PM_Project

Hittanshu Bhanderi

2023-04-23

Loading required packages

```
#install.packages('tidyverse')
library(tidyverse)
#install.packages('ggplot2')
library(ggplot2)
#install.packages('caret')
library(caret)
#install.packages('caretEnsemble')
library(caretEnsemble)
#install.packages('psych')
library(psych)
#install.packages('Amelia')
library(Amelia)
#install.packages('mice')
library(mice)
#install.packages('GGally')
library(GGally)
#install.packages('rpart')
library(rpart)
#install.packages('randomForest')
library(randomForest)
```

```
data<- read.csv("/Users/hittanshubhanderi/Downloads/train.csv")
clean_data<- read.csv("/Users/hittanshubhanderi/Downloads/clean_train.csv")
data$Loan.Status <- factor(data$Loan.Status, levels = c(0,1), labels = c("False", "True"))
clean_data$Loan.Status <- factor(clean_data$Loan.Status, levels = c(0,1), labels = c("False", "True"))
```

Studying the structure of the data

```
str(data)
```

```
## 'data.frame':    67463 obs. of  35 variables:
##  $ ID                         : int  65087372 1450153 1969101 6651430 14354669 50509046 32737431 63
##  $ Loan.Amount                : int  10000 3609 28276 11170 16890 34631 30844 20744 9299 19232 ...
##  $ Funded.Amount              : int  32236 11940 9311 6954 13226 30203 19773 10609 11238 8962 ...
##  $ Funded.Amount.Investor     : num  12329 12192 21603 17877 13540 ...
##  $ Term                       : int  59 59 59 59 59 36 59 58 59 58 ...
##  $ Batch.Enrolled             : chr  "BAT2522922" "BAT1586599" "BAT2136391" "BAT2428731" ...
##  $ Interest.Rate              : num  11.1 12.2 12.5 16.7 15 ...
##  $ Grade                      : chr  "B" "C" "F" "C" ...
```

```
##  $ Sub.Grade                   : chr  "C4" "D3" "D4" "C3" ...
##  $ Employment.Duration         : chr  "MORTGAGE" "RENT" "MORTGAGE" "MORTGAGE" ...
##  $ Home.Ownership              : num  176347 39834 91507 108287 44235 ...
##  $ Verification.Status         : chr  "Not Verified" "Source Verified" "Source Verified" "Source Veri
##  $ Payment.Plan                : chr  "n" "n" "n" "n" ...
##  $ Loan.Title                  : chr  "Debt Consolidation" "Debt consolidation" "Debt Consolidation"
##  $ Debit.to.Income             : num  16.3 15.4 28.1 18 17.2 ...
##  $ Delinquency...two.years     : int  1 0 0 1 1 3 0 0 0 1 ...
##  $ Inquires...six.months       : int  0 0 0 0 3 2 0 0 0 0 ...
##  $ Open.Account                : int  13 12 14 7 13 16 11 14 6 11 ...
##  $ Public.Record               : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Revolving.Balance           : int  24246 812 1843 13819 1544 2277 14501 13067 549 1361 ...
##  $ Revolving.Utilities         : num  74.93 78.3 2.07 67.47 85.25 ...
##  $ Total.Accounts              : int  7 13 20 12 22 20 37 33 17 30 ...
##  $ Initial.List.Status         : chr  "w" "f" "w" "w" ...
##  $ Total.Received.Interest     : num  2930 773 863 288 129 ...
##  $ Total.Received.Late.Fee     : num  0.1021 0.0362 18.7787 0.0441 19.3066 ...
##  $ Recoveries                  : num  2.498 2.377 4.316 0.107 1294.819 ...
##  $ Collection.Recovery.Fee     : num  0.794 0.975 1.02 0.75 0.369 ...
##  $ Collection.12.months.Medical: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Application.Type            : chr  "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" ...
##  $ Last.week.Pay               : int  49 109 66 39 18 32 71 87 144 9 ...
##  $ Accounts.Delinquent         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Total.Collection.Amount     : int  31 53 34 40 430 42 3388 48 26 35 ...
##  $ Total.Current.Balance       : int  311301 182610 89801 9189 126029 51252 42069 184909 68126 71650
##  $ Total.Revolving.Credit.Limit: int  6619 20885 26155 60214 22579 27480 31068 43303 7482 14871 ...
##  $ Loan.Status                 : Factor w/ 2 levels "False","True": 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(data)
```

