CSP 571 - PROJECT REPORT

Amirreza Eshragi - A20451901
Anneke Soraya Hidayat - A20406957 (Group Leader)
Namitha Venkataramanan - A20453185
Sourav Yadav - A20450418

TABLE OF CONTENTS

Traffic Accident Analysis

Abstract
This project addressed the analysis on traffic accidents in the US span from 2016 to mid 2020. Our main objective is to identify which features explain the degree of severity right after an accident happens. We want to emphasize that the traffic condition means the traffic flow post-accident. Low severity means the traffic flow is less affected, and high severity means that the traffic flow is blocked. Our data exploration shows that The South and The West region of the US has different distribution of accidents on the highway VS non-highway. The accident ratio by states also varies depending on the population and total car on the road. Our predictive analysis shows that three main features that classifies whether an accident is less severe and more severe: (1) If an accident happens on a highway, (2) If an authority broadcasts a TMC code, and (3) if there is a traffic signal(s) nearby the location.

1 Overview

Our main focus in this project is to analyze the **traffic accidents** across the US. The objective of the project is to find out the main factors that **cause** these accidents and **how** we can tackle this problem. Although our output may not be a descriptive solution, we hope that our output can be useful for government use, public interest, and to raise the awareness of the industry. We narrow down into US/North America territory since it's easier to create boundaries on our analysis into one continent.

2 Data preprocessing

2.1. Data issues

The traffic dataset contains text, categorical, binary, numerical, and timestamp values. The issues that we are facing in this dataset are the data scale, missing data, non-standardized data and data encoding. Thus, to reduce the dimension of the data and standardize so that it’s easier for our analysis, we’re performing a different approach for our data preparation which is described in this and the next chapter as well.

2.2 Missing Data

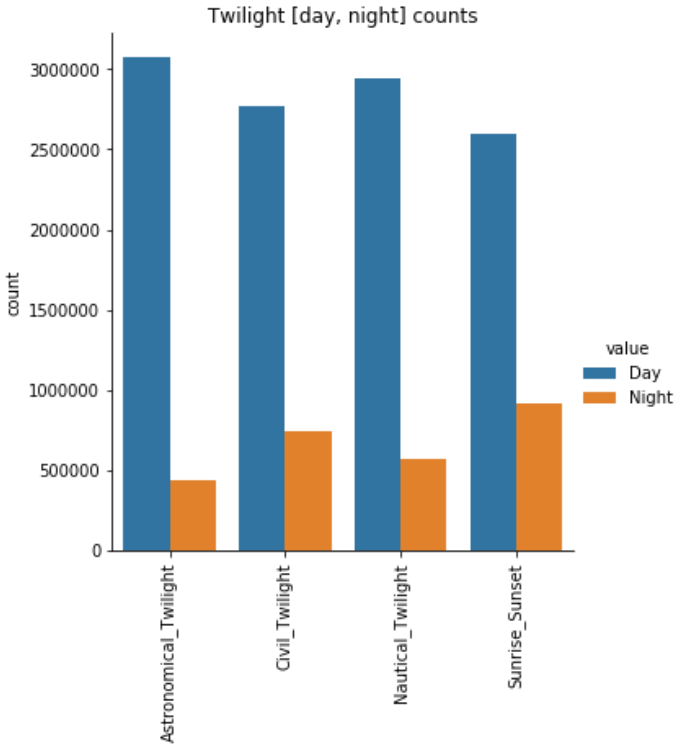
The dataset has a number of features that contain missing values. We have only taken into consideration the variables that are required for our analysis to determine our final objective which is to predict the severity of the accidents in the US. For instance, variables like ‘End_Lat’ and ‘End_Long’ have missing values when the distance of road affected by accident is very small thereby implying that the Start and End location would almost be the same, and hence this variable can be removed. TMC has 1 million missing values, which is approximately 33% of the dataset and hence we remove that variable from our analysis. For variables that have NA values, we hot encode them to be used in our analysis.

2.3 Assumptions/adjustments

Our dataset contains 49 columns that has binary, categorical, timestamp, numerical values and encoding. Thus, it gives us a huge pool of predictors with various scales and values. This project outlined as a journey to calibrate the predictor, such that we are able to interpret the result in a simple representation as possible. Although the explanatory analysis on each

assumption will be described on the next section, here's the list of assumptions and adjustments that are applicable on our project.

1. Consistency across similar features
   The features that belong under this assumption are "Civil_Twilight", "Nautical_Twilight", "Astronomical_Twilight", and "Sunrise/Sunset". These features have identical values that identify whether the traffic accident happens at day/night time. Each of the features assign the value of day/night based on the different basis, such as sunrise/sunset is based on the sun time. However, these features may not have significant influence when these values are different to each other. To simplify our dataset, we ought to get the values that have consistent value across these variables. Meaning that these four values have to match. We drop those data that have no consistency between these features. We can see that the distribution of these features are similar as shown in the chart below.



2. Drop variables that have unique values
   The feature columns under this assumption are "Source", "Country, and "ID".
3. Drop variables that have granular attributes that may share similar attributes with other variables
   The feature columns under this assumption are "Airport_Code", "Side", "Number", "Weather_Timestamp", "Timezone", "Precipitation(in)".
4. Drop values that only has one value, or NaN.
   The feature columns under this assumption are "End_Lat", "End_Lng", "Turning_Loop"

3 Data Exploration and Analysis

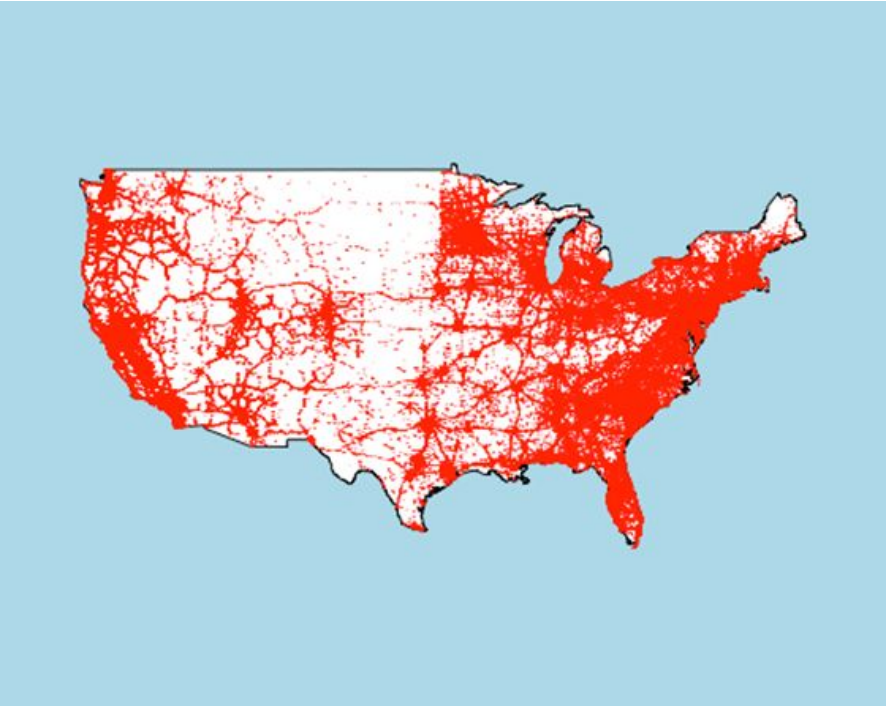
3.1 Data and Feature Exploration

This subsection covers the general data and feature exploration based on the provided predictors. Started from the general information on the spatial distribution of the traffic accidents, the distribution of the severity, and the features detail. The general information of the dataset is as follows:

Raw sample	3,513,617
Preprocessed sample	3,513,480

From here and below, we are exploring the dataset with the preprocessed samples.

