Business Analytics with R

FRAUDULENT CREDIT CARD TRANSACTION DETECTION

Presented by:
Arwin Kumar Ravi
Barath Kumar Dhanasekar
Devamsh Varma Mudunuri
Eswar Avinash Nadella
Mahadevan Ramanan
Manoj Krishna Manepalli

Problem Statement

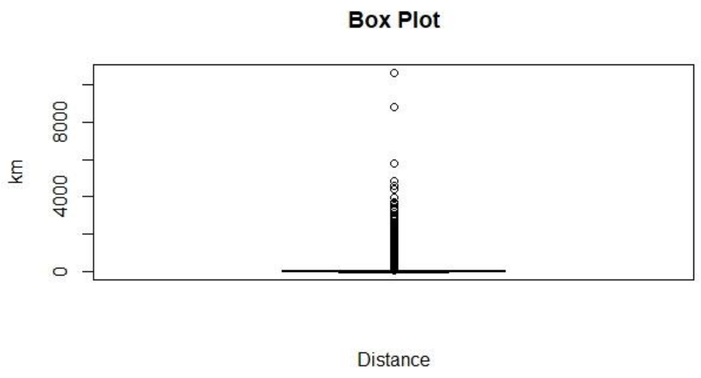
- Credit Card transactions in daily life
- Risks associated with these transactions
- Need to identify fraudulent transactions

Our Analysis and Target Variable

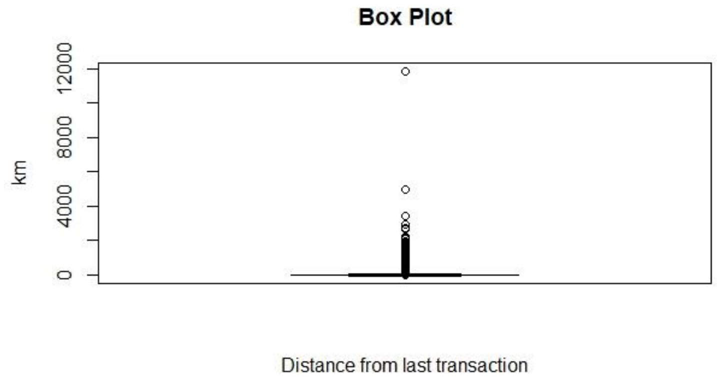
- Study and understand different variables that are involved in a credit card transaction
- Predict whether a credit card transaction is fraudulent or not
- ML algorithms used – decision tree classifier, logistic regression

Data Visualization and Exploration

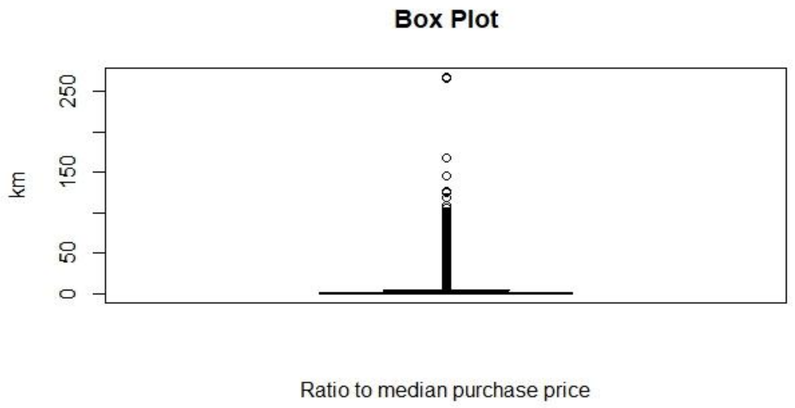
- The image shows a boxplot for numerical variables in the dataset
- Using IQR, which is a way to study the spread of the data, the upper and lower bounds are calculated by adding and subtracting 1.5 times the IQR from the median.
- IQR = Q3 – Q1, where Q1 is the 25th percentile, and Q3 is the 75th percentile



