Contents

1 Executive Summary

The increased number of vehicles on the roads and express highways has led to an increase in the number of accidents in the whole of the United States. Saving human life is important. But, a road accident is often accompanied by infrastructure loss, traffic delays, and management of resources to clear the accident spot. This project was built on the business scenario to facilitate City's Transportation Management Center (TMC) to predict the traffic delay level, post road accidents, and effectively deploy their resources. A countrywide car accident database from February 2016 to June 2020 (Moosavi, Samavatian, Parthasarathy and Ramnath, 2020) has been used as the data source. Several multivariate statistical techniques have been utilized to classify the impact of the accident on traffic delays. Predictive models were developed using ordinal logistic regression, binomial logistic regression, k-nearest neighbor, discriminant analysis, decision tree and random forest algorithms. K-Nearest Neighbor algorithm is considered as the best model for this data with high sensitivity and specificity. A conscious effort has been made in building a model with the input variables, whose data will be available when an accident is reported at TMC.

2 Introduction

United States is one the countries with busiest traffic conditions. “The level of traffic is one of the reasons leading to more traffic accidents: In 2018, there were some 12 million vehicles involved in crashes in the United States”(Wagner, 2020). Recent statistical projection of National Highway Traffic Safety Administration estimated that 36,120 people have died in motor vehicle traffic accidents in 2019. Although, total fatal crashes have declined by 1.2 percent from 2018, Federal Highway Administration reports the total vehicle miles traveled (VMT) in 2019 is increased by about 28.8 billion miles or about 0.2 percent increase from 2018 (National Center for Statistics and Analysis, 2020). This shows that people are traveling more on the roadways. One of the major impacts of higher VMT is Congestion.

In literature, congestion is believed to have very mixed effects on roadway accidents. Many studies have found a strong positive relationship between congestion and total accidents (Woo, 1957; Head, 1959). However, some researchers have found that there is U shaped function relationship between the two variables implying higher accidents at low or high congestion levels (Zhou & Sisiopiku, 1997; Martin, 2002). It has also been found that fatalities are more often at median level traffic compared to high or low traffic conditions. This implies that congestion can be viewed as a potential safety bearer but a nightmare for traffic management authorities. There is no clear consensus on the effects of congestion on vehicle crashes or vice versa (Retallack & Ostendorf, 2019). Both congestion and crashes not only incurs social cost, but also infrastructure and economic loses.

In 2010, the total economic cost of road accidents in the United States was estimated to be $242 billion. Traffic congestion caused post road accident, including travel delay and excess fuel consumption accounted for nearly $28 billion (Blincoe, Miller, Zaloshnja, & Lawrence, 2015). Minor Traffic Incidents and accidents with injuries but no fatalities have an expected event duration of up to 2 hours to clear traffic; Major Truck Accidents and Multivehicle Crashes has an event duration of 2 to 24 hours (Wallace, ITS Professional Capacity Building Program, n.d.). Hence, there is a need to study the traffic delays as a function of roadway and weather characteristics at the accident spots.

2.1 Business Scenario

Traffic Management Centers (TMC) serves as a control center to manage urban roads and highways. It is responsible to monitor traffic signals, intersections, and roads and actively strategize to alleviate congestion. They also coordinate with the local agencies during special events, emergencies, accidents, and regular day to day operations. Operators at TMC monitor the roadways through closed-circuit television (CCTV) system to inform the authorities aboutany problems. These TMC’s also have representatives from law enforcement, emergency management services(EMS), and local transit agencies for easy cooperation among themselves to efficiently manage any untoward situations.

This project would be essential for a City’s Transportation Management Center (TMC) to analyze and predict the traffic delay level post accident to efficiently manage their resources. The project would also help the state-licensed Emergency Medical Services (EMS) Dispatch centers to recognize the requirement of airlift to save the life.

2.2 Data Description

US Accidents is countrywide traffic accident dataset constituting all the accidents in 49 states of the US, from February 2016 to June 2020(Moosavi, Samavatian, Parthasarathy and Ramnath, 2020). There are nearly 3.5 million road accident records with City and state wise information about TMC code, accident severity, time when impact of accident on traffic was dismissed, GPS coordinates, length of road affected by accident, weather conditions, and nearby points of interest. The description of the variables can be found in the appendix 1.

Dataset Acknowledgments:

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019

2.3 Objective

This study focuses on two objective:

1. Perform Hypothesis testing on initial assumptions about the variables in the dataset
2. Build a predictive model using machine learning algorithms that can be used by TMC centers to effectively manage their resources

2.4 Prelimnary Assumptions

Following are the preliminary assumptions made in this study:

- Traffic delay level is correlated with the location of the accidents.
- Most of the accidents happen in adverse weather conditions and the traffic delay level is correlated with weather.
- Traffic delay level is high during AM (7:00 - 9:00) and PM (16:00 - 18:00) peak period.
- A typical weekday and weekend have a different share of traffic delay level 3 accidents.
- Temperature, Humidity and Precipitation are significant predictors of traffic delay level.

3 Data Preprocessing

The US accidents data set has a large number of incomplete and missing information. Data cleaning is done in order to address the anomalies.

3.1 Formatting

Since the data is collected for various sources, the data format is inconsistent. For the analysis, the date and time formats for Start_Time, End_Time and Weather_Timestamp have been standardized as 'YYYY-MM-DD HH:mm:ss'.

3.2 Adding a calculated field

The severity level in this data indicates the impact of an accident on traffic and not the severity of the actual accident. To avoid the confusion, severity level will be referred as **traffic delay level**. The accidents in this dataset were reported by three different sources. In order to compare the classification of the traffic delay levels among different sources, it is essential to have the duration of the traffic post-accident. A new calculated field **Traffic_Duration_min** (difference between End time when the impact of accident on traffic flow was dismissed and the accident start time) has been introduced indicating the duration of traffic post-accident in minutes. For this analysis, the Traffic duration up to six hours (360 min) has been considered.

To understand the pattern of accidents according to the time of day, day of the week and month,**STHR**, **Day** and **Month** columns have been introduced. "STHR" captures the start hour of the accident, "Day" represents the day of the week and "Month" as a categorical variable depicting the month of the accident.

Table 1: *Proportion of missing values in the data set*

	value	Proportion_Missing
TMC	649029	20.77
Number	1956163	62.59
Temperature(F)	56653	1.81
Humidity(%)	60191	1.93
Pressure(in)	47994	1.54
Visibility(mi)	65381	2.09
Wind_Speed(mph)	388896	12.44
Precipitation(in)	1691990	54.14
Weather_Times	38462	1.23

3.3 Missing values

Table 1 shows the percentage of null records of all the variables.

It was observed that End_Lat and End_long variables had nearly 70.5% of the missing values.

There are missing values in the variables depicting the weather conditions of the accidents, like Temperature, Pressure, Wind speed, Humidity, Precipitation and visibility. These missing values have been replaced with the monthly average value of the respective variables.

3.4 Dropping the Variables

- The dataset describes the US accidents. Thus, the country variable that had only single value as 'US' has been dropped.
- The End_Lat and End_long variables are also dropped considering the percent of the missing values in them.
- The temperature at the time of the accidents is already captured by 'Temperature(F)' variable. The 'Wind_Chill(F)' has been dropped to reduce redundancy in the data set.
- Similarly, the 'Sunrise_Sunset' variable captures all the information of 'Civil_Twilight', 'Nautical_Twilight' and 'Astronomical_Twilight'. The latter variables have been dropped from the dataset for this analysis.

3.5 Duplicates and Outliers

- Some records had nearly erroneous weather condition values that were identified as outliers in the dataset. These records are deleted for the analysis.
- Duplicate records were identified and deleted.

- Some of the records had discrepancies in their start and end time. These records have also been dropped from this analysis.

Table 2: *Average duration of the accident impact on traffic(in minutes)- Source vs Traffic delay categorization*

Source	1	2	3	4
Bing	39.12	80.38	36.69	59.04
MapQuest	46.40	50.44	47.48	104.26
MapQuest-Bing	55.89	60.24	52.92	93.80

4 Data Manipulation

4.1 Sources and their traffic delay level

US accidents data set has accidents reported from three different sources – MapQuest, Bing and MapQuest-Bing. From the graph below, the data set has nearly 77.2% total reported incidents by MapQuest followed by Bing which accounts for 21%. The third category MapQuest-Bing contributes less than 2% of all the records with the US accidents dataset.

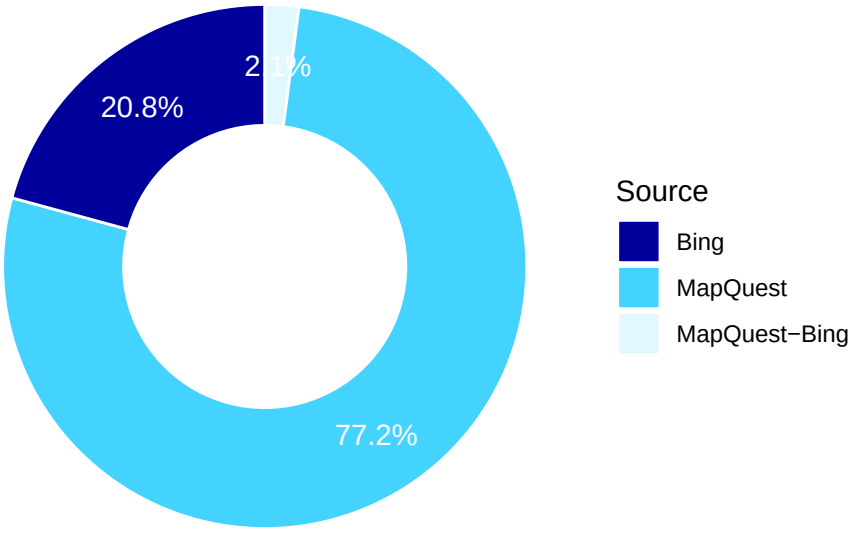


Figure 1: Sources of the accident report

These sources report traffic accidents in four traffic delay levels 1, 2, 3 and 4. Traffic delay 1 indicates the least impact on traffic (i.e. short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e. long delay).

The table 2 shows the discrepancy in traffic delay level categorization for duration of accident impact on traffic (in minutes) reported by different sources. There is a need to reclassify the traffic delay levels before proceeding with the analysis.

4.2 Weather Conditions

US accidents data set has Weather_Condition variable to depicting the conditions at the time of accidents. There are more than 20 categories defined due to the reporting of various sources.

For this analysis, the weather conditions have been redefined as ‘Clear’, ‘Partly Cloudy’, ‘Cloudy’, ‘Rain’, ‘Thunderstorm’, ‘Fog’ ,‘Other’.

5 Exploratory Data Analysis

5.1 State wise Statistics

Among the 49 states in the US, the greatest number of accidents has occurred in the state of California (800K+) followed by Texas (330K+) and Florida(250K+). California, Texas and Florida have nearly 40% of the total number of accidents in the US from February 2016 to June 2020.