3.1.1 Accident Distribution



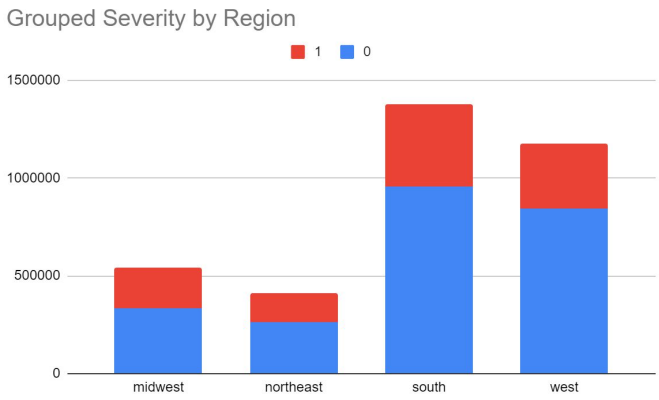
Here's showing the accident density map in the figure above. The information is based on the "Start_Lat" and "End_Lng". From the visualization above, the region that has traffic accidents are mostly concentrated on the west coast and east coast, with more dense toward the Northeast and Southwest. To illustrate the traffic accident rate in more detail, and without bias by high number of traffic accident in particular state, we are taking external information such as US population by state[1] and US car number by state[2]. This information is useful to give illustration on the traffic accident rate, to identify which state has the highest rate, not the number of accidents.

[1] https://www.infoplease.com/us/states/state-population-by-rank
[2] https://www.statista.com/statistics/196010/total-number-of-registered-automobiles-in-the-us-by-state/

	Traffic Accident		Accident Rate by Population		Accident Rate by Total number of car	
1	CA	816804	SC	0.034	SC	0.095
2	TX	329284	OR	0.021	OR	0.061
3	FL	257974	CA	0.021	UT	0.055
4	SC	173277	UT	0.016	CA	0.054
5	NC	165955	NC	0.016	NC	0.049
6	NY	160787	OK	0.015	OK	0.046
7	PA	106787	MN	0.015	LA	0.044
8	IL	99691	LA	0.013	MN	0.041
9	VA	96075	NE	0.012	TX	0.040
10	MI	95983	FL	0.012	NE	0.035

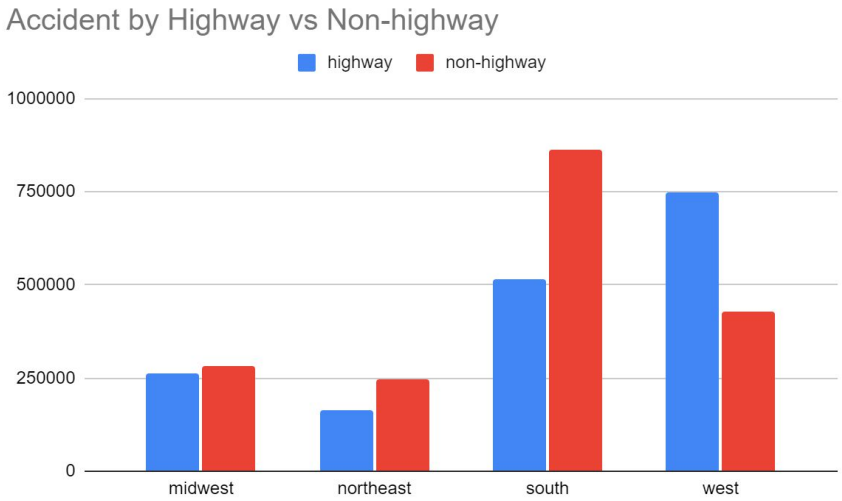
The number of the US population and car by state are given in the Appendix, along with the source and the date of the obtained survey. This number might not be accurate by the date of this project, however, the rate can give us how the number of traffic accidents is not identical with the accident rate. By the number of traffic accidents from this dataset, it is obvious that CA has the highest number of accidents. However, we shall keep in mind that CA also has a high density of population. The accident rate can give us indication which state actually has the highest *accident rate* compared to the total population and number of cars in the given state. The most *dangerous state* of traffic accidents as shown in the Table above is OR (Oregon). As we see that Oregon does not have a high number of traffic accidents and is not listed as a top 10 state with the number of traffic accidents. However, despite not having a high number of traffic accidents, it ranked 2 as the highest accident rate both by population and total number of cars. Thus, we shall keep an eye to these states for our analysis.

To go further on our data exploration, keeping the number of areas as simple as possible can help us to get more understanding at a high-level. Thus, instead of have 50 states as the region boundaries, we ought to see the density by dividing these states into four region[3]: Northeast, Midwest, South and West.



The South region by far has the highest number of accidents. This chart creates a new question to ask: How many of these traffic accidents happen on highways?

[3] https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States

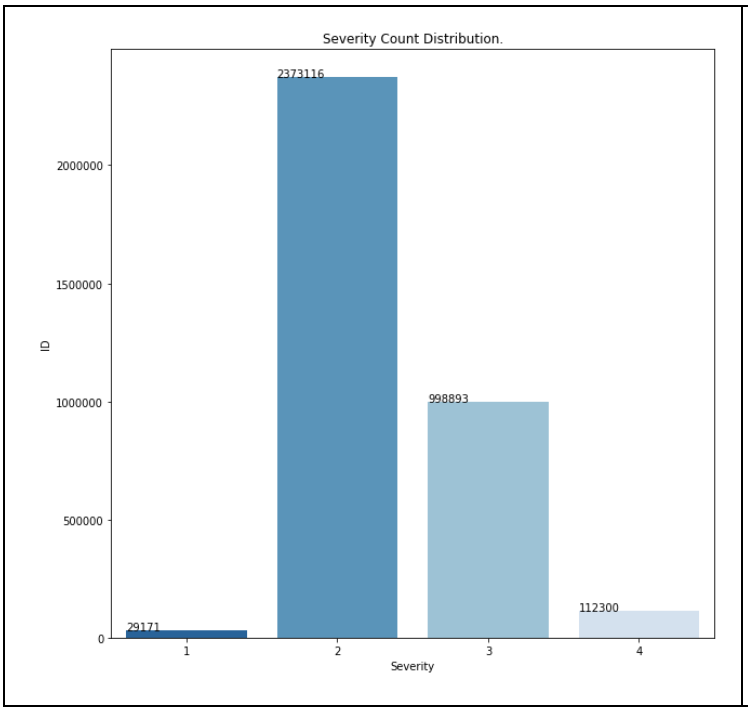


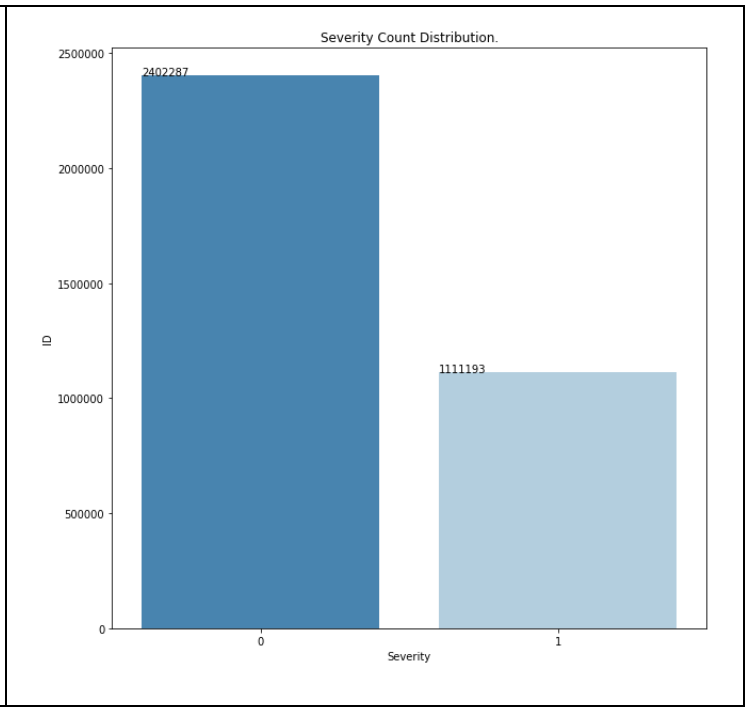
Interestingly that South and West region both have the total opposite. The Midwest and Northeast region has the same count of traffic accidents both on highways and non-highways. The South region has the most accidents on non-highways and the West region is on highways. Taking highway information into account, this can help us to separate the data between these regions.

3.1.2 Target Value : Severity

The target value for this project is the 'Severity' in which describe how bad the traffic was affected by the accident. One thing in mind that the accident in this dataset is presumably already happened. The 'severity' value does not illustrate how severe the accident, but illustrates how severe the **traffic** caused by the accident. By keeping this in mind, this can guide us and understand better on how the traffic behaves when an accident presents.

