RoadRiskAlert: A Model for US Accident Severity Prediction

1. Introduction

The major objective of this project is to analyze various factors that might lead to road accidents which in turn lead to traffic congestion and we have used the “US Accidents” dataset. The dataset consists of 2,845,342 observations covering 50 states of the United States and 47 variables. As this is a huge dataset and requires high resources for computation and processing, the data from 6 states of New England are only analyzed in this project. Some of the variables in the dataset are categorical which mentions details about road conditions, sunrise or sunset, bumps, traffic signals, etc., while other variables are numeric which mention details about temperature, wind speed, and the distance of the road closed, etc. This project aims to predict the “severity” of accidents which takes values from 1 to 4 with 4 describing the highest impact of the accident on traffic and 1 being of least impact. A prediction model needs to be created with explanatory variables that have the most impact on the variable severity.

2. Preparing the data set

2.1 Subset the dataset

To begin with, we subset the given dataset based on the US states. New England comprises of six US states namely Rhode Island, Maine, Massachusetts, Vermont, New Hampshire, and Connecticut. We filter all the observations from the New England area using *filter()* from *dplyr* library. A total of 47,029 observations were obtained and this was used for further analysis.

2.2 Removing single value columns

It is found that columns "*Country*", and "*Turning_Loop*" have only one factor level and hence they do not add any value to the prediction models. So, these columns are removed before proceeding further.

2.3 Removing rows with missing values

We know that only 33.3% of the dataset is complete and hence we look for columns with high missing values. All columns with more than 20% of data missing are removed. It is found that columns *"Number", "Precipitation.in."*, and *"Wind_Chill.F."* have high NA values and are removed. Furthermore, the dataset is filtered for only complete cases. The final cleaned dataset is stored in the object *NE_USaccidents* and it contains 43,972 observations over 42 variables.

3. Exploratory Data analysis

3.1 Categorizing the target variable

The target variable *Severity* shows the severity of the accidents. It contains the number 1 through 4, where 1 indicates the least impact on the traffic and 4 indicates a significant impact on traffic. Figure 1 illustrates the severity analysis of accidents recorded in the New England area in the last three years. There are very few incidents with low severity and most incidents are found to have severity of level 2. To simply the classification of severity, the accidents with severity 1 and 2 are grouped as "Low" and accidents with severity 3 and 4 as "High".

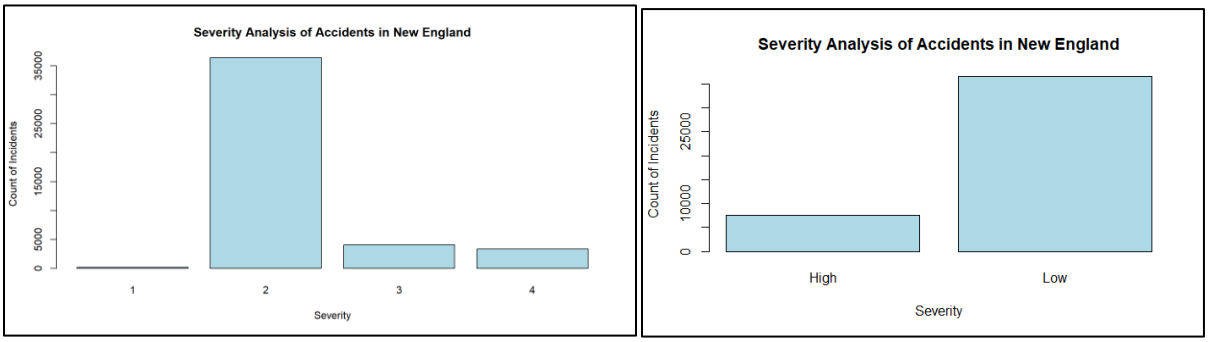


Figure 1. Bar chart illustrating the count of incidents in New England based on the Severity of the incidents.

3.2 Checking for cross-correlation

The target variable in our dataset is '*Severity*'. All other 41 variables are considered as independent variables and the correlation between these variables are checked using the *corr_cross()*. The top 15 relevant variables are removed as the multicollinearity among independent variables will result in less reliable statistical inferences. A total of 22 columns are listed and a new object is created excluding these columns. The classification models are built with this object *NE_USaccidents_1* that consists of 43,972 observations across 20 variables.

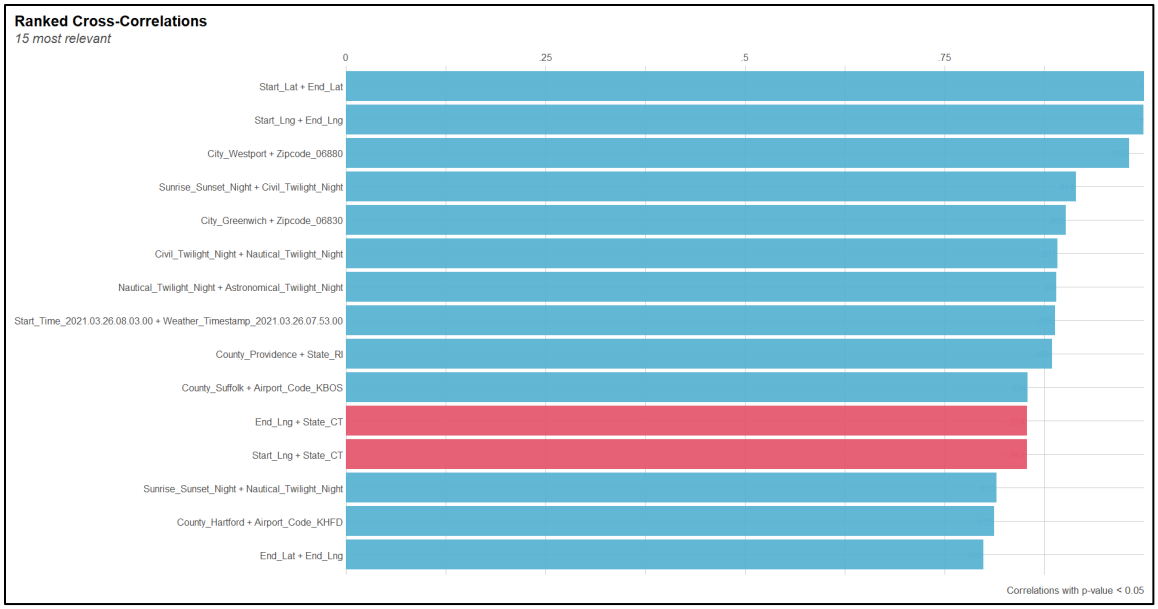


Figure 2. Top 15 most relevant cross-correlation between the explanatory variables

3.3 Impact of weather conditions on severity levels

From Figure 3, it is observed that all three features, namely Wind Speed, Humidity and Temperature have no influence individually on determining high or low severity. Both the severity levels have occurred during similar conditions of wind speed, humidity, and temperature.

