Caravan Insurance

The labelled dataset was randomly split into two subsets; training and testing datasets. All the social demographic(1:42) features were chosen as categorical.

The problem with our dataset was it contained imbalanced classes.

|  |  |
| --- | --- |
| No | Yes |
| 4236 | 264 |

The training set contained 4236 labels belonging to No and 264 labels belonging to Yes.

|  |  |
| --- | --- |
| No (probability) | Yes (probability |
| 0.94133333 | 0.05866667 |

An unbalance data was creating bias, it was creating bias towards the more common class. So first the train dataset for balanced for the modelling, the distribution was changed with sampling methods.

Rose sampling technique was used which is the hybrid method that combines under and over-sampling methods. Under and over sampling methods were also tried, but more balanced data was achieved using ROSE. Below is the output of balanced train dataset.

|  |  |
| --- | --- |
| No | Yes |
| 2295 | 2205 |

|  |  |
| --- | --- |
| No (probability) | Yes (probability |
| 0.51 | 0.49 |

After balancing the train data, decision tree model was created on the train data.

Next, we generated the details of all significant independent variables and their relative importance. Output below shows the details of selected features.

Overall

ABRAND 15.64643

ABROM 48.00310

APERSAUT 196.68819

APLEZIER 32.96883

MAUT0 29.17212

MAUT1 14.05645

MAUT2 14.61193

MBERARBG 77.47506

MBERARBO 171.52674

MBERHOOG 39.09771

MBERMIDD 27.58700

MGEMLEEF 18.91379

MGODPR 89.84165

MHHUUR 220.53181

MHKOOP 224.00245

MINK3045 87.93155

MINK4575 34.29525

MINKGEM 166.05064

MINKM30 118.31035

MKOOPKLA 52.53214

MOPLLAAG 20.00463

MOSHOOFD 55.77443

MOSTYPE 474.43235

MRELSA 39.80633

