Analysis of Car Accidents In USA

By Lior Ben David 208778225

and Sahar Ziv 312160369

Car Accidents link to code

Contents

Introduction

In our analysis we focused on car accidents data from USA.

In the US, a fatal accident occurs every 15 minutes on average.

The issue is of interest to the insurance companies and the transportation and government ministries.

In our analysis we will try to classify the severity of the injury of the drivers involved in road accidents with the help of personal data, vehicle data and the weather.

The data was taken from the Kaggle website.

Our data was collected in 2018 and includes three files:

Number of features	Number of observations	Datasets
51	48443	Acc_18
54	120230	Pers_18
87	86105	Veh_18

The feature that links all the tables is 'CASENUM' which represents the number of accidents.

For each of the data sets, they tried to complete missing values (Imputation), therefore there are identical features, and the features that have already been completed we would like to delete to remain only with the original data.

Our label feature is 'INJ_SEV' which describes the severity of the injury to a person in the crash.

Correcting the data

At first, because our research question is about drivers, we only filtered out the drivers' data from the person's dataset. Where the feature 'PER_TYP' == '1' it means that it is the observation of the driver. So now we have 85,916 observations of drivers in our data.

The second step, we merged all data sets to one data set by the feature 'CASENUM' and remove duplicate columns.

After that, we checked if our data included null values, there was not and it seems weird to us, so we checked the distribution of values for each feature.

We noticed that the missing values are displayed in the data as 9/8, 98/99, 999, etc., there for we have summarized the data on the features in a table so that we can decide whether to delete them and whether we need to fill in missing values and if so, with what digit they are represented. Then we converted all the values that represent missing values to null.
Now our data is ready to imputation. We decided to use KNN imputation because it is simple and gets very good results for imputation. (We saved our data after imputation just because it takes a little bit time to do the KNN and we didn't want to do it every time we work on our analysis)
Our label contains 7 levels after imputation:
0 No Apparent Injury
1 Possible Injury
2 Suspected Minor Injury
3 Suspected Serious Injury
4 Fatal Injury
5 Injured, Severity Unknown
6 Died Prior to Crash

We decided to combine between the levels of the label to 2 levels:

- 0 - no injury to the driver or minor
- 1 - severe injurey to the driver

After that we checked the correlations between the features and dropped variables that have a high correlation between them.

Then we perform several actions:

1. We paired our variables into numeric and categorical features.
2. We split our data to train data (70%) and test data (30%)
3. We scaled the numeric values.
4. We did One-hot Encoding on the categorical features.
5. We noticed that our data is imbalanced, there were more observations with label "0" than label "1" so we fixed this by oversampling with SMOTE so the number of observations from each category would be the same.

6. We checked the correlation of the numeric features before and after over-sampling and there was not a big difference which is good. (Appendix 1)
7. We performed feature selection with LASSO.
8. We calculated the correlation between the features and the label to know if there is feasibility for our project. We have accepted that the three most influencing variables on the label are: (appendix 2)
   1. REST_USE_3 which means that shoulder and lap belt used by this person at the time of the crash.
   2. RELJCT2_3 which means that the accident was intersection Related
   3. TRAV_SP - the speed the vehicle was traveling prior to the occurrence of the crash
9. We checked whether we have outliers in the data by fitting logisitic regression model. We noticed that the residuals are normally distributed with the mean 0. We saw that we had one outlier observation, so we dropped it. (Appendix 3)
10. Again we saved our data only to save time working on the project.

Challenges and bias in the data -

We presented the number of accidents per maker, and we saw that the number of accidents is higher to American manufacturers than non-American manufacturers.

It makes sense because our data is sampled from the USA, and it makes a choosing bias and can cause wrong conclusions.

Therefore, we decided to remove the features MAKE and MODEL.

We also saw that there are states that weren't sampled, which means that we have choosing bias in our data. We didn't have something to do with it, so we use it as is and refer to it in the conclusion.

One challenge that we had was unbalanced Data. To overcome it we over-sampled the data with SMOTE to make the label balanced.

Model Training

We trained 4 models.