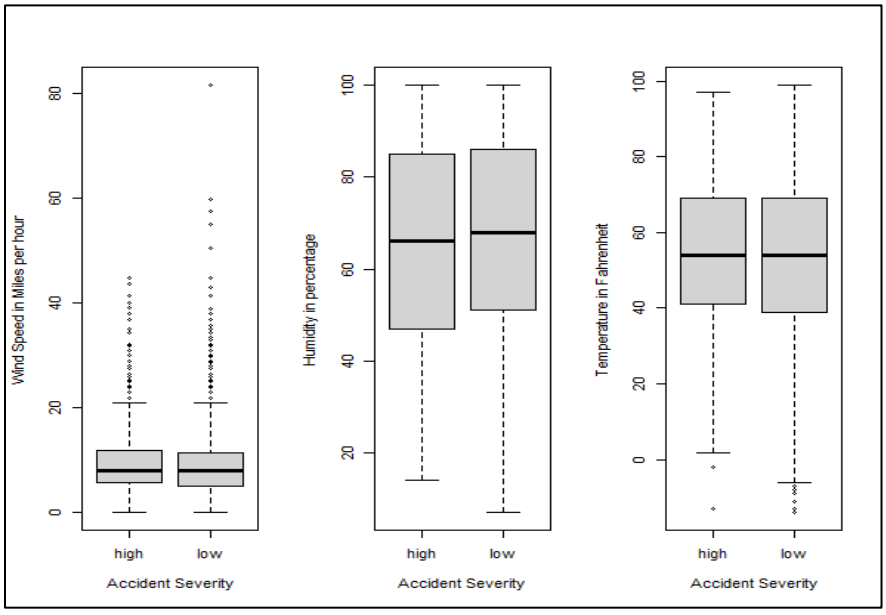


Figure 3. Boxplot showing the severity distribution across various weather factors.

3.4 Correlation of Severity with different Variables

Correlation of Severity with numeric variables

Variables	r
Wind_Speed.mph.	-0.0770510
End_Lng	-0.0537388
Start_Lng	-0.0535989
Temperature.F.	-0.0164916
Visibility.mi.	-0.0161626
Start_Lat	-0.0076956
End_Lat	-0.0076621
Pressure.in.	0.0085751
Distance.mi.	0.0311601
Humidity...	0.0468007

Correlation of Severity with other variables

Variables	r
Junction	-0.0257803
No_Exit	-0.0061204
Roundabout	-0.0033972
Bump	-0.0019614
Railway	0.0023288
Traffic_Calming	0.0059705
Sunrise_Sunset	0.0073923
Civil_Twilight	0.0077030
Astronomical_Twilight	0.0132274
Nautical_Twilight	0.0134811
Give_Way	0.0158534
Weather_Condition	0.0211557
Amenity	0.0258845
Station	0.0585162
Crossing	0.0636158
Stop	0.0727425
Traffic_Signal	0.1125122

Figure 4. Correlation of the target variable Severity with numeric (*4a.* in the left) and factor variables (*4b.* in the right)

The correlation coefficient is calculated between the target variable Severity and all other explanatory variables. There are ten numeric columns in the data set and the correlation between these numeric columns and the target variables is given by figure 4a. There are 17 columns with values that can be converted into factors. These factors are then converted into numeric values and the correlation between these columns and the response variable is calculated. Figure 4b illustrates the correlation between severity and the factor variables. We can observe that no variable has high correlation with our target variable severity.

4. Prediction Models

Since the correlation between the target variable and all other independent variables is very weak, we can't use linear regression model in our dataset. Hence, we would experiment with different classification and regression models to predict the severity of accidents in New England. The data set is sliced into two groups randomly for training and testing. The models are trained using 80% of the data (35,178 observations) and tested using the remaining 20% of the data (8,794 observations) from the dataset. The accuracies of the models are calculated and compared. The model with the highest accuracy is selected for further analysis.

4.1.1 Classification tree model

A classification tree model is built using the training data set. The model is created using *rpart()* which uses recursive partitioning to create the regression tree [Camacho, 2017]. Four major arguments (formula, data, parms, and method) are passed to *rpart* to create the model.

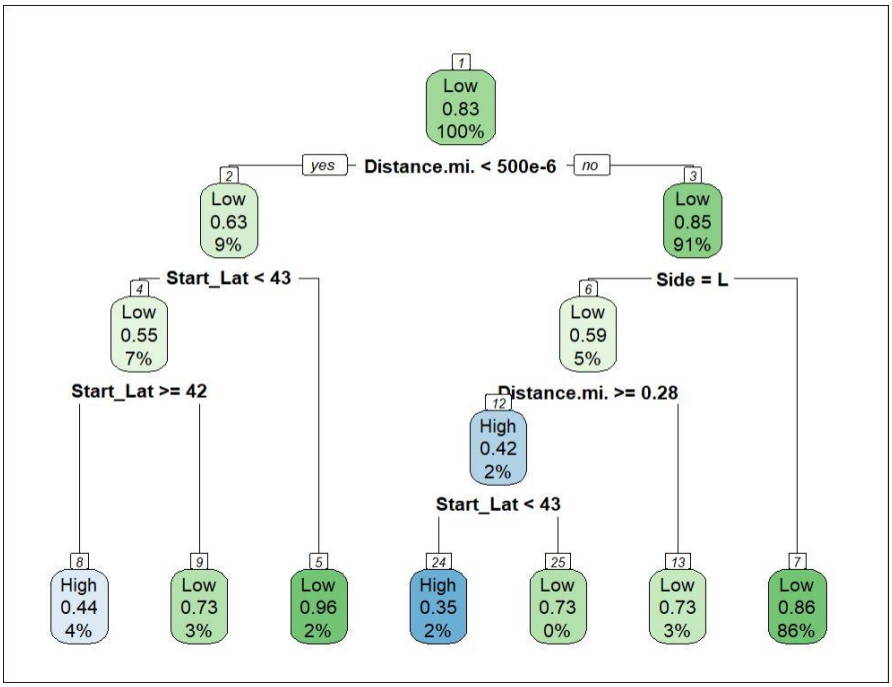


Figure 5. The Classification tree generated from the training data set

The training dataset is the data and the attribute *method* is set to "*class*" as the target variables are categorical. We are not using the loss matrix in this model. The model created with the above parameters is stored in the object 'rpart_*tree'*. Figure 5 is the visualization of the *tree* object using the function *rpart.plot()*. In the classification tree, there are a total of 13 nodes out of which 7 are terminal nodes.

From figure 5, we can see that the most significant splitter variable is the *Distance.mi* which divides the data into two groups in the ratio 1:10. The value of *Distance.mi* is significant for 91% of the data. The second most significant splitter at this stage is *Side* (relative side of the street). When *Side* is R (right), the severity of the accident is found to be 'Low' and they contribute to 86% of the total data in the training set. There is 5% of the data left with the *Side* is L (left), which is furthermore split based on *Distance.mi* and *Start_Lat.*

The 9% of the data with insignificant *Distance.mi* value is classified based on *Start_Lat.* When the values of *Start_Lat* lie within 42 and 43, the severity is predicted to be "High" and they contribute 4% of the total data in the training set. Thus, the final classification model is created with 3 variables as splitters to classify the severity of accidents as either Low or High.