MSKA 15.20386

MSKC 30.43291

MSKD 94.59939

MZFONDS 49.78426

MZPART 32.18414

PBRAND 141.88003

PBROM 76.08675

PPERSAUT 289.77294

PPLEZIER 57.90339

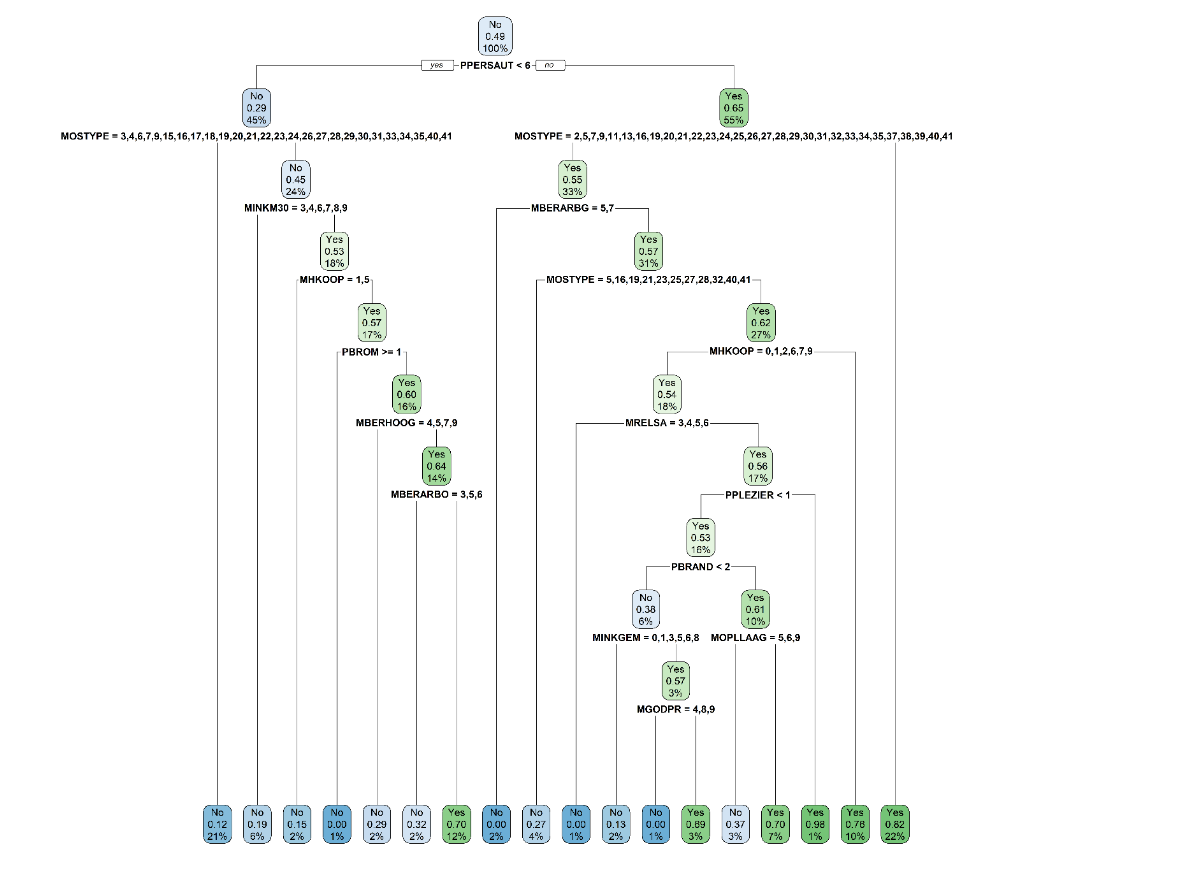
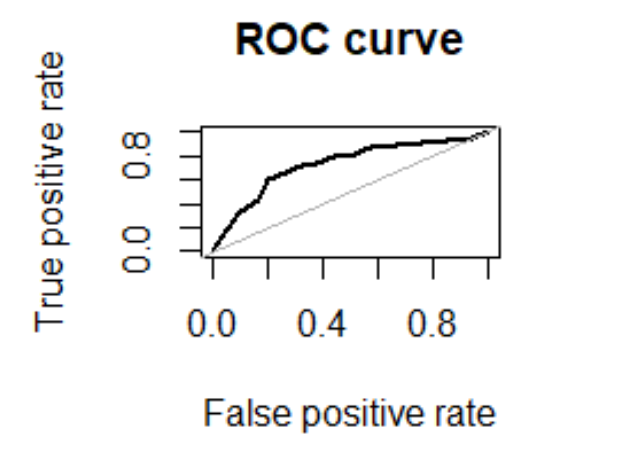


Figure : Decision Tree

Training data ROC provided the best threshold as 0.53.



Area under the curve (AUC): 0.731.

Confusion matrix for training data set(imbalanced):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Prediction |  |  |
|  |  |  |  |  |
|  | **Row Labels** | **No** | **Yes** | **Grand Total** |
| Actual | No | 1932 | 2304 | 4236 |
|  | Yes | 120 | 144 | 264 |
|  | **Grand Total** | **2052** | **2448** | **4500** |
|  |  |  |  |  |
|  |  |  |  |  |

Performance measure for training data:

Accuracy : 0.4613

Sensitivity : 0.45609

Specificity : 0.54545

Pos Pred Value : 0.94152

Neg Pred Value : 0.05882

Confusion Matrix for training data(balanced) :

Actual

Prediction

|  |  |  |  |
| --- | --- | --- | --- |
|  | No | Yes | Total |
| No | 1712 | 340 | 2052 |
| Yes | 583 | 1865 | 2448 |
| Total | 2295 | 2205 | 4500 |

Table 1: R output

Accuracy : 0.7949

Sensitivity : 0.7460

Specificity : 0.8458

Pos Pred Value : 0.8343

Neg Pred Value : 0.7618

Confusion Matrix for testing data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | Prediction | |
|  |  |  | |  |
|  | **Row Labels** | **No** | **Yes** | **Grand Total** |
| Actual | No | 873 | 365 | 1238 |
|  | Yes | 26 | 58 | 84 |
|  | **Grand Total** | **899** | **423** | **1322** |
|  |  |  |  |  |

Performance measure for testing data:

Accuracy : 0.7042

Sensitivity : 0.7052

Specificity : 0.6905

Pos Pred Value : 0.9711

Neg Pred Value : 0.1371

precision: 0.137

recall: 0.690

F: 0.114

Precision recall curve:

