Effect of weather stimuli and road points of interest in classification of traffic accident severity

Adithya Narayanan, Adit Bhargav Modi, Akhil Gaddipati

December 2019

1 Abstract

Researchers in the automobile industry have tried to build safer automobiles by improvising design and ergonomics. Yet, traffic accidents seem unavoidable. There are a wide variety of factors that can cause accidents on the road. With autonomous cars well within reach, a valuable research direction is to explore the extent of the effects of circumstantial entities, which can be handled better if we are able to establish a pattern for how cars perform in certain weather conditions and in the presence and absence of specific road patterns. The role of weather conditions in traffic accidents, coupled with understanding how different road patterns (railway crossings, speed bumps, etc.), and their relationship with accident severity is the primary direction of our foray into this field.

2 Introduction

In the year 2013, over a million people lost their lives on the road due to traffic accidents. While we have a lot of recorded data on accidents available now, thanks to modern technological advancements, using data to good avail can greatly reduce this unacceptable loss of life, and to a lesser extent, the damage sustained physically and economically.

Weather is a big contributor to traffic accidents. Conventional logic cannot refute that. The objective of this project is to try and reduce this logic to a computable expression or a series of them, which can aid in predicting the such occurrences given certain weather conditions and road conditions. As for road conditions, their role in accident severity is an area we explore deeper in an attempt to establish a relationship. Major research work that has been done in predicting accidents has revolved around the using environmental factors, regions of occurrences, their correlation with frequency and predicting accident risk. Using a vast amount of data from four different states in the US the United States over a period of three years and by analyzing the performance of several classifiers, we aim to study the strength of the relationship between our predictors and accident severity.

Accident Severity here has an unconventional definition. While conventionally severity refers to the extent of damage, human or infrastructural, in our data, accident refers to the amount of delay caused to traffic due to the occurrence of an accident. With an increase in severity (categorical from one to four), the delay the corresponding accident causes is higher.

3 Relevant Work

Traffic accidents are reducible. While that may seem like an overstatement, it can be argued that anything that can be predicted with a fair amount of accuracy can be prevented from occurring with a fair amount of success as well. Significant work has been done in such prediction of accident risk. The absolute negation of accidents is a whole different problem to solve. According to reports, the United States experienced three straight years of approximately 40,000 roadway deaths since 2015[1][2]. If the same trends of unemployment rates and vehicular miles traveled continue, by 2024, roadway deaths won't see much of a difference from these numbers[3]. The costs of fatalities and injuries caused by such accidents and the infrastructural loss they can carry with them can have a vast impact on society. In 2017 alone, the monetary losses caused by such traffic accidents was approximately $413 Billion. In 2016, the number was a percent higher [1]. Reducing traffic fatalities becomes imperative given these incentives. Here, we briefly enumerate the different approaches taken by researchers in the investigation of this field.

In a study published in 2014, the authors of [4] use Multiple Objective Particle Swarm Optimization (MOPSO) based methods to analyze accident severity based on accidents recorded in Beijing in a two year window. Unlike most other studies that use classifiers, much like the scope of this study, the MOPSO based method uses particle swarms that capture non-polynomial fits to predict targets using diverse predictive criteria by finding the best fit while traversing the data's dimensional space. The study also uses partial classification methods to handle unbalanced data which contains very few observations of fatalities to produce 'if this then that' type relationships between predictors and accident severity. They establish a relationship between dividers, a road POI not included in this study and accident severity and also establish a relationship between foul weather and severity. Both these conclusions are along the lines of what conventional logic would expect.

Using data collected from California (2006-10), the study performed by Hoon Kwon et al. has identified key predictors that are relevant to accident severity [5]. They investigate the use of a decision tree to identify dominating predictors in classifying severity along with an implementation of a naive-bayes classifier. An interesting observation of this study was that the predictors seemed to suffer from multicollinearity, which becomes an obstacle while using methods like Bayesian classifiers since they rely on independence of the predictors. Decision trees perform better in this study, since they don't rely on this assumption. Using the methods and the identified predictors above, a binary logistic classifier which classifies fatal and non-fatal accidents is built. Interestingly, the road POI 'Intersection' is one of the important factors that the study identified as a good classifying predictor. Several such classification studies have been performed in an attempt to understand accident severity and we will enumerate the concise rationale behind a few of them. Shanti et al. performed a comparative analysis of different classification algorithms- of which Random Forest is one of interest to us due to our usage of the method- and feature selection methods by studying their performance on accident severity data from the US. Out of several classifiers such as C4.5, CRT, CS-MC4 etc., Random Forest is seen to perform the best [6]. It is important to note here that this dataset accounts for approximately 450,000 observations across all of the United States. Although weather or road POIs aren't taken into account by this study, comparing the performance of Random Forests with the predictors chosen in our study and with predictors such as drug involvement, as chosen here, would be an interesting task.

Farmer, in a 2015 study, projects traffic fatality counts to approximately 34,000 in 2014 in the United States by modeling this using Linear and Poisson Regression on variables such as the year, unemployment rates, and miles travelled by vehicles [7] using data from 1990-2015. Assuming a good economic growth, miles traveled by vehicles are bound to steadily increase. Reducing fatalities which can come with the increasing use of vehicles will become a problem that rears its head again. With rapidly changing climate patterns, unexpected snow, heat waves of unseen magnitudes, and so on, it becomes imperative for understanding the role of weather stimuli in accidents again. Attempts like ours to classify accident severity have been made before, as elaborated in the passages described above. A study that details the use of different classifiers like [6] is performed by Chong et al. in [8]. To find the best classifiers, the researchers here investigated the performance of neural networks, decision trees, support vector machines, and a hybrid decision tree – neural network for predicting drivers' injury severity in head-on front impact point collisions into five classes - No Injury, Possible Injury, Non-incapacitating Injury, Incapacitating Injury, and Fatal Injury. The hybrid approach seems to be the pick of approaches when predicting across classes, while SVMs perform poorly across classes.