4.1.2 Tree Pruning

To avoid over fitting, the classification tree must be pruned. The *cptable* provides a summary of the overall fit of the model. The "*xerror*" in the cptable contains cross-validated classification error rates. We choose the CP value with a minimum *xerror* value, and it is used to prune the tree. The function *prune()* is used to prune the tree using complexity parameter (cp) as the argument. We can see that the pruned tree is the same as the original tree and hence we can say that our model is an optimal model.

4.1.3 Testing the model

The function *predict()* from the *stats* library is used to test the generated model using the 20% of the testing dataset that was created initially. The output of this function is a factor, and it is stored in an object '*rpart_tree_pred'*. A consistency table 'acc_rpart_pred' is created for this object and to better interpret the results, the confusionMatrix is applied on the table.

```
> confusionMatrix(acc_rpart_pred)
Confusion Matrix and Statistics

          rpart_tree_pred
           High  Low
      High  319 1145
      Low   225 7105

                   Accuracy : 0.8442
                     95% CI : (0.8365, 0.8517)
       No Information Rate : 0.9381
       P-Value [Acc > NIR] : 1

                      Kappa : 0.2501

     Mcnemar's Test P-Value : <2e-16

                  Sensitivity : 0.58640
                  Specificity : 0.86121
               Pos Pred Value : 0.21790
               Neg Pred Value : 0.96930
                   Prevalence : 0.06186
               Detection Rate : 0.03627
         Detection Prevalence : 0.16648
            Balanced Accuracy : 0.72380

             'Positive' Class : High
```

The confusion matrix shows the performance measurement of the classification model created [Narkhede, 2018]. The confusion matrix in figure 6 explains various statistics of the model. We can see that the accuracy of the model is only **84.42%** with many Type-1 and few Type-2 errors. The 'Positive' class in the testing data set is found to be from severity 'High' with a prevalence of 0.061. The kappa value for this model is poor and with the high type 1 and 2 errors, we can conclude that it is not the best fit for our data.

Figure 6. Confusion Matrix of testing data from model created without loss matrix

4.2 Random Forest model

In classification algorithm, Random Forest is preferred over Decision Tree because it eliminates high variance by adding randomness to the model. The algorithm works by creating multiple

decision trees using a bootstrapping method by random sampling with replacement. It uses bagging and feature randomness while building each individual tree for creating an uncorrelated forest of trees whose prediction on whole is more accurate than that of any individual tree. A bagging technique is used for the output from each decision tree. The final output is the mode of the output of all the individual decision trees.

4.2.1 Running Random Forest model

We have used *randomForest()* function which can be applied for both classification and regression algorithms. The data frame of predictors is passed in the x argument and the target variable is passed in the y argument. The argument *ntree* represents to the number of trees we want to build in the model. Since, we have used a dataset of 44,000 records we have used the standard value of 500 for ntree. The argument *mtry* represents the number of variables randomly sampled as candidates at each split. Since large mtry ensures that there is at least one strong variable in the set of mtry candidate variables (Probst, 2019), we have chosen *mtry* value as 12. Out-of-bag (OOB) error estimate which is a measure of prediction error of random forests is 12.98%

```
> ne_accidents_rf

Call:
 randomForest(x = X_train, y = Y_train, ntree = 500, mtry = 12,
    importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 12

        OOB estimate of  error rate: 12.85%
Confusion matrix:
     High   Low class.error
High 2311  3671  0.61367436
Low   848 28348  0.02904507
>
```

Figure 7. Output of random forest fit model

4.2.2 Testing the model accuracy

The *predict()* function is used to validate the Random Forest model with the testing dataset. The model is given as an input for the argument object. The test data set is passed as a parameter to the argument *newdata*. The testing data are predicted in the severity level of 'high' and 'low' by

using the Random Forest model we built using the training data. To check the accuracy of the prediction and to validate the Random Forest model, we used *confusionMatrix()* which will categorize the predictions against the actual values.

```
> confusionMatrix(randomForest_pred, NE_USaccidents_test$Severity)
Confusion Matrix and Statistics

          Reference
Prediction High  Low
      High  555  222
      Low   909 7108

               Accuracy : 0.8714
                 95% CI : (0.8642, 0.8783)
    No Information Rate : 0.8335
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4294

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.37910
            Specificity : 0.96971
         Pos Pred Value : 0.71429
         Neg Pred Value : 0.88662
             Prevalence : 0.16648
         Detection Rate : 0.06311
   Detection Prevalence : 0.08836
      Balanced Accuracy : 0.67441

       'Positive' Class : High
```

Figure 8. Confusion matrix for testing the accuracy of the Random Forest model

From figure 8, we can observe that the accuracy of the random forest model is **87.14%** and the kappa is 0.4294. So, we can say that the Random Forest is a promising model to make prediction.

4.2.3 Variable Importance in Random Forest model

The variable of importance for the Random Forest model can be obtained by using the function *importance().* Variables with high importance have a significant impact on predicting the target values. By contrast, variables with low importance might be omitted from a model, making it simpler and faster to fit and predict. Figure 9 illustrates the results of the important variables in descending order used by the generated random forest model.

```
> importance(ne_accidents_rf)
                      High        Low MeanDecreaseAccuracy MeanDecreaseGini
Start_Lat       207.828139 225.212901            280.047421     2.166006e+03
Distance.mi.    184.316802 192.185036            265.277575     1.774372e+03
Side             87.758545 135.565621            160.646121     2.476833e+02
Temperature.F.  121.675736 110.170461            143.903887     1.345052e+03
Humidity...     102.242617 103.557566            139.140333     1.155931e+03
Pressure.in.     96.891335 142.920574            166.694739     1.363853e+03
Visibility.mi.   27.207559  57.117790             66.869486     2.543126e+02
Wind_Speed.mph. 112.680805 132.763115            148.333230     9.471591e+02
Amenity           5.437446  13.685590             15.184651     1.736334e+01
Bump              0.000000   0.000000              0.000000     4.516865e-02
Give_Way          6.047074   1.700597              3.882948     1.175449e+01
Junction         40.337134  29.677309             44.622521     1.383378e+02
No_Exit           0.000000   2.702306              2.702606     3.814759e-01
Railway           4.088549   7.504395              8.477841     6.891843e+00
Roundabout        0.000000   0.000000              0.000000     3.500000e-03
Station          17.919977   5.663582             16.520528     1.403675e+01
Stop             20.511766  21.648942             27.718402     3.816977e+01
Traffic_Signal   30.775761  48.390628             50.275848     7.305845e+01
Sunrise_Sunset   32.110733  58.252806             67.052094     1.693975e+02
```

Figure 9. The output displays the variable of importance for severity levels - “high ” and “low”

The two measures of importance in the random forest are *Mean Decrease Accuracy* and *Mean Decrease Gini*. The first measure is based on how much the accuracy decreases when a particular variable is removed from the model. The second measure is based on the decrease of Gini impurity when a variable is chosen to split a node. To visualize the measures of importance, a dot chart is created using the *varImpPlot()* as shown in figure 10.