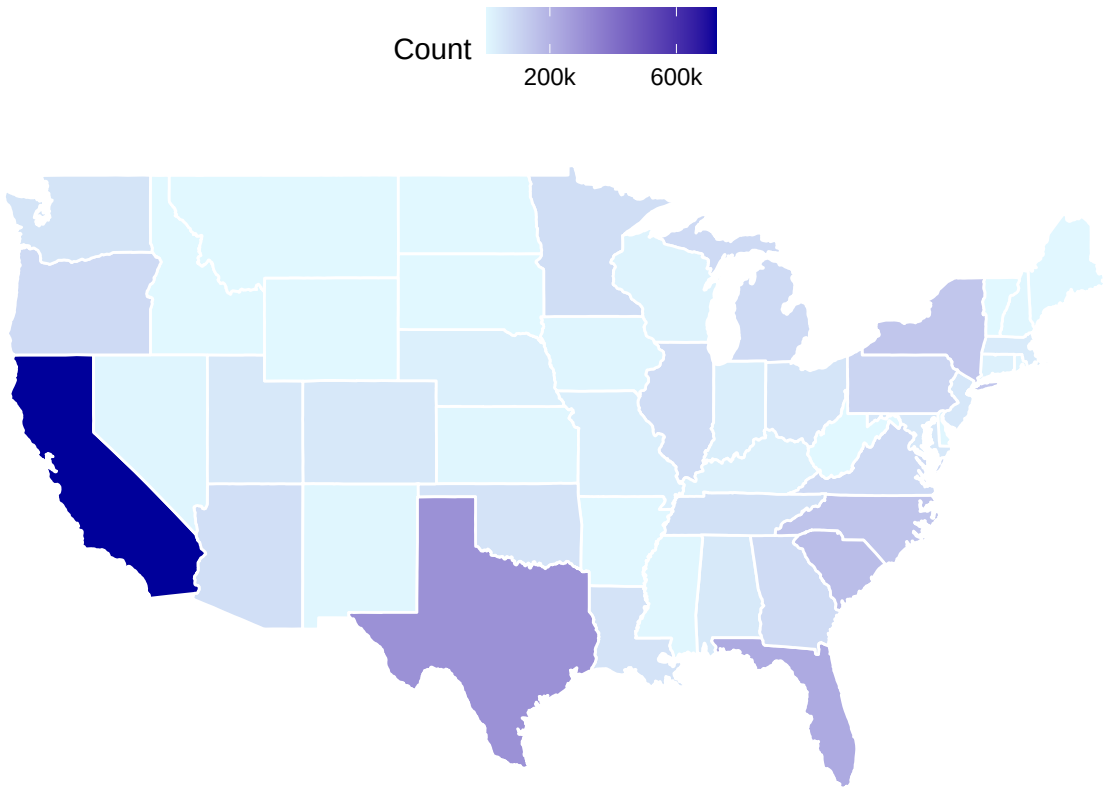


Figure 2: State wide accident statistics

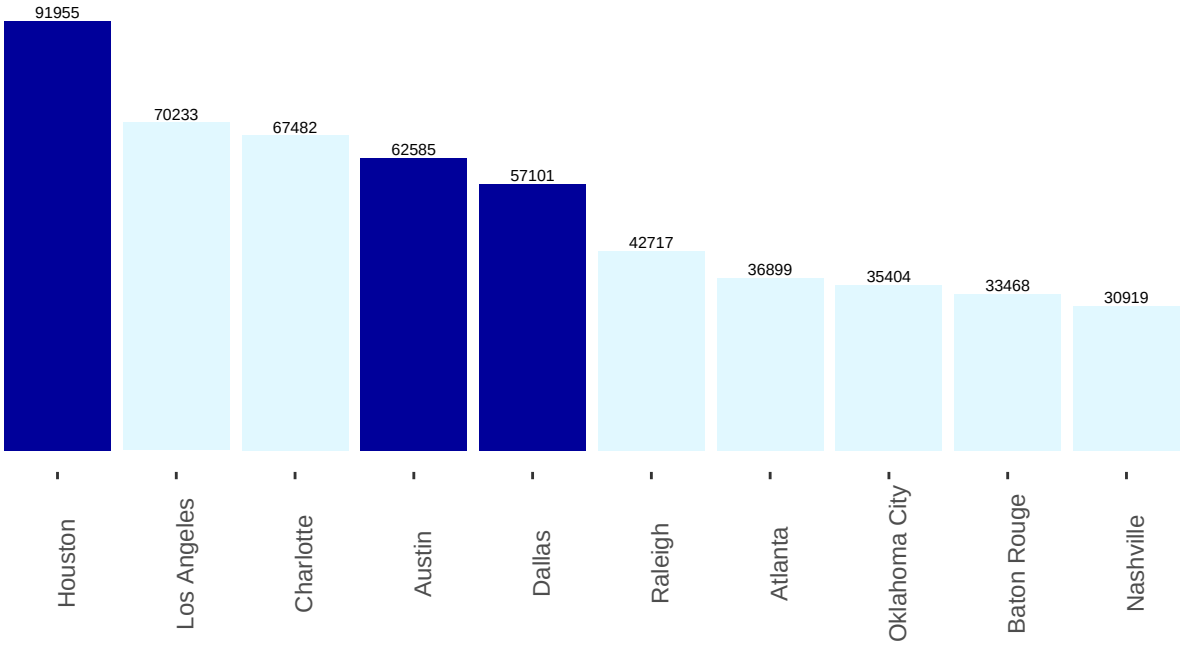


Figure 3: Number of accidents at City level

Houston, Dallas and Austin in Texas state are among top five cities with highest number of accidents records. Thus, further analysis will be focused on Accidents within Texas state from the US Accidents data set.

5.2 Texas Statistics

It is alarming to know that Texas state has not had a day without one traffic accidents since 2000 (KRW Attorneys at Law, n.d.).

Table 3: *Traffic delay level re-classification*

Traffic delay level	Time taken to clear traffic
1	Up to 30 min
2	30 min – 45 min
3	More than 45 min

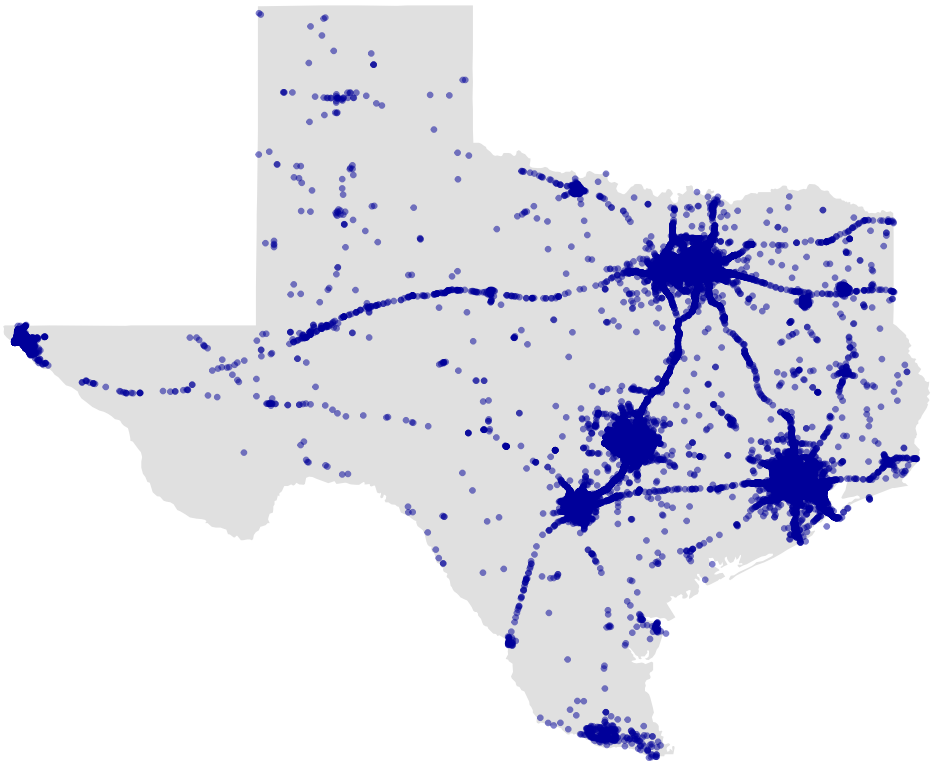


Figure 4: Accidents in Texas

From figure 4 “Accidents in Texas”, it can be inferred that the majority of the accidents in Texas state are along the Interstate Highways and major cities of El Paso, San Antonio, Austin, Houston, and Dallas.

The original data had discrepancy in traffic delay level categorization. For this analysis the traffic delay levels have been re-classified. The user defined variable Traffic_Dur_min was considered as the determining factor. Since the first quartile and the median was 30min, it was ideal to re-categorize into just three levels of severity. The new traffic delay levels were defined based on values of 33rd and 66th percentiles.

Traffic delay level 1 indicates the least impact on traffic post accident and traffic delay level 3 indicates a significant impact on traffic post road accident.

The geographic locations of accidents are overlayed on the texas map to visually analyse the relationship between traffic delay level and their locations.

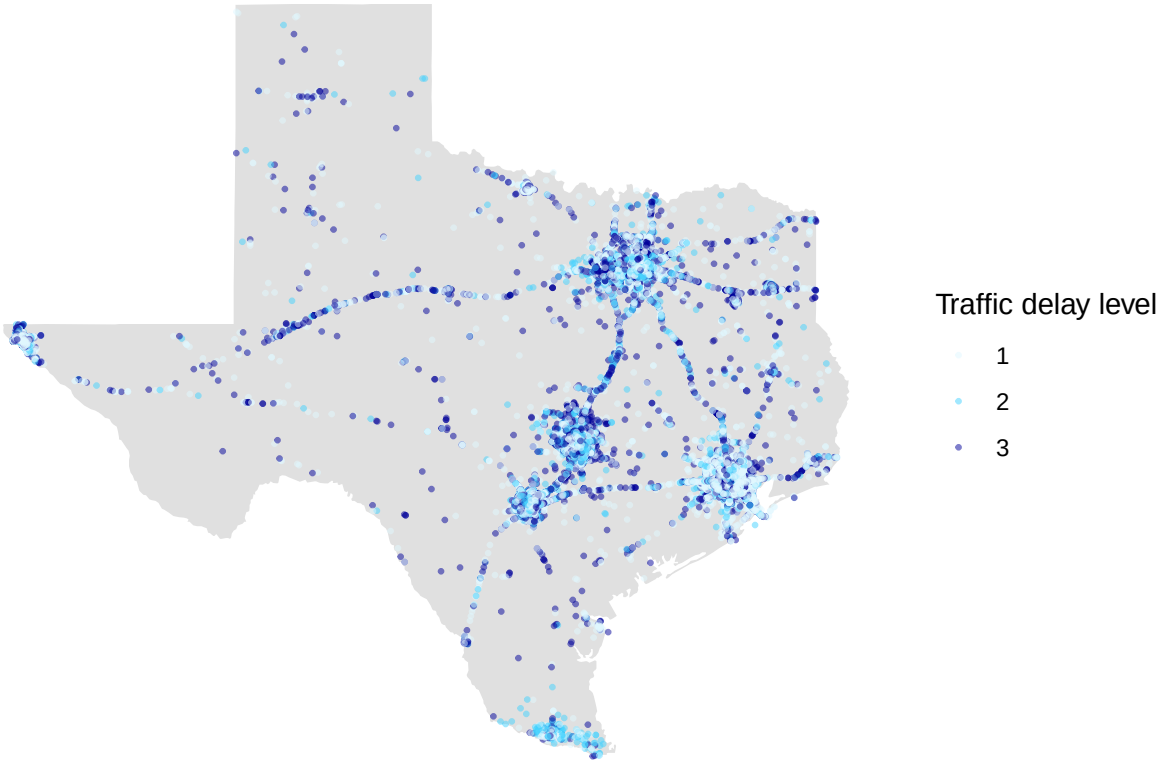


Figure 5: Texas Road Accidents with traffic delay levels

As assumed initially, it can be clearly seen that majority of the traffic delay level 3 accidents (traffic clear time more than 45min) are on the highways and toll roads. This also indicates a correlation between location of the accident and the traffic delay level. As expected, the resources of TMC and EMS would take more time to reach on these highways when compared to a location within city limits.

Frequency of accidents in Texas was plotted against various weather conditions to test the hypothesis that the road accidents are highest in adverse conditions,

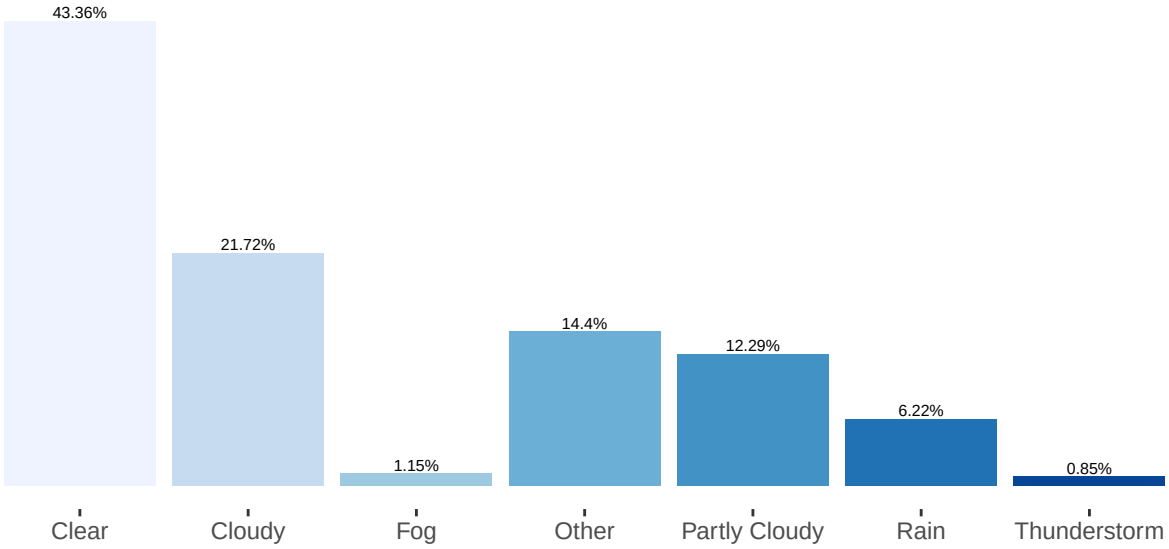


Figure 6: Number of Accidents in various weather conditions

In US, 24% of the total road accidents are weather related. (Pisano, Goodwin, & Rossetti, 2008). This holds true even in case of Texas, where 22.63% of accidents are in adverse weather conditions.

