Credit Risk Analysis

Akash Kalaranjan

2026-01-01

```
credit <- read.table(
  "https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data",
  header = FALSE
)

colnames(credit) <- c(
  "Status", "Duration", "CreditHistory", "Purpose", "CreditAmount",
  "Savings", "Employment", "InstallmentRate", "PersonalStatusSex",
  "OtherDebtors", "ResidenceDuration", "Property", "Age",
  "OtherInstallmentPlans", "Housing", "ExistingCredits",
  "Job", "NumPeopleLiable", "Telephone", "ForeignWorker", "Default"
)

head(credit)
```

```
##   Status Duration CreditHistory Purpose CreditAmount Savings Employment
## 1    A11        6           A34     A43         1169     A65        A75
## 2    A12       48           A32     A43         5951     A61        A73
## 3    A14       12           A34     A46         2096     A61        A74
## 4    A11       42           A32     A42         7882     A61        A74
## 5    A11       24           A33     A40         4870     A61        A73
## 6    A14       36           A32     A46         9055     A65        A73
##   InstallmentRate PersonalStatusSex OtherDebtors ResidenceDuration Property Age
## 1               4               A93         A101                 4     A121  67
## 2               2               A92         A101                 2     A121  22
## 3               2               A93         A101                 3     A121  49
## 4               2               A93         A103                 4     A122  45
## 5               3               A93         A101                 4     A124  53
## 6               2               A93         A101                 4     A124  35
##   OtherInstallmentPlans Housing ExistingCredits  Job NumPeopleLiable Telephone
## 1                  A143    A152               2 A173               1      A192
## 2                  A143    A152               1 A173               1      A191
## 3                  A143    A152               1 A172               2      A191
## 4                  A143    A153               1 A173               2      A191
## 5                  A143    A153               2 A173               2      A191
## 6                  A143    A153               1 A172               2      A192
##   ForeignWorker Default
## 1          A201       1
## 2          A201       2
## 3          A201       1
## 4          A201       1
## 5          A201       2
## 6          A201       1
```

```
dim(credit)
```

```
## [1] 1000   21
```

```
str(credit)
```

```
## 'data.frame':    1000 obs. of  21 variables:
##  $ Status              : chr  "A11" "A12" "A14" "A11" ...
##  $ Duration            : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ CreditHistory       : chr  "A34" "A32" "A34" "A32" ...
##  $ Purpose             : chr  "A43" "A43" "A46" "A42" ...
##  $ CreditAmount        : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ Savings             : chr  "A65" "A61" "A61" "A61" ...
##  $ Employment          : chr  "A75" "A73" "A74" "A74" ...
##  $ InstallmentRate     : int  4 2 2 2 3 2 3 2 2 4 ...
##  $ PersonalStatusSex   : chr  "A93" "A92" "A93" "A93" ...
##  $ OtherDebtors        : chr  "A101" "A101" "A101" "A103" ...
##  $ ResidenceDuration   : int  4 2 3 4 4 4 4 2 4 2 ...
##  $ Property            : chr  "A121" "A121" "A121" "A122" ...
##  $ Age                 : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ OtherInstallmentPlans: chr  "A143" "A143" "A143" "A143" ...
##  $ Housing             : chr  "A152" "A152" "A152" "A153" ...
##  $ ExistingCredits     : int  2 1 1 1 2 1 1 1 1 2 ...
##  $ Job                 : chr  "A173" "A173" "A172" "A173" ...
##  $ NumPeopleLiable     : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ Telephone           : chr  "A192" "A191" "A191" "A191" ...
