**Step-by-Step Guide for R Data Analysis – Credit Risk Dataset**

**🔹 1. Load Necessary Libraries**

r

CopyEdit

# Install (if needed) and load required packages

install.packages(c("tidyverse", "ggplot2", "caret", "corrplot", "ROCR"))

library(tidyverse)

library(caret)

library(corrplot)

library(ROCR)

**🔹 2. Import Your Dataset**

Assume your file is named financial\_risk\_dataset.csv:

r

CopyEdit

data <- read.csv("financial\_risk\_dataset.csv", stringsAsFactors = TRUE)

glimpse(data)

summary(data)

**🔹 3. Clean and Prepare the Data**

r

CopyEdit

# Handle missing values (remove or impute)

data <- na.omit(data)

# Convert Default\_Status to factor for classification

data$Default\_Status <- as.factor(data$Default\_Status)

# Optional: scale numeric features

data$Debt\_to\_Income\_Ratio <- data$Existing\_Debt / (data$Applicant\_Income + data$Coapplicant\_Income + 1)

**🔹 4. Exploratory Data Analysis (EDA)**

**Basic visualizations:**

r

CopyEdit

# Histogram of Credit Score

ggplot(data, aes(Credit\_Score)) +

geom\_histogram(fill = "skyblue", bins = 30) +

theme\_minimal() +

labs(title = "Credit Score Distribution")

# Boxplot of Loan Amount by Default Status

ggplot(data, aes(Default\_Status, Loan\_Amount)) +

geom\_boxplot(fill = "orange") +

labs(title = "Loan Amount vs. Default Status")

**Correlation matrix:**

r

CopyEdit

numeric\_data <- data %>% select\_if(is.numeric)

corrplot(cor(numeric\_data, use = "complete.obs"), method = "circle")

**🔹 5. Build a Predictive Model (e.g., Logistic Regression)**

r

CopyEdit

set.seed(123)

trainIndex <- createDataPartition(data$Default\_Status, p = 0.8, list = FALSE)

train <- data[trainIndex, ]

test <- data[-trainIndex, ]

model <- glm(Default\_Status ~ Credit\_Score + Debt\_to\_Income\_Ratio + Payment\_Delays\_6mo + Credit\_Utilization\_Ratio,

data = train, family = binomial)

summary(model)

**🔹 6. Evaluate Model Performance**

r

CopyEdit

# Predict probabilities

pred\_probs <- predict(model, test, type = "response")

pred\_class <- ifelse(pred\_probs > 0.5, 1, 0)

# Confusion matrix

confusionMatrix(as.factor(pred\_class), test$Default\_Status)

# ROC Curve

pred <- prediction(pred\_probs, test$Default\_Status)

perf <- performance(pred, "tpr", "fpr")

plot(perf, col = "blue", main = "ROC Curve")

abline(a = 0, b = 1, lty = 2, col = "gray")

**📊 Key Outputs You’ll Get:**

* Distributions of credit score and loan data
* Insights on variables most associated with default
* Logistic regression coefficients (predictive strength)
* Model accuracy and ROC curve