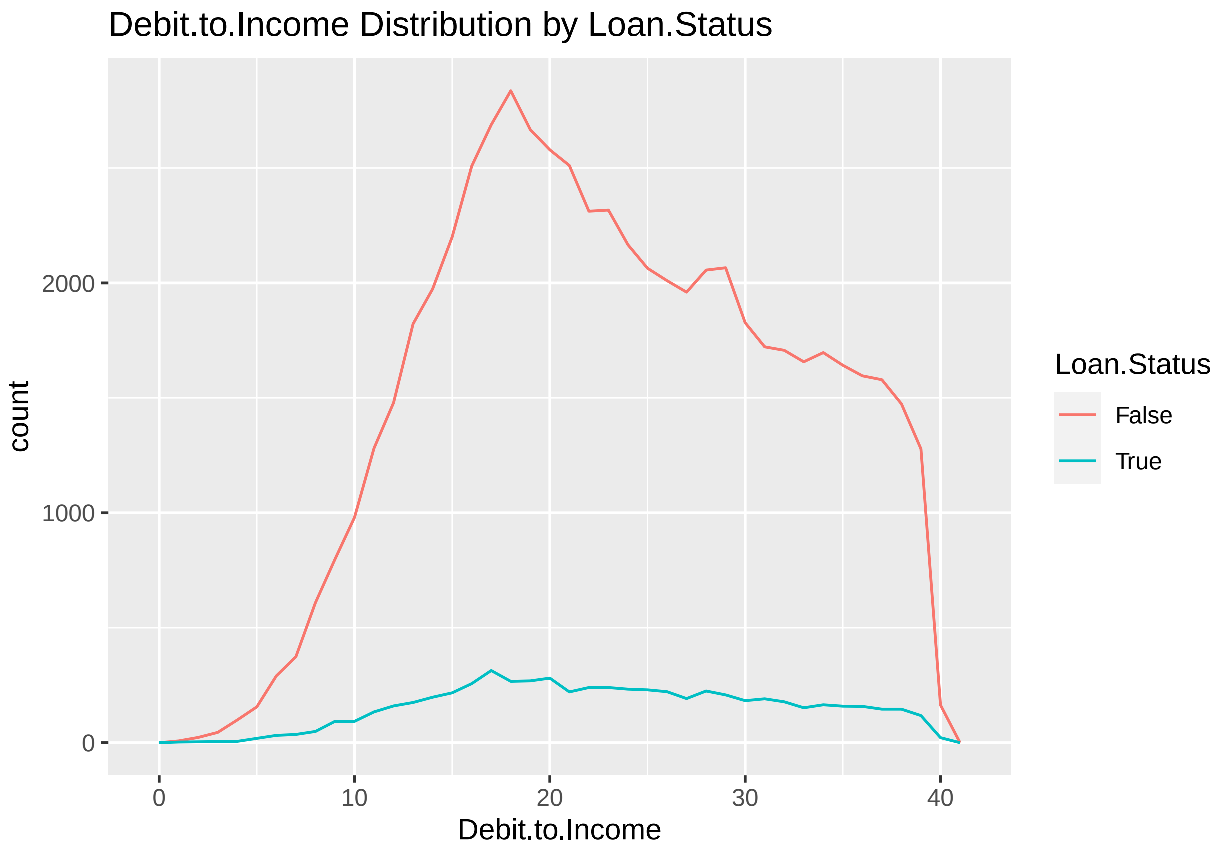
```
##         ID Loan.Amount Funded.Amount Funded.Amount.Investor Term Batch.Enrolled
## 1 65087372       10000         32236              12329.363   59     BAT2522922
## 2  1450153        3609         11940              12191.997   59     BAT1586599
## 3  1969101       28276          9311              21603.225   59     BAT2136391
## 4  6651430       11170          6954              17877.156   59     BAT2428731
## 5 14354669       16890         13226              13539.927   59     BAT5341619
## 6 50509046       34631         30203               8635.932   36     BAT4694572
##   Interest.Rate Grade Sub.Grade Employment.Duration Home.Ownership
## 1      11.13501     B        C4            MORTGAGE      176346.63
## 2      12.23756     C        D3                RENT       39833.92
## 3      12.54588     F        D4            MORTGAGE       91506.69
## 4      16.73120     C        C3            MORTGAGE      108286.58
## 5      15.00830     C        D4            MORTGAGE       44234.83
## 6      17.24699     B        G5                RENT       98957.48
##   Verification.Status Payment.Plan              Loan.Title Debit.to.Income
## 1        Not Verified            n      Debt Consolidation       16.284758
## 2     Source Verified            n      Debt consolidation       15.412409
## 3     Source Verified            n      Debt Consolidation       28.137619
## 4     Source Verified            n      Debt consolidation       18.043730
## 5     Source Verified            n Credit card refinancing       17.209886
## 6        Not Verified            n Credit card refinancing        7.914333
##   Delinquency...two.years Inquires...six.months Open.Account Public.Record
## 1                       1                     0           13             0
## 2                       0                     0           12             0
## 3                       0                     0           14             0
```

```
## 4                       1                      0            7              0
## 5                       1                      3           13              1
## 6                       3                      2           16              0
##   Revolving.Balance Revolving.Utilities Total.Accounts Initial.List.Status
## 1             24246            74.93255              7                   w
## 2               812            78.29719             13                   f
## 3              1843             2.07304             20                   w
## 4             13819            67.46795             12                   w
## 5              1544            85.25076             22                   w
## 6              2277            51.56448             20                   w
##   Total.Received.Interest Total.Received.Late.Fee   Recoveries
## 1               2929.6463              0.10205520    2.4982910
## 2                772.7694              0.03618117    2.3772148
## 3                863.3244             18.77866007    4.3162773
## 4                288.1732              0.04413137    0.1070203
## 5                129.2396             19.30664639 1294.8187510
## 6                464.8181              0.08858435    5.0435754
##   Collection.Recovery.Fee Collection.12.months.Medical Application.Type
## 1               0.7937238                            0       INDIVIDUAL
## 2               0.9748211                            0       INDIVIDUAL
## 3               1.0200750                            0       INDIVIDUAL
## 4               0.7499710                            0       INDIVIDUAL
## 5               0.3689529                            0       INDIVIDUAL
## 6               0.5816877                            0       INDIVIDUAL
##   Last.week.Pay Accounts.Delinquent Total.Collection.Amount
## 1            49                   0                      31
## 2           109                   0                      53
## 3            66                   0                      34
## 4            39                   0                      40
## 5            18                   0                     430
## 6            32                   0                      42
##   Total.Current.Balance Total.Revolving.Credit.Limit Loan.Status
## 1                311301                         6619       False
## 2                182610                        20885       False
## 3                 89801                        26155       False
## 4                  9189                        60214       False
## 5                126029                        22579       False
## 6                 51252                        27480       False
```