##  $ ForeignWorker       : chr  "A201" "A201" "A201" "A201" ...
##  $ Default             : int  1 2 1 1 2 1 1 1 1 2 ...
```

```
summary(credit)
```

```
##     Status             Duration    CreditHistory       Purpose
##  Length:1000        Min.   : 4.0   Length:1000        Length:1000
##  Class :character   1st Qu.:12.0   Class :character   Class :character
##  Mode  :character   Median :18.0   Mode  :character   Mode  :character
##                     Mean   :20.9
##                     3rd Qu.:24.0
##                     Max.   :72.0
##   CreditAmount     Savings           Employment        InstallmentRate
##  Min.   :  250   Length:1000        Length:1000        Min.   :1.000
##  1st Qu.: 1366   Class :character   Class :character   1st Qu.:2.000
##  Median : 2320   Mode  :character   Mode  :character   Median :3.000
##  Mean   : 3271                                         Mean   :2.973
##  3rd Qu.: 3972                                         3rd Qu.:4.000
##  Max.   :18424                                         Max.   :4.000
##  PersonalStatusSex  OtherDebtors       ResidenceDuration   Property
##  Length:1000        Length:1000        Min.   :1.000     Length:1000
##  Class :character   Class :character   1st Qu.:2.000     Class :character
##  Mode  :character   Mode  :character   Median :3.000     Mode  :character
##                                        Mean   :2.845
##                                        3rd Qu.:4.000
##                                        Max.   :4.000
##       Age        OtherInstallmentPlans   Housing          ExistingCredits
##  Min.   :19.00   Length:1000           Length:1000        Min.   :1.000
##  1st Qu.:27.00   Class :character      Class :character   1st Qu.:1.000
##  Median :33.00   Mode  :character      Mode  :character   Median :1.000
##  Mean   :35.55                                            Mean   :1.407
##  3rd Qu.:42.00                                            3rd Qu.:2.000
##  Max.   :75.00                                            Max.   :4.000
##      Job             NumPeopleLiable  Telephone         ForeignWorker
##  Length:1000        Min.   :1.000   Length:1000        Length:1000
##  Class :character   1st Qu.:1.000   Class :character   Class :character
##  Mode  :character   Median :1.000   Mode  :character   Mode  :character
##                     Mean   :1.155
##                     3rd Qu.:1.000
##                     Max.   :2.000
##     Default
##  Min.   :1.0
##  1st Qu.:1.0
##  Median :1.0
##  Mean   :1.3
##  3rd Qu.:2.0
##  Max.   :2.0
```

```
credit[sapply(credit, is.character)] <-
  lapply(credit[sapply(credit, is.character)], factor)

str(credit)
```

```
## 'data.frame':	1000 obs. of  21 variables:
##  $ Status              : Factor w/ 4 levels "A11","A12","A13",..: 1 2 4 1 1 4 4 2 4 2 ...
##  $ Duration            : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ CreditHistory       : Factor w/ 5 levels "A30","A31","A32",..: 5 3 5 3 4 3 3 3 3 5 ...
##  $ Purpose             : Factor w/ 10 levels "A40","A41","A410",..: 5 5 8 4 1 8 4 2 5 1 ...
##  $ CreditAmount        : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ Savings             : Factor w/ 5 levels "A61","A62","A63",..: 5 1 1 1 1 5 3 1 4 1 ...
##  $ Employment          : Factor w/ 5 levels "A71","A72","A73",..: 5 3 4 4 3 3 5 3 4 1 ...
##  $ InstallmentRate     : int  4 2 2 2 3 2 3 2 2 4 ...
##  $ PersonalStatusSex   : Factor w/ 4 levels "A91","A92","A93",..: 3 2 3 3 3 3 3 3 1 4 ...
##  $ OtherDebtors        : Factor w/ 3 levels "A101","A102",..: 1 1 1 3 1 1 1 1 1 1 ...
##  $ ResidenceDuration   : int  4 2 3 4 4 4 4 2 4 2 ...
##  $ Property            : Factor w/ 4 levels "A121","A122",..: 1 1 1 2 4 4 2 3 1 3 ...
##  $ Age                 : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ OtherInstallmentPlans: Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Housing             : Factor w/ 3 levels "A151","A152",..: 2 2 2 3 3 3 2 1 2 2 ...
##  $ ExistingCredits     : int  2 1 1 1 2 1 1 1 1 2 ...
##  $ Job                 : Factor w/ 4 levels "A171","A172",..: 3 3 2 3 3 2 3 4 2 4 ...
##  $ NumPeopleLiable     : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ Telephone           : Factor w/ 2 levels "A191","A192": 2 1 1 1 1 2 1 2 1 1 ...
##  $ ForeignWorker       : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Default             : int  1 2 1 1 2 1 1 1 1 2 ...