1. **Logistic Regression** with L1 penalty ("Lasso") – We chose this model because it can easily predict a binary label.
2. **K-Nearest Neighbors** - We chose this model because it is simple and can predict well and does not require any prior preparation of the data.
3. **Classification Tree** - We chose a classification tree to check if it classifies the data well without scaling
4. **Random Forest -** We know that this model has good classification results because it uses classification from many trees and thus reduces its error.

All the model's tuning parameters were estimated in 5-fold Cross-Validation and we will show the results of the best estimators.

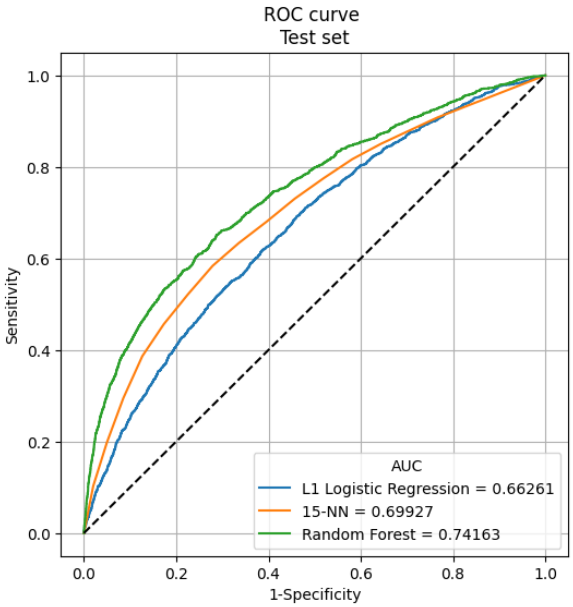
We will compare the method by Accuracy, Sensitivity, Specificity and AUC of ROC curve.

Results and conclusions

Results –

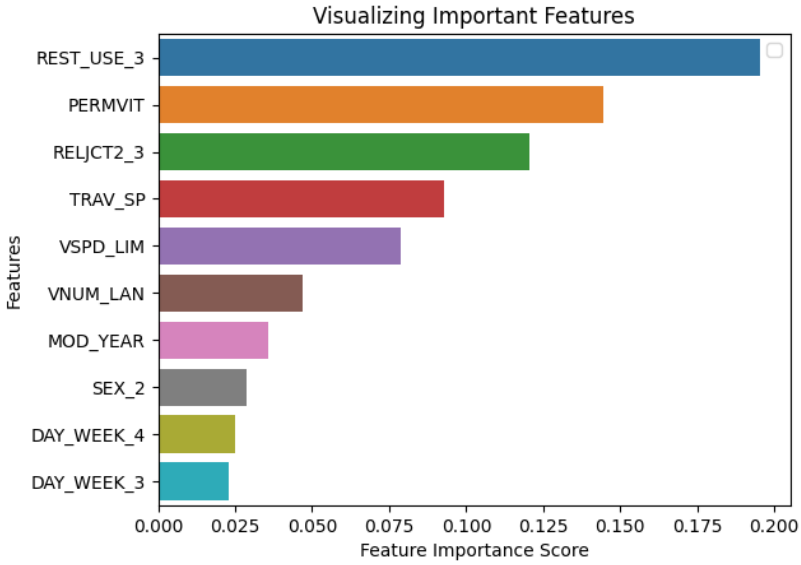
Model	Accuracy	Sensitivity	Specificity
Logistic regression	0.85998	0.2604	0.89677
KNN	0.66262	0.63356	0.6644
Random Forest	0.83371	0.49329	0.8546
Classification tree	0.942		

Although the accuracy of the classification tree is very high, we saw that it classified almost all observations into a certain group, so it is not a good model for our data. (appendix 4)



We can see that the Random Forest model is the best model because its AUC is the biggest.