Data Visualization

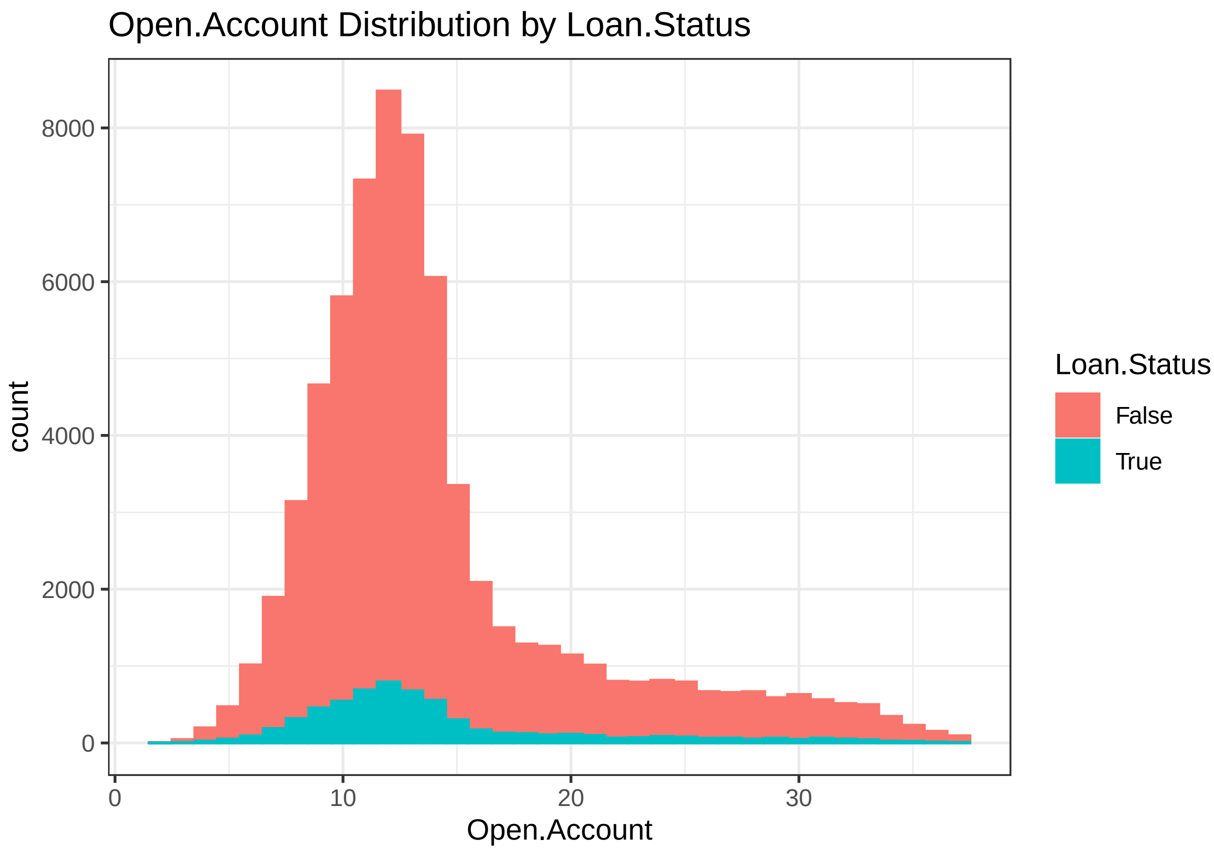
Visual 1

```
ggplot(data, aes(Debit.to.Income, colour = Loan.Status)) + geom_freqpoly(binwidth = 1) + labs(title="De
```



