City of Seattle 911 Call Data

516 Final Report Team 4

Maggie Du, Kali Legg, John Minorchio, Brett Penfold, and Kunpei Peng

Objective

Ensuring that the City of Seattle spends their fire department funds in a responsible and defendable way is becoming increasingly important. Decisions on where money should be allocated can no longer depend on the status quo and instead must be made using data-driven decisions. By using past 911 call and weather data, our team will attempt to build useful models that will help the City of Seattle Fire Departments predict when and where calls will occur so that they may more effectively allocate funds and resources to the communities and situations that show the most need according to empirical data.

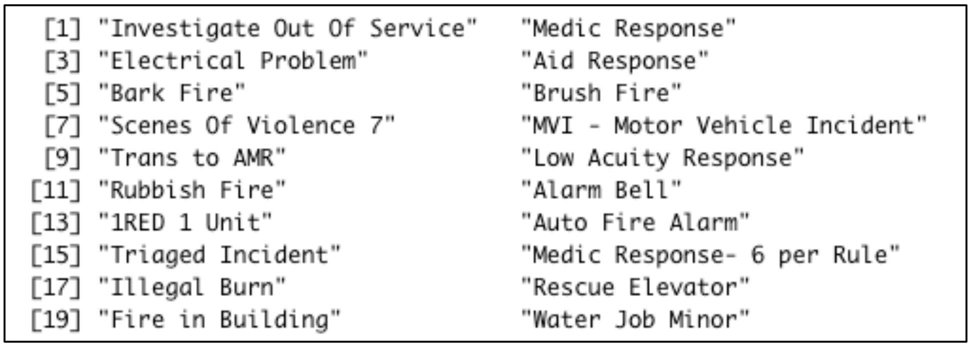
Goal: Predict the probability that the Seattle Fire Department will receive a high-volume of calls based on location, air temperature, and road temperature.

Data

The following datasets were obtained through the City of Seattle Open Data Program and were used to assist us in our analyses.

Seattle Real Time Fire 911 Calls

This dataset includes a record of every 911 call dispatched to the Seattle Fire Department including the location of the incident, date and time of the call, and the type of response. This dataset provides approximately 17 years' worth of calls from January 1, 2004 to December 31, 2020. The whole dataset includes 232 unique types of responses (Figure 1).



```
 [1] "Investigate Out Of Service"   "Medic Response"
 [3] "Electrical Problem"           "Aid Response"
 [5] "Bark Fire"                    "Brush Fire"
 [7] "Scenes Of Violence 7"         "MVI - Motor Vehicle Incident"
 [9] "Trans to AMR"                 "Low Acuity Response"
[11] "Rubbish Fire"                 "Alarm Bell"
[13] "1RED 1 Unit"                  "Auto Fire Alarm"
[15] "Triaged Incident"             "Medic Response- 6 per Rule"
[17] "Illegal Burn"                 "Rescue Elevator"
[19] "Fire in Building"             "Water Job Minor"
```

Figure 1 Example of call types included in the Seattle Real Time Fire 911 Data dataset including 20 of 232 unique call types.

Road Weather Information Stations

This data set includes data points of road surface temperature and air temperature derived from eight sensor stations located on bridges and surface streets throughout the City of Seattle. The sensor stations collect road and air temperatures every second and averages those values into temperature readings per minute for each station. The dataset we are using includes approximately 1.5 years of data from September 17, 2019 to March 9, 2021.

Independent Variables

After reviewing the datasets we established that we would use the following variables to help predict the probability that fire will receive a high volume of calls.

- Road Weather Information Stations Date Time: Time and date of the sensor reading.
- Road Weather Information Stations Latitude and Longitude
- Road Weather Information Stations Road Surface Temperature
- Road Weather Information Stations Air Temperature
- Seattle Real Time Fire 911 Date Time: Time and date the fire 911 call was received.
- Seattle Real Time Fire 911 Latitude and Longitude

Methods

Data Clean Up

Before any models could be run, the following data clean up actions were taken on each data set and variable specified above.