```

```
credit$Default <- factor(
  credit$Default,
  levels = c(1, 2),
  labels = c("Good", "Bad")
)

table(credit$Default)
```

```
##
## Good  Bad
##  700  300
```

```
prop.table(table(credit$Default))
```

```
##
## Good  Bad
##  0.7  0.3
```

```
# Logistic Regression Setup
# Step 1: Check the structure one more time
str(credit)
```

```
## 'data.frame':    1000 obs. of  21 variables:
##  $ Status              : Factor w/ 4 levels "A11","A12","A13",..: 1 2 4 1 1 4 4 2 4 2 ...
##  $ Duration            : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ CreditHistory       : Factor w/ 5 levels "A30","A31","A32",..: 5 3 5 3 4 3 3 3 3 5 ...
##  $ Purpose             : Factor w/ 10 levels "A40","A41","A410",..: 5 5 8 4 1 8 4 2 5 1 ...
##  $ CreditAmount        : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ Savings             : Factor w/ 5 levels "A61","A62","A63",..: 5 1 1 1 1 5 3 1 4 1 ...
##  $ Employment          : Factor w/ 5 levels "A71","A72","A73",..: 5 3 4 4 3 3 5 3 4 1 ...
##  $ InstallmentRate     : int  4 2 2 2 3 2 3 2 2 4 ...
##  $ PersonalStatusSex   : Factor w/ 4 levels "A91","A92","A93",..: 3 2 3 3 3 3 3 3 1 4 ...
##  $ OtherDebtors        : Factor w/ 3 levels "A101","A102",..: 1 1 1 3 1 1 1 1 1 1 ...
##  $ ResidenceDuration   : int  4 2 3 4 4 4 4 2 4 2 ...
##  $ Property            : Factor w/ 4 levels "A121","A122",..: 1 1 1 2 4 4 2 3 1 3 ...
##  $ Age                 : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ OtherInstallmentPlans: Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Housing             : Factor w/ 3 levels "A151","A152",..: 2 2 2 3 3 3 2 1 2 2 ...
##  $ ExistingCredits     : int  2 1 1 1 2 1 1 1 1 2 ...
##  $ Job                 : Factor w/ 4 levels "A171","A172",..: 3 3 2 3 3 2 3 4 2 4 ...
##  $ NumPeopleLiable     : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ Telephone           : Factor w/ 2 levels "A191","A192": 2 1 1 1 1 2 1 2 1 1 ...
##  $ ForeignWorker       : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Default             : Factor w/ 2 levels "Good","Bad": 1 2 1 1 2 1 1 1 1 2 ...
```

```
# Step 2: Fit logistic regression model using all variables
model <- glm(Default ~ ., data = credit, family = binomial)

# Step 3: Inspect results
summary(model)
```

```
##
## Call:
## glm(formula = Default ~ ., family = binomial, data = credit)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 4.005e-01  1.084e+00   0.369 0.711869