Severity has four integer values: 1,2,3 and 4. 1 indicates that the traffic is not affected by the accident (or less affected), and 4 indicates that the traffic is highly affected by the accident (may require further action from the authority). Below is the distribution of the severity



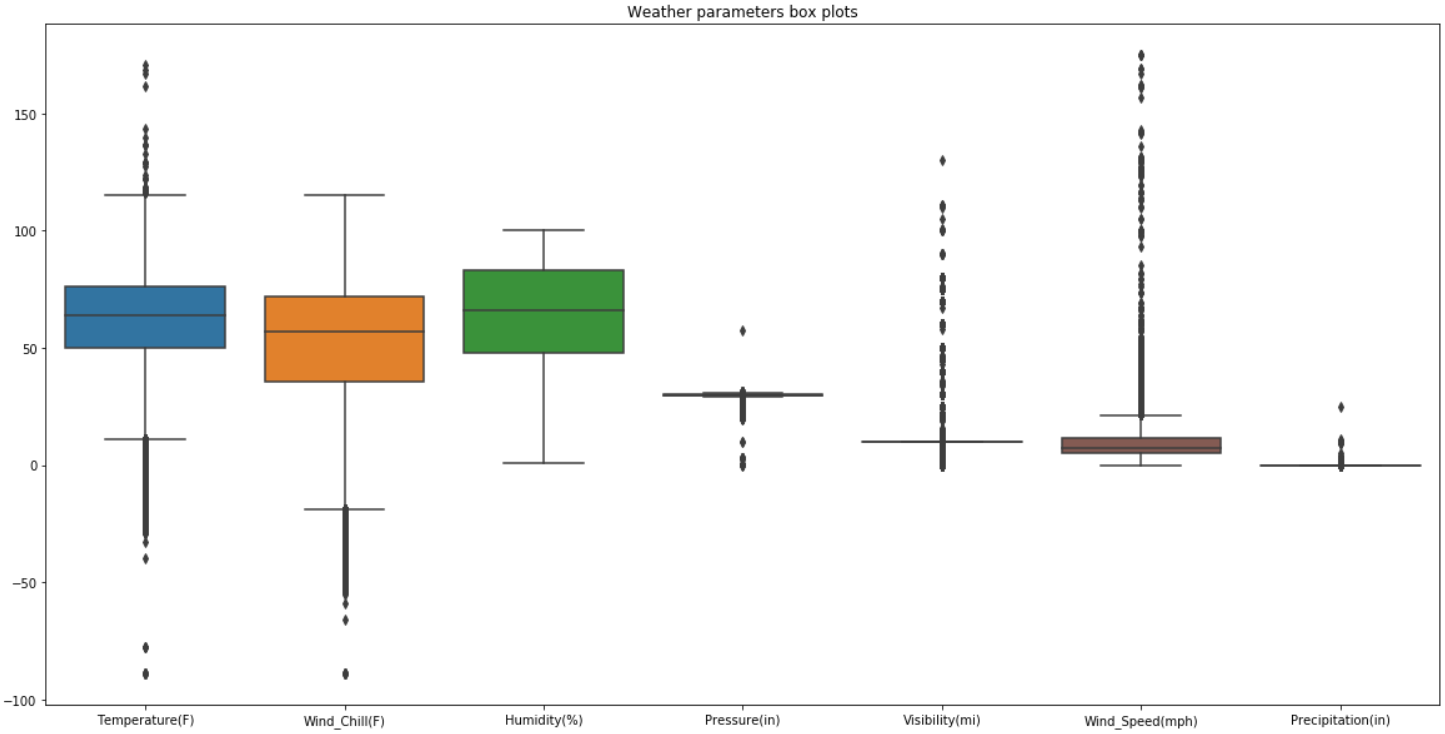


From the illustration above, we can see that 67% of the data has severity of 2. Most of the accidents somewhat affect the traffic situation in general. With severity 1 and 4 are the lowest among the class. Meanwhile it is important to classify the traffic severity as four categories, for our analysis purposes and feature analysis, we will map these severity categories into binary classes of 0 and 1. The main reason behind this is that the multinomial regression provided by R does not handle multi-class regression as we expected. The ideal situation is to classify the category as "1 vs the rest", "2 vs the rest", and so on. However, the multinomial regression in R works by taking one category as reference and holding it against the other categories, for example taking severity "1" as reference, it runs the regression as "1 vs 2", "1 vs 3", and "1 vs 4". This scenario is not ideal for our analysis since the feature importance of severity 1 will not be calculated properly.

To tackle this problem, and also for our analysis purposes, we are mapping the severity classes of "1 and 2" as 0 and "3 and 4" as 1. After mapping the severity we have 68% of the data belongs to 0 zero class and 32% belongs to 1 class. Overall the data distribution follows the normal distribution skewed to the right.

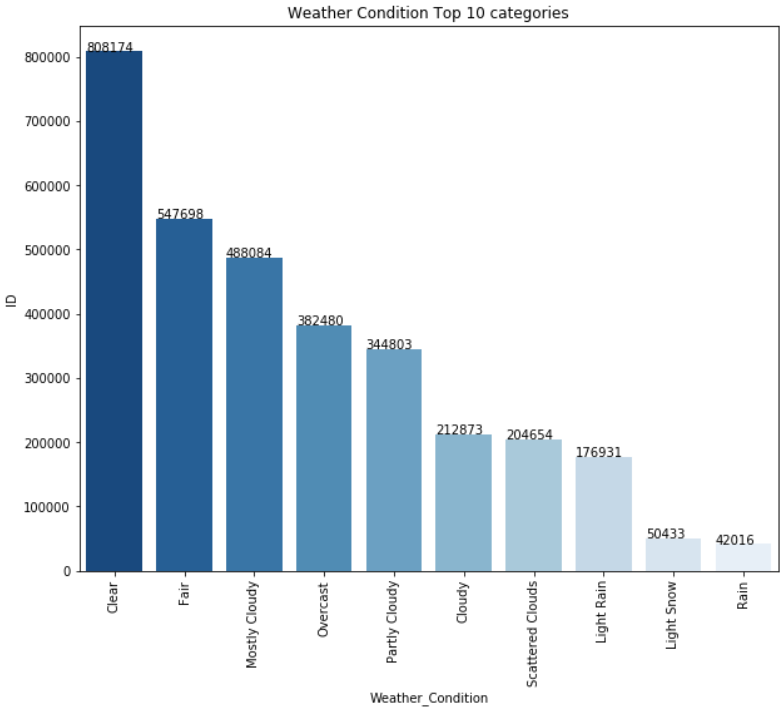
3.1.3 Weather Features

Weather features are considered as close features through the severity since it may highly correlate with the traffic condition given an accident occured. Most of the weather features are numerical variables except for Weather_Condition. First let's observe the weather features: Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Validity(mi), Wind_Speed(mph), and Precipitation(in). Below is the boxplot that illustrates the scale of each weather features.



It is important to note that Wind_Speed(mph) has outliers that are above 200 mph. To standardize the range, we exclude those samples with extreme values of Wind_Speed(mph). In the case of Precipitation, the 1st, 2nd and 3rd quartile lie in the similar range. Although the scale might vary across the weather features, we decided to discard Precipitation from the weather features.

Quick analysis on Temperature(F) and Wind_Chill(F) are having a similar scale, since the unit is in Fahrenheit. The support for this analysis will be provided in the PCA section.