- Any records that returned a NULL or blank value were removed.
- Any duplicated rows were removed using the duplicate function.
- Any field that represented a date and/or time was converted to a date-time data type using Lubridate package.
- Combined latitude and longitude strings were separated into latitude and longitude variables.
- Resolution of weather data was reduced from every hour to every 6 hours.
- A subset of the Seattle Real Time Fire 911 data was created which included only the range of dates of the Road Weather Information Stations dataset
- New fields titled morning, afternoon, evening, and night were calculated from each datasets' date-time field.
- A new field titled holiday was added that determined if a call was received during one of 17 major US holidays.
- Outliers were removed from the Seattle Real Time Fire 911 data, such as abnormally high call volume (ex. calls during the George Floyd protests).

Exploratory Analysis

We used the table function to observe the number of unique values under each variable columns. We also used the summary function to observe the summary statistics of each variable used.

Temperature	Min	Max	Mean
Road Surface	-3.34	111.10	53.12
Air	-3.28	107.10	51.42

Table 1 Road Weather Information Station temperature summary

The Seattle Real Time Fire 911 Call dataset includes over 1.5 million records however, because we wanted to use the Road Weather Information Stations data with the Seattle Fire 911 Call data, we were only able to use calls from dates that were included in both datasets. Further, to reduce the number of records to better suit our hardware constraints, we only used data from May 26, 2020 to March 9, 2021, reducing the number of records to 6752.

For the Seattle Real Time Fire 911 Call we used the latitude and longitude coordinates of each call and compared it the latitude and longitude coordinates of the eight sensor stations (Appendix 1) and assigned each call to its closest sensor station that we refer to as Location in our analysis. The distance was calculated between the latitude/longitude pairs using the Geopshere package. The Seattle Real Time Fire 911 Call data does not classify the number of calls that it considers high-volume. So, we calculated the number of calls received every 6 hours and established a threshold to classify each call time-period as either high or low volume (See Appendix 2). The median number of calls was six calls per 6-hour period (Table 2). Given this information we established that a time-period with seven or more calls would be considered high-volume for this analysis.

Summary	# of calls
Min	0
Max	2772
Mean	8.73
Std Deviation	34.94
Median	6

Table 2 Summary of number of calls per 6-hour time period per sensor station area.

Clustering

Because our dataset was relatively large, we felt that k-mean clustering would be more efficient compared to using hierarchical clustering. However, we had to determine the proper number of clusters to use. Before we could begin clustering the data it was important to normalize the variables first because distance is highly influenced by the scale of variables. If the data is on a different scale, it can lead to false inferences. For example, geographic distance and temperature are not binary variables and are on two different scales, and so their values would dominate and skew the results. Through experimenting with different numbers of clusters and examining the results for interpretability, so determined that six clusters was most useful for this analysis. Through interpretation we identified and labeled the following clusters (See Appendix 3):

Seattle Real Time Fire 911 Call Clusters:

- Cluster 1: Mid temp, afternoon at Location 7
- Cluster 2: Mid temp, any time of day at Location 6
- Cluster 3: High temp, non- holiday, later in the month at Location 2
- Cluster 4: High-mid temp, evening at all Locations except Location 2
- Cluster 5: Mid temp, any time of day at Location 4
- Cluster 6: Low temp, any time of day at Location 5
- Cluster 7: Mid temp, mornings and nights

Classification

We determined that a classification tree would be the ideal classification model for this analysis. A classification tree algorithm auto selects the best variables for splitting using Gini index or entropy. Because we were unsure which variables would be most relevant to predict arrest and high-volume calls, we wanted to use a classification tree model. We also knew that we had large datasets and therefore were not constrained by the amount of call data we had. Additionally, the application of our analyses would likely be shared with an audience less familiar with analytical modelling. A classification tree would be more useful and interpretable by stakeholders.

In order to assess the effectiveness of our models we needed to create testing and training datasets. We created our testing and training datasets by splitting the Seattle Real Fire 911 dataset with 70% of the records and 30% of the records being included in the training and testing data sets respectively. By using the sample.split function we were able to ensure that the testing and training datasets had approximately the same proportions of high-volume calls.