Theofilatos[9] investigates road accident likelihood by using road traffic and weather data from urban arterial roads in the major city of Athens, Greece. This study identifies significant variables using methods such as Random Forests and then uses Bayesian Logistic regression and different types of logit models to study the causes of accident likelihood and severity. Interestingly, the regression model seemed to suggest that accident likelihood is impacted by the traffic patterns, but severity seemed to produce mixed results, at odds with other studies which seem to suggest otherwise (much like what conventional logic would dictate). The most interesting pattern in this study however, is that the methods mentioned above for the region chosen, seem to suggest that weather did not impact either severity or likelihood, which automatically makes this study a feature of interest. The study seems to marvel at this conclusion as well, since Greece tends to face adverse weather, and speculates that this could be due to users adapting to adverse weather, or that data for accidents under such adverse weather is sparse. They also speculate that the relationships between weather and the responses, are probably not entirely described by linear models and functions.

To alleviate concerns of dependencies being taken into account with stochastic methods (such as regression methods), the study performed in [10] uses palm probabilities and no other modeling methods to analyse accidents in Finland. They take into account, different kinds of roads and the weather conditions that prevail over the traffic during and outside accident observations, thereby accounting for behavioral impact of traffic (even outside what is captured by the observations) on accidents (eg: time spent on certain types of roads). Both road type and weather conditions were found to have significant impact on accident severity in ways dictated by general sense (eg: increase in risk with snow is sharp as opposed to no form of precipitation). Relationships established by studies between weather and accidents don't seem to have changed much over time, although the analysis has obtained sophistication. A 1978 study by Sherretz et al. [11], finds that linearity describes the relationship between rainfall and accidents fairly accurately. However, the study itself questions its scalability since it accounted for only a niche region (seven cities in Illinois). In their 2011 paper, Savolainen et al, discuss the problems with analyzing such datasets and some alternative ways to study them outside the ones discussed above like binary classifications, probit modeling, logit models which adopt a markovian approach, and neural networks, and discusses their levels of sophistication in either selecting the right predictors or their predictive performance [12]. Several studies employ the topics discussed by[12], but on diverse data, such as motorcycle crash data [13] using multinomial logit models, and rural freeway crashes [14].

The Montella et al [15] study analyses on two wheeler crashes in Italy uses classification trees and rules discovery. Rules discovery help establish a relationship between influencing factors and outcomes like the if this then that method described in [4]. Alignment of the roads studied here, at crash sites, which include subsets such as intersection, curved roads and so on, were observed to be related to weather conditions, with sometimes conventionally good weather also accounting towards crashes given other conditions like rural curved roads. A tree based modeling approach done on Slovenian traffic injury severity [16] shows the pitfalls of such a classification approach, since the tree model's performance suffers due to low variance and high bias. This study however, also includes the role of human error in such accidents and pin points them as a very important factor in the causation of such accidents. This remains an essential observation, since automation can completely reduce human error, leading to the reduction of accidents using predictive studies like ours and inferring better solutions from them.

In our study, we use data from the dataset titled US Accidents. Traffic data is often rarely well documented across a vast array of metrics which include region (city, state, county etc.), combined with recorded environmental conditions (temperature, wind speed, visibility etc.) and points of interest near the accident zone (stop signs, road crossings etc.). Several studies prior to this suffer from their localized approach, or lack of data, and tend to use over simplified methods, and this study aims to alleviate these concerns by employing the US Accidents dataset, specifically curated for this purpose. The study which led to the curation of this dataset uses such high dimensional and large data, and a deep neural network, called DAP, Deep Accident Prediction, to predict accident risks in real time to time durations as fine as 15 minutes in a 5 sq.km radius with a few other variables included in its scope. The paper finds patterns which can govern how we design cities and traffic flow networks [18]. Real world data, much like the US accidents data we use, is sparse and unbalanced. Data is said to be unbalanced if one target observation class dominates others or if one such class is not dominant enough to be classified given some predictors since it doesn't appear enough. In our study's dataset, high severity incidents are bound to be less featured, as dictated by conventional logic, given their rare occurrences. Such datasets affect the performance of predictive stochastic models, since not enough data is available to model the lesser represented classes as well as the ones that dominate. The Data Science community often uses simple techniques such as undersampling (discounting some dominant observations to match observation numbers) and oversampling (generating observations similar to the underrepresented observations) [17]. In [19], Chawla discusses the different methods to alleviate the issues caused by unbalanced data. He discusses methods such as synthetic Minority Oversampling Technique (SMOTE), which is an oversampling technique supported by [17] as well. He also discusses ensemble based methods like boosting and their combinations with SMOTE to aid in better predictive results by identifying best predictors while ensuring class balance.

A few key observations from our study of past work in this field are noted below:

- If predictions are to be carried out, using a wider variety of models might yield better prediction results as different mechanisms work for different kinds of data. Finding the right model is important.
- The dataset can have a lot of data which might be irrelevant to our objective. Finding the right observations and the right variables to satisfy our objectives is key. Using a lot of redundant variables, or unnecessary and irrelevant predictors will only add noise to our predictions.
- The studies performed above majorly side with one conclusion. That there exists a relationship between weather, road conditions, and human involvement, and road accidents. Our modeling will have to account for this and see if we can or cannot side with this hypothesis since contradictory results have been found, possibly due to a city's culture or behaviour to adverse weather, as seen in the Athens urban accidents observations in [17].

4 Data Exploration and Visualization

For the purposes of our study, we aim to use data covering several states from the US Accidents Dataset [22]. The data compilation has been carried out by the researchers of [5]. It has been pooled in for multiple sources and documents approximately 2.2 million incidents of road accidents in the United States from March 2016-April 2019. It accounts for 49 states, thereby ensuring a fairly comprehensive weather and road points of interest coverage. The variables of the dataset are comprehensively described in a table form in the appendix. To get a better idea of how the dataset looks visually, it's structure and the distribution and coherence if it's variables, we conducted an EDA using the R package DataExplorer.

For the purposes of understanding the distribution of data before we subset the data for different states, we explore the entire dataset. we discarded a few variables such as Source, TMC, Start and End Times, and Start and End Latitudes and Longitudes, Twilights, and geographical variables, due to their anticipated inability to contribute to an initial intuitive understanding of the data. While a lot of these might be important predictors of accident severity, we decided to use primarily weather related variables and road conditions related variables and accident severity to keep the predictors withing the scope of this study- to analyze the relationship between accident severity and the weather stimuli and road POIs.