## StatusA12                  -3.749e-01  2.179e-01  -1.720 0.085400 .
## StatusA13                  -9.657e-01  3.692e-01  -2.616 0.008905 **
## StatusA14                  -1.712e+00  2.322e-01  -7.373 1.66e-13 ***
## Duration                    2.786e-02  9.296e-03   2.997 0.002724 **
## CreditHistoryA31            1.434e-01  5.489e-01   0.261 0.793921
## CreditHistoryA32           -5.861e-01  4.305e-01  -1.362 0.173348
## CreditHistoryA33           -8.532e-01  4.717e-01  -1.809 0.070470 .
## CreditHistoryA34           -1.436e+00  4.399e-01  -3.264 0.001099 **
## PurposeA41                 -1.666e+00  3.743e-01  -4.452 8.51e-06 ***
## PurposeA410                -1.489e+00  7.764e-01  -1.918 0.055163 .
## PurposeA42                 -7.916e-01  2.610e-01  -3.033 0.002421 **
## PurposeA43                 -8.916e-01  2.471e-01  -3.609 0.000308 ***
## PurposeA44                 -5.228e-01  7.623e-01  -0.686 0.492831
## PurposeA45                 -2.164e-01  5.500e-01  -0.393 0.694000
## PurposeA46                  3.628e-02  3.965e-01   0.092 0.927082
## PurposeA48                 -2.059e+00  1.212e+00  -1.699 0.089297 .
## PurposeA49                 -7.401e-01  3.339e-01  -2.216 0.026668 *
## CreditAmount                1.283e-04  4.444e-05   2.887 0.003894 **
## SavingsA62                 -3.577e-01  2.861e-01  -1.250 0.211130
## SavingsA63                 -3.761e-01  4.011e-01  -0.938 0.348476
## SavingsA64                 -1.339e+00  5.249e-01  -2.551 0.010729 *
## SavingsA65                 -9.467e-01  2.625e-01  -3.607 0.000310 ***
## EmploymentA72              -6.691e-02  4.270e-01  -0.157 0.875475
## EmploymentA73              -1.828e-01  4.105e-01  -0.445 0.656049
## EmploymentA74              -8.310e-01  4.455e-01  -1.866 0.062110 .
## EmploymentA75              -2.766e-01  4.134e-01  -0.669 0.503410
## InstallmentRate             3.301e-01  8.828e-02   3.739 0.000185 ***
## PersonalStatusSexA92       -2.755e-01  3.865e-01  -0.713 0.476040
## PersonalStatusSexA93       -8.161e-01  3.799e-01  -2.148 0.031718 *
## PersonalStatusSexA94       -3.671e-01  4.537e-01  -0.809 0.418448
## OtherDebtorsA102            4.360e-01  4.101e-01   1.063 0.287700
## OtherDebtorsA103           -9.786e-01  4.243e-01  -2.307 0.021072 *
## ResidenceDuration           4.776e-03  8.641e-02   0.055 0.955920
## PropertyA122                2.814e-01  2.534e-01   1.111 0.266630
## PropertyA123                1.945e-01  2.360e-01   0.824 0.409743
## PropertyA124                7.304e-01  4.245e-01   1.721 0.085308 .