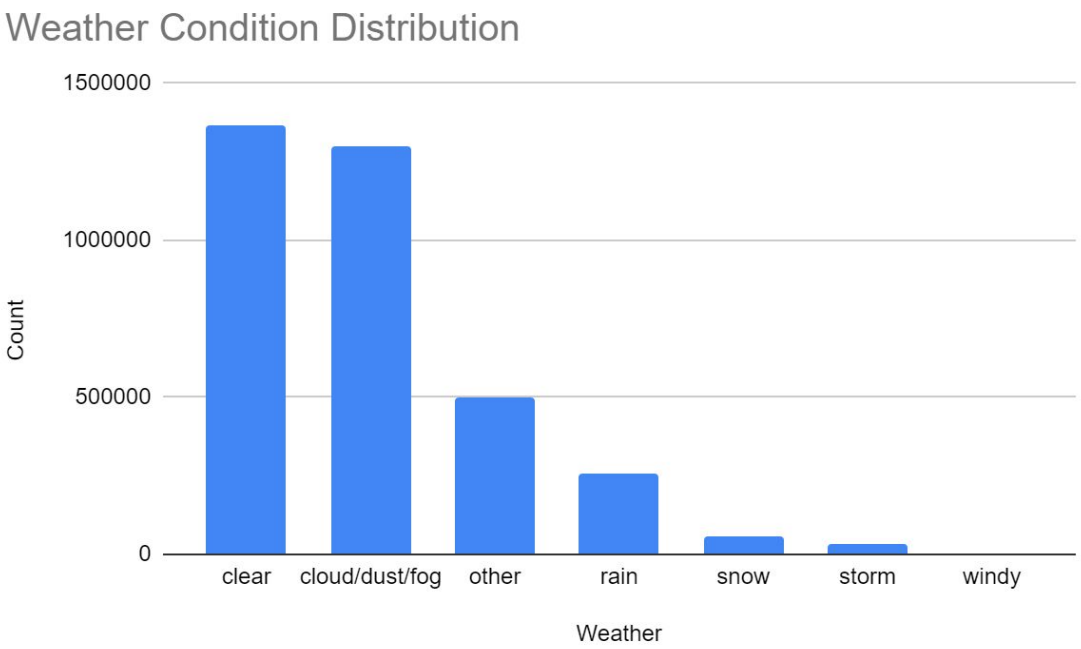
One of the tricky variables to preprocess is Weather_Condition. Weather_Condition has 127 unique strings. This feature is similar to description, except it describes the weather in a phrase. Some snippets on the Weather_Condition value are given below.

Overcast
Light Snow
Light Snow
Scattered Clouds
Overcast
Overcast
Partly Cloudy
Clear
Light Snow
Overcast

To simplify the variance of the weather condition, we preprocessed the weather condition by bucket the values into smaller categories. We apply a simple rule described as follow:

```
If [“storm”, “thunder”, “smoke”, “tornado” in Weather_Condition:
        Assign “storm”
elif [“clear”, “fair”] in Weather_Condition:
        Assign “clear”
Elif [“rain”, “drizzle”] in Weather_Condition:
        Assign “rain”
Elif [“cloud”, “dust”, “fog”] in Weather_Condition:
        Assign “cloud/dust/fog”
Elif [“snow”, “ice”] in Weather_Condition:
        Assign “snow”
Elif [“wind”] in Weather_Condition:
        Assign “windy
Else:
        Assign “other”
```

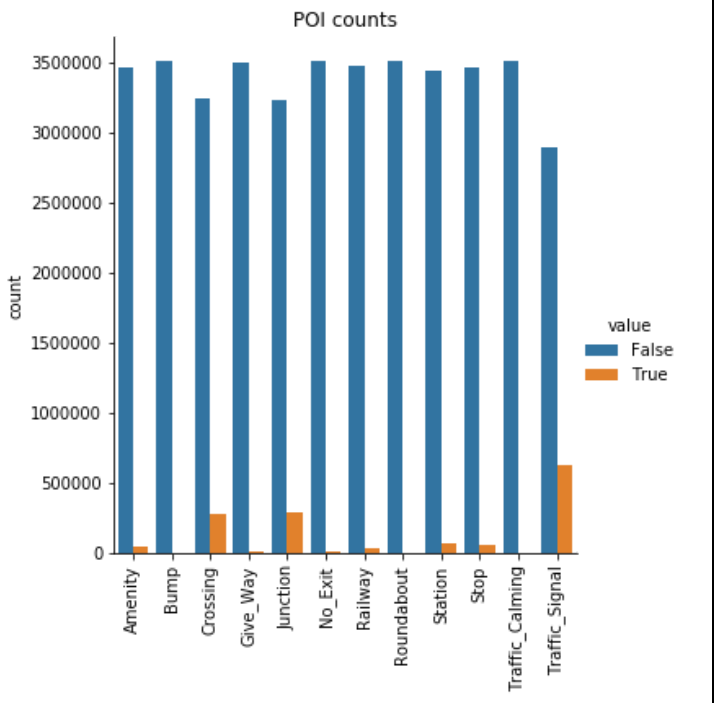
We ended up having seven of values in which we can treat as factors

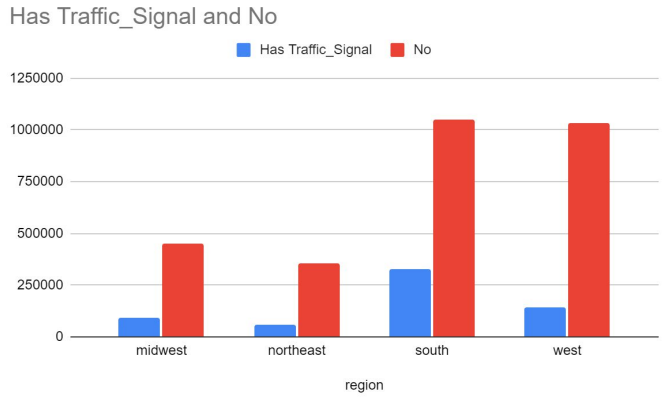


Most of the traffic accidents in the US mostly happen on “clear” and “cloud” weather. From this data, it is feasible to see that if we separate the data with this feature, such that those accidents that do not happen on “clear” or “cloud”, it may give information for the extreme severity of the traffic accident.

3.1.4 POI

POI (Point of Interest) are features that have binary values. Each of POI features indicates whether POI exists near the location of the traffic accident.





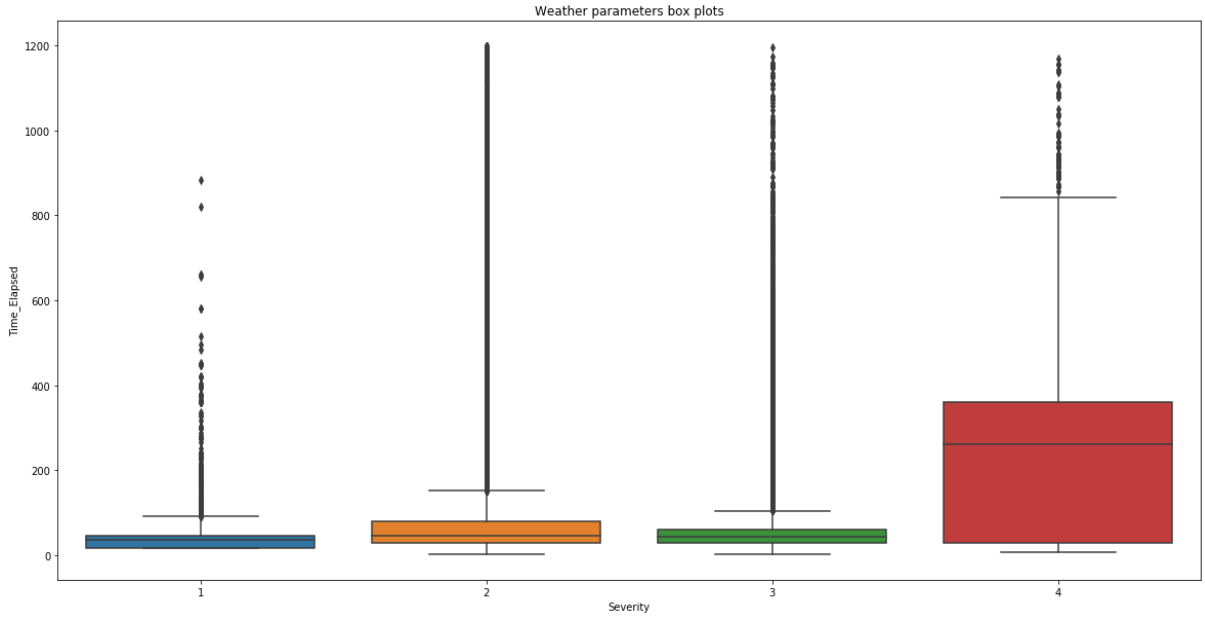
In the chart above, we can see that these features are sparse, that most of the traffic accidents that happen does not have POI located nearby. Among these features, Traffic_Signal, Junction, and Crossing are the most common POI. The chart on the right hand side shows that the number of traffic accidents that has Traffic_Signal nearby follows the accidents distribution. From this output, we can conclude that for predictive or analysis modelling, the POI features may not be a good predictor. Instead, POI can be used as a grouping variable for further analysis.