Results

Classification models were ran on both the unclustered and clustered data in order to predict a high volume of calls for each location and time-period (1 = call volume ≥ 7, 0 = call volume < 7). The explanatory variables included the road surface temperature, air temperature, closest sensor location, time of day, whether it was a holiday, and whether it was a weekend.

The results of the model are compared with a baseline model that just considers assigning each time-period/location as low volume. The accuracy of this baseline model is 3651/6752 or 54%.

Unclustered Data

The output of the model and confusion matrix is provided in Figure 2 and Table 3 respectively. The accuracy of this model is 85.4%, which is a significant improvement to the baseline model. The main consideration for the model in making predictions is location.

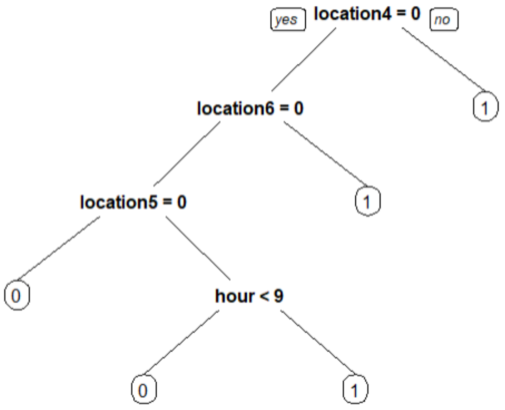


Figure 2 Decision Tree

	Predict 0	Predict 1
Actual 0	1002	93
Actual 1	202	728

Table 3 Confusion Matrix of Unclustered Model

Clustered Data

The confusion matrix and accuracy of the clustered models is provided in Table 4. The overall accuracy is 85.7%, which is similar to the accuracy of the model without clustering. The output the decision tree for each cluster is provided in Appendix 4.

		Predict 0	Predict 1	Accuracy
Cluster 1	**Actual 0**	275	8	86.14%
	Actual 1	39	17	
Cluster 2	**Actual 0**	0	62	75.30%
	Actual 1	0	189	
Cluster 3	**Actual 0**	0	1	87.50%
	Actual 1	0	7	
Cluster 4	**Actual 0**	205	5	86.76%
	Actual 1	51	162	
Cluster 5	**Actual 0**	8	2	98.39%
	Actual 1	2	237	
Cluster 6	**Actual 0**	67	53	71.11%
	Actual 1	40	162	
Cluster 7	**Actual 0**	407	2	93.78%
	Actual 1	25	0	

Table 4 Confusion Matrix of Clusters and their Accuracy

XGBoost

An XGBoost model was also trained on the unclustered data in order to evaluate if an alternative classification model could provide improved results. The confusion matrix for the predictions on the test dataset is provided in Table 5. The accuracy of this model is 86.2%, which is slightly improved as compared to the decision tree.

XGB	Predict 0	Predict 1
Actual 0	999	96
Actual 1	182	748

Table 5 Confusion Matrix of XGBoost Model

Conclusion

Based on our findings from this model, we can infer that location is a predominant factor in determining call volumes throughout the day.   As we can also see in the decision trees in Appendix 4, weather and road conditions did not have as strong of an influence as hypothesized. In most clusters, the weather-related variables such as Air and Road Temperature were not seen at all; if they were, they were toward the bottom of the tree.

Application

Our recommendation is to use the results of this analysis to make high-level decisions about where to allocate fire and police resources and make those decisions transparent to the community.

By identifying what temperatures and locations will increase the likelihood of high-volume calls, when the Seattle Fire Department begins to evaluate their budget, they can allocate resources seasonally and towards certain communities instead of evenly distributing resources throughout the year and city. By evaluating resource allocation this way, the Seattle Fire Department may be able to avoid understaffing during certain times of the year and areas and overstaffing during others. Additionally, this model can be used to help identify preventative measures that can be taken to reduce the probability of high-volume calls. For instance, if certain areas of the City of Seattle are more likely to receive a high volume of calls during lower temperatures, the City of Seattle can attempt to target social programs towards those communities.