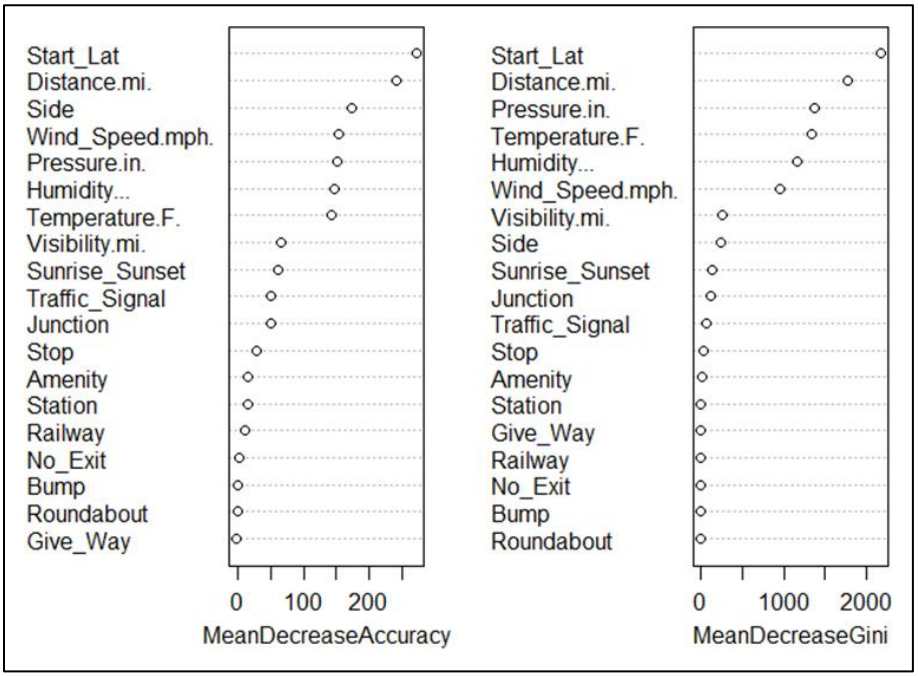


Figure 10. Dotchart of variable of importance

From figures 9 and 10, we can conclude that the random forest model considers latitude, distance, side, temperature, pressure, humidity, wind speed and visibility as the variables of high importance for predicting the severity of the accident. Similarly, the point of interest (POI) attributes such as Amenity, Bump, Give Way, Junction, No Exit, Railway, Roundabout, Station, and Stop are the variables of low importance for predicting the severity of the accident.

4.3 Gradient Boost Method

We now try the gradient boost method to reduce the bias errors which helps to determine which predictor variables are highly important. The train and test data will remain the same. Using all the 20 predictor variables, we created the first GBM_1 model using *train()* function and method = gbm. The cross-validation training control method has been used with number of repeats being 10. Scaling and centering of the data is done by using preProcess arguments.

```
> summary(GBM_1, las = 1)
                                  var     rel.inf
Start_Lat                   Start_Lat 33.58156827
Distance.mi.             Distance.mi. 24.45685465
SideR                           SideR 13.21969655
Wind_Speed.mph.       Wind_Speed.mph. 10.76393744
Temperature.F.         Temperature.F.  8.57279307
Traffic_SignalTrue Traffic_SignalTrue  3.51601579
Humidity...               Humidity...  2.28424274
Pressure.in.             Pressure.in.  1.40427459
JunctionTrue             JunctionTrue  0.72860498
StationTrue               StationTrue  0.43418996
Sunrise_SunsetNight Sunrise_SunsetNight 0.42897930
Visibility.mi.         Visibility.mi.  0.30631785
Give_WayTrue             Give_WayTrue  0.21906377
AmenityTrue               AmenityTrue  0.04389581
RailwayTrue               RailwayTrue  0.03956523
BumpTrue                     BumpTrue  0.00000000
No_ExitTrue               No_ExitTrue  0.00000000
RoundaboutTrue         RoundaboutTrue  0.00000000
```

Figure 11. Summary of GBM_1 model

Based on the above summary of the GBM_1 model, it is found that few predictor variables namely Roudabout, No Exit, Bump, Amenity and Railway have 0 relative influence on the target variable severity. Hence, we remove those variables and try another model GBM_2 with reduced variables. The test data is then used on the model to predict the severity. Confusion matrix is created to check the predictions and its accuracy level.

```
> summary(GBM_2, las = 1)
                                 var    rel.inf
Start_Lat                   Start_Lat 33.8896221
Distance.mi.             Distance.mi. 25.8636275
SideR                           SideR 13.4033766
Wind_Speed.mph.     Wind_Speed.mph. 10.5502832
Temperature.F.       Temperature.F.  9.3077999
Traffic_SignalTrue Traffic_SignalTrue  3.3998737
Humidity...               Humidity...  2.3312559
StationTrue               StationTrue  0.4546938
Sunrise_SunsetNight Sunrise_SunsetNight  0.3693782
Give_WayTrue           Give_WayTrue  0.2728271
JunctionTrue           JunctionTrue  0.1572621
AmenityTrue             AmenityTrue  0.0000000
```

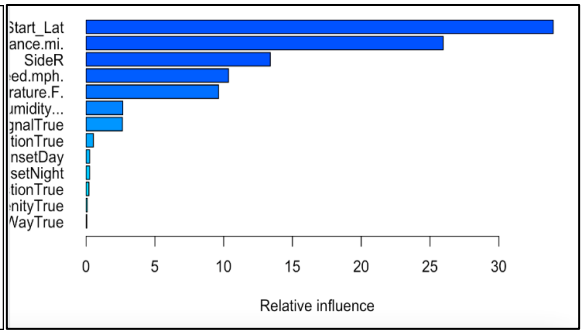


Figure 12. Summary of GBM model 2 showing relative influence of predictor variables

The relative influence of the predictor variables can be identified by using *summary ()* of the second gbm model (GBM_2). Figure 12 shows that the variables Start_Lat, Distance.mi., Side(R), Wind_Speed.mph, and Temperature.F are having relatively high influence on severity when compared to other variables.

```
> confusionMatrix(GBM_pred_2, NE_USaccidents_test$Severity)
Confusion Matrix and Statistics

          Reference
Prediction High  Low
      High  261  133
      Low  1203 7197

               Accuracy : 0.8481
                 95% CI : (0.8404, 0.8555)
    No Information Rate : 0.8335
    P-Value [Acc > NIR] : 0.0001118

                  Kappa : 0.2263

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.17828
            Specificity : 0.98186
         Pos Pred Value : 0.66244
         Neg Pred Value : 0.85679
             Prevalence : 0.16648
         Detection Rate : 0.02968
   Detection Prevalence : 0.04480
      Balanced Accuracy : 0.58007

       'Positive' Class : High
```

Figure 13. Confusion matrix of GBM model 2

From Figure 13, we can observe the number of predictions for severity levels. There are 116 false positives for severity low predicted as high and there are 1203 false negatives that predicted actual high severity as low severity levels. This model gives us an **85%** accuracy in prediction at a 95% confidence interval. Though the accuracy is 85%, we see that there are huge misclassification errors. Though we have an 85% accuracy, we can observe that the Kappa

Statistic is 0.2263 which indicates that there is a poor agreement between the observed accuracy and the expected accuracy at a random chance. Hence, we cannot say that this model is a better fit.