Visual 2

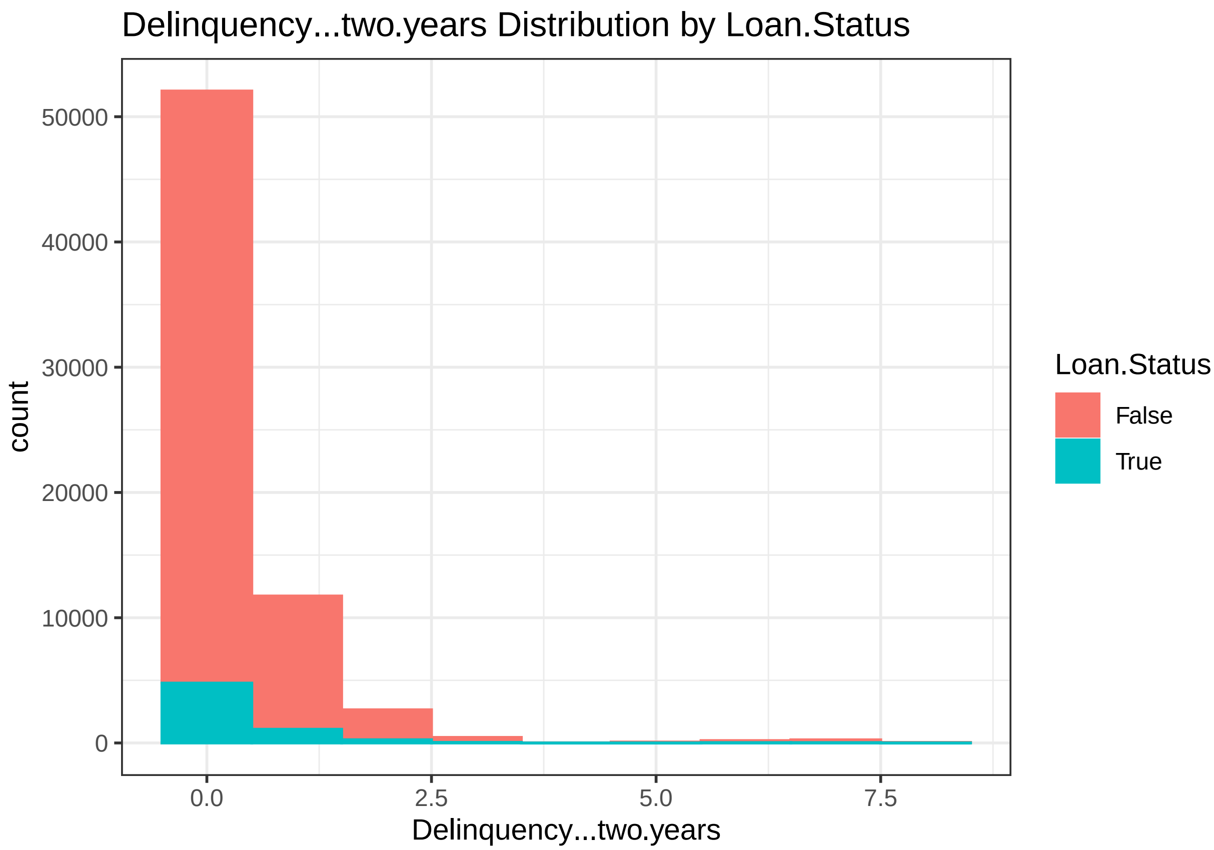
```
c <- ggplot(data, aes(x=Open.Account, fill=Loan.Status, color=Loan.Status)) + geom_histogram(binwidth =
c + theme_bw()
```



