

```
Data
credit_data          800 obs. of 32 variables
credit_data_enc…     800 obs. of 32 variables
CreditRisk_Data      800 obs. of 32 variables
CreditRisk_Data…     1 obs. of 1 variable
CreditRisk_Veri…     200 obs. of 32 variables
model_logit          List of 30
roc_curve            List of 15
test_data            240 obs. of 32 variables
train_data           560 obs. of 32 variables
trainIndex           int [1:560, 1] 4 5 6 8 10 11 12 13 14 16 …
Values
pred_classes         Named num [1:240] 0 0 0 0 0 1 0 1 0 0 ...
pred_probs           Named num [1:240] 0.0266 0.4332 0.0332 0.04…
```

```
   Detection Prevalence : 0.7500
      Balanced Accuracy : 0.6885

       'Positive' Class : 0

> # ROC Curve and AUC
> roc_curve <- roc(test_data$DEFAULT, pred_probs)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> # Plot ROC
> plot(roc_curve, main = "ROC Curve for Logistic Regression", col = "blue")
> # AUC
> auc(roc_curve)
Area under the curve: 0.8168
>
```

Model Performance Metrics: Additional evaluation metrics such as sensitivity, specificity, precision, negative predictive value, detection prevalence, and balanced accuracy were calculated to give a fuller picture of model performance. These metrics offer insight into the model's strengths and weaknesses when predicting defaults versus non-defaults. The performance metrics output seen in *Figure 4*, shows a sensitivity of 86.31%, meaning the model correctly identifies a high proportion of non-defaults, and a specificity of 51.39%, indicating moderate ability to correctly identify defaults. The balanced accuracy of 68.85% reflects a reasonable overall model performance across both classes, while the positive predictive value (80.56%) and negative predictive value (61.67%) further detail the model's reliability in classification outcomes.

Figure 4: metrics table showing sensitivity, specificity, etc.



```
Mcnemar's Test P-Value : 0.14863

           Sensitivity : 0.8631
           Specificity : 0.5139
        Pos Pred Value : 0.8056
        Neg Pred Value : 0.6167
            Prevalence : 0.7000
        Detection Rate : 0.6042
  Detection Prevalence : 0.7500
     Balanced Accuracy : 0.6885

       'Positive' Class : 0
```

CRISP-DM Evaluation Phase: Model Fit Evaluation

Here we begin the evaluation phase by examining the key model fit statistics to assess the performance and reliability of the final logistic regression model. The logistic regression model produced an AIC of 568.36, indicating a reasonable balance between model fit and complexity. Performance metrics, including a balanced accuracy of 68.85% and an ROC-AUC score of 0.8168, suggest the model performs well in distinguishing between defaults and non-defaults.

Overfitting is monitored by evaluating the model's performance on a separate test dataset, ensuring that results were consistent outside the training data.

CRISP-DM Evaluation Phase: Areas of Concern

One area of concern was the moderate class imbalance, where fewer default cases could increase the risk of false negatives, meaning some risky applicants might be incorrectly classified as safe. While most predictors showed meaningful significance, a few variables had higher p-values, and slight multicollinearity among financial features was observed, although not severe enough to warrant removal. Additionally, logistic regression assumes a linear relationship between the predictors and the log-odds of the outcome, which was considered reasonable based on the exploratory data analysis.

CRISP-DM Evaluation Phase: Model Performance Evaluation and Selection Justification

The logistic regression model here was selected as the final model based on its strong overall performance, interpretability, and alignment with business objectives. During model development, key fit statistics such as the Akaike Information Criterion (AIC) were monitored, with the final model achieving an AIC of 568.36, demonstrating an effective balance between complexity and fit. In the evaluation phase, performance metrics including balanced accuracy, sensitivity, specificity, and ROC-AUC score confirmed the model's ability to reliably predict loan defaults and non-defaults. Overfitting concerns were addressed by validating the model's results on a separate test dataset, with consistent outcomes observed.

CRISP-DM Evaluation Phase: Model Results and Final Evaluation

Model results confirmed that the logistic regression model reliably distinguishes between default and non-default applicants. These findings validated the analytic plan and demonstrated strong predictive strength without the need for additional modeling adjustments. Refinements throughout modeling were made, such as variable selection, feature scaling, and hyperparameter tuning, which helped optimize the model's generalization ability. The final model is now positioned for deployment, with strong alignment to the original business objective of improving credit risk assessment through transparent methods.

References

1. Bobbitt, Z. (2021, September 29). *How to perform logistic regression in R (step-by-step).* Statology. https://www.statology.org/logistic-regression-in-r/
2. Chumbar, S. (2023, September 24). *The CRISP-DM process: A comprehensive guide.* Medium. https://medium.com/@shawn.chumbar/the-crisp-dm-process-a-comprehensive-guide-4d893aecb151
3. CRISP-DM help overview. (n.d.). *IBM Documentation.* https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview
4. DAT 690 Milestone One Guidelines and Rubric. (n.d.). *Southern New Hampshire University (SNHU).* https://learn.snhu.edu/d2l/le/content/1893958/viewContent/40146445/View
5. How to explain the ROC AUC score and ROC curve? (n.d.). *EvidentlyAI.* https://www.evidentlyai.com/classification-metrics/explain-roc-curve
6. Okorie, G., Udeh, C., Adaga, E., DaraOjimba, O., & Oriekhoe, O. (2024). Ethical considerations in data collection and analysis: A review. *International Journal of Applied Research in Social Sciences, 6*(1), 1–22. https://doi.org/10.51594/ijarss.v6i1.688
7. Ray, S. (2025, April 4). *Top 10 machine learning algorithms you must know.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
8. Research guides: Data analytics — Identifying data needs. (n.d.). *Southern New Hampshire University.* https://libguides.snhu.edu/c.php?g=934243&p=7157165

9. Sharma, R. (2024, November 28). *Data cleaning techniques: Learn simple & effective ways to clean data.* upGrad blog. https://www.upgrad.com/blog/data-cleaning-techniques/