We then split the dataset into two subsets, one taking into account weather related predictors, such as Temperature(F), Wind Chill(F), Humidity, Pressure(in), Visibility(mi), Wind Direction, Wind Speed(mph), Precipitation(in), and Weather Condition, and the other containing road POI related predictors such as Amenity, Bump, Crossing, Give Way, Junction, No Exit, Railway, Roundabout, Station, Stop, Traffic Calming, Traffic Signal, and Turning Loop. Each of these data subsets also include the severity variable since modeling is done to predict severity of accidents with respect to these predictors mentioned.

Analyzing the structure of the weather Subset, we notice that most of these variables are numeric, with a few categorical variables. The structure of the data in this subset is described in the plot below.

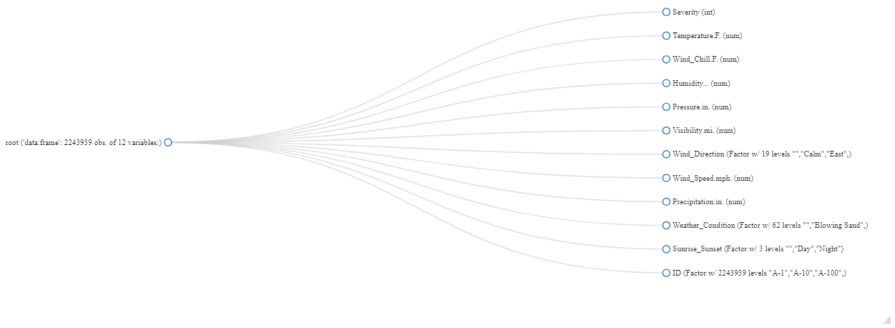
root ('data.frame': 2243939 obs. of 12 variables:)
Severity (int)
Temperature.F. (num)
Wind_Chill.F. (num)
Humidity... (num)
Pressure.in. (num)
Visibility.mi. (num)
Wind_Direction (Factor w/ 19 levels "","Calm","East",)
Wind_Speed.mph. (num)
Precipitation.in. (num)
Weather_Condition (Factor w/ 62 levels "","Blowing Sand",)
Sunrise_Sunset (Factor w/ 3 levels "","Day","Night")
ID (Factor w/ 2243939 levels "A-1","A-10","A-100",)

Figure 1: Weather Variables Structure

Unlike the former, the Road POI subset is primarily Binomial factors, i.e., true or false variables. The structure of this subset can be noted below.

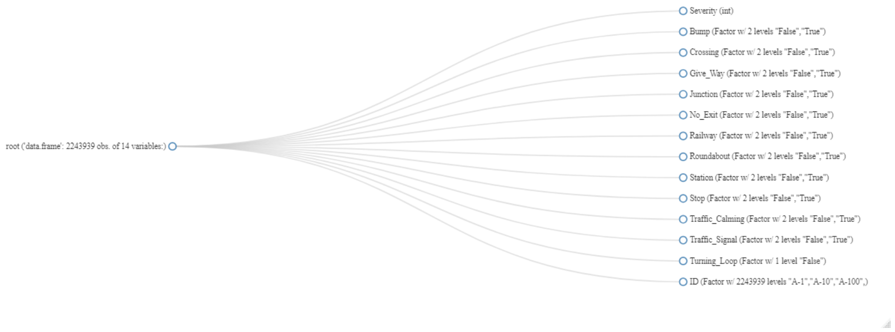
root ('data.frame': 2243939 obs. of 14 variables:)
Severity (int)
Bump (Factor w/ 2 levels "False","True")
Crossing (Factor w/ 2 levels "False","True")
Give_Way (Factor w/ 2 levels "False","True")
Junction (Factor w/ 2 levels "False","True")
No_Exit (Factor w/ 2 levels "False","True")
Railway (Factor w/ 2 levels "False","True")
Roundabout (Factor w/ 2 levels "False","True")
Station (Factor w/ 2 levels "False","True")
Stop (Factor w/ 2 levels "False","True")
Traffic_Calming (Factor w/ 2 levels "False","True")
Traffic_Signal (Factor w/ 2 levels "False","True")
Turning_Loop (Factor w/ 1 level "False")
ID (Factor w/ 2243939 levels "A-1","A-10","A-100",)

Figure 2: Road POI Variables Structure

The next step would be to study how much of the data is missing. The missing data has been quantified with the plots below.

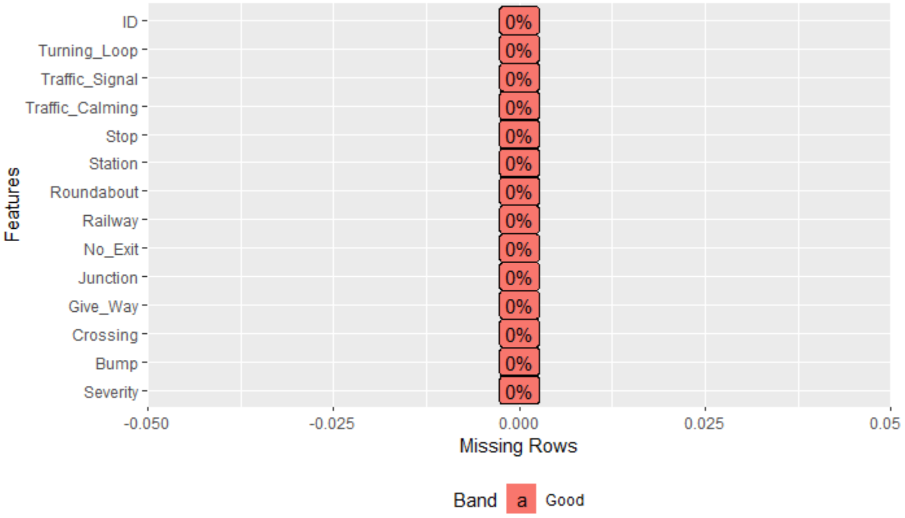

Figure 3: Percentage of Missing Values in Road POI Variables

The missing points in the weather data and their observations are described below. Since there are variables with significant amounts of data missing, these variables are discarded, and for all other variables, just the observations with missing values are removed from the data before the final modeling.