#

Visual 3

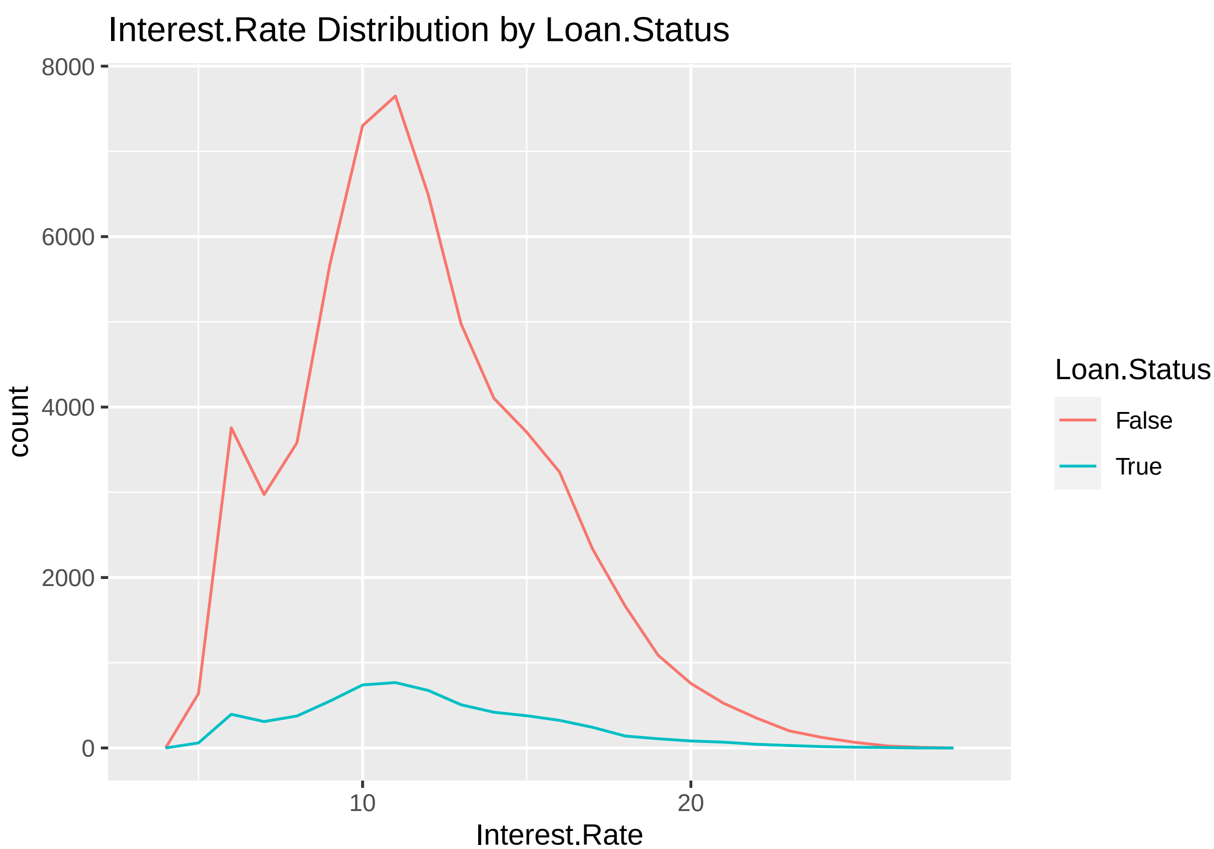
```
c <- ggplot(data, aes(x=Delinquency...two.years, fill=Loan.Status, color=Loan.Status)) + geom_histogram
c + theme_bw()
```