The highest number of accidents have occurred during the clear weather condition. It is surprising to note that 77.37% of the total accidents are in clear, partly cloudy and cloudy weather conditions. This could possibly mean that people avoid driving or drive more carefully during adverse weather conditions like rain, thunderstorm etc.

To test the Null Hypothesis that the traffic delay post the accident is dependent on the weather condition, Pearson's chi-squared test is carried out. The goodness of the fit of the data, as well as the check for dependency of the two categorical variables can be understood.

```
##
##  Pearson's Chi-squared test
##
## data:  Chi_weather.severity
## X-squared = 9508.4, df = 12, p-value < 0.00000000000000022
```

```
##
##                        1          2          3
##   Clear         67624.163 24728.6173 36377.2197
##   Cloudy        33877.221 12388.1287 18223.6504
##   Fog            1797.638   657.3552   967.0073
```

```
##   Other         22454.710  8211.1765 12079.1133
##   Partly Cloudy 19173.575  7011.3400 10314.0848
##   Rain           9694.740  3545.1459  5215.1137
##   Thunderstorm   1326.953   485.2364   713.8107

##
##                     1     2     3
##   Clear         69396 24935 34399
##   Cloudy        28188 11356 24945
##   Fog            1684   501  1237
##   Other         28273  9079  5393
##   Partly Cloudy 17976  7266 11257
##   Rain           9404  3404  5647
##   Thunderstorm   1028   486  1012
```

Since the associated p-value is less than significance level of 5% , we do not reject the null hypothesis. In a study conducted in 2009, the effect of weather conditions on the daily traffic intensities were examined. It tested if all the weather condition impacted in a similar way or not. The conclusions indicated that extreme weathers such as snowfall, rainfall, and increased wind speed had a diminishing impact on the traffic intensity, while increased temperatures on open and sunny days had a positive impact on the traffic intensity measured over a specific segment of the road (Cools, Moons, & Wets, 2009). It can be concluded that the traffic delay level is dependent on the weather conditions.

There is a significant difference between the observed values and expected values across the different traffic delay levels. The contingency table demonstrates the distribution of traffic delay levels across different Weather Conditions. As seen from figure 6, most of the accidents takes place in clear weather conditions.

The distribution of accidents by time period is considered to test the assumption that traffic delay level is high during AM (7:00 - 9:00) and PM (16:00 - 18:00) peak periods within a day.

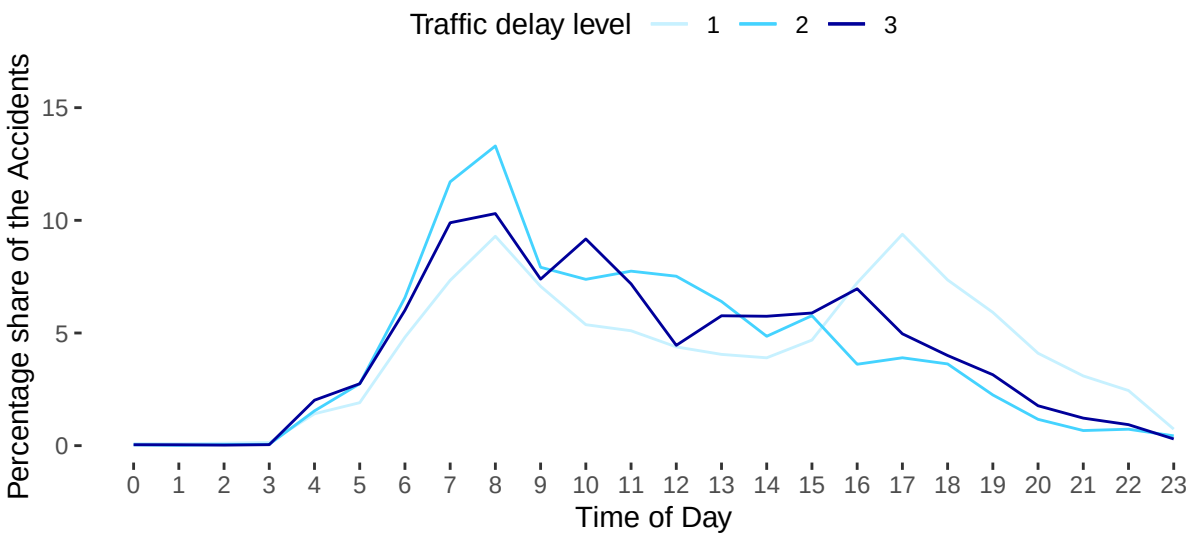


Figure 7: Texas Road Accidents by time of day

The percentage of the accidents is highest for all three traffic delay levels during AM peak period. This can be related to the fact that a greater number of vehicles are on road for either going to workplace or dropping kids to school. Traffic delay level 2 is highest during AM peak which indicates that the traffic post accidents mostly clears up within 30 min to 45 min.

A second peak can be observed in traffic delay level 1 and 3 during PM peak (16:00 - 18:00). The traffic delay level 3 around 16:00 hours might be due to trips taken back from school or work and shopping trips.

In this dataset, 89.6% of accidents have been reported during a weekday and 10.4% of accidents have been reported on weekends. The bar chart below represents the percentage share of accidents for different severity levels on each day of the week.

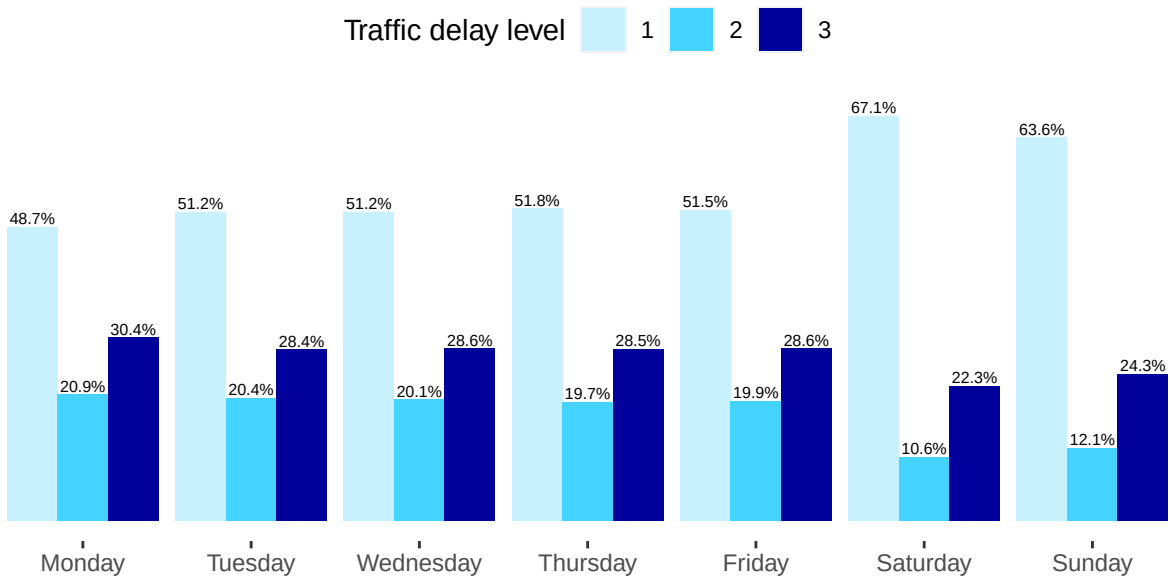


Figure 8: Texas Road Accidents by Weekdays

As seen from Figure 8, percentage share of traffic delay level 1 and 3 are similar across all the weekdays. However, the traffic delay level 1 is higher on weekends compared to weekdays. Unlike traffic delay 1, the traffic delay 2 has less percentage share on weekends when compared to weekdays. The traffic delay 3 is similar across weekdays and weekends.

```
##
##  Pearson's Chi-squared test
##
## data:  Chi_day.severity
## X-squared = 2901.7, df = 12, p-value < 0.00000000000000022
```

```
##
##                     1         2         3
##   Monday    25137.508  9192.215 13522.277
##   Tuesday   28772.708 10521.524 15477.768
##   Wednesday 28729.632 10505.772 15454.596
##   Thursday  27871.262 10191.886 14992.851
##   Friday    28637.701 10472.155 15405.144
##   Saturday  10303.058  3767.594  5542.348
##   Sunday     6497.131  2375.853  3495.016
```

```
##
##                 1     2     3
##   Monday    23298 10011 14543
##   Tuesday   28052 11159 15561
```

```
##    Wednesday 28013 11014 15663
##    Thursday  27481 10434 15141
##    Friday    28077 10829 15609
##    Saturday  13157  2085  4371
##    Sunday     7871  1495  3002
```

From the contingency table it is observed that the number of accidents happen more on the weekdays than on the weekends. This is possible as the number of vehicles on roads is higher on weekdays compared to weekends.

Further, the percentage share of delay levels on an average weekday and weekend was analyzed to test the hypothesis that the delay level 3 is different on a typical weekday compared to typical weekend.

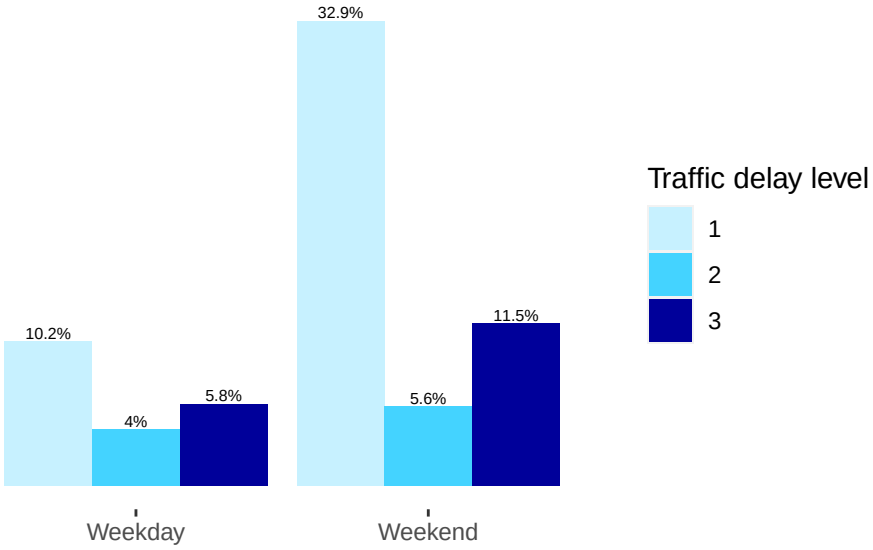


Figure 9: Texas Road Accidents on Weekday v/s weekend

The proportion of accidents with traffic delay 3 (more than 45min to clear traffic post-accident) is higher on a typical weekend when compared to weekday. Prop test is also conducted to know if proportion of number of traffic delay 3 accidents on weekdays and weekends are statistically differing.

```
##
##      Weekday Weekend
##    1  134921   21028
##    2   53447    3580
##    3   76517    7373
```

```
##
##        Weekday    Weekend
##   1 0.45448451 0.07083331
##   2 0.18003746 0.01205931
##   3 0.25774929 0.02483612
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  [ out of (TableDelayWeekend out of TableDelayWeekend[3, 1] + TableDelayWeekend
## X-squared = 56990, df = 1, p-value < 0.00000000000000022
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.08599177 0.08982377
## sample estimates:
##         p
## 0.0878889
```

The p - value is less than the level of significance 0.05. Therefore, we reject the hypothesis that the proportion of accidents with traffic delay level 3 is equal on weekdays and on weekends.

The test of proportions as used as the test statistic, to understand if the observed proportion is different from our initial expected proportion.

6 Empirical Analysis

This project is aimed to facilitate City's Transportation Management Center (TMC) to predict traffic delay level (categorical data) post road accidents and effectively deploy their resources.

The summary of traffic delay levels showed that nearly 52% of the total observation belonged to traffic delay level 1. To ensures that each delay level within the total accident population receives proper representation within the sample, for this analysis, the observations are classified as train data and test data using a stratified random sample method. The training data consists of 80% of the observations randomly selected from each delay level. The rest 20% of the observations are considered as test and validation data.