4.4 Improving the model accuracy – Random Forest Model

From the above analysis of all three models, we found that the random forest model has produced the highest accuracy of prediction with 87% accuracy. Next, we wanted to test if the model is statistically significant. Hence, we used *rf.significance()* function from "*rfUtilities*" package to test the model significance at 0.05 significance level

```
> rf.perm
Number of permutations:  99
p-value:  0
Model signifiant at p = 0
         Model OOB error:  0.1290332
         Random OOB error:  0.1696132
         min random global error: 0.1691024
         max random global error:  0.1702091
         min random within class error: 0.9964688
         max random within class error:  0.9964688
```

Figure 14. Random Forest model significance test

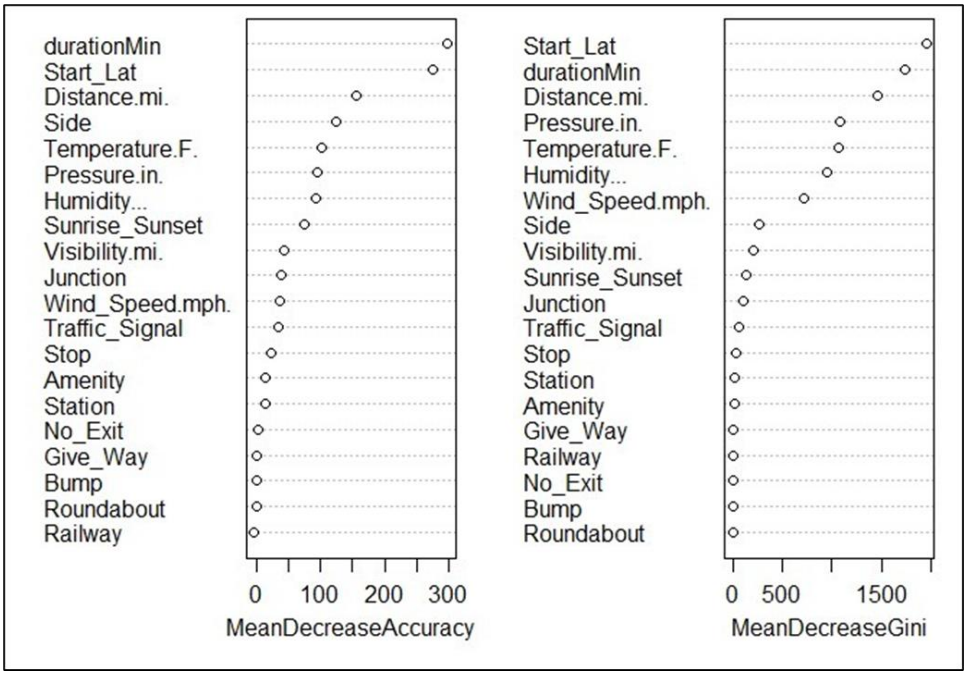
From figure 14, it is evident that the p-value is lesser than 0.05, and hence the model is tested to be significant for predicting the severity level. Next, the features of the model are revisited to improve its accuracy. A new calculated field "Duration" is created based on the difference between the start time and end time of the accident and added as a predictor variable to the training data set. The model is rebuilt with this new training set and then used to predict the test data.

Figure 15 illustrates the confusion matrix of the random forest model after including duration and we can observe that the accuracy of the model has increased from **87% to 88%**. The misclassification of high severity as low is more dangerous than predicting low severity as high. We can see a significant increase in the reduction in misclassification of high severity as low from 909 to 791. The kappa value is also found to be improved.

```
> confusionMatrix(randomForest_pred, NE_USaccidents_test$Severity)
Confusion Matrix and Statistics

          Reference
Prediction High  Low
      High  673  263
      Low   791 7067

               Accuracy : 0.8801
                 95% CI : (0.8732, 0.8869)
    No Information Rate : 0.8335
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4953

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.45970
            Specificity : 0.96412
         Pos Pred Value : 0.71902
         Neg Pred Value : 0.89934
             Prevalence : 0.16648
         Detection Rate : 0.07653
   Detection Prevalence : 0.10644
      Balanced Accuracy : 0.71191

       'Positive' Class : High
```

Figure 15. Confusion matrix for the improved Random Forest model

Also, to explore more on the explanatory variables used in this model, the graph of variables of importance is created as shown in figure 16. It is observed that the variable "Duration" has the highest influence on the response variable "Severity". With all these, we can conclude that this prediction model is the best of all the models for this dataset.



,

Figure 16. Dotchart of variable of importance

4.5 Analysis of the top 5 variables of importance

From all the above models, it is observed that the top variables that influence the severity of the road accidents are Duration, Start_Lat, Distance.mi, Side, and Wind_Speed.mph.

Duration – The top variable of importance to predict severity is duration. Hence, we analyzed the frequency of duration of accidents and their severity levels.

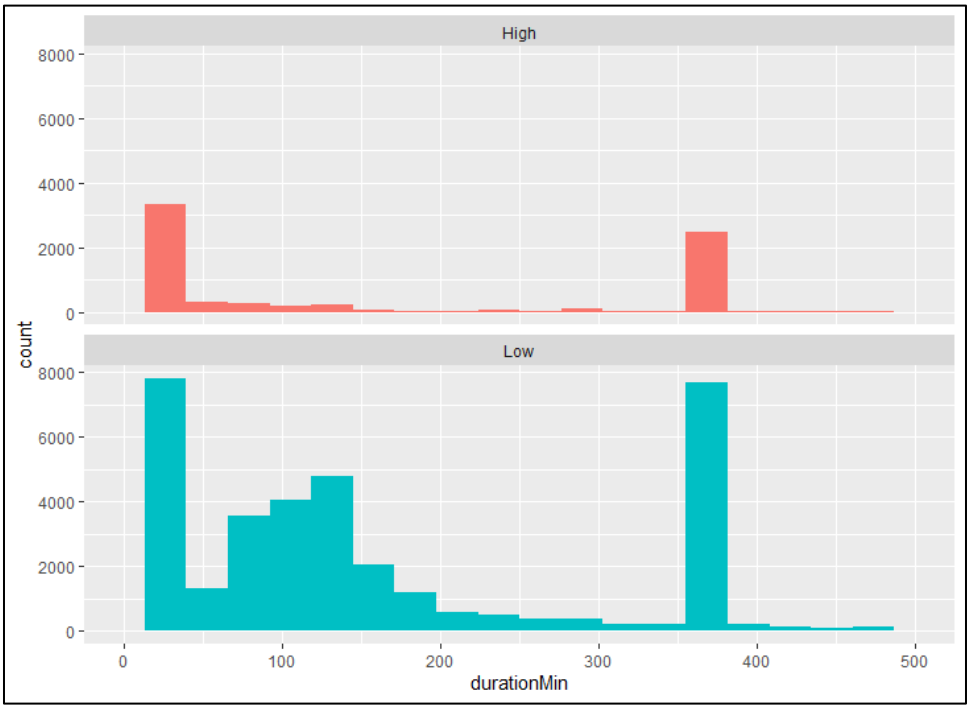


Figure 17. Histogram of duration (in minutes) for the severity levels 'High' and 'Low'

Many accidents have lasted for 30 minutes (½ hour) and 360 minutes (6 hours) in both high and low severity levels equally (figure 17). This shows that duration has no high correlation with severity. The high duration of the incident could be due to various reasons like delay in tow truck driver clearing the vehicles, waiting for clearance from insurance companies, weather conditions, etc. Hence, severity is not solely dependent on the duration of the accident.