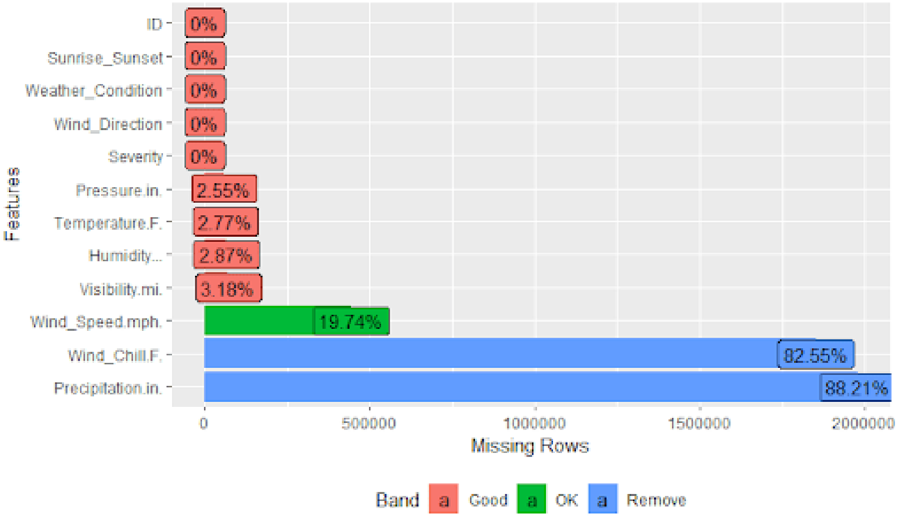

Figure 4: Percentage of Missing Values in Weather related variables

Lastly, we try to plot the distributions of the variables, in two forms, density plots for the continuous metrics and bar plots for frequencies of categorical data.

Several interesting trends can be seen below. Pressure seems to be constant, centered around the same value, while precipitation seems to be perennially low. Wind chill during recorded accidents favors positive, yet low temperatures, while wind speed seems to be long tailing as speed increases. Recorded temperature seems to be very normally distributed between 0 and 100, indicating a fair amount of observations in all conditions while humidity is significant fairly around most values observed, but favors the higher numbers.

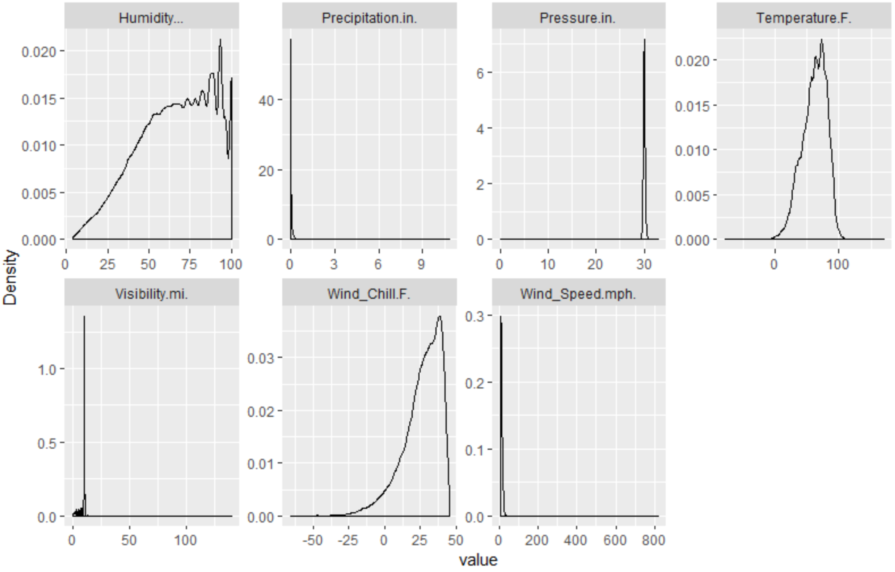

Figure 5: Density Plots of Continuous Weather Related Variables

A fair bit of diversity in observations exists in the categorical weather related variables. Categorical weather related variables exhibit larger diversity compared to other categorical variables. The sample plots follow.

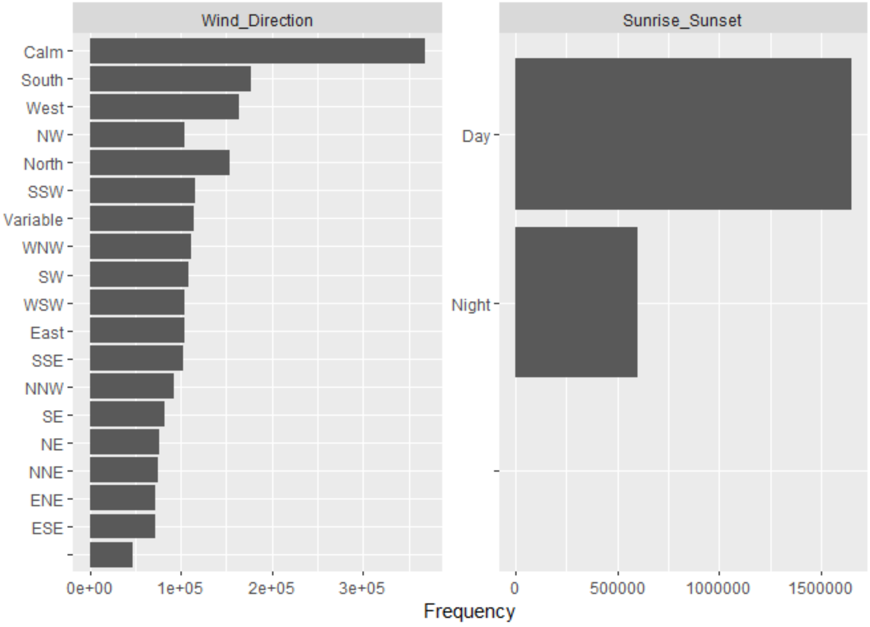

Figure 6: Bar Charts depicting the frequency of Categorical Weather Related Variables

Plotting the bar charts of Road POIs are not of particular interest since they are sparse and do not add value visually. It is not common to expect POIs to be present a significant amount of times at accident sites, especially when the dataset is this large. Plotting the bar charts corroborates this theory, and is therefore left out of this visualization section.