6.1 Stepwise regression model

A conscious effort has been made to build a model with the input variables, whose data will be available at the moment an accident has been reported. The variables manual chosen are "Start_Lat", "Start_Lng", "ST", "Bump", "Crossing", "Give_Way", "Junction", "No_Exit", "Railway", "Roundabout", "Station", "Stop", "Traffic_Calming", "Traffic_Signal", "Sunrise_Sunset", "Day" , "Weather_New", "Precipitation", "Temperature", "WindSpeed", "Humidity" , "Pressure" , "STHR" (time of day) , "Weekend".

After running the stepwise regression, the model with the highest adjusted R squared was selected as the best model. The detailed results of the stepwise regression can be referred to in the appendix 2. The model 26 has the highest adjusted r square value(0.1746969). The predictor variables used in this model are Start_Lat, Start_Lng, ST, Crossing, Give_Away, Junction, Railway, Roundabout, Station, Stop, Traffic_Signal, Sunrise_Sunset, Precipitation, Temperature, Windspeed, Humidity, Pressure, STHR, and Weekend.

6.2 Ordinal Multiclass Classification

6.2.1 Ordinal logistic regression

Since the traffic delay level is ordinal data, Ordinal logistic regression was the first choice to predict an ordinal dependent variable given one or more independent variables. The variables of the best stepwise regression model were used as the predictors for the response variable in the ordinal logistic model.

```
## Call:
## polr(formula = Delay_new ~ Start_Lat + Start_Lng + Crossing +
##     Give_Way + Junction + Railway + Roundabout + Station + Stop +
##     Traffic_Signal + Sunrise_Sunset + Precipitation + Temperature +
##     WindSpeed + Humidity + Pressure + Weekend, data = train.data,
##     Hess = T)
##
## Coefficients:
##                          Value Std. Error   t value
## Start_Lat          -0.0081288 0.00302590    -2.686
```

```
## Start_Lng            0.0899306 0.00193366    46.508
## Crossing1            0.0406533 0.01656992     2.453
## Give_Way1           -0.1987373 0.05245878    -3.788
## Junction1           -0.5317516 0.02287767   -23.243
## Railway1            -0.1045626 0.04337202    -2.411
## Roundabout1          1.4633808 0.00004657 31426.193
## Station1             0.2522004 0.03101153     8.132
## Stop1               -0.1246336 0.03031751    -4.111
## Traffic_Signal1      0.0405278 0.00951026     4.261
## Sunrise_SunsetNight -0.5158101 0.01057776   -48.764
## Precipitation       -2.4238798 0.13198114   -18.365
## Temperature         -0.0095715 0.00027260   -35.112
## WindSpeed           -0.0245361 0.00095433   -25.710
## Humidity             0.0008483 0.00020018     4.238
## Pressure            -1.4001418 0.00593684  -235.840
## WeekendWeekend      -0.5084784 0.01416845   -35.888
##
## Intercepts:
##     Value       Std. Error  t value
## 1|2    -51.7294      0.0039 -13142.4073
## 2|3    -50.8138      0.0055  -9275.2207
##
## Residual Deviance: 454765.55
## AIC: 454803.55
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2     3
##          1 28949  9936 10632
##          2     0     0     0
##          3  2241  1469  6146
##
## Overall Statistics
##
##                Accuracy : 0.5911
##                  95% CI : (0.5871, 0.5951)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.206
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
## Statistics by Class:
```

Table 4: **Number of Observations before and after SMOTE**

Delay.Levels	Before.Over.Sampling	After.Over.Sampling
1	124759	124759
2	45622	119426
3	67112	148270

```
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.9282   0.0000   0.3663
## Specificity            0.2702   1.0000   0.9129
## Pos Pred Value         0.5846      NaN   0.6236
## Neg Pred Value         0.7726   0.8079   0.7853
## Prevalence             0.5253   0.1921   0.2826
## Detection Rate         0.4876   0.0000   0.1035
## Detection Prevalence   0.8340   0.0000   0.1660
## Balanced Accuracy      0.5992   0.5000   0.6396
```

The accuracy of this ordinal logistic model is approximately 59%. But the model was not able to predict class 2 i.e. Delay level 2 records. In the main Texas dataset, Traffic delay level 1 is the dominant class with more than 52% of records. Hence, one of the possible reasons for this model to ignore the delay level 2 prediction was the lack of traffic delay level 2 observations in the training dataset.

6.2.2 Ordinal logistic regression with SMOTE

To even out the number of records across all three traffic delay levels, oversampling for traffic delay level 2 and 3 observations were carried out using the SMOTE technique for the training data. The SMOTE technique can be applied to only two classes. For this project, this technique was applied to the training dataset of traffic delay levels 2 and 3 only.

```
##   Var1  Freq
## 1    2 34216
## 2    3 50334
```

```
##   Var1  Freq
## 1    2 11406
## 2    3 16778
```

```
##      1      3      2
## 124759 119426 148270
```

After performing SMOTE, the number of observations for traffic delay level 2 and 3 has increased as seen from table 4. The test/validation dataset is kept untainted.

```
##      1      3      2
## 124759 119426 148270
```

```
## Call:
## polr(formula = Delay_new ~ Start_Lat + Start_Lng + Crossing +
##     Give_Way + Junction + Railway + Roundabout + Station + Stop +
##     Traffic_Signal + Sunrise_Sunset + Precipitation + Temperature +
##     WindSpeed + Humidity + Pressure + Weekend, data = train.data,
##     Hess = T)
##
## Coefficients:
##                           Value Std. Error   t value
## Start_Lat           -0.0081288 0.00302590    -2.686
## Start_Lng            0.0899306 0.00193366    46.508
## Crossing1            0.0406533 0.01656992     2.453
## Give_Way1           -0.1987373 0.05245878    -3.788
## Junction1           -0.5317516 0.02287767   -23.243
## Railway1            -0.1045626 0.04337202    -2.411
## Roundabout1          1.4633808 0.00004657 31426.193
## Station1             0.2522004 0.03101153     8.132
## Stop1               -0.1246336 0.03031751    -4.111
## Traffic_Signal1      0.0405278 0.00951026     4.261
## Sunrise_SunsetNight -0.5158101 0.01057776   -48.764
## Precipitation       -2.4238798 0.13198114   -18.365
## Temperature         -0.0095715 0.00027260   -35.112
## WindSpeed           -0.0245361 0.00095433   -25.710
## Humidity             0.0008483 0.00020018     4.238
## Pressure            -1.4001418 0.00593684  -235.840
## WeekendWeekend      -0.5084784 0.01416845   -35.888
##
## Intercepts:
##     Value       Std. Error  t value
## 1|2    -51.7294      0.0039 -13142.4073
## 2|3    -50.8138      0.0055  -9275.2207
##
## Residual Deviance: 454765.55
## AIC: 454803.55
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2     3
```

```
##          1 22091  8167 11287
##          2  5520  2042  3584
##          3  3579  1196  1907
##
## Overall Statistics
##
##                Accuracy : 0.4386
##                  95% CI : (0.4346, 0.4426)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0056
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.7083  0.17904  0.11366
## Specificity            0.3097  0.81021  0.88790
## Pos Pred Value         0.5317  0.18320  0.28539
## Neg Pred Value         0.4896  0.80586  0.71777
## Prevalence             0.5253  0.19209  0.28259
## Detection Rate         0.3721  0.03439  0.03212
## Detection Prevalence   0.6997  0.18773  0.11254
## Balanced Accuracy      0.5090  0.49463  0.50078
```

Even with oversampling and even distribution of observations across categories, the overall performance of the model was not improved significantly. When compared to the previous ordinal logistic model, there is an improvement in predicting delay level 2 observations. The overall accuracy has plummeted to 39%, which is not acceptable.

Since Traffic_Dur_min was the key variable to recategorize the traffic delay levels, the distribution of Traffic_Dur_min was closely looked at.

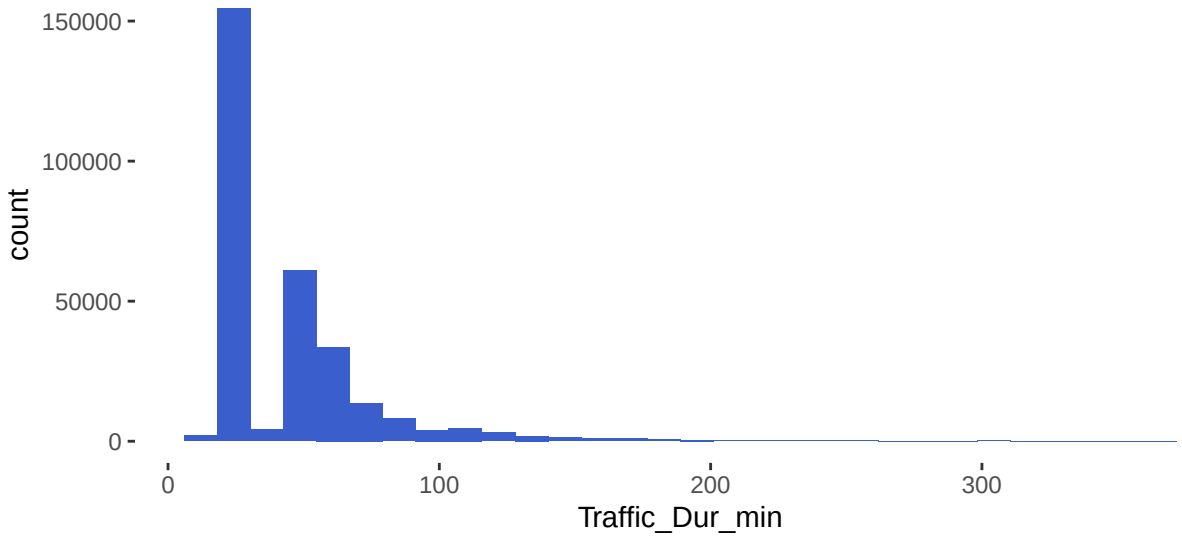


Figure 10: Distribution of traffic duration in minutes

Since the histogram is bi-modal, it was realized that a binary classification model might be a better fit for this dataset.

6.3 Binary classificication

The Traffic_Dur_min variable has a median of 30min and the mean around 45min. Hence, two scenarios were considered for the preliminary analysis. In the first scenario, accidents with Traffic_Dur_min less than 30min were classified as traffic delay level 1 and more than 30min as traffic delay level 2. In the second scenario, a 45min split was applied to reclassify records as traffic delay levels 1 and 2.

After the accuracy, sensitivity and the specificity obtained by the binary logistic model of 45min split was better when compared to 30min split of delay level classification(Summary of 30 min split model can be referred in the appendix 4). Therefore, 45min split re-classification was considered for this analysis.

