Logistic Regression

Kevin Akhondo

2024-07-15

## Loading Libraries

library(tidyr)  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.3

library(corrr)  
library(recipes)

## Loading required package: 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

##   
## Attaching package: 'recipes'

## The following object is masked from 'package:stats':  
##   
## step

library(parsnip)  
library(yardstick)  
library(rsample)

## Data Importation

#Import the data  
loan\_data <- read.csv("C:\\Users\\KEVINS\\OneDrive\\Desktop\\TABLEAU\\Loan\_default.csv")  
#View the first 5 observations  
head(loan\_data,5)

## LoanID Age Income LoanAmount CreditScore MonthsEmployed NumCreditLines  
## 1 I38PQUQS96 56 85994 50587 520 80 4  
## 2 HPSK72WA7R 69 50432 124440 458 15 1  
## 3 C1OZ6DPJ8Y 46 84208 129188 451 26 3  
## 4 V2KKSFM3UN 32 31713 44799 743 0 3  
## 5 EY08JDHTZP 60 20437 9139 633 8 4  
## InterestRate LoanTerm DTIRatio Education EmploymentType MaritalStatus  
## 1 15.23 36 0.44 Bachelor's Full-time Divorced  
## 2 4.81 60 0.68 Master's Full-time Married  
## 3 21.17 24 0.31 Master's Unemployed Divorced  
## 4 7.07 24 0.23 High School Full-time Married  
## 5 6.51 48 0.73 Bachelor's Unemployed Divorced  
## HasMortgage HasDependents LoanPurpose HasCoSigner Default  
## 1 Yes Yes Other Yes 0  
## 2 No No Other Yes 0  
## 3 Yes Yes Auto No 1  
## 4 No No Business No 0  
## 5 No Yes Auto No 0

#Skim through the data to check the number of missing values  
skimr::skim(loan\_data) # There are no missing values/nulls

Data summary

|  |  |
| --- | --- |
| Name | loan\_data |
| Number of rows | 255347 |
| Number of columns | 18 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 8 |
| numeric | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LoanID | 0 | 1 | 10 | 10 | 0 | 255347 | 0 |
| Education | 0 | 1 | 3 | 11 | 0 | 4 | 0 |
| EmploymentType | 0 | 1 | 9 | 13 | 0 | 4 | 0 |
| MaritalStatus | 0 | 1 | 6 | 8 | 0 | 3 | 0 |
| HasMortgage | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| HasDependents | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| LoanPurpose | 0 | 1 | 4 | 9 | 0 | 5 | 0 |
| HasCoSigner | 0 | 1 | 2 | 3 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 0 | 1 | 43.50 | 14.99 | 18.0 | 31.00 | 43.00 | 56.00 | 69.0 | ▇▇▇▇▇ |
| Income | 0 | 1 | 82499.30 | 38963.01 | 15000.0 | 48825.50 | 82466.00 | 116219.00 | 149999.0 | ▇▇▇▇▇ |
| LoanAmount | 0 | 1 | 127578.87 | 70840.71 | 5000.0 | 66156.00 | 127556.00 | 188985.00 | 249999.0 | ▇▇▇▇▇ |
| CreditScore | 0 | 1 | 574.26 | 158.90 | 300.0 | 437.00 | 574.00 | 712.00 | 849.0 | ▇▇▇▇▇ |
| MonthsEmployed | 0 | 1 | 59.54 | 34.64 | 0.0 | 30.00 | 60.00 | 90.00 | 119.0 | ▇▇▇▇▇ |
| NumCreditLines | 0 | 1 | 2.50 | 1.12 | 1.0 | 2.00 | 2.00 | 3.00 | 4.0 | ▇▇▁▇▇ |
| InterestRate | 0 | 1 | 13.49 | 6.64 | 2.0 | 7.77 | 13.46 | 19.25 | 25.0 | ▇▇▇▇▇ |
| LoanTerm | 0 | 1 | 36.03 | 16.97 | 12.0 | 24.00 | 36.00 | 48.00 | 60.0 | ▇▇▇▇▇ |
| DTIRatio | 0 | 1 | 0.50 | 0.23 | 0.1 | 0.30 | 0.50 | 0.70 | 0.9 | ▇▇▇▇▇ |
| Default | 0 | 1 | 0.12 | 0.32 | 0.0 | 0.00 | 0.00 | 0.00 | 1.0 | ▇▁▁▁▁ |

## Data Split into train and test data

data\_split <- initial\_split(data = loan\_data)  
#Train dataset  
train <- training(data\_split)  
  
#Test dataset  
test <- testing(data\_split)  
  
#Skim through the test and train dataset  
skimr::skim(train)