The overall observation count of each of the classes is of interest, to observe if the data is balanced. The bar chart below depicts the quantitative proportions of each of the categories of the severity variable. It can be observed that one particular category dominates the others, depicting a need to balance the observations to ensure the best fit of the models that are described in further sections.

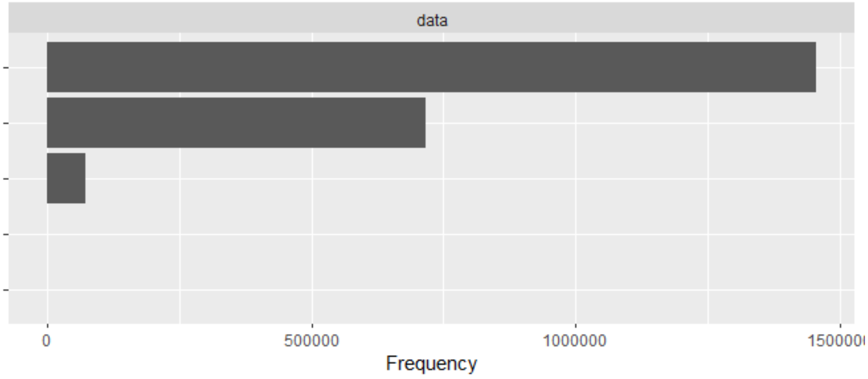

Figure 7: Bar Chart depicting the frequency of the observed severity values

4.1 Subsetting the Data

The initial data spanned 49 states in the US and covered almost all of the US. It consisted of 2.25 Million observations. To reduce the computational effort, and yet obtain a fair spread of predictors, we rationally choose four states from different parts of the US and analyze their performance in the scope of this study individually.

The four states chosen for this purpose are Washington, Arizona, Pennsylvania, and Florida. As depicted in the map that follows, these states are situated in different corners of the geography of the country and are also exposed to vastly different and widely diverse weather conditions. By studying these states individually, we hope to obtain a big picture understanding of how the combination of features we have selected performs in classifying accident severity.

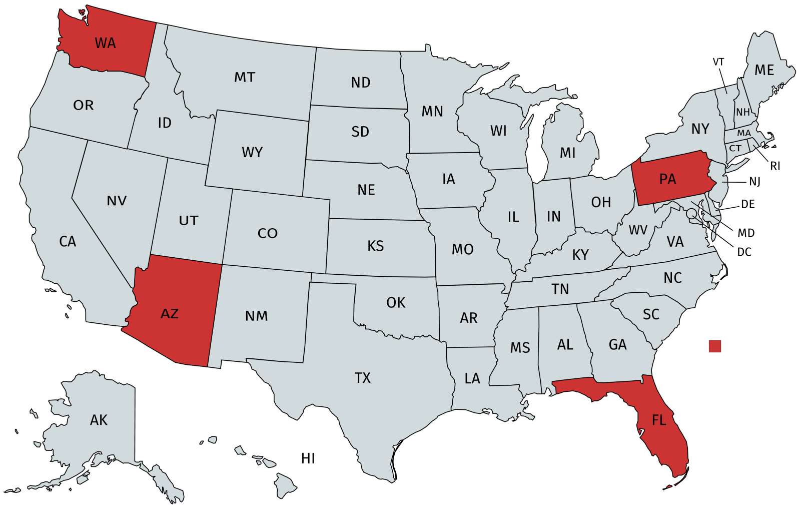

Figure 8: Map depicting the states of US included in the analyses

4.2 Variable Selection

The initial data contained 49 variables that included Weather related variables, Road POIs, Start and End Times of the Accident, Start and End Latitudes and Longitudes for each accident, source from which the data was obtained and many more. The appendix houses a list that describes each of the variables included in the dataset initially. For the scope of this study, we only needed weather related variables and Road POIs. To direct our focus towards these, we discard any variables that do not fall in either category. A categorical variable that had over sixty factor levels was also nullified since such variables do not allow for ease of model building. Since our dataset is fairly large, even post-processing, we retain a significant number of observations. The final set of weather related variables included in the analyses were: Temperature, Wind Direction, Wind Speed, Humidity, Pressure and Visibility. Similarly, the road POIs that found a place in our analyses were: Amenity, Crossing, Station, Stop, Give Way, Traffic Calming, Junction, No Exit, Traffic Signal, Railway Crossing, Bump, and Roundabout.

5 Research Methods

Several different approaches can be taken to classify accident severity into the different classes that encompass it. Here, we submit a brief review of the conventionally accepted classification techniques that this study incorporates.

5.1 Supervised Learning

Supervised learning is the process of predicting the value of a target (commonly called response) given a set of predictors using a mathematical relationship established between them by modeling this using data that establishes the trends between them. The way the target value behaves with respect to changes in the values of predictors is used to build a mathematical model that can estimate the target given a set of inputs(values that the predictors can assume) that the model hasn't encountered before. Such predicted outputs (response) are generally depicted by the variable y and the input values of the predictors are depicted by the variables $x_1, x_2..x_n$ for n predictors (also referred to as dimensions).

The relationship can be summarized by the modeling $y = f(x)$ where a mathematical function of the predictors give the predicted output y. There is an irreducible error component ϵ to them, but since this component is generally irrelevant to out purposes of understanding and also largely inconsequential due to its non-tamperable nature, we will ignore it's presence through this discussion for ease of understanding. Errors which cause a difference between predicted and actual response values due to noise in the data are are generally irreducible.

A model's performance is often tested by metrics such as the Mean Square Error (MSE) or Accuracy (a receiver operating characteristic - ROC) . In such metrics, the difference between the actual value of the response variable and the value predicted by the model is calculated (by testing the model on data for which the response values are known). Supervised learning models, when being built, aim to minimize this prediction error by fitting successively better models. Since this error can be minimized, such errors are called reducible errors. Supervised learning methods gradually fit better and better models to a given dataset (used for training-called training data) by minimizing the reducible error each time the relationship is modeled (by carrying out a round of predictions on the training data and checking for errors), until a final relationship is established which can be used for prediction outside the training data.