The following algorithms have been considered for the binary classification:

- Binary logistic regression
- Discriminant Analysis
- Decision Tree
- K-Nearest Neighbors
- Random Forest

6.3.1 Binary logistic regression

The binary logistic regression is the most basic algorithm for a classification problem. For all the following algorithms, 45min traffic delay level classification is considered. This model is also used to test the hypothesis that Temperature, Humidity, and Precipitation are significant predictors of traffic delay level.

```
##
## Call:
## glm(formula = Delay_new ~ Start_Lat + Start_Lng + Crossing +
##     Give_Way + Junction + Railway + Roundabout + Station + Stop +
##     Traffic_Signal + Sunrise_Sunset + Weather_New + Precipitation +
##     Temperature + WindSpeed + Humidity + Pressure + STHR + Weekend,
##     family = "binomial", data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4511  -0.7911  -0.5960   1.0640   4.3548
##
## Coefficients:
##                            Estimate Std. Error  z value             Pr(>|z|)
## (Intercept)              68.8821992  0.7008744   98.280 < 0.0000000000000002
## Start_Lat                -0.1180196  0.0039048  -30.224 < 0.0000000000000002
```

```
## Start_Lng                 0.1865630  0.0038312   48.696 < 0.0000000000000002
## Crossing1                -0.0064030  0.0198685   -0.322                0.7472
## Give_Way1                -0.1137574  0.0605268   -1.879                0.0602
## Junction1                -0.5886129  0.0290372  -20.271 < 0.0000000000000002
## Railway1                  0.0037362  0.0496154    0.075                0.9400
## Roundabout1               1.8360486  0.7670706    2.394                0.0167
## Station1                  0.3351978  0.0365413    9.173 < 0.0000000000000002
## Stop1                    -0.0608314  0.0349236   -1.742                0.0815
## Traffic_Signal1           0.0667622  0.0114839    5.814         0.00000000612
## Sunrise_SunsetNight      -0.2690419  0.0132933  -20.239 < 0.0000000000000002
## Weather_NewCloudy         0.4471681  0.0130293   34.320 < 0.0000000000000002
## Weather_NewFog            0.2232719  0.0436108    5.120         0.00000030610
## Weather_NewOther         -0.6149452  0.0197411  -31.150 < 0.0000000000000002
## Weather_NewPartly Cloudy  0.2643968  0.0156525   16.892 < 0.0000000000000002
## Weather_NewRain           0.4372943  0.0233647   18.716 < 0.0000000000000002
## Weather_NewThunderstorm   1.0827066  0.0535799   20.207 < 0.0000000000000002
## Precipitation            -4.9456180  0.2229013  -22.187 < 0.0000000000000002
## Temperature              -0.0098901  0.0003788  -26.110 < 0.0000000000000002
## WindSpeed                -0.0259236  0.0011739  -22.083 < 0.0000000000000002
## Humidity                 -0.0056164  0.0003205  -17.526 < 0.0000000000000002
## Pressure                 -1.5506837  0.0140847 -110.097 < 0.0000000000000002
## STHR                     -0.0469886  0.0012566  -37.393 < 0.0000000000000002
## WeekendWeekend           -0.1671543  0.0169859   -9.841 < 0.0000000000000002
##
## (Intercept)              ***
## Start_Lat                ***
## Start_Lng                ***
## Crossing1
## Give_Way1                .
## Junction1                ***
## Railway1
## Roundabout1              *
## Station1                 ***
## Stop1                    .
## Traffic_Signal1          ***
## Sunrise_SunsetNight      ***
## Weather_NewCloudy        ***
## Weather_NewFog           ***
## Weather_NewOther         ***
## Weather_NewPartly Cloudy ***
## Weather_NewRain          ***
## Weather_NewThunderstorm  ***
## Precipitation            ***
## Temperature              ***
## WindSpeed                ***
```

```
## Humidity                 ***
## Pressure                 ***
## STHR                     ***
## WeekendWeekend           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 282796  on 237492  degrees of freedom
## Residual deviance: 250760  on 237468  degrees of freedom
## AIC: 250810
##
## Number of Fisher Scoring iterations: 5
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 40670 13105
##          2  1925  3673
##
##                Accuracy : 0.7469
##                  95% CI : (0.7433, 0.7503)
##     No Information Rate : 0.7174
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.2177
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.21892
##             Specificity : 0.95481
##          Pos Pred Value : 0.65613
##          Neg Pred Value : 0.75630
##              Prevalence : 0.28259
##          Detection Rate : 0.06186
##    Detection Prevalence : 0.09429
##       Balanced Accuracy : 0.58686
##
##        'Positive' Class : 2
##
```

The associated p-value for Temperature, Humidity, and Precipitation is less than the significance level of

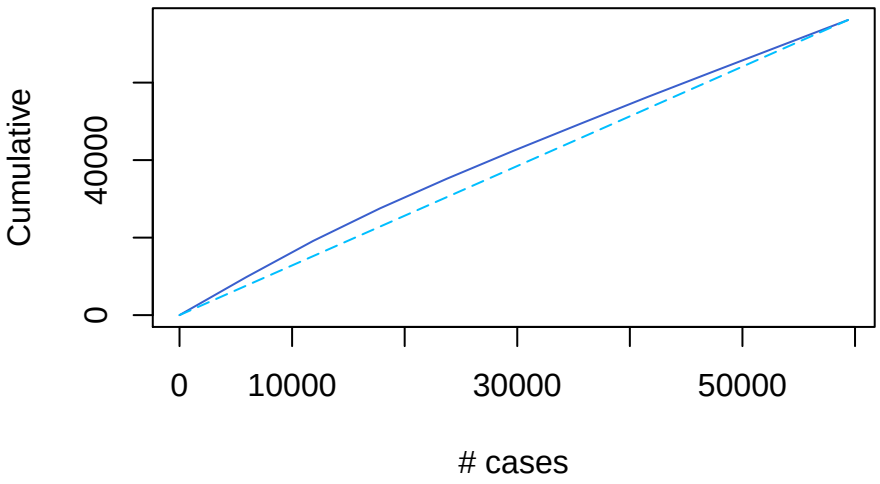
0.05. Thus, we reject the null hypothesis that these variables have no effect on Traffic delay level. There is enough evidence to conclude that with 95% confidence, Temperature, Humidity, and Precipitation are significant predictors of traffic delay level.

As seen in the exploratory data analysis, weekend(weekend and weekday), weather, STHR (Time of day), and Start_Lat & Start_Lng (denoting the location of the accident) variables are also significant predictors of delay level.

The traffic delay level 2 (impact to clear traffic more than 45min) is considered as the class of interest. Therefore, various probability threshold levels were considered for delay level 2 (refer to appendix 5). After considering sensitivity and specificity, along with accuracy, a probability threshold of 0.25 was considered.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 25526  4012
##          2 17069 12766
##
##                Accuracy : 0.6449
##                  95% CI : (0.6411, 0.6488)
##     No Information Rate : 0.7174
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2914
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.7609
##             Specificity : 0.5993
##          Pos Pred Value : 0.4279
##          Neg Pred Value : 0.8642
##              Prevalence : 0.2826
##          Detection Rate : 0.2150
##    Detection Prevalence : 0.5025
##       Balanced Accuracy : 0.6801
##
##        'Positive' Class : 2
##
```

After applying a threshold of 0.25, as seen from the confusion matrix, more number of observations are classified as delay level 2 correctly. There is also an increase in the Sensitivity from 21%(threshold of 0.5) to 76% and a decrease in Specificity from 95% to 60%.



The lift chart shows that the model has gained a lift compared to the base model. In the first 20,000 cases, the model can cumulatively identify nearly 6,000 more traffic delay level 2 cases compared to benchmark model.

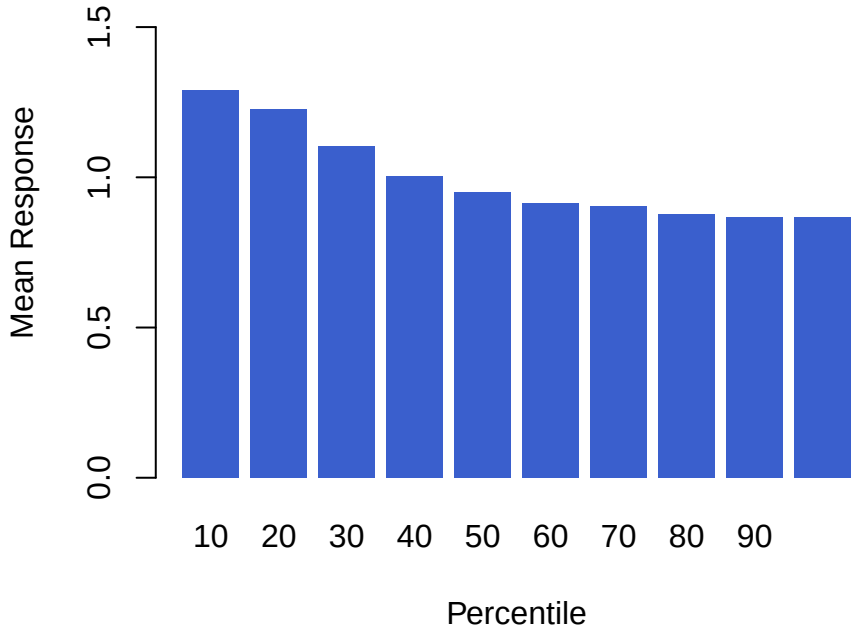


Figure 11: Decile - Lift chart with the logistic model

The model is likely to identify a delay level 2 accidents in the first 3 deciles (30%) of the decile chart, when the observations are ranked by their propensities. Model can perform nearly 1.22X times better on average in the top three deciles than the random classification model. Since the decile chart is exhibiting better staircase decile, it can be understood that the model is performing it's best standard.

6.3.2 Discriminant Analysis

The discriminant analysis was considered as it is a model-based approach to classification. The output of a discriminant analysis procedure generates estimated “classification functions,” which are translated into classifications or propensities.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 40998 13692
##          2  1597  3086
##
##                Accuracy : 0.7425
##                  95% CI : (0.739, 0.746)
##     No Information Rate : 0.7174
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.1874
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.18393
##             Specificity : 0.96251
##          Pos Pred Value : 0.65898
##          Neg Pred Value : 0.74964
##              Prevalence : 0.28259
##          Detection Rate : 0.05198
##    Detection Prevalence : 0.07887
##       Balanced Accuracy : 0.57322
##
##        'Positive' Class : 2
##
```

Sensitivity is only 0.18393 and Specificity is 0.96251 with an Accuracy of 74.25% considering the class of interest as ‘delay level 2’. This cutoff threshold needs to be reduced to increase sensitivity.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 26944  4402
```