103,631 outliers in the column Distance from home



124,367 outliers in the column Distance from last transaction

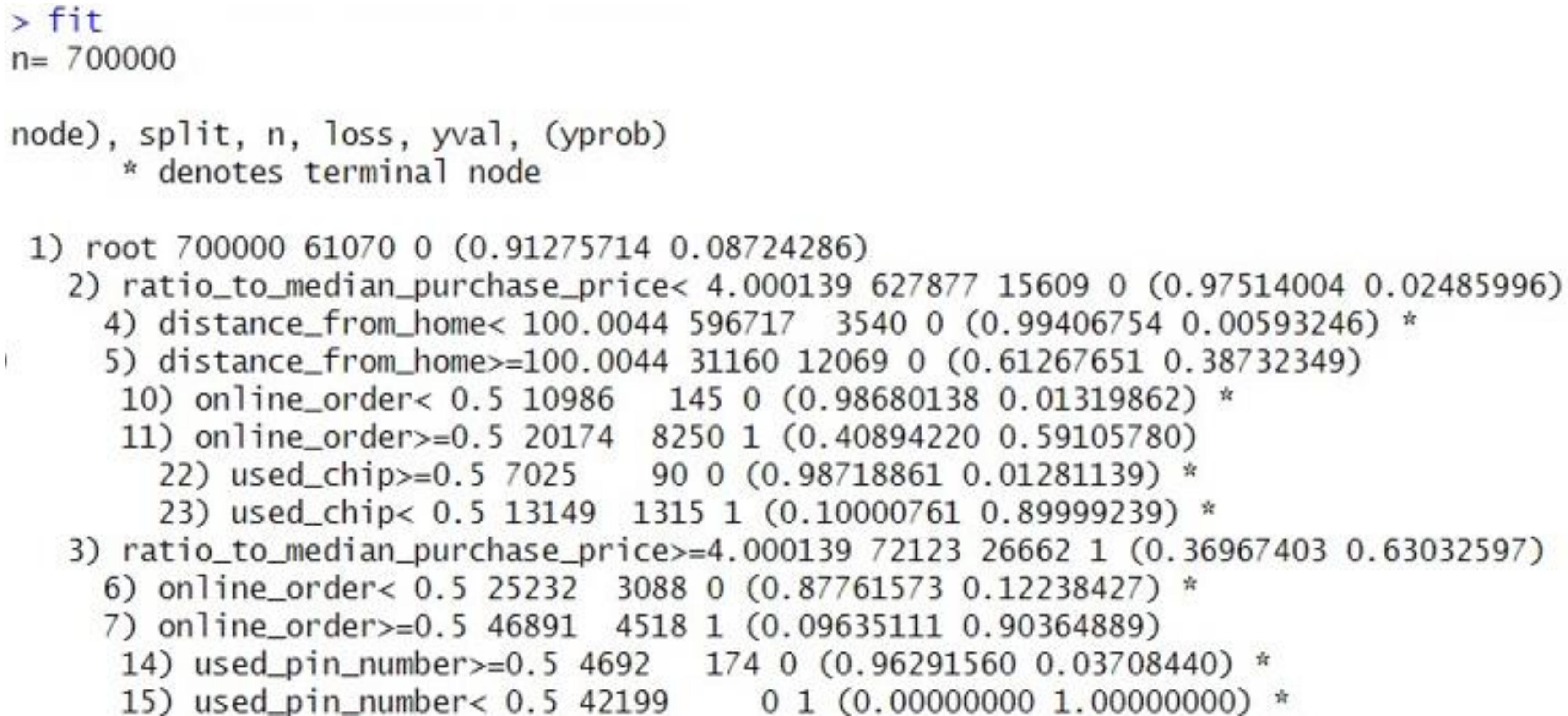


84,386 outliers in the column Ratio to median purchase price

Decision Tree Classifier

We have trained the model based on the following split – 70% train, 30% test

Basic Results



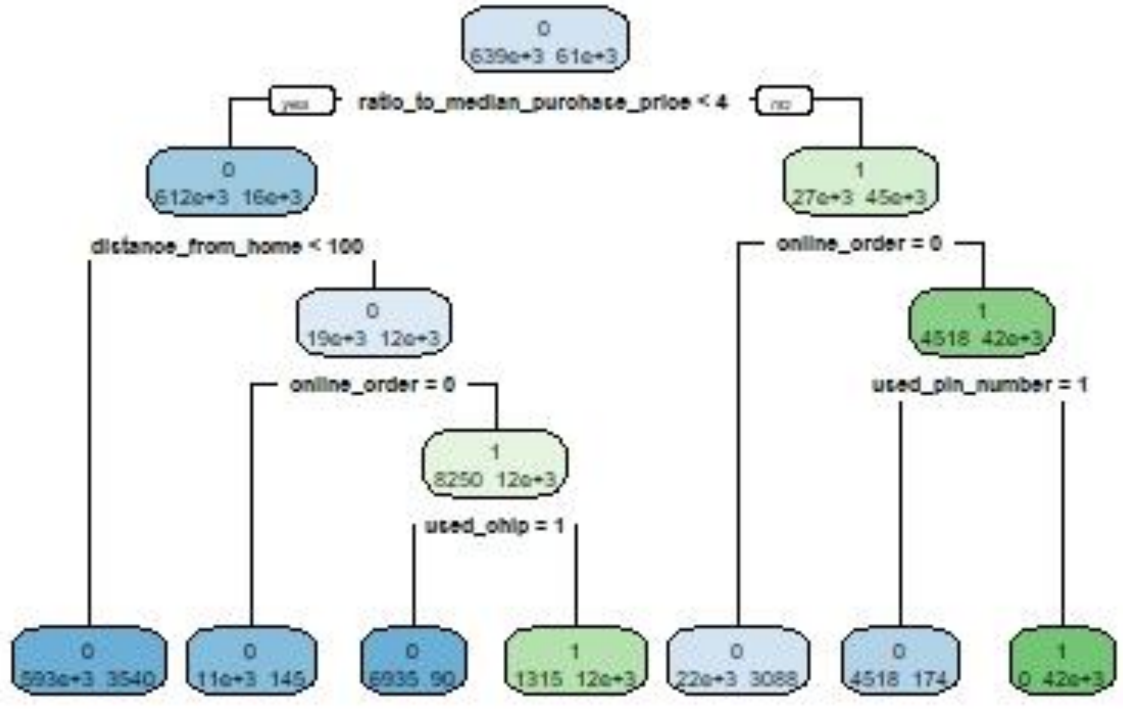
```
> fit
n= 700000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 700000 61070 0 (0.91275714 0.08724286)
   2) ratio_to_median_purchase_price< 4.000139 627877 15609 0 (0.97514004 0.02485996)
     4) distance_from_home< 100.0044 596717  3540 0 (0.99406754 0.00593246) *
     5) distance_from_home>=100.0044 31160 12069 0 (0.61267651 0.38732349)
      10) online_order< 0.5 10986   145 0 (0.98680138 0.01319862) *
      11) online_order>=0.5 20174  8250 1 (0.40894220 0.59105780)
        22) used_chip>=0.5 7025    90 0 (0.98718861 0.01281139) *
        23) used_chip< 0.5 13149  1315 1 (0.10000761 0.89999239) *
   3) ratio_to_median_purchase_price>=4.000139 72123 26662 1 (0.36967403 0.63032597)
     6) online_order< 0.5 25232  3088 0 (0.87761573 0.12238427) *
     7) online_order>=0.5 46891  4518 1 (0.09635111 0.90364889)
      14) used_pin_number>=0.5 4692   174 0 (0.96291560 0.03708440) *
      15) used_pin_number< 0.5 42199     0 1 (0.00000000 1.00000000) *
```

Decision Tree Classifier

Tree Interpretation – Test Dataset

True Positive (TP): Pred Pos & Actual Pos
False Positive (FP): Pred Pos & Actual Neg
True Negative (TN): Pred Neg & Actual Neg
False Negative (FN): Pred Neg & Actual Pos