This are the most important features:



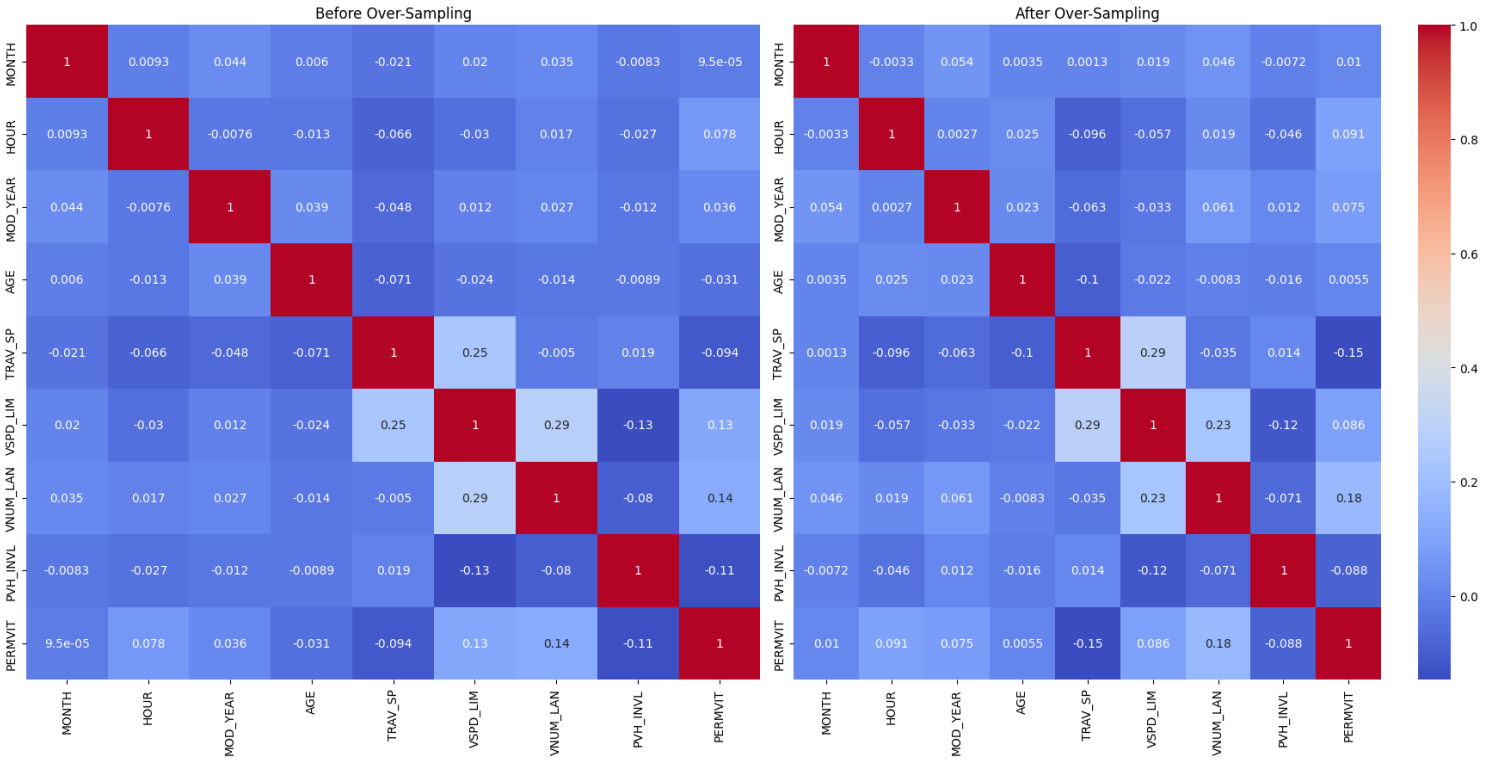
1. REST_USE_3 which means that shoulder and lap belt used by this person at the time of the crash.
2. PERMVIT – which is the number of persons in motor vehicles.
3. RELJCT2_3 which means that the accident was intersection Related.
4. TRAV_SP - the speed the vehicle was traveling prior to the occurrence of the crash.