```
##          2 15651 12376
##
##                Accuracy : 0.6623
##                  95% CI : (0.6584, 0.6661)
##     No Information Rate : 0.7174
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3077
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.7376
##             Specificity : 0.6326
##          Pos Pred Value : 0.4416
##          Neg Pred Value : 0.8596
##              Prevalence : 0.2826
##          Detection Rate : 0.2084
##    Detection Prevalence : 0.4720
##       Balanced Accuracy : 0.6851
##
##        'Positive' Class : 2
##
```

After applying a threshold of 0.25, there is also an increase in the Sensitivity from 18% to 73%. The accuracy of this model is 66%.

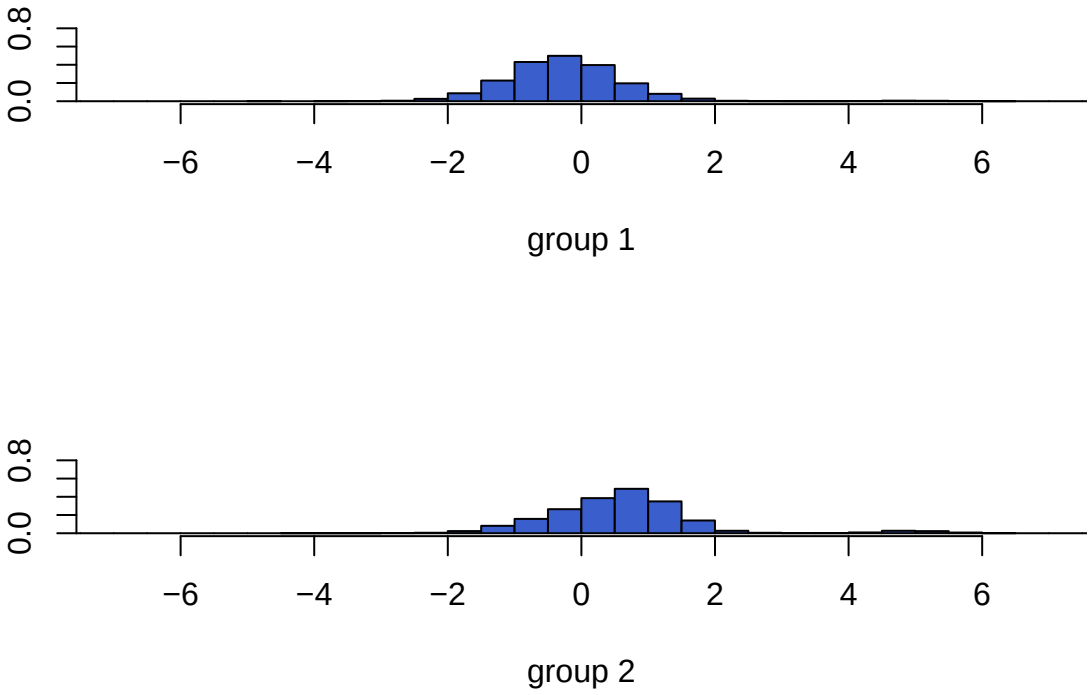


Figure 12: LDA plots for train data

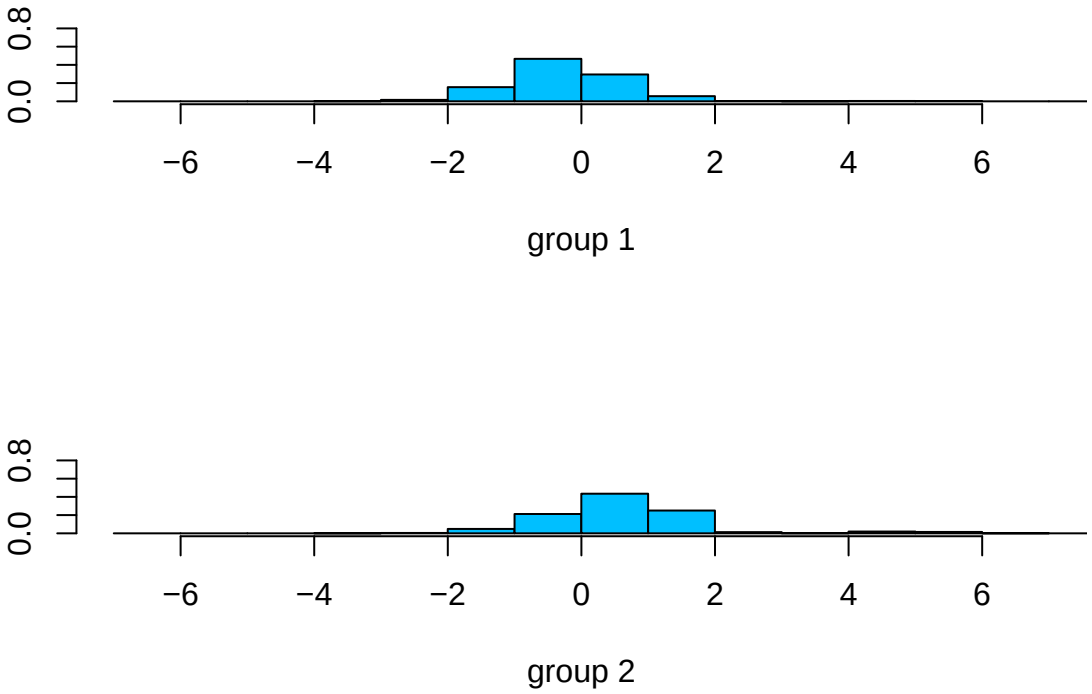


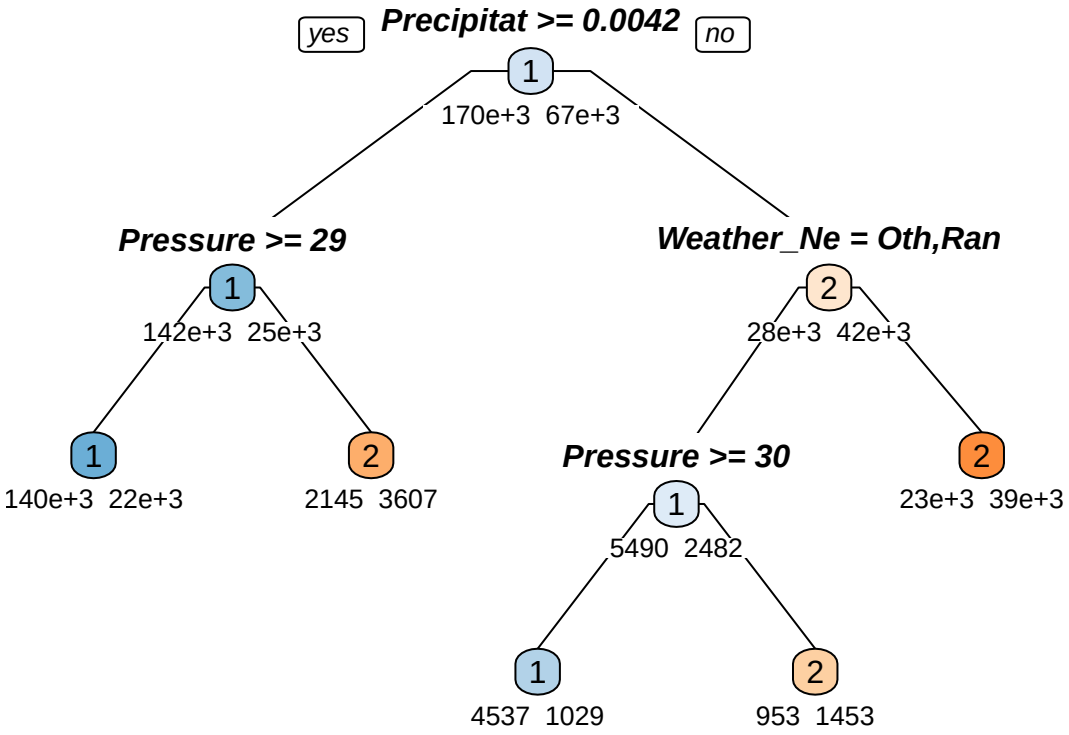
Figure 13: LDA plots for test data

Precipitation, Roundabout, and pressure LD scores(refer appendix 6) have the highest impact on the separation line.

In both the LDA plots of training and validation data, the score of less than 0 are mostly classified as traffic delay level 1, and score greater than 0 are mostly classified as traffic delay level 2. Though there are some misclassifications, scores hold true for the majority of the observations. The LDA has created separation among the two classes.

6.3.3 Decision Tree

Decision tree is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. Decision tree works on both classification and regression problems and also works on both categorical and continuous input output variable.




```
##  Delay_new
##       0.13 when Precipitation >= 0.0042 & Pressure >= 29
##       0.18 when Precipitation <  0.0042 & Pressure >= 30 & Weather_New is
##       0.60 when Precipitation <  0.0042 & Pressure <  30 & Weather_New is
##       0.63 when Precipitation >= 0.0042 & Pressure <  29
##       0.63 when Precipitation <  0.0042                  & Weather_New is Clear or Cl
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 36111  5566
##          2  6484 11212
##
##                Accuracy : 0.797
##                  95% CI : (0.7938, 0.8003)
##     No Information Rate : 0.7174
##     P-Value [Acc > NIR] : < 0.00000000000000022
```

```
##
##                   Kappa : 0.5076
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.6683
##             Specificity : 0.8478
##          Pos Pred Value : 0.6336
##          Neg Pred Value : 0.8664
##              Prevalence : 0.2826
##          Detection Rate : 0.1888
##    Detection Prevalence : 0.2980
##       Balanced Accuracy : 0.7580
##
##        'Positive' Class : 2
##
```

Decision tree is performing better compared to the binary logistic regression model used above. This model has accuracy of 0.72 and sensitivity and specificity of 0.79 and 0.65 respectively.

6.3.4 K-Nearest Neighbours (KNN)

This supervised machine learning algorithm was considered to classify and predict traffic delay levels. The method relies on finding “similar” records in the training data. These “neighbors” are then used to derive a classification or prediction for the new record.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 41179  4661
##          2  1416 12117
##
##                Accuracy : 0.8976
##                  95% CI : (0.8952, 0.9001)
##     No Information Rate : 0.7174
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.7318
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.7222
##             Specificity : 0.9668
##          Pos Pred Value : 0.8954
##          Neg Pred Value : 0.8983
##              Prevalence : 0.2826
##          Detection Rate : 0.2041
##    Detection Prevalence : 0.2279
##       Balanced Accuracy : 0.8445
##
##        'Positive' Class : 2
##
```

The KNN model has a sensitivity of 72% and an accuracy of 90%. Since the training dataset is large, and each class is characterized by multiple combinations of predictor values, this KNN model has performed well. The time required to run this algorithm might be more when compared to other models as it computes distances from the entire set of training records only at the time of prediction.

6.3.5 Random Forest

In the random forest approach, a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. In classification, it uses the majority vote of the outcome variable and in the regression model, it considers the average of the output variable.

The Random Forest model has a sensitivity of 69% and an accuracy of 83%. The run time for this model is more as it decorrelates several trees which are generated on the different bootstrapped samples from the training Data and then averages the trees and reduces the Variance and avoids Over fit in Trees.

Table 5: *Summary of all the Algorithms*

Classifier	Sensitivity	Specificity	Accuracy
Binary logistic regression	0.76	0.60	64.48%
Discriminant Analysis	0.74	0.63	66.26%
Decision Tree	0.79	0.65	71.79%
K-Nearest Neighbors	0.96	0.74	89.78%
Random Forest	0.69	0.89	83.24%

7 Conclusions

It is more important to correctly identify the accidents that have a tremendous impact on traffic time. With delay level 2 (high impact on traffic) as the class of interest, Sensitivity, specificity, and accuracy of all the models were compared. The results are summarized in table 5.

KNN is considered as the **best model** with 72% sensitivity, 96% specificity, and 89.7% accuracy. The variables considered are Start_Lat, Start_Lng, Crossing, Give_Way, Junction, Railway, Roundabout, Station, Stop, Traffic_Signal', Day_Night, Precipitation, Pressure, Temperature, WindSpeed, Humidity, Weekend. A conscious effort has been made in building a model with the input variables, whose data will be available when an accident is reported at TMC.

If two accidents are reported at a given point in time, Transportation Management Center (TMC) can allot its resources based on the traffic delay levels. Thus, there is a need to identify both delay level 1 and delay level 2 accidents correctly i.e; a model with both high sensitivity and specificity along with accuracy. In case of high delay level, Emergency Medical Services can take a call on the requirement of airlift instead of an ambulance by road to save a life.

Different versions of this model can be created for each city level TMC for better prediction. This type of prediction model can help reduce traffic delay time and in turn, reducing the congestion costs, including travel delay and excess fuel consumption.

8 References

Blincoe, L., Miller, T. R., Zaloshnja, E., & Lawrence, B. A. (2015). *The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised)*. Washington, DC: National Highway Traffic Safety Administration.

Cools, M., Moons, E., & Wets, G. (2009). Assessing the Impact of Weather on Traffic Intensity. *Weather, Climate and Society*, 2, 60-68.

Head, A. J. (1959). Predicting traffic accidents from roadway elements on urban extensions of state highways. *Highway Research Board Bulletin, 208*, 45-63. National Research Council (USA), Highway Research Board.

KRW Attorneys at Law. (n.d.). *Texas Car Accident Statistics*. Retrieved 11 03, 2020, from KRW Attorneys at Law: https://www.krwlawyers.com/san-antonio-personal-injury-lawyer/automotive-accidents/car-accidents/texas-car-accident-statistics/

Martin, J.-L. (2002). Relationship between crash rate and hourly traffic flow on interurban motorways. *Accident Analysis & Prevention, 34(5)*, 619-629.

National Center for Statistics and Analysis. (2020, May). Early Estimate of Motor Vehicle Traffic Fatalities in 2019. *Traffic Safety Facts(DOT HS 812 946)*. National Highway Traffic Safety Administration.

Pisano, P. A., Goodwin, L. C., & Rossetti, M. A. (2008). U.S. highway crashes in adverse road weather conditions. 24th Conference on IIPS.

Retallack, A. E., & Ostendorf, B. (2019, September). Current Understanding of the Effects of Congestion on Traffic Accidents. *Int J Environ Res Public Health*.

U.S. Department of Transportation. (2017, February 1). *Regional Concept for Transportation Operations*. Retrieved September 07, 2020, from Federal Highway Administration: https://ops.fhwa.dot.gov/publications/rctoprimer/prim0701.htm

Wagner, I. (2020, July 24). *Road accidents in the United States - Statistics & Facts*. Retrieved September 07, 2020, from Statista: https://www.statista.com/topics/3708/road-accidents-in-the-us/

Wallace, C. E. (n.d.). *ITS Professional Capacity Building Program*. Retrieved September 7, 2020, from United States Department of Transportation: https://www.pcb.its.dot.gov/eprimer/module4.aspx#fn4

Woo, J. C. (1957). Correlation of Accident Rates and Roadway Factors. Purdue University.

Zhou, M., & Sisiopiku, V. P. (1997). Relationship between volume-to-capacity ratios and accident rates. *Transportation research record*, 1581, 47-52.

Dataset Acknowledgments:

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. *"A Countrywide Traffic Accident Dataset."*, arXiv preprint arXiv:1906.05409 (2019).