3.1.5 Traffic Features

The traffic features cover TMC, Distance(mi) and Time_Elapsed. TMC (Traffic Message Channel) covers the radio code that exchanges between authorities to describe the traffic situation. The TMC code that is included in this dataset is provided on the appendix. Overall, there are 33% of data points that have missing value on TMC. There is no way for us to map or impute those values into TMC code, thus we won't use our predictive model. However, the use of TMC code can provide us with an illustration of the accident itself. The description on each code has similar and lots of overlapping values. However these information can help us to analyse the traffic condition even further. Some samples of the TMC code are shown below for illustration.

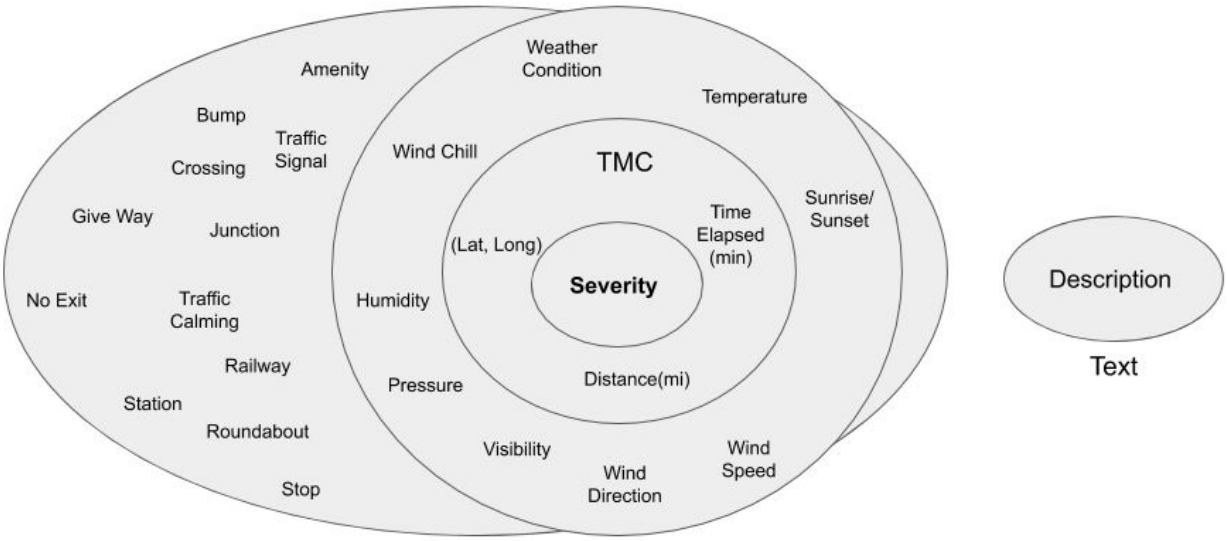
TMC Code	Description	Freq
201:	accidents	2080341
241:	(Q) accident(s). Right lane blocked	249852
245:	(Q) accident(s). Two lanes blocked	40338

We attempted to categorize the TMC codes with hope we can extract some useful information on the accidents. However due to a large proportion of missing value, we pass this feature analysis and move to the next traffic feature.



3.2 Feature selection and Analysis

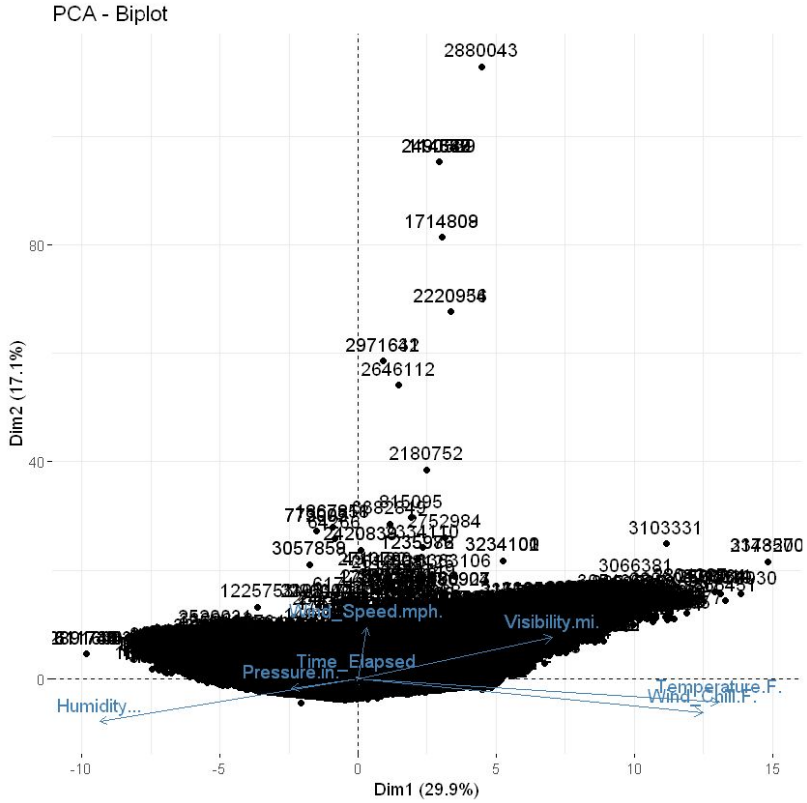
In general, we can sketch our traffic accidents hierarchy as the figure below.



We have Severity as the target variable located at the center, followed by the traffic features, weather features, POI, and description. The text description is grouped as independent since it does not carry information to predict severity, but only as additional variables that *explain* the severity. Thus, the description feature won't be included in our model.

3.2.1 PCA

To see the correlation between numerical features on the whole dataset, we performed PCA

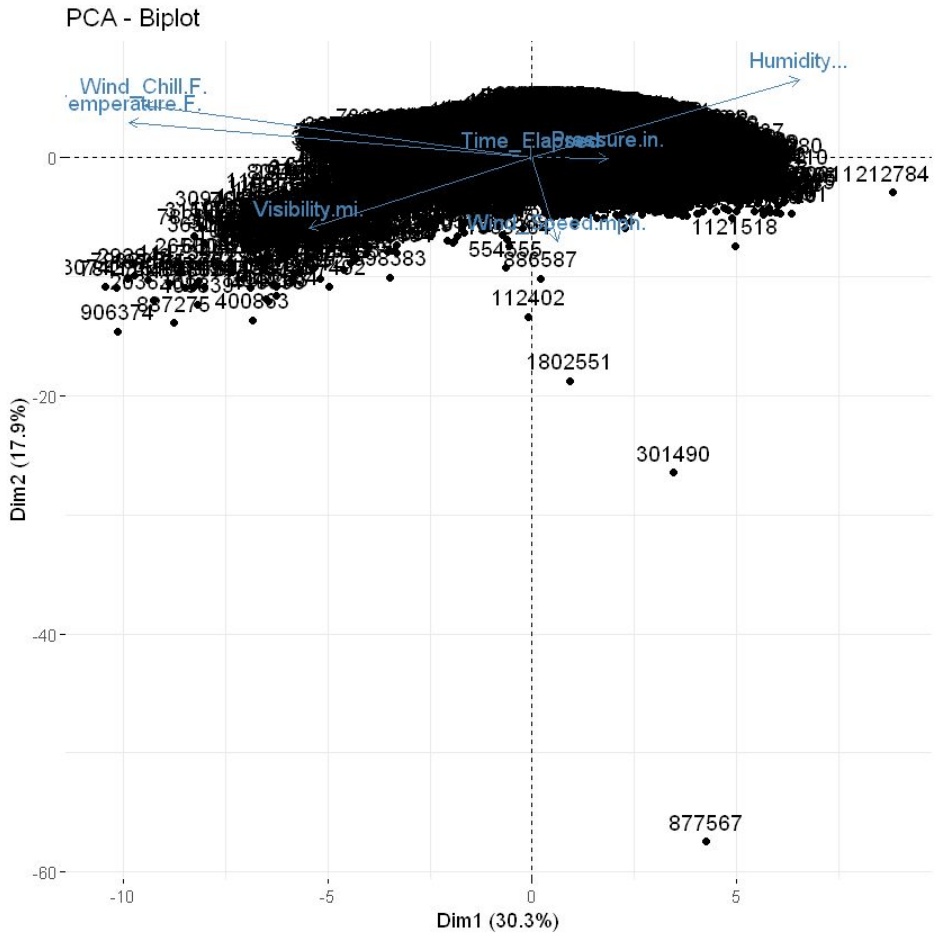


As shown in the figure above, the first principal component has 30% variation explained. Temperature and Wind chill are strongly correlated with each other based on the principal component direction. In general, both of the features should correlate with each other since the temperature happens in the air and wind_chill will have a minimum difference. Other features that have negative correlation are humidity and visibility, as shown the opposite direction in the biplot. By far, Temperature, Wind_Chill, Humidity and Visibility strongly influence the first principal component since it ranges across the first axis. On the other hand, Wind speed influences the second principal component as the direction is across the second axis.