Start_Lat (Location) – Latitude and longitude are used to represent geolocation. It is observed that location has the highest relative influence on the severity and hence with this information, we can classify the high severity accident-prone roadways in the New England region. We created a heatmap of the locations only with high severity.

From the map in figure 18, we can see that Connecticut has the highest accidents with high severity, the second highest in Massachusetts especially around the Boston area, and the

third highest being Rhode Island. The states of Maine, New Hampshire, and Vermont have very few high-severity accidents. This could be since the first three states have metro cities with more commercials, organizations, and heavy traffic. People from neighborhoods of the metro cities commute to these locations regularly for work.

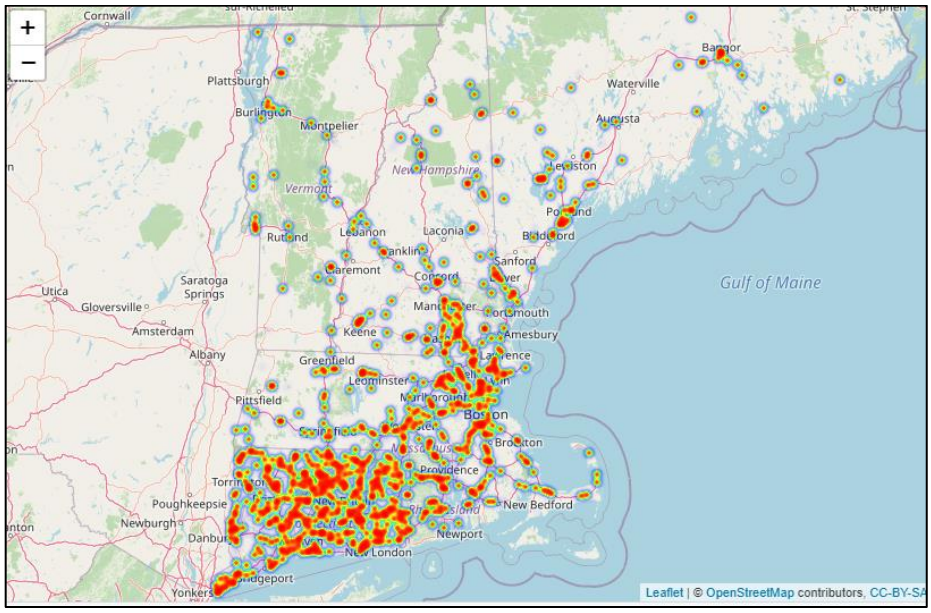


Figure 18. Heat Map showing high severity locations in New England

Distance – Recorded in miles is the distance of traffic congestion due to an accident. This has a highly positive correlation with the severity of an accident.

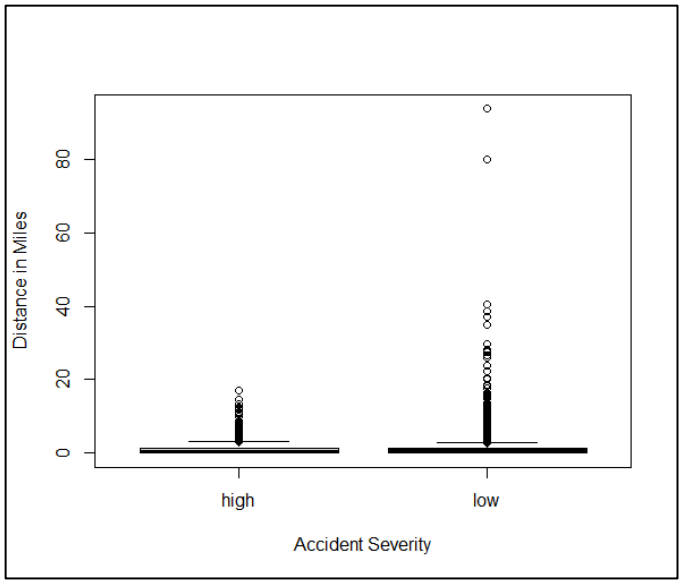


Figure 19. Boxplot showing distribution of severity levels w.r.t to the distance

From the above accident severity versus distance plot (figure 19), we can observe that most of the accidents the length of the road extent affected by the accident is between 0 to 3 miles as the mean and median value of the distance are 0.9921 and 0.6020 miles, respectively. For both high and low severity, the distance due to traffic congestion is the same with some extreme outlier found for accidents with low severity.

Side – Represents the relative side of the street and has an influence on the severity of the accident. From EDA (Exploratory Data Analysis), we observe that most accidents were located on the right side of the driver. This might be because in the USA, left lanes are reserved for faster-moving vehicles and turning left. If there happens to be a slow driver on the left lane, then it might force the driver in the vehicle behind to overtake on the right side which is designated for slow lanes and the driver's blind spot from the car proves to be evident for resulting in more accidents on the right side.

Temperature – The weather conditions do affect safe driving. The visualization below shows occurrence of accident severity for various temperature (F) among the states of New England region. In Massachusetts, most of the high severity accidents occur in sub-zero temperatures.

From figure 20, we can see that for the states Maine and New Hampshire, we could hardly observe any high severity accidents. For the states Connecticut, Rhode Island and Vermont, high severity accidents are recorded when the temperature is above 75°F. Surprisingly, for these 3 states no accidents were recorded at sub-zero temperatures.

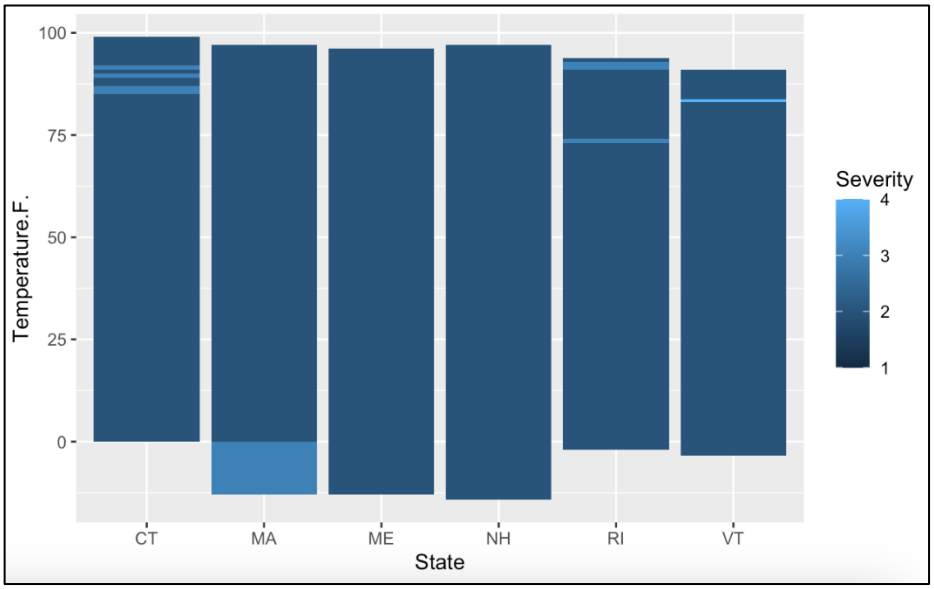


Figure 20. Bar plot showing severity levels in New England states across different temperatures

Year of accident – The number of accidents observed every year based on the severity levels was analyzed. The below figure shows the frequency of accidents across the years 2016 – 2022.