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. *"Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights."* In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019

9 Appendices

9.1 *Appendix 1: Description of variables in US Accidents data set*

X.	Attribute	Description
1	ID	This is a unique identifier of the accident record.
2	Source	Indicates source of the accident report (i.e. the API which reported the accident.).
3	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more de
4	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least i
5	Start_Time	Shows start time of the accident in local time zone.
6	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact
7	Start_Lat	Shows latitude in GPS coordinate of the start point.
8	Start_Lng	Shows longitude in GPS coordinate of the start point.
9	End_Lat	Shows latitude in GPS coordinate of the end point.
10	End_Lng	Shows longitude in GPS coordinate of the end point.
11	Distance(mi)	The length of the road extent affected by the accident.
12	Description	Shows natural language description of the accident.
13	Number	Shows the street number in address field.
14	Street	Shows the street name in address field.
15	Side	Shows the relative side of the street (Right/Left) in address field.
16	City	Shows the city in address field.
17	County	Shows the county in address field.
18	State	Shows the state in address field.
19	Zipcode	Shows the zip code in address field.
20	Country	Shows the country in address field.
21	Timezone	Shows time zone based on the location of the accident
22	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the acciden
23	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).
24	Temperature(F)	Shows the temperature (in Fahrenheit).
25	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
26	Humidity(%)	Shows the humidity (in percentage).
27	Pressure(in)	Shows the air pressure (in inches).
28	Visibility(mi)	Shows visibility (in miles).
29	Wind_Direction	Shows wind direction.
30	Wind_Speed(mph)	Shows wind speed (in miles per hour).
31	Precipitation(in)	Shows precipitation amount in inches, if there is any.
32	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
33	Amenity	A POI annotation which indicates presence of amenity in a nearby location.
34	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.
35	Crossing	A POI annotation which indicates presence of crossing in a nearby location.
36	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.
37	Junction	A POI annotation which indicates presence of junction in a nearby location.
38	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.
39	Railway	A POI annotation which indicates presence of railway in a nearby location.
40	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.
41	Station	A POI annotation which indicates presence of station in a nearby location.
42	Stop	A POI annotation which indicates presence of stop in a nearby location.
43	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.
44	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.

9.2 *Appendix 2: Summary of Stepwise regression for variable selection*

```
##              adjr2 V1 Start_Lat Start_Lng ST Bump1 Crossing1 Give_Way1
## 1  ( 1 )  0.1006401                       *
## 2  ( 1 )  0.1244059                       *
## 3  ( 1 )  0.1398449                       *
## 4  ( 1 )  0.1470183                       *
## 5  ( 1 )  0.1512102                       *
## 6  ( 1 )  0.1544673                       *
## 7  ( 1 )  0.1572497                     *  *
## 8  ( 1 )  0.1588682                     *  *
## 9  ( 1 )  0.1599917                     *  *
## 10  ( 1 ) 0.1609213                     *  *
## 11  ( 1 ) 0.1616951                     *  *
## 12  ( 1 ) 0.1622497                     *  *
## 13  ( 1 ) 0.1627514                     *  *
## 14  ( 1 ) 0.1632575                     *  *
## 15  ( 1 ) 0.1637429                     *  *
## 16  ( 1 ) 0.1641019                     *  *
## 17  ( 1 ) 0.1188773  *         *        *  *     *         *         *
## 18  ( 1 ) 0.1645957  *                  *  *
## 19  ( 1 ) 0.1244587  *         *        *  *     *         *         *
## 20  ( 1 ) 0.1647053  *         *        *  *
## 21  ( 1 ) 0.1647376  *         *        *  *                          *
## 22  ( 1 ) 0.1647638  *         *        *  *                          *
## 23  ( 1 ) 0.1647833  *         *        *  *                          *
## 24  ( 1 ) 0.1648061  *         *        *  *                          *
## 25  ( 1 ) 0.1648202  *         *        *  *                          *
## 26  ( 1 ) 0.1648363  *         *        *  *               *         *
##           Junction1 No_Exit1 Railway1 Roundabout1 Station1 Stop1
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )
## 5  ( 1 )
## 6  ( 1 )          *
## 7  ( 1 )          *
## 8  ( 1 )          *
## 9  ( 1 )          *
## 10  ( 1 )         *
## 11  ( 1 )         *
## 12  ( 1 )         *                                 *
## 13  ( 1 )         *                                 *
```

```
## 14  ( 1 )         *                                   *
## 15  ( 1 )         *                                   *
## 16  ( 1 )         *                                   *
## 17  ( 1 )         *        *        *          *       *     *
## 18  ( 1 )         *                                   *
## 19  ( 1 )         *        *        *          *       *     *
## 20  ( 1 )         *                                   *     *
## 21  ( 1 )         *                                   *     *
## 22  ( 1 )         *                             *     *     *
## 23  ( 1 )         *                             *     *     *
## 24  ( 1 )         *                             *     *     *
## 25  ( 1 )         *                             *     *     *
## 26  ( 1 )         *                  *          *     *     *
##           Traffic_Calming1 Traffic_Signal1 Sunrise_SunsetNight
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )
## 5  ( 1 )                                                  *
## 6  ( 1 )                                                  *
## 7  ( 1 )                                                  *
## 8  ( 1 )                                                  *
## 9  ( 1 )                                                  *
## 10  ( 1 )                                                 *
## 11  ( 1 )                                                 *
## 12  ( 1 )                                                 *
## 13  ( 1 )                                                 *
## 14  ( 1 )                                                 *
## 15  ( 1 )                                                 *
## 16  ( 1 )                          *                      *
## 17  ( 1 )               *          *                      *
## 18  ( 1 )                          *                      *
## 19  ( 1 )               *          *                      *
## 20  ( 1 )                          *                      *
## 21  ( 1 )                          *                      *
## 22  ( 1 )                          *                      *
## 23  ( 1 )                          *                      *
## 24  ( 1 )                          *                      *
## 25  ( 1 )                          *                      *
## 26  ( 1 )                          *                      *
##           Weather_NewCloudy Weather_NewFog Weather_NewOther
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )                                          *
```

```
## 5  ( 1 )                                                *
## 6  ( 1 )                                                *
## 7  ( 1 )                                                *
## 8  ( 1 )                                                *
## 9  ( 1 )                  *                             *
## 10  ( 1 )                 *                             *
## 11  ( 1 )                 *                             *
## 12  ( 1 )                 *                             *
## 13  ( 1 )                 *                             *
## 14  ( 1 )                 *                             *
## 15  ( 1 )                 *                             *
## 16  ( 1 )                 *                             *
## 17  ( 1 )                 *
## 18  ( 1 )                 *                             *
## 19  ( 1 )                 *             *               *
## 20  ( 1 )                 *                             *
## 21  ( 1 )                 *                             *
## 22  ( 1 )                 *                             *
## 23  ( 1 )                 *                             *
## 24  ( 1 )                 *                             *
## 25  ( 1 )                 *             *               *
## 26  ( 1 )                 *                             *
##           Weather_NewPartly.Cloudy Weather_NewRain Weather_NewThunderstorm
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )
## 5  ( 1 )
## 6  ( 1 )
## 7  ( 1 )
## 8  ( 1 )
## 9  ( 1 )
## 10  ( 1 )
## 11  ( 1 )
## 12  ( 1 )
## 13  ( 1 )                       *
## 14  ( 1 )                       *                                 *
## 15  ( 1 )                       *                                 *
## 16  ( 1 )                       *                                 *
## 17  ( 1 )
## 18  ( 1 )                       *             *                   *
## 19  ( 1 )
## 20  ( 1 )                       *             *                   *
## 21  ( 1 )                       *             *                   *
## 22  ( 1 )                       *             *                   *
```

```
## 23  ( 1 )                        *              *                        *
## 24  ( 1 )                        *              *                        *
## 25  ( 1 )                        *              *                        *
## 26  ( 1 )                        *              *                        *
##           DayTuesday DayWednesday DayThursday DayFriday DaySaturday DaySunday
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )
## 5  ( 1 )
## 6  ( 1 )
## 7  ( 1 )
## 8  ( 1 )
## 9  ( 1 )
## 10  ( 1 )
## 11  ( 1 )
## 12  ( 1 )
## 13  ( 1 )
## 14  ( 1 )
## 15  ( 1 )
## 16  ( 1 )
## 17  ( 1 )
## 18  ( 1 )
## 19  ( 1 )
## 20  ( 1 )
## 21  ( 1 )
## 22  ( 1 )
## 23  ( 1 )          *
## 24  ( 1 )          *                        *
## 25  ( 1 )          *                        *
## 26  ( 1 )          *                        *
##           Precipitation Temperature WindSpeed Humidity Pressure STHR
## 1  ( 1 )
## 2  ( 1 )                                                      *
## 3  ( 1 )                                                      *    *
## 4  ( 1 )                                                      *    *
## 5  ( 1 )                                                      *    *
## 6  ( 1 )                                                      *    *
## 7  ( 1 )                                                      *    *
## 8  ( 1 )                                                      *    *
## 9  ( 1 )                                                      *    *
## 10  ( 1 )                                 *                   *    *
## 11  ( 1 )             *                   *                   *    *
## 12  ( 1 )             *                   *                   *    *
## 13  ( 1 )             *                   *                   *    *
```

```
## 14  ( 1 )              *                      *                *    *
## 15  ( 1 )              *                      *        *       *    *
## 16  ( 1 )              *                      *        *       *    *
## 17  ( 1 )
## 18  ( 1 )              *                      *        *       *    *
## 19  ( 1 )
## 20  ( 1 )              *                      *        *       *    *
## 21  ( 1 )              *                      *        *       *    *
## 22  ( 1 )              *                      *        *       *    *
## 23  ( 1 )              *                      *        *       *    *
## 24  ( 1 )              *                      *        *       *    *
## 25  ( 1 )              *                      *        *       *    *
## 26  ( 1 )              *                      *        *       *    *
##           WeekendWeekend
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )
## 5  ( 1 )
## 6  ( 1 )
## 7  ( 1 )
## 8  ( 1 )               *
## 9  ( 1 )               *
## 10  ( 1 )              *
## 11  ( 1 )              *
## 12  ( 1 )              *
## 13  ( 1 )              *
## 14  ( 1 )              *
## 15  ( 1 )              *
## 16  ( 1 )              *
## 17  ( 1 )
## 18  ( 1 )              *
## 19  ( 1 )
## 20  ( 1 )              *
## 21  ( 1 )              *
## 22  ( 1 )              *
## 23  ( 1 )              *
## 24  ( 1 )              *
## 25  ( 1 )              *
## 26  ( 1 )              *
```

9.3 *Appendix 3: Multinomial logistic regression (3 class classification)*

```
## # weights:  26 (25 variable)
## initial  value 164617.603353
## iter  10 value 134309.167766
## iter  20 value 132273.454686
## iter  30 value 125381.503468
## final  value 125380.130723
## converged
```

```
## Call:
## multinom(formula = Delay_new ~ Start_Lat + Start_Lng + Crossing +
##     Give_Way + Junction + Railway + Roundabout + Station + Stop +
##     Traffic_Signal + Sunrise_Sunset + Weather_New + Precipitation +
##     Temperature + WindSpeed + Humidity + Pressure + STHR + Weekend,
##     data = train.data)
##
## Coefficients:
##                               Values    Std. Err.
## (Intercept)              68.880693970 0.0101344867
## Start_Lat                -0.118018604 0.0037936065
## Start_Lng                 0.186553464 0.0023669892
## Crossing1                -0.006401572 0.0198658110
## Give_Way1                -0.113798806 0.0605024613
## Junction1                -0.588594025 0.0289955698
## Railway1                  0.003668284 0.0496050281
## Roundabout1               1.839661277 0.0002842737
## Station1                  0.335156635 0.0365211068
## Stop1                    -0.060761409 0.0348678623
## Traffic_Signal1           0.066759812 0.0114833997
## Sunrise_SunsetNight      -0.269035633 0.0132890997
## Weather_NewCloudy         0.447129981 0.0130289336
## Weather_NewFog            0.223122043 0.0436094388
## Weather_NewOther         -0.614908564 0.0195903755
## Weather_NewPartly Cloudy  0.264376043 0.0156239809
## Weather_NewRain           0.437242740 0.0233326000
## Weather_NewThunderstorm   1.082551393 0.0535515260
## Precipitation            -4.945022089 0.2205700702
## Temperature              -0.009890181 0.0003592572
## WindSpeed                -0.025921286 0.0011702878
## Humidity                 -0.005615676 0.0003086938
## Pressure                 -1.550668548 0.0069310055
## STHR                     -0.046986722 0.0012543330
## WeekendWeekend           -0.167149630 0.0169807728
```

```
##
## Residual Deviance: 250760.3
## AIC: 250810.3
```