## Age                        -1.454e-02  9.222e-03  -1.576 0.114982
## OtherInstallmentPlansA142 -1.232e-01  4.119e-01  -0.299 0.764878
## OtherInstallmentPlansA143 -6.463e-01  2.391e-01  -2.703 0.006871 **
## HousingA152                -4.436e-01  2.347e-01  -1.890 0.058715 .
## HousingA153                -6.839e-01  4.770e-01  -1.434 0.151657
## ExistingCredits             2.721e-01  1.895e-01   1.436 0.151109
## JobA172                     5.361e-01  6.796e-01   0.789 0.430160
## JobA173                     5.547e-01  6.549e-01   0.847 0.397015
## JobA174                     4.795e-01  6.623e-01   0.724 0.469086
```

```
## NumPeopleLiable            2.647e-01  2.492e-01   1.062 0.288249
## TelephoneA192             -3.000e-01  2.013e-01  -1.491 0.136060
## ForeignWorkerA202         -1.392e+00  6.258e-01  -2.225 0.026095 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.73  on 999  degrees of freedom
## Residual deviance:  895.82  on 951  degrees of freedom
## AIC: 993.82
##
## Number of Fisher Scoring iterations: 5
```

```
# Predicted probabilities for each loan
pred_prob <- predict(model, type = "response")

# Look at the first 10
head(pred_prob, 10)
```

```
##          1          2          3          4          5          6          7
## 0.03523168 0.63226241 0.02806240 0.25180213 0.75200112 0.26233560 0.06890966
##          8          9         10
## 0.28779903 0.01146434 0.73998613
```

```
set.seed(123)

n <- nrow(credit)
train_idx <- sample(seq_len(n), size = 0.7 * n)

train <- credit[train_idx, ]
test  <- credit[-train_idx, ]
```

```
model <- glm(
  Default ~ Duration + CreditAmount + Age + InstallmentRate + ExistingCredits,
  data = train,
  family = binomial
)

summary(model)
```

```
##
## Call:
## glm(formula = Default ~ Duration + CreditAmount + Age + InstallmentRate +
##     ExistingCredits, family = binomial, data = train)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.593e+00  4.394e-01  -3.625 0.000289 ***
## Duration         2.486e-02  8.806e-03   2.823 0.004751 **
## CreditAmount     6.191e-05  3.852e-05   1.607 0.108046