Work to be continued –

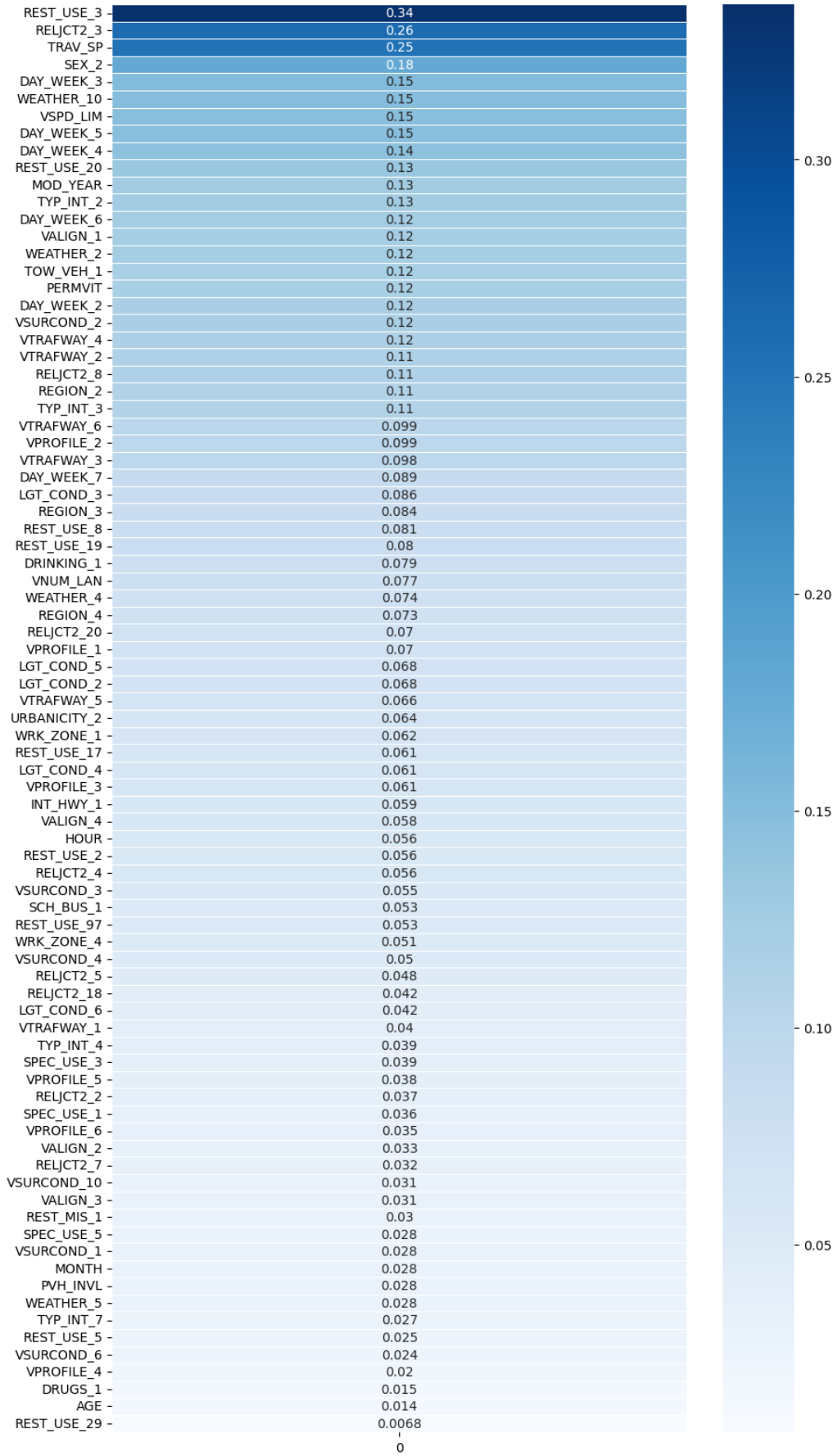
In the future, the performance of these models can also be evaluated on data from other years or other places.