In this research project, we will be using Supervised Learning methods to predict severity of traffic accidents given certain weather conditions and road points of interests.To ensure continuity in comparing the different models, they will all be compared based on their accuracy rates. Among other ROC characteristics such as specificity and sensitivity, accuracy is the most relevant to this study. The process of how we obtain this measure is depicted in the methods section.

5.1.1 Bias Variance trade-off

Bias is a metric that depicts how closely a model fits a training dataset. Variance is a metric that depicts the performance of a model when exposed to different datasets. A model that very accurately captures the relationships in a training dataset might not perform well when exposed to unseen data. Such a model is said to have high bias and low variance, and such a scenario is commonly known as overfitting (the model overfits the training data, giving high prediction accuracy when exposed to the same data on which it is modeled, but captures the general relationships between the predictors and response poorly in general, owing to the low prediction accuracy when tested elsewhere). High variance and low bias can be defined by complementing these definitions. In supervised learning, a good model often balances bias and variance (trade-off) to capture the relationship between predictors and response variables accurately, while also maintaining a plausible prediction accuracy out of sample (whether an accuracy is good is determined by the data and the relationship and is subjective, although often obvious).

5.1.2 Classification and Regression

When supervised learning models predict responses that are categorical (finite number of categories), they are said to be classifying data. For instance, a model that predicts whether or not a patient has cancer is a classifier. Models where the predicted value is continuous and generally indicative of the magnitude of what it represents are called regression models. For instance, predicting the value of temperature given certain other weather conditions would be a regression problem.

5.2 Logistic Regression

Logistic Regression models are generally used as two-class classifiers (eg: Yes/No, True/False etc.) but can be extended conceptually to classify more. The two classes can be attributed with the numeric values 0 and 1. Logistic Regression models establish a linear relationship between the log $\frac{p(y)}{1-p(y)}$ response (called the logit form), where $p(y)$ is the probability of class y being the target and the predictors x_{1-n} in n-dimensions. The following equation depicts this modeling, where W_i (where $i = 0$ implies the intercept) can be interpreted as the weight the predictor x_i carries in depicting the left hand side of the equation.

$$\log \frac{p(y)}{1-p(y)} = W_0 + \sum_{i=1}^{n} W_i x_i$$

This can be modified to yield a $P(y)$ value that can be depicted in a sigmoid curve form.

$$p(y) = \frac{1}{1 - e^{-z}}$$

where z depicts the right hand side of the first equation.

This formulation ensures that all predicted $p(y)$ values for the given input predictors are between zero and one, and values that are greater than a set threshold (by testing highest accuracy on training data) can be attributed to the class depicted with 1 and the others can be attributed to the class corresponding to 0. By using the logit form, it is ensured that the values of $p(y)$ are bound between 0 and 1- a necessity to maintain logical consistency while predicting the probability of a class.

Logistic regression is traditionally used as a method to set the benchmark of accuracy for classification problems and will occupy the same role here. If any of the models discussed further perform poorer or even comparable to logistic regression, the former will be considered ineffective due to the inherent ease of computing and interpretability advantages of Logistic regression.

5.3 Multivariate Adaptive Regression Splines (MARS)

Assuming linearity in the relationship between predictors and responses is often untrue in real world data. Better predictive accuracy can be obtained while modeling the relationship as functions of higher order polynomials. But with polynomial relationships, interpretability of models is often an unfair expectation. To build an interpretable model and account for non linearity, MARS models split the predictor-response space into sections at different knots (also known as cut points) in a way to fit different linear or polynomial models to the now different sections using the concept of hinge functions. At every cut point, a new relationship is modeled and the previous one ends. Instead of one relationship depicting the entire model, multiple relationships exist across different sections of the data. However, to enhance interpretability, the relationship between the response and predictor is still depicted as a single equation modeled using hinge functions. A MARS Equation follows the format:

$$y = i_0 + \sum_{i=1}^{n} W_i(h(x_i))$$

Where each W_i depicts weights in the same way as earlier and each $h(x)$ (called basis function) depicts a hinge function that models interactions between different linear relationships at the cut point. Basis functions are of the form

$max(0, x - k)$ or $max(0, k - x)$ where the knot for variable x is at the value of k. A basis function can also be a constant or a product of more than one basis function, which depicts the interaction between the two variables that are contained in the two functions. Beginning with just a single constant term in its function, MARS greedily adds hinge functions that lead to the largest reduction in the reducible error and terminates when the change in this metric is negligible. Following the construction of this model, pruning (like in tree based models) is carried out to enable better variance. Pruning can be carried out using Generalized Cross Validation, that balances the fit of model (and thereby the bias) and the complexity (which is generally inversely proportional to variance)[24].

MARS models are used to classify injury severity in this project. The hinge functions will assist in modeling the non-linear relationships between the given predictors encompassing environmental stimuli and road points of interest, if any.

5.4 Classification Tree

Decision Trees are non parametric mechanisms to arrive at continuous or categorical predicted values of a response or target, by passing the input predictors through a tree like structure that captures the relationship between the two. At each node of the tree, a predictor of interest is tested to satisfy a precondition and is allowed to pass through one of two edges that emanate from it. If a node exists such that no edges depart from it, these are called leaves and are the final predicted value of the target. When the predicted value is continuous (generally an average of all continuous values in the leaf node's predictor space) then the tree is said to be a regression tree, while if the output is a categorical value, it is said to be a classification tree. For this study, we build classification trees[25].

Not all predictors are used to build nodes, and generally the predictors which have the biggest influence on the target value based on their variations find a place in the tree. In tree building algorithms, a root node is picked such that a binary split of the predictor housed in the root node influences the classification most significantly. Building further from here, whenever a binary split of a predictor aids the classification process based on the position of a new node is added until a leaf node is reached, from where tree building recursively continues from the point where the branching eventually led to the current leaf node. Reduction in misclassification rate can be used as a binary splitting criterion. Misclassification rate is defined as ratio of observations classified wrongly [26][27].

Trees are often built to closely fit a data (overfit) and are then pruned. Several pruning approaches, such as using GCV to balance the fit and variance can be used.