## Age             -8.697e-03  7.914e-03  -1.099 0.271813
## InstallmentRate  2.282e-01  8.561e-02   2.666 0.007683 **
## ExistingCredits -3.149e-01  1.638e-01  -1.923 0.054526 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 844.80  on 699  degrees of freedom
## Residual deviance: 803.71  on 694  degrees of freedom
## AIC: 815.71
##
## Number of Fisher Scoring iterations: 4
```

```
probs <- predict(model, newdata = test, type = "response")

pred <- ifelse(probs > 0.5, "Bad", "Good")
pred <- factor(pred, levels = c("Good", "Bad"))

table(pred, test$Default)
```

```
##
## pred   Good Bad
##   Good  201  89
##   Bad     3   7
```

```
mean(pred == test$Default)
```

```
## [1] 0.6933333
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.5.2
```

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

```
##
## Attaching package: 'pROC'
```

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

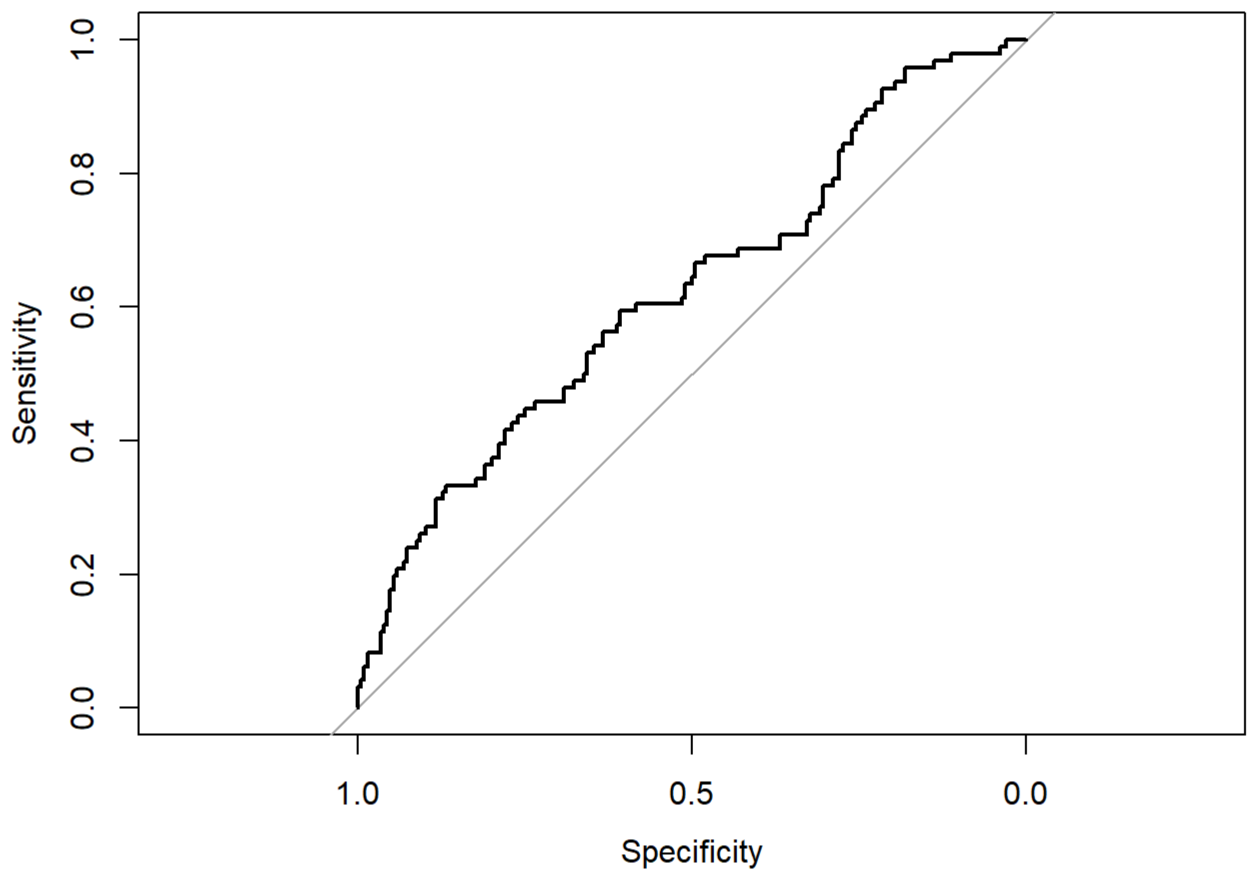
```
# make sure Default is a factor
test$Default <- factor(test$Default, levels = c("Good", "Bad"))

# probs = predicted probability of "Bad"
roc_obj <- roc(
  response = test$Default,
  predictor = probs,
  levels = c("Good", "Bad"),
  direction = "<"
)

auc(roc_obj)
```

```
## Area under the curve: 0.6242
```

```
plot(roc_obj)
```




```
model_base <- glm(
  Default ~ Duration + CreditAmount + Age +
            InstallmentRate + ExistingCredits,
  data = train,
  family = binomial
)

probs_base <- predict(model_base, test, type = "response")
```

```
probs_full <- predict(model, newdata = test, type = "response")
```

```
library(pROC)

roc_base <- roc(test$Default, probs_base, levels = c("Good", "Bad"))
```

```
## Setting direction: controls < cases
```

```
roc_full <- roc(test$Default, probs_full, levels = c("Good", "Bad"))
```

```
## Setting direction: controls < cases
```

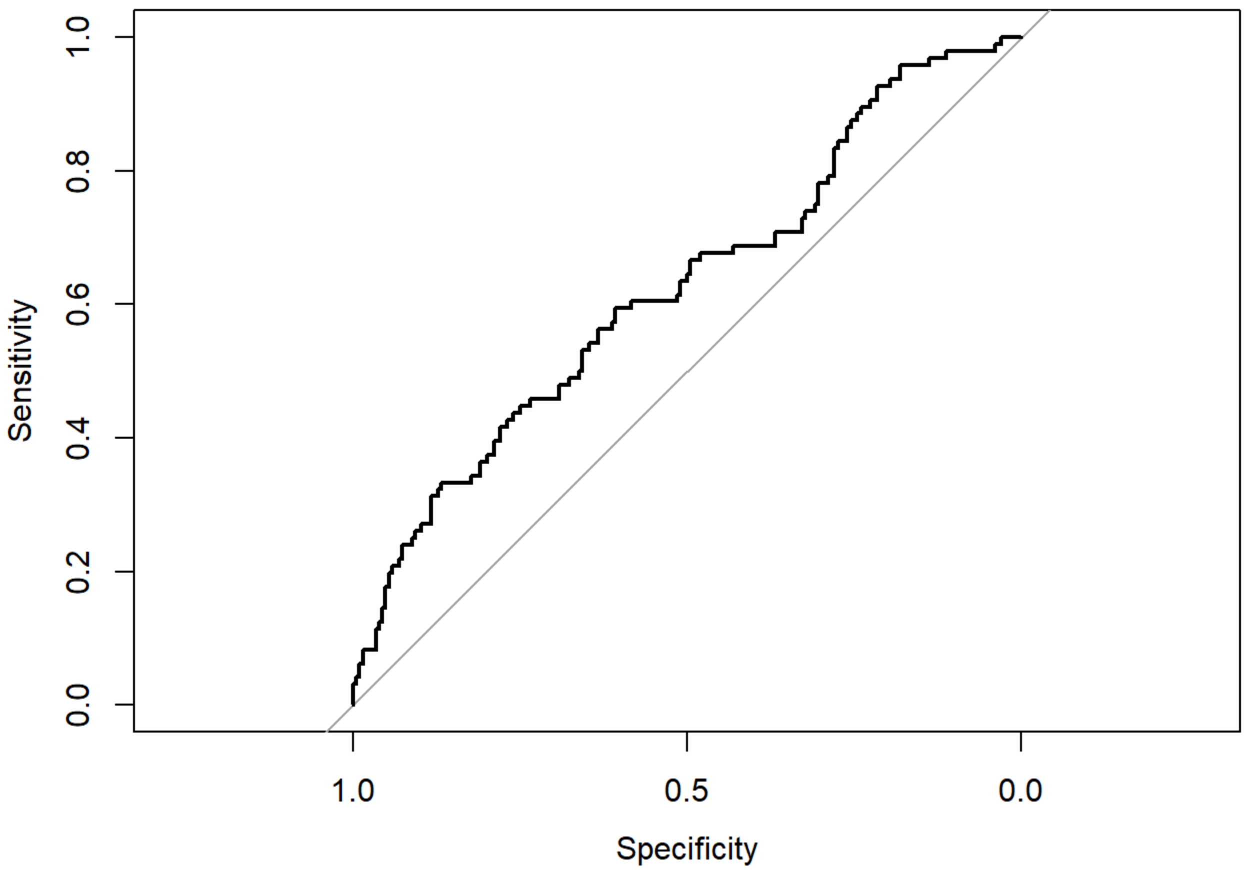
```
auc(roc_base)
```

```
## Area under the curve: 0.6242
```

```
auc(roc_full)
```

```
## Area under the curve: 0.6242
```

```
plot(roc_full)
```



In this project, a logistic regression model was developed to estimate the probability of default using the German Credit dataset. The model achieved an AUC of approximately 0.62, indicating limited but non-trivial discriminatory power between good and bad credit outcomes.