4 Model Implementation

This section shows the model implementation on the traffic dataset. In the beginning of our data exploration, we ought to build the model on the whole dataset with 3 Million of inputs. Despite the data partition of training and test, our personal computer can't handle the massive computation by the data size. Suppose that we have three millions samples, and 10 features on our data frame. The computation needed is at least 10-fold of the sample size. Thus, we have our lower bound of computation by $\Omega(N \times p)$ where N is the total samples and p is the number of predictors.

There is no general rule of thumb on how many data samples can fit into an 8GB RAM computer, however, with an approximate number of samples by 300,000 and we have a lower bound of 3 Million, it is enough to run the model. We opt to choose representative samples from the general population by an upper bound of 300,000 samples. Given four different regions in the US that we assigned into the dataset, we draw 50,000 samples from each region. It yields 200,000 samples as our sample size. To see the representativeness of the samples, we looked at the PCA on the same features on the whole dataset to evaluate how similar the samples behave as the general population.



The biplot shows the same feature response as the whole dataset. Thus, under this assumption, we implement models to do inference analysis on which predictors have significant influence on the severity of the traffic accident.

4.1 Model detail

Our target is to classify the severity of the traffic accident mapped into binary response. To evaluate the predictors importance we perform dimensional reduction to see each feature interaction along with how much of the data that can be explained in the lower dimension. Next, we are using glm with binomial family, glmnet for applying feature selection on lasso, decision tree to interpret the rule based, and finally SVM. Our approach on modeling is to see feature selection from different models' perspectives. Glm and glmnet with Lasso are useful to see the predictors / coefficients z-values on the chance of null hypothesis is correct. Decision tree is useful to see the rule path and SVM for performance comparison. For training/testing purposes, we are splitting the dataset into training and testing set by the ratio of 80:20. Each of sampling and data split are using random seed of 42 to ensure the reproducibility.

4.2 Performance Metric and Result

Since we are fitting the model against binary response, we are extracting the confusion matrix as our performance metrics. Among the available variables, we are reporting only the accuracy both on training and testing split. The reason to select this metric is because it is the easiest metric to interpret. We are under the assumption that the class probability threshold is 0.5 and there is always one of two classes as the target. We are not performing event detection in the context that one class is more important than the other. Instead, both classes have meaning and a given input always belongs either to the severity 0 (1 or 2) or 1 (3 or 4).
Below is the full result of our inference modelling.

	Models	# Param	Param		Accuracy	
					Training	Testing
1	GLM	12	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.", "Visibility.mi.", "Wind_Direction",	"Wind_Speed.mph.", "Sunrise_Sunset", "is_highway", "region", "Time_Elapsed", "Weather"	0.6939	0.6917
2		12	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.", "Visibility.mi.", "Wind_Speed.mph.",	"Sunrise_Sunset", "is_highway", "region", "Weather" "Traffic_Signal", "Junction"	0.6974	0.6994
3		11	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.", "Visibility.mi.", "Wind_Speed.mph.",	"Sunrise_Sunset", "is_highway", "region", "Weather" "Traffic_Signal", "Junction"	0.6971	0.699

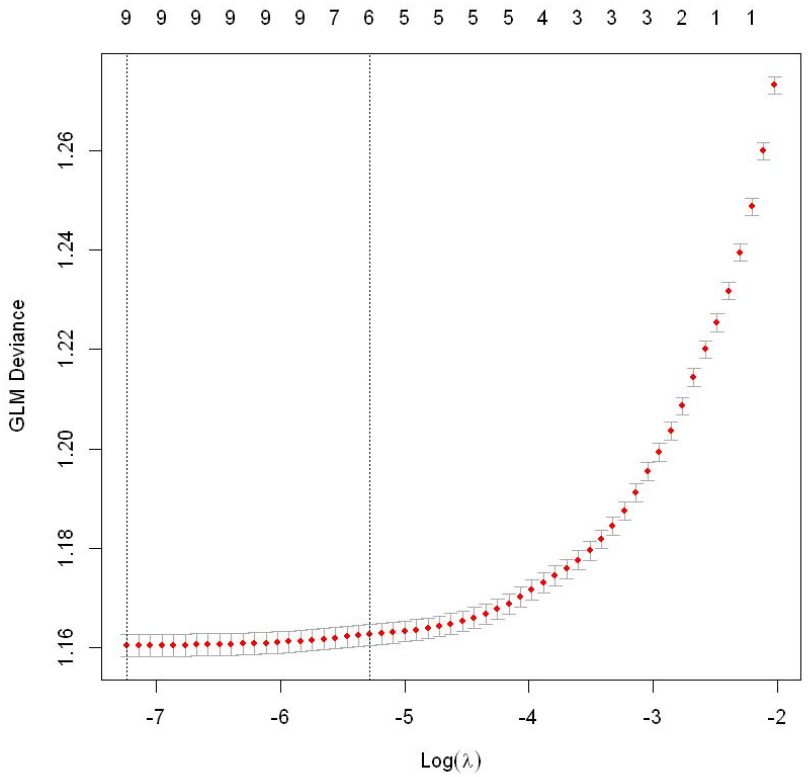
4		11	"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset", "Distance.mi"	"is_highway", "region", "weather" "Traffic_Signal" "TMC."	0.7334	0.7338
5		10	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.", "Visibility.mi.",	"Wind_Speed.mph.", "Sunrise_Sunset", "is_highway", "region", "Weather"	0.6923	0.6915
6		10	"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset",	"is_highway", "region", "weather" "Traffic_Signal" "Distance.mi."	0.6979	0.6991
7		8	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.",	"Wind_Speed.mph.", "Sunrise_Sunset", "is_highway", "region",	0.6954	0.6964
8		9	"Temperature.F.", "Humidity...", "Pressure.in.", "Wind_Speed.mph.", "Sunrise_Sunset",	"is_highway", "region", "weather" "Traffic_Signal" "Distance.mi."	0.6977	0.6979
9		8	"Temperature.F.", "Pressure.in.", "Wind_Speed.mph.", "Sunrise_Sunset",	"is_highway", "region", "weather" "Traffic_Signal	0.6962	0.6971
10	GLM + Lasso	9	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.", "Visibility.mi.",	"Wind_Speed.mph.", "Distance.mi.", "is_highway", "Traffic_Signal"	0.6768	0.6768
11	GLM	5	"Pressure.in.", Wind_Speed.mph.", "Traffic_Signal", "is_highway", "Distance.mi."	-	0.6718	0.6746
12	Decision Tree	11	"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset", "Distance.mi"	"is_highway", "region", "weather" "Traffic_Signal" "TMC."	0.7351	0.7344
13		3	"Traffic_Signal" "is_highway", "TMC"	-	0.7351	0.7344

14	*SVM**	*11*	*"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset", "Distance.mi"*	*"is_highway", "region", "weather" "Traffic_Signal" "TMC."*	*0.7259*	*0.5721*

The pipeline of our model is started with GLM models with different combinations of features. By looking at the feature's z value at each of the model's summary, we leave one feature out and in. By the combination that we tried and see it's importance value, GLM at line 4th is the best performance among GLM models. Thus, this model's features become the basis of the features that we are applying on other models. The significant difference with this particular model is the TMC code feature. Our previous data exploratory analysis indicates that TMC code may not affect the traffic condition. However GLM model shows the opposite. It is the most important feature in the model, despite having 33% of the samples missing. The depth of analysis will be provided in the discussion section.