Summary
TP = 24632
FP = 689
TN = 272978
FN = 1701



Decision Tree Classifier

Confusion matrix of train dataset -

Actual		
Predicted	**0**	**1**
0	637615	7037
1	1315	54033

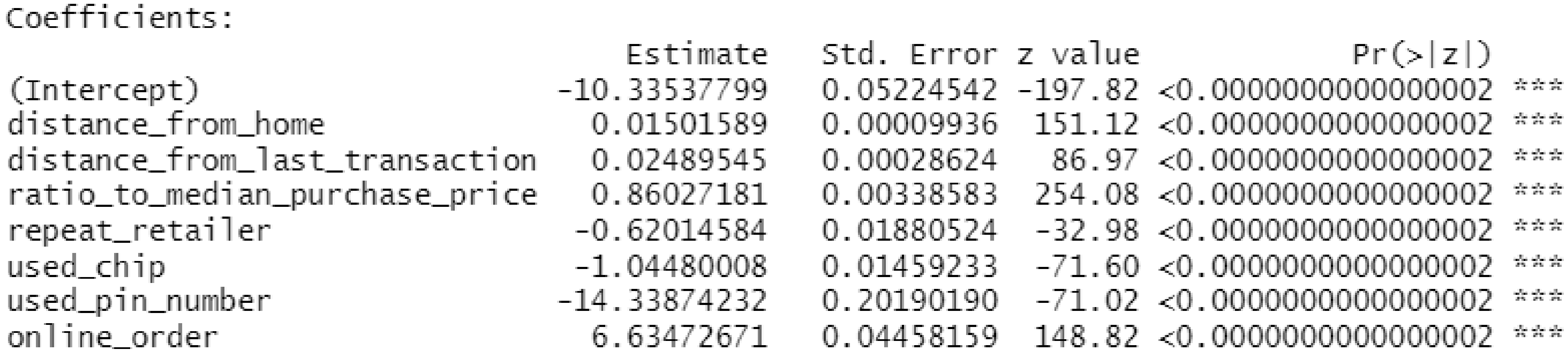
Confusion matrix of test dataset -

Actual		
Predicted	**0**	**1**
0	273122	3121
1	545	23212

- Performance of the train dataset
  - Accuracy = 0.992
  - Error Rate = 0.008
  - TPR = 0.935
  - FNR = 0.065
  - TNR = 0.997
  - FPR = 0.003
- Performance of the test dataset
  - Accuracy = 0.988
  - Error Rate = 0.012
  - TPR = 0.88
  - FNR = 0.12
  - TNR = 0.998
  - FPR = 0.002

Logistic Regression

- We have trained the model based on the following split – 70% train, 30% test
- Summary of the logistic regression performed –



```
Coefficients:
                                 Estimate   Std. Error z value           Pr(>|z|)    
(Intercept)                  -10.33537799   0.05224542 -197.82 <0.0000000000000002 ***
distance_from_home             0.01501589   0.00009936  151.12 <0.0000000000000002 ***
distance_from_last_transaction  0.02489545   0.00028624   86.97 <0.0000000000000002 ***
ratio_to_median_purchase_price  0.86027181   0.00338583  254.08 <0.0000000000000002 ***
repeat_retailer               -0.62014584   0.01880524  -32.98 <0.0000000000000002 ***
used_chip                     -1.04480008   0.01459233  -71.60 <0.0000000000000002 ***
used_pin_number              -14.33874232   0.20190190  -71.02 <0.0000000000000002 ***
online_order                   6.63472671   0.04458159  148.82 <0.0000000000000002 ***
```

- Interpretations –
  1. An increase in distance_from_home by a mile, provided all other values are kept constant will increase the odds of credit card transaction being fraudulent by 1.51%
  2. An increase in distance_from_last_transaction by a mile, provided all other values are kept constant will increase the odds of credit card transaction being fraudulent by 2.5%
  3. Provided all other values are constant, If the ratio of the purchase price to the median price increases by a unit, the odds of credit card transaction being fraudulent increases by 136%

Logistic Regression - Performance

Confusion matrix of train dataset -

Actual		
Predicted	**0**	**1**
0	634573	24281
1	4357	36789

Confusion matrix of test dataset -

Actual		
Predicted	**0**	**1**
0	271813	10712
1	1854	15621

- **Performance of the train dataset**
  - Accuracy = 0.96
  - Error Rate = 0.04
  - TPR = 0.60
  - FNR = 0.40
  - TNR = 0.99
  - FPR = 0.01
- **Performance of the test dataset**
  - Accuracy = 0.96
  - Error Rate = 0.04
  - TPR = 0.59
  - FNR = 0.41
  - TNR = 0.99
  - FPR = 0.01

Logistic Regression - Performance

Confusion matrix of train dataset -

	Actual	
Predicted	**0**	**1**
0	634573	24281
1	4357	36789

Confusion matrix of test dataset -

	Actual	
Predicted	**0**	**1**
0	271813	10712
1	1854	15621

- **Performance of the train dataset**
  - Accuracy = 0.96
  - Error Rate = 0.04
  - TPR = 0.60
  - FNR = 0.40
  - TNR = 0.99
  - FPR = 0.01
- **Performance of the test dataset**
  - Accuracy = 0.96
  - Error Rate = 0.04
  - TPR = 0.59
  - FNR = 0.41
  - TNR = 0.99
  - FPR = 0.01