Future Improvements

It’s important to note that the models performed in this analysis should be re-evaluated and adjusted continuously. As more call and weather data becomes available, the model should be recalibrated and rerun to keep up with changes in trends and behaviors in call and weather conditions.

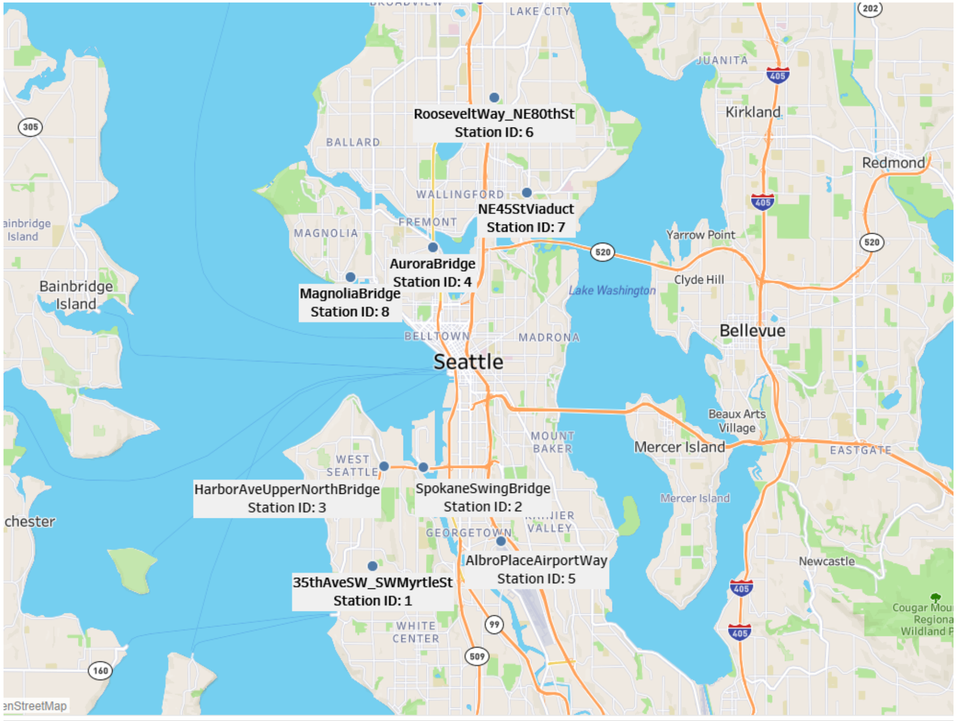
One recommendation to improve upon this analysis is to expand the type of weather data included in the analysis. Air and road temperature is a narrow assessment of the overall weather conditions of a region. Variables such as average inches of precipitation, wind speeds, and air quality could all help make the models more accurate and reliable. Additionally, the data could be grouped in a more geographically appropriate way than by distance to sensor stations. For instance, using Fire Department service areas or specific business districts to estimate call volume could have more useful applications.

Another important factor to consider when estimating the probability of high-volume calls would be population density. Some areas within the City of Seattle are likely to have higher call volumes because the population density is higher. Therefore, it would be advantageous to normalize the call data based on information about population density. Moreover, information regarding the fire station's staffing levels at a point in time could help lead to a model that could recommend increasing or decreasing staff based on the frequency of calls and incidents.