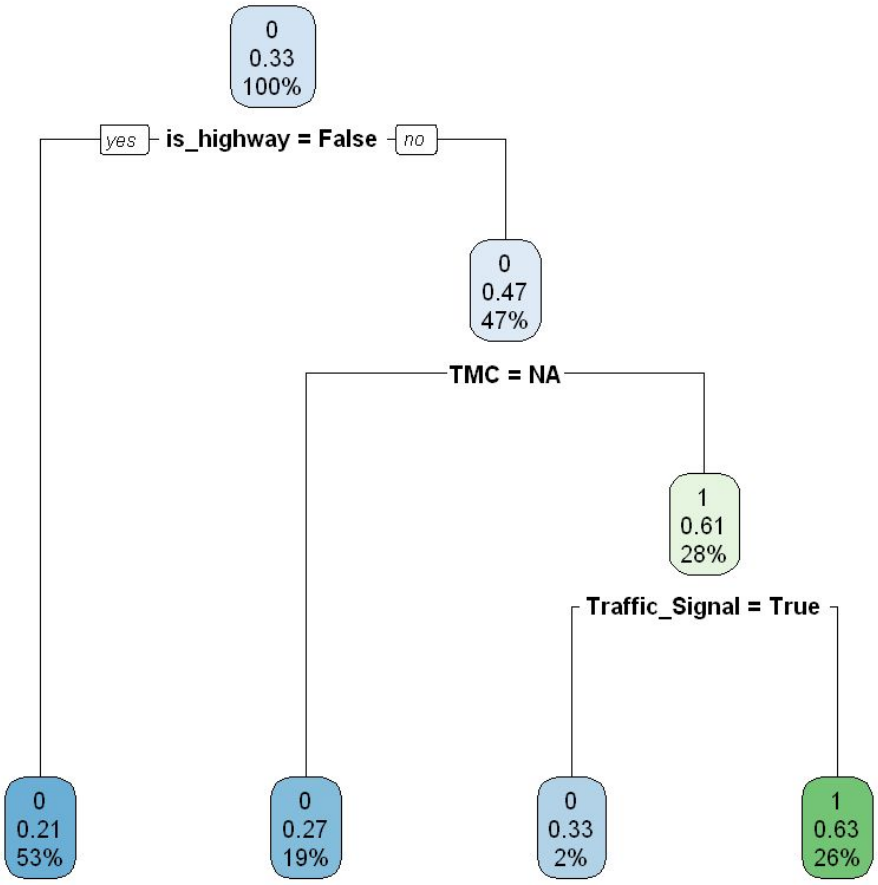
To support the analysis by GLM, we apply lasso with hope that we can do feature selection even further with L1 distance. The summary output from our lasso constraint with cross validation to select the best lambda is as follows.

(Intercept) -3.869037259
Temperature.F. .
Wind_Chill.F. .
Humidity... .
Pressure.in. 0.081668916
Visibility.mi. .
Wind_Speed.mph. 0.008904345
Distance.mi. 0.111850245
is_highway 0.969244253
Traffic_Signal -0.743849349



On the left hand side of the table shows the coefficient outputs from the GLM with the best lambda value obtained from cross validation. The best lambda value shown in the graph on the right hand side. As we expected on the Lasso constraint, it shrinks the coefficient value as well as feature selection, by zero-out some of less significant features. It leaves us five important features as above. Note that from GLM with Lasso, it shows that is_highway and Traffic_Signal as the highest with absolute value. More evidence on these features will be shown on the next model.

Refer to the result table model line 12 and 13 are shown for decision tree. We trained the decision trees with features from GLM model line 4. Surprisingly, it performs better than GLM. The output of the decision trees are as follows.



Despite having 11 features passed as an input, the output of the decision tree showed three major features on its split. Before jumping into analysis, we also tried to train the decision tree *only with* three features shown on the nodes above. The results of the trees are identical with the tree above. This indicates that with only three features we are able to separate the data! The Occam's razor principle is applicable in this sense. So far, the decision tree model provides both performance and the interpretability of rule based visualization of the model, with accuracy of 73%.

As for additional comparison, we applied SVM to the data with several constraints due to the SVM exhaustive algorithm by calculating the distance and finding the maximum margin:

1. Sample 10,000 data points from both training and testing dataset. Thus we have overall 20,000 samples for this setting
2. Only applicable as model comparison but cannot taken into account since the comparison is not an apple-to-apple comparison.

Set SVM aside from other models, we can see that with a minimum number of data, it can classify the training dataset up to 72% accuracy only with 10,000 samples. However, this approach might overfit the training data, since the testing performance dropped to 57% accuracy. It's unclear whether we need to use the exact amount of training data or the support vectors are only overfit to the training data. The reason for this constraint is solely because of the computation limit that we encounter while running this approach.

5 Discussion

Perhaps our biggest discovery is towards the decision tree. We observed each feature and did exploratory analysis on the whole dataset. Looking at the data distribution, target distribution, scale, outliers, and one dimensional statistical output, it is trivial to see that weather features most likely indicate the separation of the severity (0 and 1) the best among all features. Time_Elapsed which indicates how long the traffic blocked last until it comes back to normal also did not affect the prediction very well. Overall the statistical analysis on each feature distribution somehow did not see a promising trajectory. All we can get from the data is that by the region, South and West have an opposite distribution on the severity. This might be influenced by the terrain and the population in each region. Unfortunately, we don't have further information to derive a conclusion.

Linear models such as GLM and SVM with linear kernel are useful for feature analysis. It is not our objective to create a powerful predictive model with high accuracy, however, to perform inference analysis to which predictors that explain data the best. From our model analysis output, shows that there are a small number of important features that separate the severity the best. Those features are shown from our decision tree

1. **Is_highway**: boolean
2. **TMC**: factors, including NA as factor.
3. **Traffic_Signal**: boolean

Here's our discussion on why these features are important. When an accident happens on a highway, there is only one way to go since there is no other traffic direction that can be taken to avoid the accident. Furthermore, the traffic severity can increase significantly since the speed limit is higher than non-highway, then there is a chance that another accident can happen. Second is TMC code. At first, we did not think that TMC has significant importance and should be excluded from the dataset. Also taking into account that 33% of the data is missing for TMC code. However, after observing the full list of TMC code[4], we agreed that it may contribute to the severity. To see if our hypothesis is correct, we factorized the TMC code and added NA as one of the factors. It turns out that TMC code carries more important information. From what we explore from the data, TMC usually exists when a higher severity of traffic accident happens. Despite having every traffic condition listed as the code, practically no authorities will broadcast TMC code if it's not necessary. Given that the severity is 0 (1 or 2), then the traffic condition might not be serious that needed immediate attention. Thus, the authority will not use TMC code to inform through the radio channel. On the contrary, when a more severe accident happens thus blocking the way, it is important for an authority to inform other surrounding officers to solve the situation immediately. The last is traffic_signal. The reason behind this is that when a traffic signal exists nearby an accident, it is most likely that the authority will block the way and revert the traffic direction. Also take in consideration that if there is a traffic signal, then there probably exists an intersection, since it is possible that an intersection has no traffic signal. On the other hand, under assumption that there may not be an intersection given a traffic signal does not exist, then the traffic condition probably worse.

Through this project, we learned the fact that analysis by observation and by model can result as opposite to each other. However, next time want to classify whether a give traffic accident is more severe or less severe, we can ask ourselves:

1. Does it happen on the highway?
2. Is the authority broadcasting a TMC code?
3. No traffic signal around?

Then 63% chance that the traffic accident is more severe (see decision tree).

[4] https://wiki.openstreetmap.org/wiki/TMC/Event_Code_List

6 Limitation and Conclusion

In this project we explore the traffic accident dataset across the United States that span from 2016 until mid 2020. The target of this dataset is the severity of the traffic condition after an accident happens. The location and the aftermath of the accident can affect on how the next traffic flow would be. It is a serious problem and requires immediate attention to smooth the traffic. Especially United States transportation happens mostly on the road. Even worse if an accident happens on a busy highway that carries a lot of cargo across states. Our biggest limitation on this project is the computational resources since all preprocessing happens in a personal computer. Given a large dataset, 49 feature columns and various types of features (numerical, unique, categories, boolean, timestamp), it requires more time on data exploration compared to modeling.
Under the assumption of random sampling that represents the general population, we conducted analysis through a classification model. Our final result indicates that among default features from the dataset, top-3 features are enough to separate the severity into two categories with a significant performance.