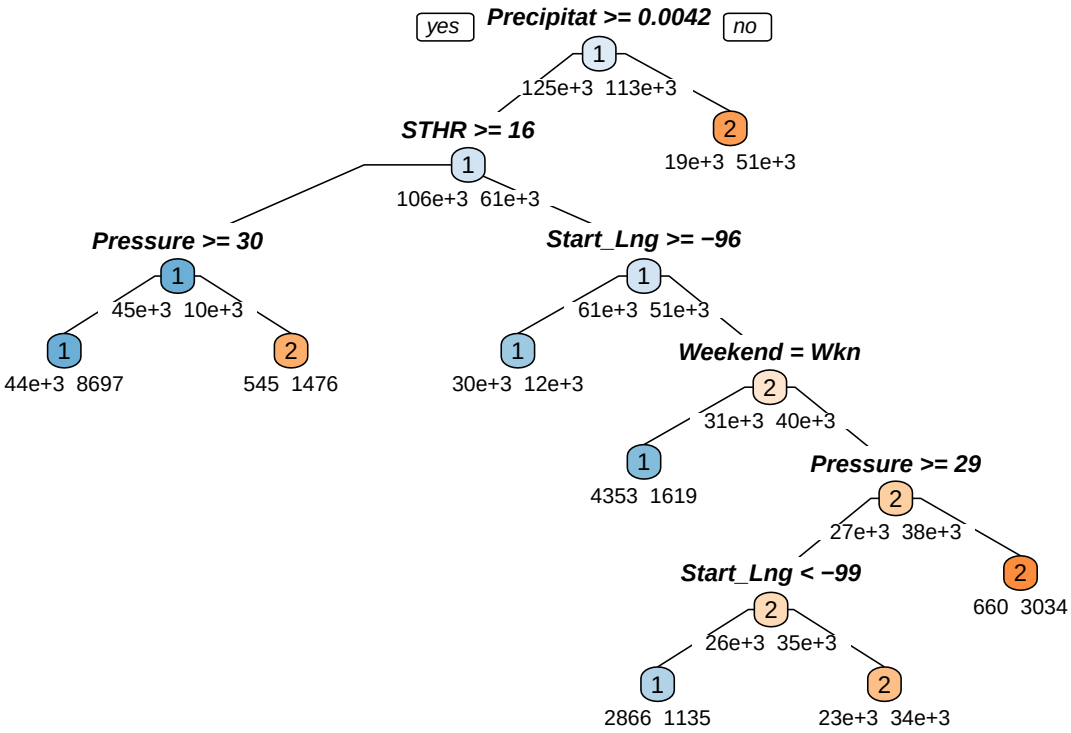
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 40670 13106
##          2  1925  3672
##
##                Accuracy : 0.7468
##                  95% CI : (0.7433, 0.7503)
##     No Information Rate : 0.7174
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.2176
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.9548
##             Specificity : 0.2189
##          Pos Pred Value : 0.7563
##          Neg Pred Value : 0.6561
##              Prevalence : 0.7174
##          Detection Rate : 0.6850
##    Detection Prevalence : 0.9057
##       Balanced Accuracy : 0.5868
##
##        'Positive' Class : 1
##
```

9.4 *Appendix 4: Logistic regression with 30min reclassification split*

```
##
## Call:
## glm(formula = Delay_new ~ Start_Lat + Start_Lng + Crossing +
##     Give_Way + Junction + Railway + Roundabout + Station + Stop +
##     Traffic_Signal + Sunrise_Sunset + Weather_New + Precipitation +
##     Temperature + WindSpeed + Humidity + Pressure + STHR + Weekend,
##     family = "binomial", data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7121  -1.0512  -0.5973   1.1151   2.5454
##
## Coefficients:
##                             Estimate Std. Error z value             Pr(>|z|)
## (Intercept)              46.4124249  0.5760397  80.572 < 0.0000000000000002 ***
## Start_Lat                 0.0230351  0.0034417   6.693      0.0000000000219 ***
## Start_Lng                 0.0570821  0.0026162  21.819 < 0.0000000000000002 ***
## Crossing1                 0.0147090  0.0180735   0.814              0.41574
## Give_Way1                -0.1626731  0.0557249  -2.919              0.00351 **
## Junction1                -0.5734824  0.0241428 -23.754 < 0.0000000000000002 ***
## Railway1                 -0.1456117  0.0461099  -3.158              0.00159 **
## Roundabout1               1.2994421  0.7657917   1.697              0.08972 .
## Station1                  0.3519817  0.0345996  10.173 < 0.0000000000000002 ***
## Stop1                    -0.1933161  0.0319021  -6.060      0.0000000013641 ***
## Traffic_Signal1           0.0491969  0.0103210   4.767      0.0000018730101 ***
## Sunrise_SunsetNight      -0.3813533  0.0118660 -32.138 < 0.0000000000000002 ***
## Weather_NewCloudy         0.2907879  0.0121859  23.863 < 0.0000000000000002 ***
## Weather_NewFog           -0.0466866  0.0415091  -1.125              0.26070
## Weather_NewOther         -0.2884068  0.0151901 -18.987 < 0.0000000000000002 ***
## Weather_NewPartly Cloudy  0.1944644  0.0141768  13.717 < 0.0000000000000002 ***
## Weather_NewRain           0.2082383  0.0213252   9.765 < 0.0000000000000002 ***
## Weather_NewThunderstorm   0.7354400  0.0524036  14.034 < 0.0000000000000002 ***
## Precipitation            -1.5524255  0.1198483 -12.953 < 0.0000000000000002 ***
## Temperature              -0.0055741  0.0003377 -16.505 < 0.0000000000000002 ***
## WindSpeed                -0.0156688  0.0010482 -14.949 < 0.0000000000000002 ***
## Humidity                 -0.0066752  0.0002852 -23.405 < 0.0000000000000002 ***
## Pressure                 -1.3247763  0.0144736 -91.530 < 0.0000000000000002 ***
## STHR                     -0.0892818  0.0011190 -79.788 < 0.0000000000000002 ***
## WeekendWeekend           -0.4316261  0.0150704 -28.641 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 328626  on 237492  degrees of freedom
## Residual deviance: 298146  on 237468  degrees of freedom
## AIC: 298196
##
## Number of Fisher Scoring iterations: 5
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1  4680  1026
##          2 26510 27157
##
##                Accuracy : 0.5362
##                  95% CI : (0.5322, 0.5402)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : 0.00000005322
##
##                   Kappa : 0.1089
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.9636
##             Specificity : 0.1500
##          Pos Pred Value : 0.5060
##          Neg Pred Value : 0.8202
##              Prevalence : 0.4747
##          Detection Rate : 0.4574
##    Detection Prevalence : 0.9039
##       Balanced Accuracy : 0.5568
##
##        'Positive' Class : 2
##
```




```
##  Delay_new
##       0.17 when Precipitation >= 0.0042 & STHR >= 16 & Pressure >= 30
##       0.27 when Precipitation >= 0.0042 & STHR <  16                  & Start_Lng <
##       0.28 when Precipitation >= 0.0042 & STHR <  16                  & Start_Lng >=
##       0.28 when Precipitation >= 0.0042 & STHR <  16 & Pressure >= 29 & Start_Lng <
##       0.59 when Precipitation >= 0.0042 & STHR <  16 & Pressure >= 29 & Start_Lng is
##       0.73 when Precipitation >= 0.0042 & STHR >= 16 & Pressure <  30
##       0.73 when Precipitation <  0.0042
##       0.82 when Precipitation >= 0.0042 & STHR <  16 & Pressure <  29 & Start_Lng <
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 20312  5857
##          2 10878 22326
##
##                Accuracy : 0.7181
##                  95% CI : (0.7145, 0.7218)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.4396
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
```

```
##             Sensitivity : 0.7922
##             Specificity : 0.6512
##          Pos Pred Value : 0.6724
##          Neg Pred Value : 0.7762
##              Prevalence : 0.4747
##          Detection Rate : 0.3760
##    Detection Prevalence : 0.5592
##       Balanced Accuracy : 0.7217
##
##        'Positive' Class : 2
##
```

9.5 *Appendix 5: Various Cut off levels for Binary Logistic regression*

Cutoff level 0.15

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1  1075   216
##          2 30115 27967
##
##                Accuracy : 0.4891
##                  95% CI : (0.4851, 0.4932)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0255
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.99234
##             Specificity : 0.03447
##          Pos Pred Value : 0.48151
##          Neg Pred Value : 0.83269
##              Prevalence : 0.47468
##          Detection Rate : 0.47104
##    Detection Prevalence : 0.97826
##       Balanced Accuracy : 0.51340
##
##        'Positive' Class : 2
##
```

Cutoff level 0.20

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1  2675   527
##          2 28515 27656
##
##                Accuracy : 0.5109
##                  95% CI : (0.5068, 0.5149)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : 1
```

```
##
##                  Kappa : 0.064
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.98130
##             Specificity : 0.08576
##          Pos Pred Value : 0.49235
##          Neg Pred Value : 0.83542
##              Prevalence : 0.47468
##          Detection Rate : 0.46580
##    Detection Prevalence : 0.94607
##       Balanced Accuracy : 0.53353
##
##        'Positive' Class : 2
##
```

Cutoff level 0.35

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1  9954  3224
##          2 21236 24959
##
##                Accuracy : 0.588
##                  95% CI : (0.5841, 0.592)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.1986
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.8856
##             Specificity : 0.3191
##          Pos Pred Value : 0.5403
##          Neg Pred Value : 0.7553
##              Prevalence : 0.4747
##          Detection Rate : 0.4204
##    Detection Prevalence : 0.7780
##       Balanced Accuracy : 0.6024
##
##        'Positive' Class : 2
##
```

9.6 *Appendix 6: Coefficients of Linear Discriminants*

```
##                                LD1
## Start_Lat               -0.125663480
## Start_Lng                0.188481643
## Crossing1               -0.014343310
## Give_Way1               -0.138446694
## Junction1               -0.626160197
## Railway1                 0.016362794
## Roundabout1              2.256790580
## Station1                 0.424351524
## Stop1                   -0.079487444
## Traffic_Signal1          0.079728192
## Sunrise_SunsetNight     -0.277037162
## Weather_NewCloudy        0.608175668
## Weather_NewFog           0.309954468
## Weather_NewOther        -0.601124531
## Weather_NewPartly Cloudy 0.325525883
## Weather_NewRain          0.473077679
## Weather_NewThunderstorm  1.313792612
## Precipitation           -3.555815011
## Temperature             -0.011963168
## WindSpeed               -0.031778731
## Humidity                -0.005668312
## Pressure                -1.763471214
## STHR                    -0.052878879
## WeekendWeekend          -0.179360813
```

The coefficient of linear discriminants gives us the weightage of each of the variable. The LD's formed here are used to create a separation line between the classes (delay level 1 and 2).

The coefficients of linear discriminants help us to understand the impact of variables on the separation line. The higher weight variables have more impact. They contribute to the transformation of original data set into a common scale. The coefficient are multipied by the actual variables to get the linear discriminant scores.

In this case, the separation line formula would be given by, Y = -0.12566 (Start_Lat) + 0.18848 (Start_Lng) - 0.01434 (crossing) .. . . -0.17936 (Weekend)

Y = Coeff of LD1 * Variable 1

9.7 *Appendix 7: Various Cut off levels for LDA*

Cutoff level 0.15

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1  7907  4629
##          2 23283 23554
##
##                Accuracy : 0.5299
##                  95% CI : (0.5259, 0.5339)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : 0.01309
##
##                   Kappa : 0.0865
##
##  Mcnemar's Test P-Value : < 0.0000000000000002
##
##             Sensitivity : 0.8358
##             Specificity : 0.2535
##          Pos Pred Value : 0.5029
##          Neg Pred Value : 0.6307
##              Prevalence : 0.4747
##          Detection Rate : 0.3967
##    Detection Prevalence : 0.7889
##       Balanced Accuracy : 0.5446
##
##        'Positive' Class : 2
##
```

Cutoff level 0.22

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 16230  9928
##          2 14960 18255
##
##                Accuracy : 0.5808
##                  95% CI : (0.5768, 0.5848)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : < 0.00000000000000022
```

```
##
##                  Kappa : 0.1667
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.6477
##             Specificity : 0.5204
##          Pos Pred Value : 0.5496
##          Neg Pred Value : 0.6205
##              Prevalence : 0.4747
##          Detection Rate : 0.3075
##    Detection Prevalence : 0.5594
##       Balanced Accuracy : 0.5840
##
##        'Positive' Class : 2
##
```

Cutoff level 0.30

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     1     2
##          1 23053 15566
##          2  8137 12617
##
##                Accuracy : 0.6008
##                  95% CI : (0.5968, 0.6047)
##     No Information Rate : 0.5253
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.1892
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.4477
##             Specificity : 0.7391
##          Pos Pred Value : 0.6079
##          Neg Pred Value : 0.5969
##              Prevalence : 0.4747
##          Detection Rate : 0.2125
##    Detection Prevalence : 0.3496
##       Balanced Accuracy : 0.5934
##
##        'Positive' Class : 2
##
```