Appendices

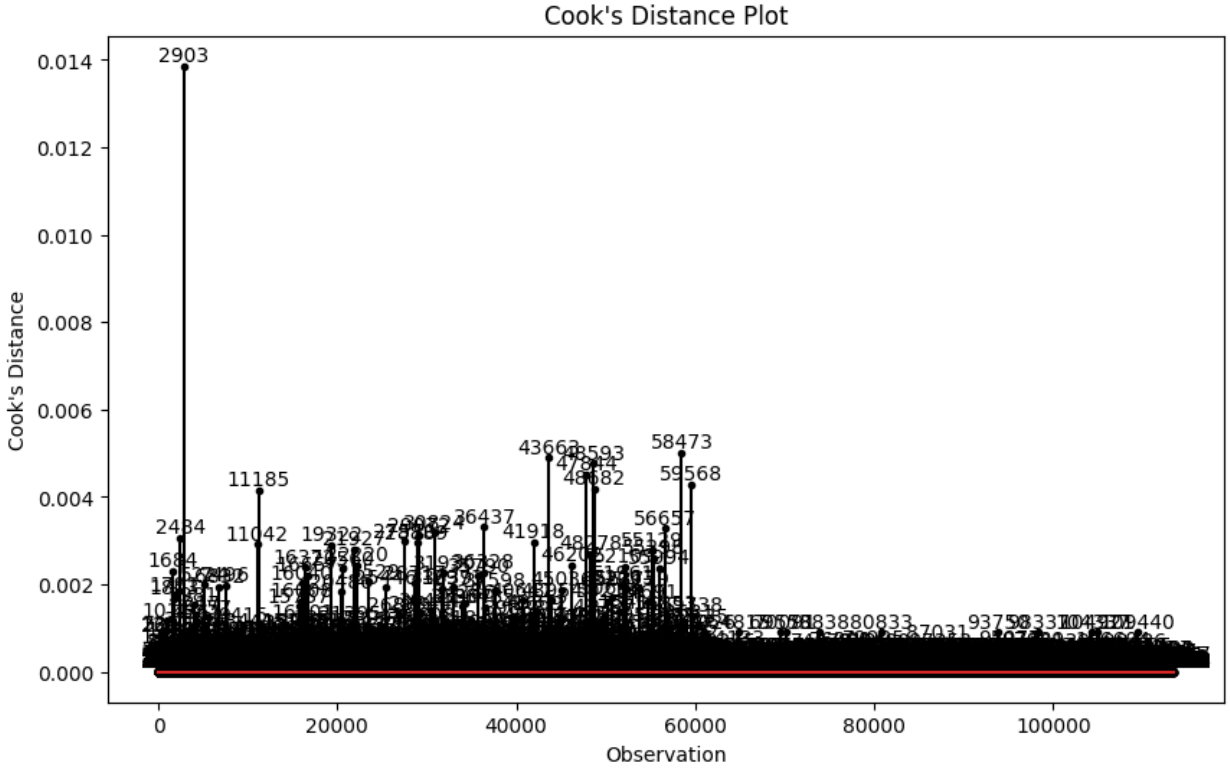
Appendix 1 –



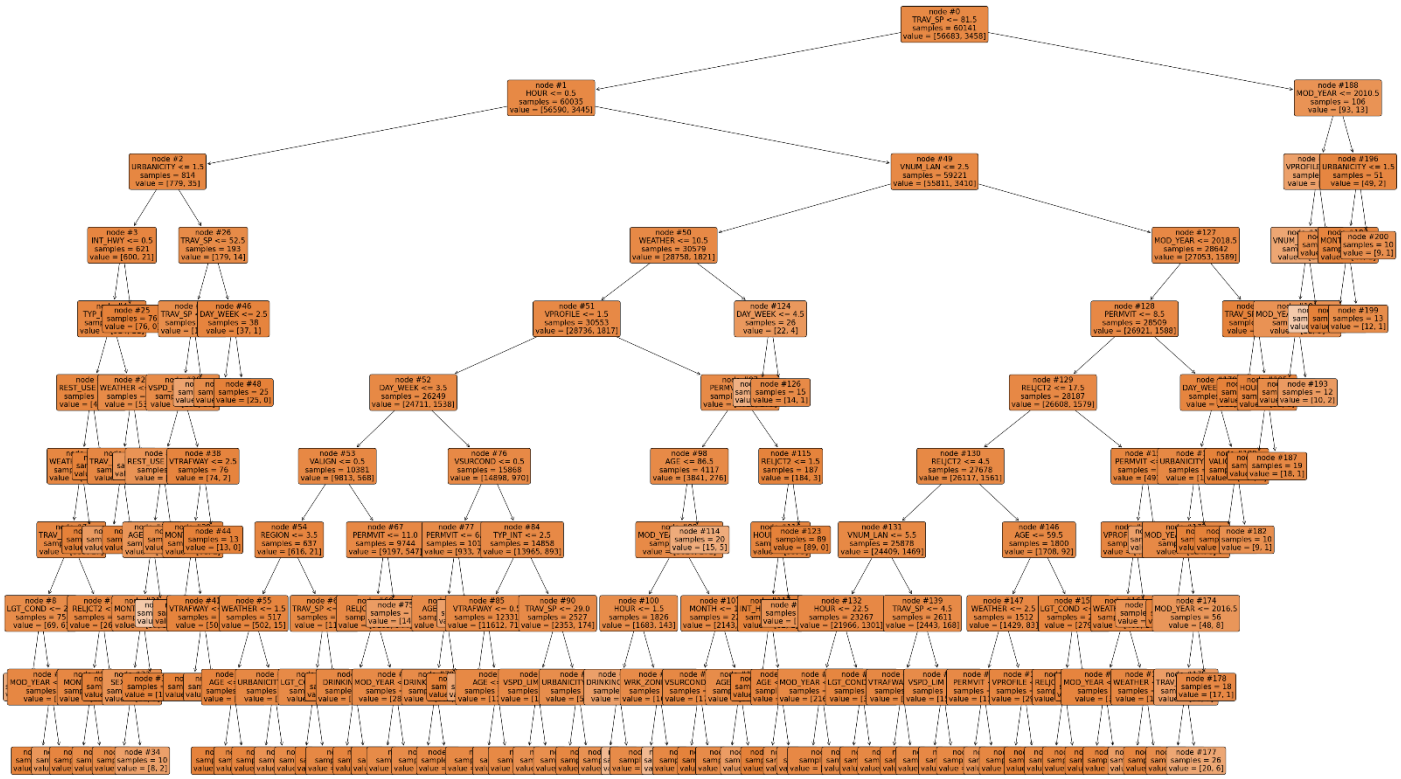
Appendix 2 –


REST_USE_3 0.34
RELJCT2_3 0.26
TRAV_SP 0.25
SEX_2 0.18
DAY_WEEK_3 0.15
WEATHER_10 0.15
VSPD_LIM 0.15
DAY_WEEK_5 0.15
DAY_WEEK_4 0.14
REST_USE_20 0.13
MOD_YEAR 0.13
TYP_INT_2 0.13
DAY_WEEK_6 0.12
VALIGN_1 0.12
WEATHER_2 0.12
TOW_VEH_1 0.12
PERMVIT 0.12
DAY_WEEK_2 0.12
VSURCOND_2 0.12
VTRAFWAY_4 0.12
VTRAFWAY_2 0.11
RELJCT2_8 0.11
REGION_2 0.11
TYP_INT_3 0.11
VTRAFWAY_6 0.099
VPROFILE_2 0.099
VTRAFWAY_3 0.098
DAY_WEEK_7 0.089
LGT_COND_3 0.086
REGION_3 0.084
REST_USE_8 0.081
REST_USE_19 0.08
DRINKING_1 0.079
VNUM_LAN 0.077
WEATHER_4 0.074
REGION_4 0.073
RELJCT2_20 0.07
VPROFILE_1 0.07
LGT_COND_5 0.068