#

Visual 4

```
ggplot(data, aes(Interest.Rate, colour = Loan.Status)) + geom_freqpoly(binwidth = 1) + labs(title="Inte
```



Building a model

Split data into training and test data sets

```
indxTrain <- createDataPartition(y = data$Loan.Status,p = 0.75,list = FALSE)
training <- data[indxTrain,]
testing <- data[-indxTrain,] #Check dimensions of the split
```

```
prop.table(table(data$Loan.Status)) * 100
```

```
##
##     False      True
## 90.749003  9.250997
```

```
prop.table(table(training$Loan.Status)) * 100
```

```
##
##     False      True
## 90.748646  9.251354
```

```
prop.table(table(testing$Loan.Status)) * 100
```

```
##
##     False      True
## 90.750074  9.249926
```

Create objects x which holds the predictor variables and y which holds the response variables

```
x = training[,-35]
y = training$Loan.Status
```

```
library(e1071)
```

```
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))
```

```
model
```

```
## Naive Bayes
##
## 50598 samples
##    34 predictor
##     2 classes: 'False', 'True'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 45538, 45538, 45538, 45539, 45539, 45537, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa
##   FALSE            NaN           NaN
##    TRUE      0.9065971  -0.001417927
##
## Tuning parameter 'fL' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were fL = 0, usekernel = TRUE and adjust
##  = 1.
```

```
Predict <- predict(model,newdata = testing )
```

Get the confusion matrix to see accuracy value and other parameter values

```
confusionMatrix(Predict, testing$Loan.Status )
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction False  True
##      False 15294  1558
##      True     11     2
##
##                Accuracy : 0.907
##                  95% CI : (0.9025, 0.9113)
##     No Information Rate : 0.9075
##     P-Value [Acc > NIR] : 0.601
##
```

```
##                  Kappa : 0.001
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.999281
##             Specificity : 0.001282
##          Pos Pred Value : 0.907548
##          Neg Pred Value : 0.153846
##              Prevalence : 0.907501
##          Detection Rate : 0.906849
##    Detection Prevalence : 0.999229
##       Balanced Accuracy : 0.500282
##
##        'Positive' Class : False
##
```

R Output Explanation:-

The output is a confusion matrix and statistics for a machine learning model. The confusion matrix shows the number of true positives, true negatives, false positives, and false negatives for the model.

In this case, the model has made 15295 correct predictions that loans will be paid back (true negatives) and 0 correct predictions that loans will default (true positives). However, the model has also made 10 incorrect predictions that loans will be paid back when they actually default (false negatives), and 1560 incorrect predictions that loans will default when they actually will be paid back (false positives).

The statistics include accuracy, sensitivity, specificity, positive predictive value, negative predictive value, prevalence, and balanced accuracy.

Accuracy is the proportion of correct predictions out of all predictions. In this case, the accuracy is 0.9069, which means that the model is correct in its predictions 90.69% of the time.

Sensitivity (also called recall) is the proportion of true positives out of all actual positives. In this case, the sensitivity is 0.9993, which means that the model is able to correctly identify 99.93% of loans that will default.

Specificity is the proportion of true negatives out of all actual negatives. In this case, the specificity is 0.0000, which means that the model is not able to correctly identify any loans that will be paid back.

Positive predictive value (PPV) is the proportion of true positives out of all predicted positives. In this case, the PPV is 0.9074, which means that when the model predicts that a loan will default, it is correct 90.74% of the time.

Negative predictive value (NPV) is the proportion of true negatives out of all predicted negatives. In this case, the NPV is 0.9864, which means that when the model predicts that a loan will be paid back, it is correct 98.64% of the time.

Prevalence is the proportion of actual positives in the dataset. In this case, the prevalence is 0.9075, which means that 90.75% of loans in the dataset will actually default.

Balanced accuracy is the average of sensitivity and specificity. In this case, the balanced accuracy is 0.4997, which means that the model is not able to distinguish between loans that will default and those that will be paid back.

The output also shows the Kappa value, which is a measure of agreement between the model's predictions and the actual outcomes, and the p-value for Mcnemar's Test, which is a statistical test to determine if the model's errors are significantly different between false positives and false negatives. Finally, the output indicates that the "positive" class is "False", which means that the model is focused on predicting loans that will be paid back.

The model in this case appears to be very poor, as it is not able to correctly identify any of the loans that will be paid back and falsely identifies many loans as defaults. The low specificity and negative predictive value indicate that the model is not helpful in identifying loans that are likely to be paid back.