Reference

[1] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath, “A countrywide traffic accident dataset,” arXiv preprint arXiv:1906.05409, 2019.
[2] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. 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, pp. 33–42, 2019.
[3] C. Gutierrez-Osorio and C. Pedraza, “Modern data sources and techniques for analysis and forecast of road accidents: a review,” Journal of traffic and transportation engineering (English edition), 2020.
[4] X. Xiong, L. Chen, and J. Liang, “A new framework of vehicle collision prediction by combining svm and hmm,” IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 3, pp. 699–710, 2017.
[5] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, “A deep learning approach to the citywide traffic accident risk prediction,” in 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3346–3351, IEEE, 2018.
[6] S. Kumar and D. Toshniwal, “A data mining framework to analyze road accident data,” Journal of Big Data, vol. 2, no. 1, p. 26, 2015.
[7] “Real-time traffic accident prediction.” https://github.com/swdev1202/Traffic-Accident-Prediction. Accessed: 2020-09-27.
[8] “Road traffic accident prediction web application.” https://github.com/meraldoantonio/AccidentPredic tor. Accessed: 2020-09-27.
[9] “Uk road safety: Traffic accidents and vehicles.” https://www.kaggle.com/tsiaras/uk-road-safety-acci dents-and-vehicles. Accessed: 2020-09-27.
[10] “Us-accidents: A countrywide traffic accident dataset.” https://smoosavi.org/datasets/us_accidents. Accessed: 2020-09-27.
[11] “Us accidents (3.5 million records).” https://www.kaggle.com/sobhanmoosavi/us-accidents . Accessed: 2020-09-26.

Appendix

Appendix 1 US Population and Car by State

State	population	car
AL	4,903,185	2,161,212
AK	731,545	731,545
AZ	7,278,717	2,391,772
AR	3,017,852	921,161
CA	39,512,223	15,065,827
CO	5,758,736	1,798,177
CT	3,565,287	1,306,709
DE	973,764	433,363
FL	21,477,737	7,966,091
GA	10,617,423	3,557,469
HI	1,415,872	509,492
ID	1,787,065	598,774
IL	12,671,821	4,477,763
IN	6,732,219	2,248,870
IA	3,155,070	1,242,219
KS	2,913,314	975,171
KY	4,467,673	1,721,942
LA	4,648,794	1,389,249
ME	1,344,212	390,506
MD	6,045,680	1,922,463
MA	6,949,503	2,182,530
MI	9,986,857	3,023,940
MN	5,639,632	1,976,525
MS	2,976,149	825,338
MO	6,137,428	2,102,216
MT	1,068,778	452,845
NE	1,934,408	683,020
NV	3,080,156	1,073,760
NH	1,359,711	506,959
NJ	8,882,190	2,754,253
NM	1,096,829	655,766
NY	19,453,561	4,712,779

NC	10,488,084	3,393,781
ND	762,062	240,048
OH	11,689,100	4,603,594
OK	3,956,971	1,296,218
OR	4,217,737	1,488,623
PA	12,801,989	4,424,183
RI	1,059,361	412,255
SC	5,148,714	1,830,186
SD	884,659	358,859
TN	6,833,174	2,285,329
TX	28,995,881	8,248,322
UT	3,205,958	937,421
VT	623,989	218,302
VA	8,535,519	3,267,735
WA	8,614,893	2,964,939
WV	1,792,147	560,118
WI	5,822,434	2,087,518
WY	578,759	203,546
DC	705,749	209,723

Appendix 2 US Region

US state region by Census

- Region 1: Northeast
    - Division 1: New England (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont)
    - Division 2: Mid-Atlantic (New Jersey, New York, and Pennsylvania)
- Region 2: Midwest (Prior to June 1984, the Midwest Region was designated as the North Central Region.)[7]
    - Division 3: East North Central (Illinois, Indiana, Michigan, Ohio, and Wisconsin)
    - Division 4: West North Central (Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota)
- Region 3: South
    - Division 5: South Atlantic (Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, District of Columbia, and West Virginia)
    - Division 6: East South Central (Alabama, Kentucky, Mississippi, and Tennessee)
    - Division 7: West South Central (Arkansas, Louisiana, Oklahoma, and Texas)
- Region 4: West
    - Division 8: Mountain (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming)
    - Division 9: Pacific (Alaska, California, Hawaii, Oregon, and Washington)

Appendix 3 TMC code

Full list of TMC code that exist in the dataset

TMC Code	Description	Freq
201:	accidents	2080341
241:	(Q) accident(s). Right lane blocked	249852
245:	(Q) accident(s). Two lanes blocked	40338
229:	(Q) accident(s). Slow traffic	22932
203:	multi-vehicle accident (involving Q vehicles)	17639
222:	(Q) accident(s). Queuing traffic	13154
244:	(Q) accident(s). Hard shoulder blocked	12185
406:	(Q th) entry slip road closed	11109
246:	(Q) accident(s). Three lanes blocked	7118
343:	(Q) earlier accident(s)	6930
202:	(Q) serious accident(s)	6298
247:	accident. Delays (Q)	4775
236:	(Q) accident(s). Heavy traffic	2121
206:	(Q) fuel spillage accident(s)	1274
248:	accident. Delays (Q) expected	1025
339:	(Q) jackknifed trailer(s)	920
341:	(Q) jackknifed articulated lorr(y/ies	592
336:	(Q) oil spillage accident(s)	89
200:	multi vehicle pile up. Delays (Q)	66
239:	(Q) accident(s). Traffic building up	54
351:	(Q) accident(s) in roadworks area	6

Appendix 4 Variables

Full list of US Traffic Accident dataset's feature description. Attribute is the feature name, Description is the detail of the feature explanation, and nullable whether the feature can contain null / NA

#	ATTRIBUTE	DESCRIPTION	NULLABLE
1	ID	This is a unique identifier of the accident record.	No
2	Source	Indicates source of the accident report (i.e. the API which reported the accident.).	No
3	TMC	A traffic accident may have a **Traffic Message Channel (TMC)** code which provides more detailed description of the event.	Yes
4	Severity	Shows the severity of the accident, a number between 1 and 4, where 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).	No
5	Start_Time	Shows start time of the accident in local time zone.	No
6	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	No
7	Start_Lat	Shows latitude in GPS coordinate of the start point.	No
8	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
9	End_Lat	Shows latitude in GPS coordinate of the end point.	Yes
10	End_Lng	Shows longitude in GPS coordinate of the end point.	Yes
11	Distance(mi)	The length of the road extent affected by the accident.	No
12	Description	Shows natural language description of the accident.	No
13	Number	Shows the street number in address field.	Yes
14	Street	Shows the street name in address field.	Yes
15	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
16	City	Shows the city in address field.	Yes
17	County	Shows the county in address field.	Yes
18	State	Shows the state in address field.	Yes
19	Zipcode	Shows the zipcode in address field.	Yes
20	Country	Shows the country in address field.	Yes
21	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Yes
22	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Yes
23	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	Yes
24	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes
25	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	Yes
26	Humidity(%)	Shows the humidity (in percentage).	Yes
27	Pressure(in)	Shows the air pressure (in inches).	Yes
28	Visibility(mi)	Shows visibility (in miles).	Yes
29	Wind_Direction	Shows wind direction.	Yes
30	Wind_Speed(mph)	Shows wind speed (in miles per hour).	Yes
31	Precipitation(in)	Shows precipitation amount in inches, if there is any.	Yes
32	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
33	Amenity	A **POI** annotation which indicates presence of **amenity** in a nearby location.	No
34	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No
35	Crossing	A POI annotation which indicates presence of **crossing** in a nearby location.	No
36	Give_Way	A POI annotation which indicates presence of **give_way** in a nearby location.	No

37	Junction	A POI annotation which indicates presence of **junction** in a nearby location.	No
38	No_Exit	A POI annotation which indicates presence of **no_exit** in a nearby location.	No
39	Railway	A POI annotation which indicates presence of **railway** in a nearby location.	No
40	Roundabout	A POI annotation which indicates presence of **roundabout** in a nearby location.	No
41	Station	A POI annotation which indicates presence of **station** in a nearby location.	No
42	Stop	A POI annotation which indicates presence of **stop** in a nearby location.	No
43	Traffic_Calming	A POI annotation which indicates presence of **traffic_calming** in a nearby location.	No
44	Traffic_Signal	A POI annotation which indicates presence of **traffic_signal** in a nearby location.	No
45	Turning_Loop	A POI annotation which indicates presence of **turning_loop** in a nearby location.	No
46	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Yes
47	Civil_Twilight	Shows the period of day (i.e. day or night) based on **civil twilight**.	Yes
48	Nautical_Twilight	Shows the period of day (i.e. day or night) based on **nautical twilight**.	Yes
49	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on **astronomical twilight**.	Yes