LGT_COND_2 0.068
VTRAFWAY_5 0.066
URBANICITY_2 0.064
WRK_ZONE_1 0.062
REST_USE_17 0.061
LGT_COND_4 0.061
VPROFILE_3 0.061
INT_HWY_1 0.059
VALIGN_4 0.058
HOUR 0.056
REST_USE_2 0.056
RELJCT2_4 0.056
VSURCOND_3 0.055
SCH_BUS_1 0.053
REST_USE_97 0.053
WRK_ZONE_4 0.051
VSURCOND_4 0.05
RELJCT2_5 0.048
RELJCT2_18 0.042
LGT_COND_6 0.042
VTRAFWAY_1 0.04
TYP_INT_4 0.039
SPEC_USE_3 0.039
VPROFILE_5 0.038
RELJCT2_2 0.037
SPEC_USE_1 0.036
VPROFILE_6 0.035
VALIGN_2 0.033
RELJCT2_7 0.032
VSURCOND_10 0.031
VALIGN_3 0.031
REST_MIS_1 0.03
SPEC_USE_5 0.028
VSURCOND_1 0.028
MONTH 0.028
PVH_INVL 0.028
WEATHER_5 0.028
TYP_INT_7 0.027
REST_USE_5 0.025
VSURCOND_6 0.024
VPROFILE_4 0.02
DRUGS_1 0.015
AGE 0.014
REST_USE_29 0.0068
0
0.30
0.25
0.20
0.15
0.10
0.05