Regression Trees can be explained numerically using the equation:

$$y = i_0 + \sum_{i=1}^{n} W_i(R_i)$$

Where R_i is the Rth partition of the predictor space. With small tweaks, this equation's underlying logic can be used for classification trees too. This makes the interpretability of tree based models similar to how MARS models or Logistic Models are interpreted.

5.5 Bagging and Boosting

Since trees are often overfit, their applicability to unseen data can be questionable. This concern can be alleviated a bit by using a method called Bagging or Bootstrapping Aggregation. In this method, multiple trees are created using multiple datasets of similar structure and the mode of their classification outputs is considered the final predicted output. Since a k-way split to build k trees would leave very less data for each tree to be built on, bagging randomly samples a fixed large number of observations k times, from the original dataset by replacing after every sampling. This way, each tree is built on randomized yet significant quantity of data. For each tree, the error can be calculated by testing its accuracy on the observations that were not used to build that particular tree. This measure is called Out of Bag Error. Tree models are built by minimizing this metric. Bagged Trees are an instance of ensemble learning where results from multiple similar (but not same) models produce predicted outputs which are averaged out to produce the final output. [28]

Boosting is an ensemble method that utilizes trees too. Much like bagging, boosting too builds several trees from the training data, but instead of using arbitrary subsets, it fits a weak classification tree of low depth to the data. By penalising the output of this tree using a tuning parameter (learning rate, like ridge regression), it builds a cumulative model with an ensemble of trees in which it successively adds shrunken versions of these trees. Each time a tree is built and a penalty is applied, the overall model sees a small improvement in predictive accuracy.[29][31]

The additive modeling equation is

$$f = f + \lambda f_b$$

where f denotes the cumulative model, and f_b is the tree at the bth iteration. λ denotes the learning parameter. The residuals are updated at every iteration using

$$r = r - \lambda r_b$$

where b again denotes the bth iteration.The final model is defined by

$$f = \sum_{b=1}^{B} \lambda f_b$$

5.6 Random Forest

Random Forest uses the ensemble learning method too, except the variance reduction is carried out by restricting the number of predictors considered. Multiple bagged tree construction cycles are carried out where prior to each cycle, a set of m predictors out of a total of n are separated out and a large number of trees of diverse depths are constructed with the bagging approach with just one added restriction. Each of these trees can include splits at only the variables belonging to a maximum of m predictors. Conceptually, out of bag error and the mode of output as the final class are applicable to random forests as well, since they mimic the functioning of bagged classification trees. For Classification problems, the size of m is made to be equal to the square root of the number of dimensions approximately. This is helpful in constructing trees that yield lower variance, since their structures are now coerced to be different due to the restriction on which variables to use during splitting. [30]

All tree model performances can be evaluated using their accuracy while classifying the testing set as well. In our study, Random Forest will be one of the ensemble learning methods used to model the relationship between the predictors and severity to enable a better understanding of impact of individual predictors on the response. Since Random Forest is a method that does not use all predictors to model the relationship at once, it can unearth insights about impactful parameters and partial dependencies.

5.7 Model Building

After shortlisting four states, as described in the previous section, the next course of action was to implement the listed models on each of them to observe their ability to classify potential accidents.

A description of the average delays involved in each of the levels of accident severity in the original dataset is tabulated here.

Variable of Interest	**Severity**	**No. of Observations**
Level 1	2 Minutes and 30 seconds	814
Level 2	3 Minutes and 15 seconds	1455524
Level 3	8 Minutes	715582
Level 4	18 Minutes	72002

Table 1: Severity Table

To reduce the problem to a more logical level of understanding from a driver/commuter standpoint, we club the first three of the levels of severity together and call them "not significant". We adjudge level four to be "significant". The reason for this reclassification if driven by the simple rationale that the gulf between levels three and four is massive compared to the other differences and that eighteen minutes of delay caused due to an accident can be conventionally perceived as a far greater delay than eight. Thus, the terms of reclassification, based on the significance of delay find their place in our data.

A brief look at the data suggests that there is prevalent class imbalance in the data across the levels of severity across each of the states. For each state, post reclassification, we observe that if we undersample the majority class (not significant), we still retain a significant number of observations (in the thousands). Oversampling on the other hand creates a significant amount of non-independent observations since new observations from the underrepresented class are generated based on existing ones. Hence, we decided to stick with this method. In undersampling, we algorithmically sample roughly the same number of observations from the dominating class to make the dataset balanced. Post balancing, the number of observations retained for each of the four states considered is tabulated below.

	Unbalanced		**Balanced**	
States	**Non Significant**	**Significant**	**Non Significant**	**Significant**
WA	38998	1368	1353	1368
FL	147426	4167	4113	4167
AZ	31783	1803	1793	1803
PA	53751	3201	3188	3201

Table 2: Number of Observations before and after balancing the data

We then implemented each of the models discussed in the preceding subsections with different levels of tuning as described by the flowchart in the figure that follows. For each state, the process was carried out independently and identically. A training set for each state with 65% of the observations was sampled and the rest of the data was used for testing the predictive accuracy of the model.