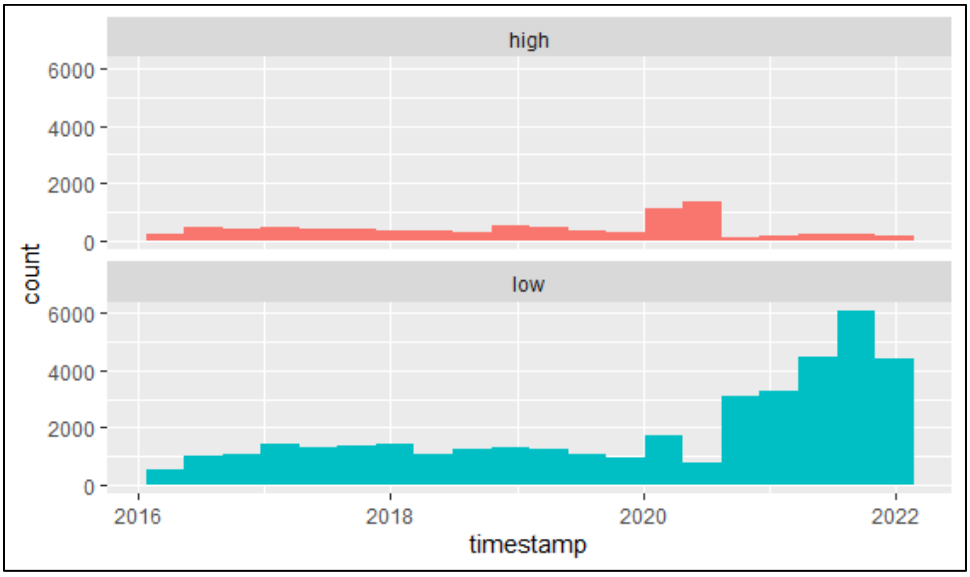


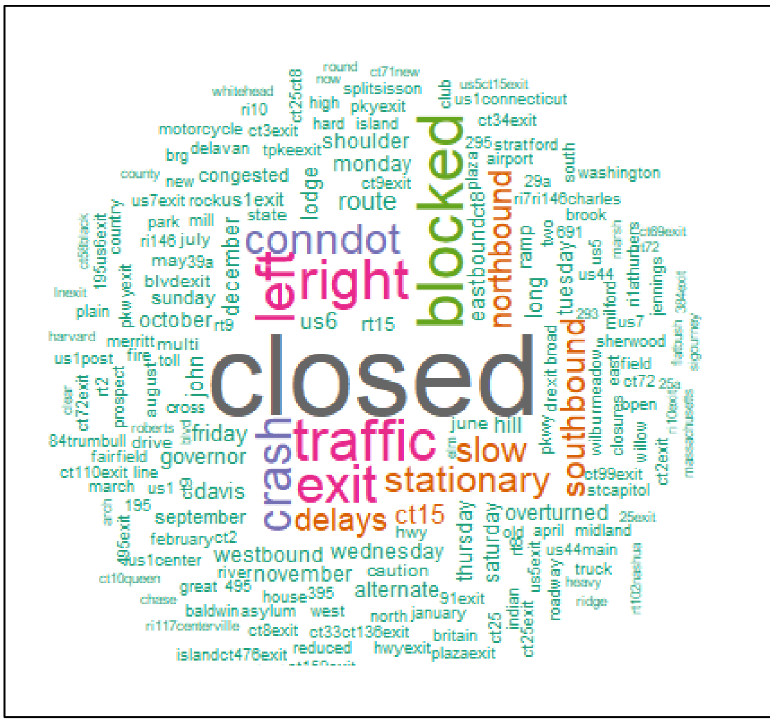
Figure 21. Frequency of accidents over the last 5 years in New England categorized by severity

The above graph (figure 21) clearly shows that the number of accidents increased between the years 2020 and 2021. During the year 2020, we see a sudden surge in high severity accidents which have again decreased later second half of 2020 but the number of low severity accidents have been increasing since 2020. This was an interesting observation for us because we expected the overall accidents to be lower during the pandemic and lockdown.

This intrigued us to explore the reason for the increased accidents during the pandemic years. The National Highway Traffic Safety Administration (NHTSA) of the Department of Transportation reported that traffic fatalities increased 18.4 percent in the first half of 2021 compared to the first half of 2020 (Taken from gao.gov, 2022). According to a few experts, this driving behavior is likely to reflect widespread psychological feelings like isolation, loneliness, and depression. It is also said by a researcher that, along with high road accidents there are also other pandemic trends like the skyrocketing sales of alcohol, drug overdoses, and increase in homicides. This could also be a reason for road accidents that many people would have been driving under the influence of alcohol or drugs, not wearing seat belts etc.

5. Text Mining

The variable "Description" in the data set gives a unique description for every incident. Text mining is performed on this unstructured variable to extract meaningful information and identify any patterns with respect to the severity of the accidents. The aim is to identify if there are any specific locations that has high severity accidents, patterns on the days or weeks or months when the accidents occur, any impact due to the traffic or the road closures on the severity, etc. The package "*tm*" is used to perform text mining and "*wordcloud*" library is used for generating a cloud of the most frequently used words.




```
> head(d_ne, 25)
```

	word	freq
closed	closed	16223
blocked	blocked	8250
traffic	traffic	7351
right	right	6870
left	left	6667
exit	exit	6386
crash	crash	5581
conndot	conndot	4550
stationary	stationary	3776
slow	slow	3458
southbound	southbound	3046
northbound	northbound	2966
delays	delays	2905
ct15	ct15	2168
route	route	1608
friday	friday	1494
long	long	1455
december	december	1425
john	john	1420
governor	governor	1384
lodge	lodge	1366
davis	davis	1360
us6	us6	1352
eastbound	eastbound	1349
westbound	westbound	1280

Figure 22. Word cloud representation of high-frequency words (15a - on the left); Tabular view of top 25 frequently seen words (15b - on the right)

Figures 22a and 22b shows that the words *closed, blocked,* and *traffic* are the top three words that have been frequently mentioned, indicating the types of impacts caused on the road as a result of accidents. The words *right* and *left* refer to the lanes where the accident occurred. The right lane is more frequently used than the left, indicating that more accidents occur in the right lane. The term "*conndot*" refers to the 'Connecticut Transportation Department', which means that most of the accidents occurred in Connecticut. Next, we see that the words *northbound* and *southbound* are frequently used, indicating that those exits are more likely to be involved in an accident than *eastbound* and *westbound* exits. We were intrigued to see names like "John," "David," and "governor" appear frequently in accident descriptions, and we wanted

to learn more about what those words meant. We discovered that those words came from the "Governor John Davis Lodge Turnpike," a 128-mile-long controlled-access highway in Connecticut where the vast majority of the accidents occurred. Months such as *October, November,* and *December* are frequently used, indicating that the fall and early winter months saw a higher number of accidents.