Key risk drivers such as loan duration and installment burden showed statistically significant positive relationships with default, while borrower age and prior credit exposure exhibited weaker protective effects. Although the model’s predictive performance is insufficient for automated credit approval, it serves as a transparent baseline consistent with traditional scorecard approaches.

Given the limited feature set and the age of the dataset, this level of performance is expected. In practice, such a model would be suitable for preliminary risk screening or as a benchmark against more complex models, supporting downstream manual review or advanced machine-learning–based decision systems.

The coefficient for CreditAmount is positive, indicating that higher loan amounts are associated with an increased probability of default. This aligns with real-world credit risk, as larger loans place greater financial strain on borrowers and increase repayment risk.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

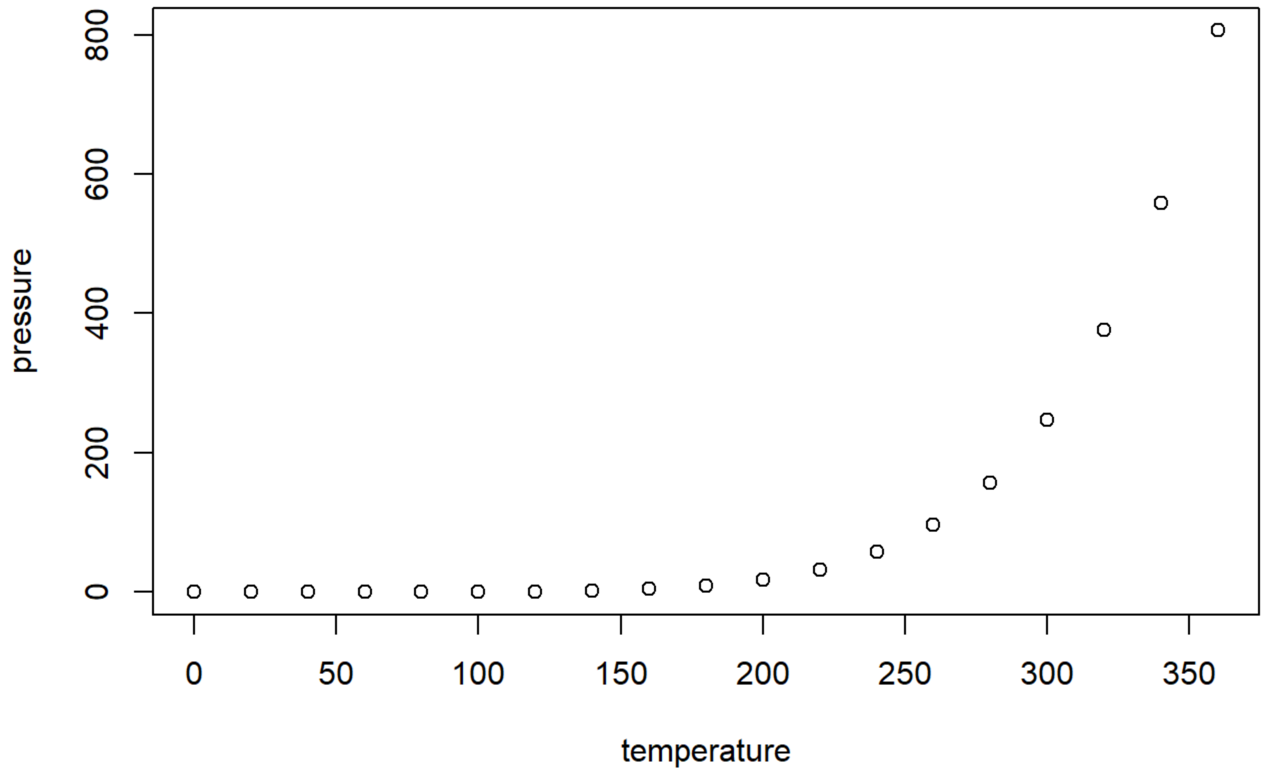
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.