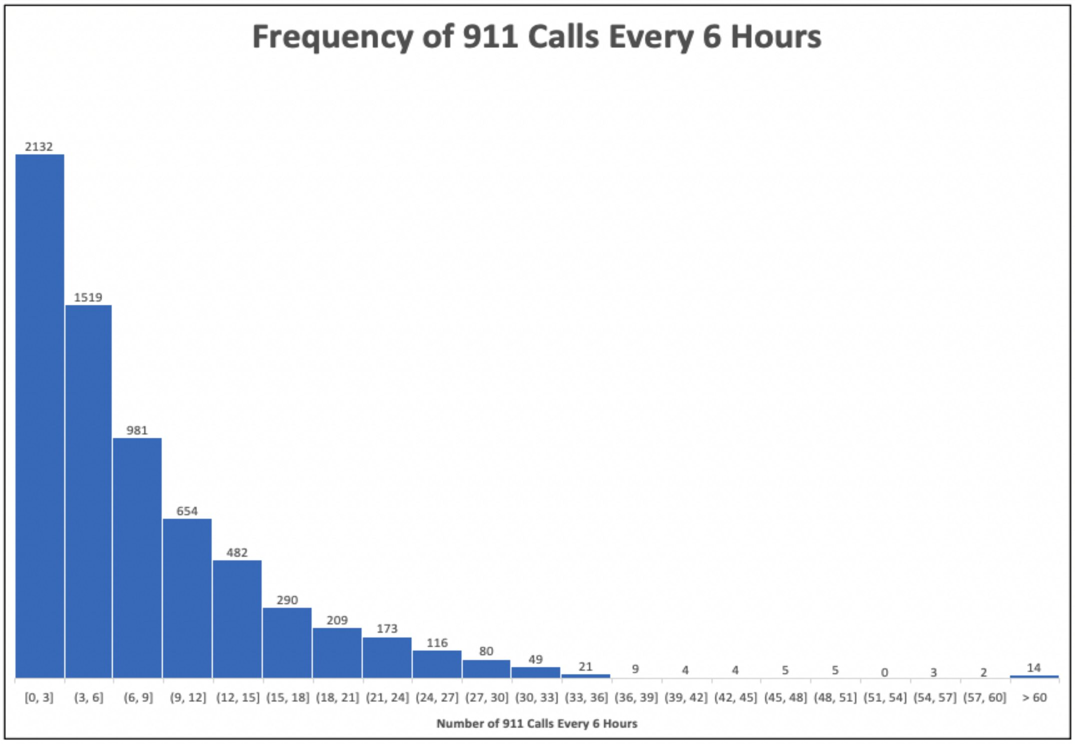
Additionally, when trying to determine the number of clusters to use it could be advantageous to first run a hierarchical clustering model to establish the ideal number of clusters. Then, re-run the k-mean clustering model with the number of clusters produced from the hierarchical clustering. Determining the number of clusters to use can be considered a subjective process in clustering models but this is one way that could help improve the analysis.

Appendix

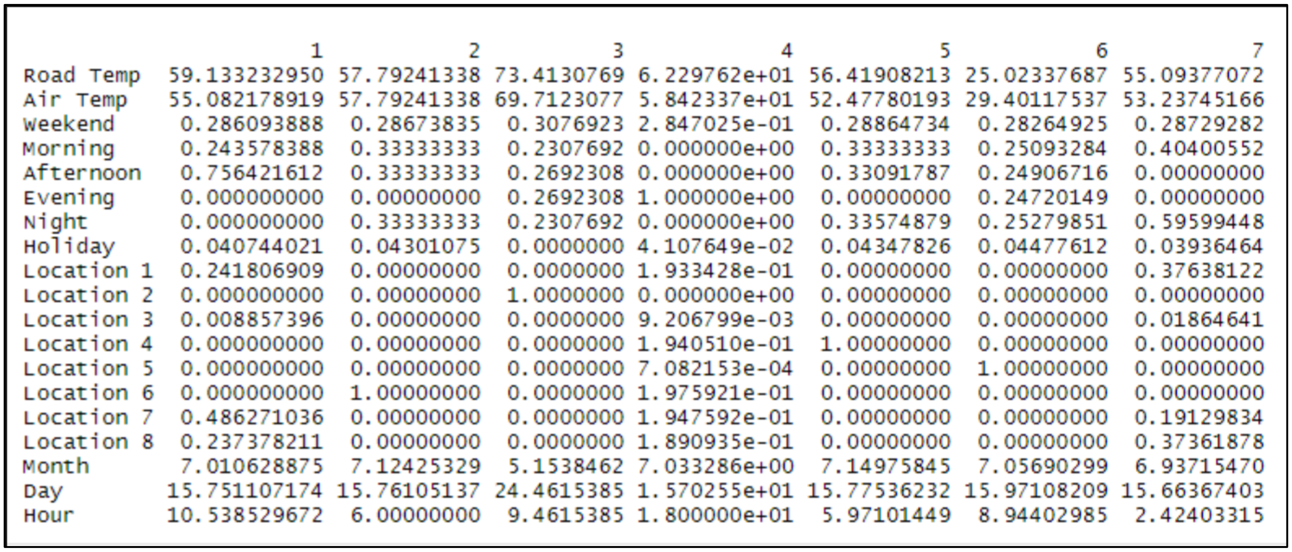
Appendix 1 Weather Sensor Station Locations. Note: Station ID below is represented by the variable called Location in our analysis.