Appendix 3 –


Cook's Distance Plot
Cook's Distance
Observation
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
20000
40000
60000
80000
100000
2903
11185
2434
11042
43662
58473
59568
56657
26437
41918

Appendix 4 –


node #0
TRAV_SP <= 81.5
samples = 60141
value = [56683, 3458]
node #1
HOUR <= 0.5
samples = 60035
value = [56590, 3445]
node #188
MOD_YEAR <= 2010.5
samples = 106
value = [93, 13]
node #2
URBANICITY <= 1.5
samples = 814
value = [779, 35]
node #49
VNUM_LAN <= 2.5
samples = 59221
value = [55811, 3410]
node #196
URBANICITY <= 1.5
samples = 51
value = [49, 2]
node #3
INT_HWY <= 0.5
samples = 621
value = [600, 21]
node #26
TRAV_SP <= 52.5
samples = 193
value = [179, 14]
node #50
WEATHER <= 10.5
samples = 30579
value = [28758, 1821]
node #127
MOD_YEAR <= 2018.5
samples = 28642
value = [27053, 1589]
node #51
VPROFILE <= 1.5
samples = 30553
value = [28736, 1817]
node #124
DAY_WEEK <= 4.5
samples = 26
value = [22, 4]
node #128
PERMVIT <= 8.5
samples = 28509
value = [26921, 1588]
node #52
DAY_WEEK <= 3.5
samples = 26249
value = [24711, 1538]
node #129
RELICT2 <= 17.5
samples = 28187
value = [26608, 1579]
node #53
VALIGN <= 0.5
samples = 10381
value = [9813, 568]
node #76
VSURCOND <= 0.5
samples = 15868
value = [14898, 970]
node #98
AGE <= 86.5
samples = 4117
value = [3841, 276]
node #115
RELICT2 <= 1.5
samples = 187
value = [184, 3]
node #130
RELICT2 <= 4.5
samples = 27678
value = [26117, 1561]
node #54
REGION <= 3.5
samples = 637
value = [616, 21]
node #67
PERMVIT <= 11.0
samples = 9744
value = [9197, 547]
node #77
PERMVIT <= 6
samples = 103
value = [93, 7]
node #84
TYP_INT <= 2.5
samples = 14856
value = [13965, 891]
node #114
samples = 20
value = [15, 5]
node #123
samples = 89
value = [89, 0]
node #131
VNUM_LAN <= 5.5
samples = 25878
value = [24409, 1469]
node #146
AGE <= 59.5
samples = 1800
value = [1708, 92]
node #182
samples = 10
value = [9, 1]
node #8
LGT_COND <= 2
samples = 75
value = [69, 6]
node #41
VTRAFWAY <= 0.5
samples = 517
value = [502, 15]
node #55
WEATHER <= 1.5
samples = 517
value = [502, 15]
node #85
VTRAFWAY <= 0.5
samples = 12331
value = [11612, 719]
node #90
TRAV_SP <= 29.0
samples = 2527
value = [2353, 174]
node #100
HOUR <= 1.5
samples = 1826
value = [1683, 143]
node #132
HOUR <= 22.5
samples = 23267
value = [21966, 1301]
node #139
TRAV_SP <= 4.5
samples = 2611
value = [2443, 168]
node #147
WEATHER <= 2.5
samples = 1512
value = [1429, 83]
node #174
MOD_YEAR <= 2016.5
samples = 56
value = [48, 8]