Data summary

|  |  |
| --- | --- |
| Name | train |
| Number of rows | 191510 |
| Number of columns | 18 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 8 |
| numeric | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LoanID | 0 | 1 | 10 | 10 | 0 | 191510 | 0 |
| Education | 0 | 1 | 3 | 11 | 0 | 4 | 0 |
| EmploymentType | 0 | 1 | 9 | 13 | 0 | 4 | 0 |
| MaritalStatus | 0 | 1 | 6 | 8 | 0 | 3 | 0 |
| HasMortgage | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| HasDependents | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| LoanPurpose | 0 | 1 | 4 | 9 | 0 | 5 | 0 |
| HasCoSigner | 0 | 1 | 2 | 3 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 0 | 1 | 43.50 | 14.99 | 18.0 | 31.00 | 43.00 | 56.00 | 69.0 | ▇▇▇▇▇ |
| Income | 0 | 1 | 82421.98 | 38952.66 | 15000.0 | 48745.00 | 82305.50 | 116165.75 | 149999.0 | ▇▇▇▇▇ |
| LoanAmount | 0 | 1 | 127562.30 | 70879.36 | 5001.0 | 66020.50 | 127624.00 | 188994.25 | 249999.0 | ▇▇▇▇▇ |
| CreditScore | 0 | 1 | 574.25 | 158.81 | 300.0 | 437.00 | 574.00 | 712.00 | 849.0 | ▇▇▇▇▇ |
| MonthsEmployed | 0 | 1 | 59.54 | 34.67 | 0.0 | 30.00 | 60.00 | 90.00 | 119.0 | ▇▇▇▇▇ |
| NumCreditLines | 0 | 1 | 2.50 | 1.12 | 1.0 | 2.00 | 2.00 | 3.00 | 4.0 | ▇▇▁▇▇ |
| InterestRate | 0 | 1 | 13.49 | 6.64 | 2.0 | 7.76 | 13.45 | 19.25 | 25.0 | ▇▇▇▇▇ |
| LoanTerm | 0 | 1 | 36.02 | 16.98 | 12.0 | 24.00 | 36.00 | 48.00 | 60.0 | ▇▇▇▇▇ |
| DTIRatio | 0 | 1 | 0.50 | 0.23 | 0.1 | 0.30 | 0.50 | 0.70 | 0.9 | ▇▇▇▇▇ |
| Default | 0 | 1 | 0.12 | 0.32 | 0.0 | 0.00 | 0.00 | 0.00 | 1.0 | ▇▁▁▁▁ |

skimr::skim(test)

Data summary

|  |  |
| --- | --- |
| Name | test |
| Number of rows | 63837 |
| Number of columns | 18 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 8 |
| numeric | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LoanID | 0 | 1 | 10 | 10 | 0 | 63837 | 0 |
| Education | 0 | 1 | 3 | 11 | 0 | 4 | 0 |
| EmploymentType | 0 | 1 | 9 | 13 | 0 | 4 | 0 |
| MaritalStatus | 0 | 1 | 6 | 8 | 0 | 3 | 0 |
| HasMortgage | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| HasDependents | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| LoanPurpose | 0 | 1 | 4 | 9 | 0 | 5 | 0 |
| HasCoSigner | 0 | 1 | 2 | 3 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 0 | 1 | 43.49 | 15.00 | 18.0 | 31.0 | 44.00 | 56.00 | 69.0 | ▇▇▇▇▇ |
| Income | 0 | 1 | 82731.27 | 38993.44 | 15002.0 | 49112.0 | 82896.00 | 116380.00 | 149997.0 | ▇▇▇▇▇ |
| LoanAmount | 0 | 1 | 127628.56 | 70725.15 | 5000.0 | 66613.0 | 127334.00 | 188931.00 | 249972.0 | ▇▇▇▇▇ |
| CreditScore | 0 | 1 | 574.30 | 159.17 | 300.0 | 437.0 | 573.00 | 712.00 | 849.0 | ▇▇▇▇▇ |
| MonthsEmployed | 0 | 1 | 59.56 | 34.57 | 0.0 | 30.0 | 60.00 | 90.00 | 119.0 | ▇▇▇▇▇ |
| NumCreditLines | 0 | 1 | 2.50 | 1.12 | 1.0 | 2.0 | 2.00 | 4.00 | 4.0 | ▇▇▁▇▇ |
| InterestRate | 0 | 1 | 13.50 | 6.62 | 2.0 | 7.8 | 13.47 | 19.23 | 25.0 | ▇▇▇▇▇ |
| LoanTerm | 0 | 1 | 36.06 | 16.94 | 12.0 | 24.0 | 36.00 | 48.00 | 60.0 | ▇▇▇▇▇ |
| DTIRatio | 0 | 1 | 0.50 | 0.23 | 0.1 | 0.3 | 0.50 | 0.70 | 0.9 | ▇▇▇▇▇ |
| Default | 0 | 1 | 0.12 | 0.32 | 0.0 | 0.0 | 0.00 | 0.00 | 1.0 | ▇▁▁▁▁ |