Appendix 2 Histogram of 911 Call Frequency Every 6 Hours



Appendix 3 K-Means Clustering



```
                      1           2          3            4           5           6           7
Road Temp  59.133232950 57.79241338 73.4130769 6.229762e+01 56.41908213 25.02337687 55.09377072
Air Temp   55.082178919 57.79241338 69.7123077 5.842337e+01 52.47780193 29.40117537 53.23745166
Weekend     0.286093888  0.28673835  0.3076923 2.847025e-01  0.28864734  0.28264925  0.28729282
Morning     0.243578388  0.33333333  0.2307692 0.000000e+00  0.33333333  0.25093284  0.40400552
Afternoon   0.756421612  0.33333333  0.2692308 0.000000e+00  0.33091787  0.24906716  0.00000000
Evening     0.000000000  0.00000000  0.2692308 1.000000e+00  0.00000000  0.24720149  0.00000000
Night       0.000000000  0.33333333  0.2307692 0.000000e+00  0.33574879  0.25279851  0.59599448
Holiday     0.040744021  0.04301075  0.0000000 4.107649e-02  0.04347826  0.04477612  0.03936464
Location 1  0.241806909  0.00000000  0.0000000 1.933428e-01  0.00000000  0.00000000  0.37638122
Location 2  0.000000000  0.00000000  1.0000000 0.000000e+00  0.00000000  0.00000000  0.00000000
Location 3  0.008857396  0.00000000  0.0000000 9.206799e-03  0.00000000  0.00000000  0.01864641
Location 4  0.000000000  0.00000000  0.0000000 1.940510e-01  1.00000000  0.00000000  0.00000000
Location 5  0.000000000  0.00000000  0.0000000 7.082153e-04  0.00000000  1.00000000  0.00000000
Location 6  0.000000000  1.00000000  0.0000000 1.975921e-01  0.00000000  0.00000000  0.00000000
Location 7  0.486271036  0.00000000  0.0000000 1.947592e-01  0.00000000  0.00000000  0.19129834
Location 8  0.237378211  0.00000000  0.0000000 1.890935e-01  0.00000000  0.00000000  0.37361878
Month       7.010628875  7.12425329  5.1538462 7.033286e+00  7.14975845  7.05690299  6.93715470
Day        15.751107174 15.76105137 24.4615385 1.570255e+01 15.77536232 15.97108209 15.66367403
Hour       10.538529672  6.00000000  9.4615385 1.800000e+01  5.97101449  8.94402985  2.42403315
```

Appendix 4 Cluster Decision Trees

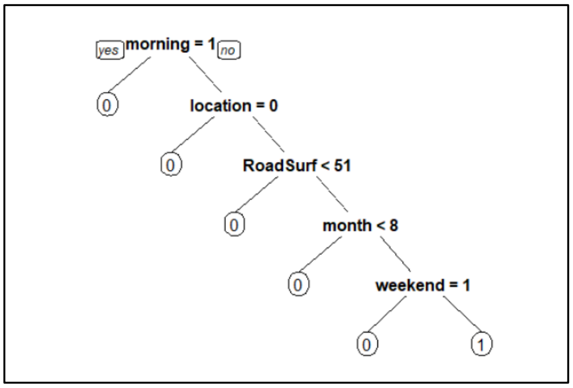


Figure 3 Cluster 1 classification tree results

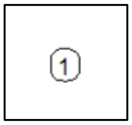


Figure 4 Cluster 2 classification tree results (always predicts high volume calls)