Figure 23 depicts the most frequently used words in low and high severity levels. The regression line in red separates the words that were exclusively used in a particular severity level, and words that fall on or near the line were equally used in both descriptions of both severity levels. It is observed that the type of road accident impacts like “traffic” and “delays” lead to low severity levels. The word “closed” means that the road has been completely closed for commute and the word “reduced” means the number of lanes was temporarily reduced due to an accident. These two words “closed” and “reduced” are exclusively used for high severity which is more obvious than minor impacts like “traffic” and “delays”. Most accidents of both severity levels occur near exits. It is also clear that the words "hills," "valley," and "farm" are used with high severity as they could be popular tourist destinations or area that are difficult to clear in case of an accident.

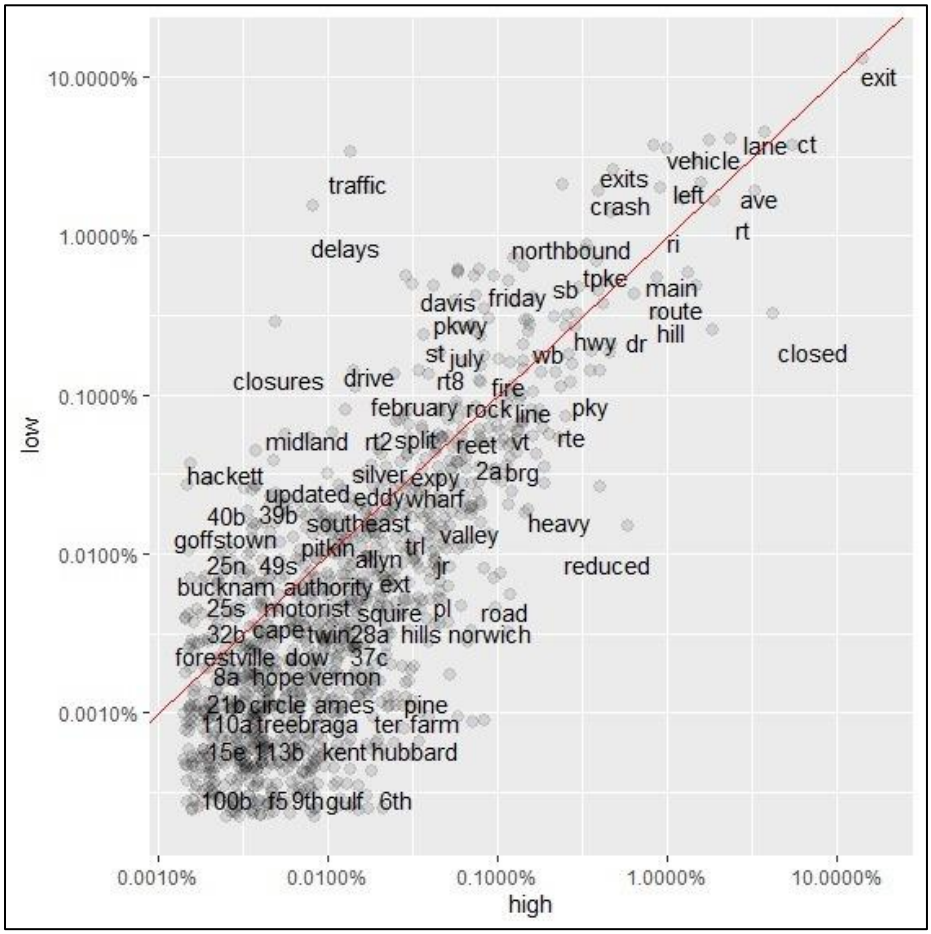


Figure 23. High-frequency words classified for low and high severity accidents

6. Recommendations from Data mining

Based on the above findings we can recommend that more traffic control cameras be placed in locations with high severity. It is recommended to have a safe ride while travelling in those accident-prone zones. The accidents can be reduced by installing warning sign boards and prescribing speed limits on the identified accident-prone zones. To avoid traffic congestion due to an accident, it is recommended to suggest people use live traffic data with their GPS navigation. Clearly mentioning the entry and exit points in highways helps the driver to make quick decisions. Educating drivers about proper lane usage like the right lane for cruising and the left lane for passing and the overtaking rules.

In addition, weather conditions do affect safe driving and it is advised to check for any adverse weather forecast before planning for a road drive. The government can also recommend the use of public transportation by people whenever possible while commuting to work, which might avoid heavy traffic congestion and avoid last-minute rash driving.

7. Conclusion

This analysis aim was to identify the highly influential variable that impacts the severity of accidents with which we can predict the accident severity in the future. It is found that the features such as *Start_Lat* which is the location of the accident, *Distance.mi.* which given the distance between the start and end of a roadblock due to accidents, *Side* which tells the side of accidents (right/left), *Wind_Speed.mph,* and *Temperature* are having a relatively high influence on the severity when compared to other variables. It is also found that the duration of the day has no significant impact on severity though the accidents are higher in daytime than in the nighttime. From the three prediction models, we can conclude that the Random Forest model is better compared to the other two models. The accuracy of the random forest model is **88%** and the kappa is 0.49 (moderate agreement).

From the above analysis, we identified that the hotspot locations in New England with high severity levels are Connecticut, Massachusetts, and Rhode Island. Though these three states together have reported more accidents, the "**Governor John Davis Lodge Turnpike**" highway in Connecticut alone has reported nearly 10% of the total accidents in New England.

As we discovered that driver's behavioral trends can be a cause of an increase in the number of accidents and affect their severity, we would like to collect additional data on driver details such as age, vision, and behavioral factors such as high stress, anxiety, the influence of alcohol, drugs, and so on, to study further, which may help in more accurate predictions of the severity levels of road accidents. Severity could also be impacted by the severity of injuries due to an accident.

7. Reference

Camacho, A. (2017). *Classification tree using rpart (100% Accuracy).* Kaggle. https://www.kaggle.com/code/jhuno137/classification-tree-using-rpart-100-accuracy/report

Narkhede, S. (2018, May 9). *Understanding Confusion Matrix.* Toward Data Science, https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Probst, P., Wright, M. N., & Boulesteix, A. (2019). *Hyperparameters and tuning strategies for random forest.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3). https://doi.org/10.1002/widm.1301

During COVID-19, Road Fatalities Increased and Transit Ridership Dipped. (2022, January 25). U.S. GAO. https://www.gao.gov/blog/during-covid-19-road-fatalities-increased-and-transit-ridership-dipped