## Prepare a Data Preprocessing Recipe

#Creating a recipe  
data\_rec <- recipe(Default~.,train)%>%  
 step\_mutate(Default = ifelse(Default==0,"Not.Defaulted","Defaulted"))%>%  
 step\_string2factor(Default)%>%  
 step\_mutate(HasMortgage = ifelse(HasMortgage=="Yes",1,0))%>%#Enoding if client has a mortgate, Yes with 1 and if not ,0  
 step\_mutate(HasDependents = ifelse(HasDependents=="Yes",1,0))%>% #Encoding if the client has dependents, yes with 1 and no with 0  
 step\_mutate(HasCoSigner = ifelse(HasCoSigner=="Yes",1,0))%>% #Encoding if the the client has cosigner yes with 1 and no with 0  
 step\_mutate(MaritalStatus = ifelse(MaritalStatus == "Single",1,ifelse(MaritalStatus=="Married",2,3)))%>% #Encoding marriage status  
 step\_mutate(EmploymentType = recode(EmploymentType,"Full-time"=1, "Part-time"=2,"Self-employed"=3,"Unemployed"=4))%>% #Encoding employment type  
 step\_mutate(LoanPurpose = recode(LoanPurpose,"Auto"=1,"Business"=2,"Education"=3,"Home"=4,"Other"=5))%>%  
 step\_mutate(Education = recode(Education,"High School"=1, "Bachelor's"=2, "Master's"=3,"PhD"=4))%>%  
 step\_rm(LoanID)%>%  
 step\_dummy(all\_nominal(),-all\_outcomes())%>%  
 step\_zv(all\_predictors())%>%  
 step\_center(all\_predictors(),-all\_nominal())%>%  
 step\_scale(all\_predictors(),-all\_nominal())  
#Preparing the data  
data\_prep <- data\_rec%>%  
 prep()

## Fitting Logistic Regression

#Fitting the model  
fitted\_model <- logistic\_reg()%>%  
 set\_engine("glm")%>%  
 set\_mode("classification")%>%  
 fit(Default~.,data = bake(data\_prep,train))  
print(fitted\_model)

## parsnip model object  
##   
##   
## Call: stats::glm(formula = Default ~ ., family = stats::binomial, data = data)  
##   
## Coefficients:  
## (Intercept) Age Income LoanAmount CreditScore   
## 2.378115 0.589778 0.340326 -0.298664 0.125875   
## MonthsEmployed NumCreditLines InterestRate LoanTerm DTIRatio   
## 0.338393 -0.098577 -0.458934 -0.001517 -0.072216   
## Education EmploymentType MaritalStatus HasMortgage HasDependents   
## 0.103579 -0.145905 -0.033488 0.085192 0.130138   
## LoanPurpose HasCoSigner   
## 0.031727 0.132711   
##   
## Degrees of Freedom: 191509 Total (i.e. Null); 191493 Residual  
## Null Deviance: 137400   
## Residual Deviance: 121100 AIC: 121100

#Making predictions  
predictions<-fitted\_model%>%  
 predict(new\_data = bake(data\_prep,test))%>%  
 bind\_cols(  
 bake(data\_prep,test)%>%  
 select(Default))

#Construction of confusion matrix  
predictions%>%  
 conf\_mat(Default,.pred\_class)

## Truth  
## Prediction Defaulted Not.Defaulted  
## Defaulted 256 142  
## Not.Defaulted 7206 56233

## Performance Metrics

### Accuracy

#Accuracy  
predictions%>%  
 metrics(Default,.pred\_class)%>%  
 select(-.estimator)%>%  
 filter(.metric == "accuracy")

## # A tibble: 1 × 2  
## .metric .estimate  
## <chr> <dbl>  
## 1 accuracy 0.885

### Precision and Recall

#Precision  
predictions%>%  
 precision(Default,.pred\_class)

## # A tibble: 1 × 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 precision binary 0.643

#Recalll  
predictions%>%  
 recall(Default,.pred\_class)

## # A tibble: 1 × 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 recall binary 0.0343

### F-Score

predictions%>%  
 f\_meas(Default,.pred\_class)

## # A tibble: 1 × 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 f\_meas binary 0.0651

### ROC Curve

#Start by computing ROC data  
roc\_data <- predictions%>%  
 mutate(pred.prob=ifelse(.pred\_class=="Not.Defaulted",0,1))%>% #Convert the .pred\_class to numeric and name it pred.prob  
 select(-.pred\_class)%>% #Exclude the .pred\_class  
 roc\_curve(truth = Default,pred.prob)  
#Plot the ROC Curve  
ggplot(roc\_data,aes(x=1-specificity,y=sensitivity))+  
 geom\_path(color="blue")+  
 geom\_abline(linetype = "dashed",color = "red")+  
 labs(title = "ROC Curve", x = "1-specificity",y = "Sensitivity")