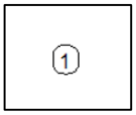


Figure 5 Cluster 3 classification tree results (always predicts high volume calls)

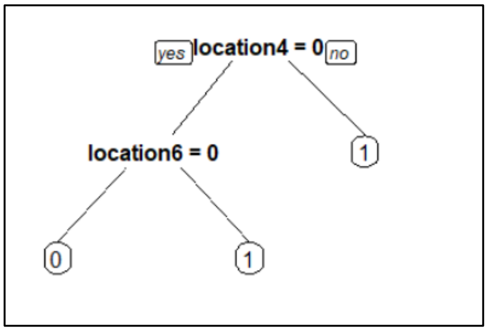


Figure 6 Cluster 4 classification tree results

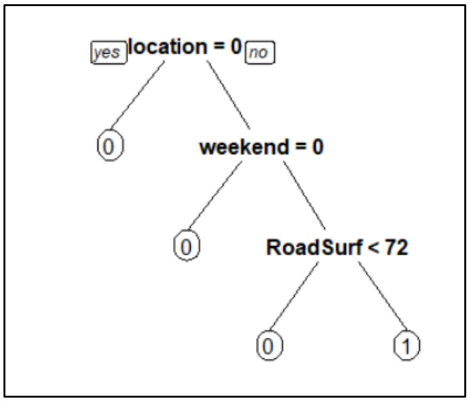


Figure 7 Cluster 5 classification tree results

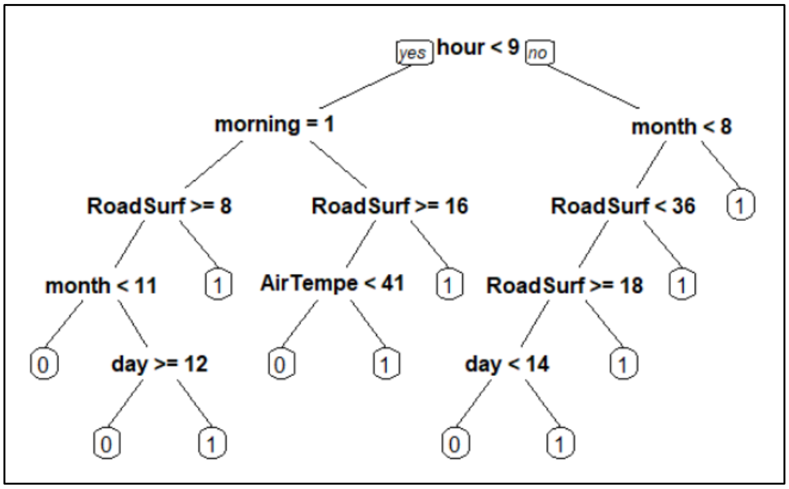


Figure 8 Cluster 6 classification tree results

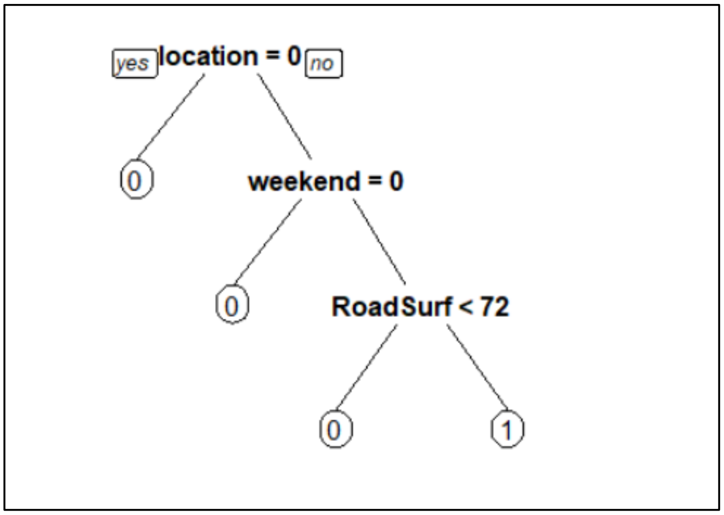


Figure 9 Cluster 7 classification tree results