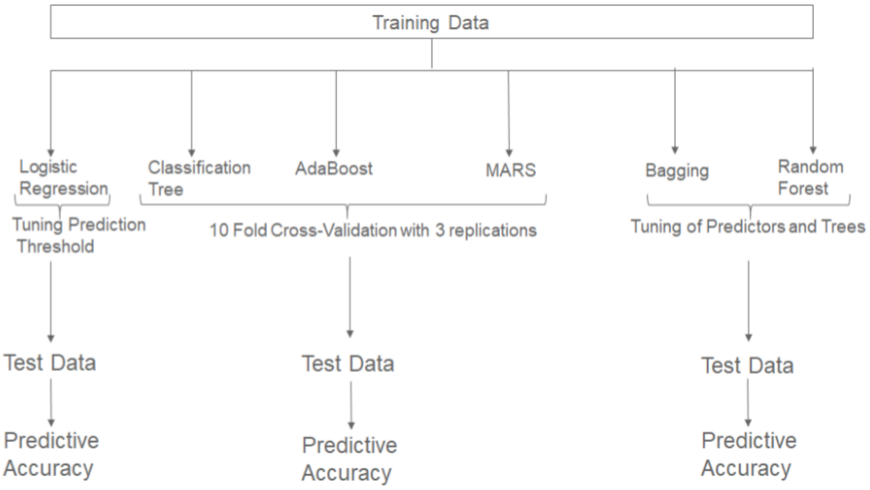

Figure 9: Diagram depicting the model building process

Like in all binomial classification problems, we relied on Logistic Regression to set the benchmark for classification accuracy. For each of the states, we tune the classifying threshold to obtain the highest prediction accuracy on the testing data. We then implemented a Classification Tree, an AdaBoost model (our boosting algorithm), and a MARS model, with 10 Fold Cross Validation and three such replications. Expanded grid searches were performed to tune the parameters for the models wherever applicable. For instance, multiple learning rates and depths were used to tune the AdaBoost model. The number of predictors and trees used for Random Forest and Bagging and the prediction threshold used for the logit model are tabulated in the appendix. To further test ensemble models, we trained and tested bagging and random forest models on the data and tuned the parameters of number of trees and number of predictors, respectively, using out of bag error. The out-of-sample predictive accuracy of each of these modes is discussed in the succeeding section.

6 Results

Each of these models did not deliver a very different commentary from the other, although some were marginally better in terms of their predictive accuracy. The accuracy measure is calculated using the formula

$$Accuracy = \frac{(True\ Positive + True\ Negatives)}{Observations}$$

where $TruePositives$, $TrueNegatives$, and $Observations$ denote the number of each of these entities respectively. Here, the values of True Positives, True Negatives and the total values follow the definition of a standard confusion matrix. A tabulation of just the accuracy measures for each of the models implemented for each follows. The accuracy measure stems from subtracting the mean error from 1, where mean error is defined by the mean of number of wrong classifications of the test data.

It can be observed that across states, all accuracies do not tend to vary much. The range of accuracies were from the late 50s to the late 60s in percentages. In general, improvements over the benchmark set by the logit model were observed, but the improvements were not significant enough for the mathematical relationship between the combination of predictors to be considered strong. We also plotted the classification trees (post cross validation and multiple replications) for these states and noticed very few predictors manifesting on most trees and a poor prediction rate. Washington delivered a bigger tree with several predictors and a large number of terminal nodes manifesting on it, but the prediction accuracy was barely better than the other states we looked at.

State	**Model**	**Accuracy**
Florida	Classification Tree	0.5783299
	Bagging	0.6066253
	Random Forest	0.6045549
	AdaBoost	0.6090407
	MARS	0.6080055
	Logistic Regression	0.6011042
Pennsylvania	Classification Tree	0.5958873
	Bagging	0.7027269
	Random Forest	0.6937863
	AdaBoost	0.6428252
	MARS	0.606616
	Logistic Regression	0.606169
Washington	Classification Tree	0.5414481
	Bagging	0.6505771
	Random Forest	0.6463799
	AdaBoost	0.5844701
	MARS	0.5466946
	Logistic Regression	0.5414481
Arizona	Classification Tree	0.5671168
	Bagging	0.6536934
	Random Forest	0.6521048
	AdaBoost	0.5949166
	MARS	0.6155679
	Logistic Regression	0.5782367

Table 3: Out of Sample Predictive Accuracies

This tabulation too firmly points us in the direction of these predictors not being good classifiers of accident severity. The tree we plotted for the state of Florida is represented for reference. The trees plotted for the other states are placed in the appendix. The tree plotted for Washington is too big for it to be represented in this document, and will be enclosed separately as a portable document file.

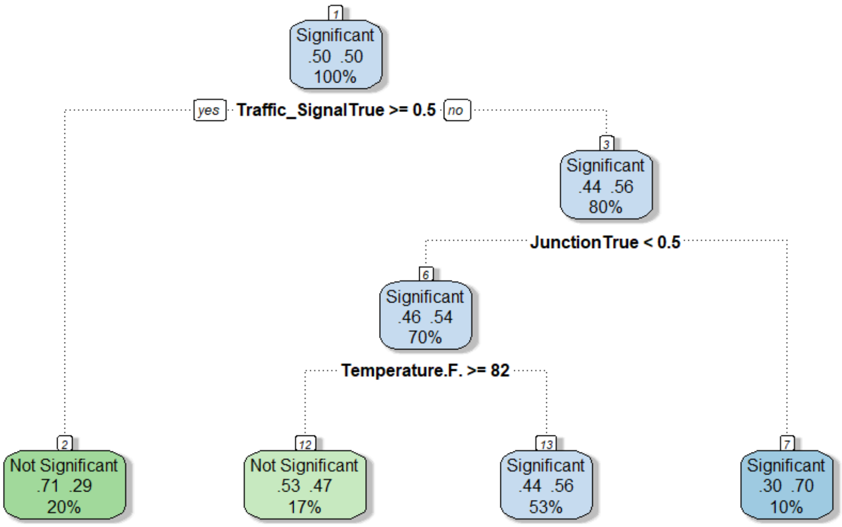

Figure 10: Recursively Partitioned Classification Tree to classify accident Severity for the state of Florida

7 Inference and Discussion

The attempt to classify accident severity using these particular variables was done to try and observe if there exists a strong or any relationship between them. It has been established by several studies that weather stimuli has significant impact on all things related to accidents. In an attempt to try and establish a link between the systemic road conditions (Road POIs) along with weather on such accidents and their severity, we attempted this study. Studying just Road POIs and their impact on accident severity seems a little myopic due to the sparsity such data can exhibit- as observed in the exploratory sections- due to the inherent poor rationale backing such a connection without accounting for any contextual relationship surrounding it. To create the context, we include weather variables. For instance, if a strong relationship had been established between our predictors and target, we were aiming towards making hypotheses like accident severity tended to be significant in intersections during low visibility due to the multi directional traffic compounding not only impact but also cleanup time. However, this was not to be. The variable importance plots plotted using the Random Forest cultivated for these states points to the same conclusion as well, with weather variables generally contributing more to accuracy than road POIs, in line with past studies and our present study. We portray the variable importance obtained from the state of Florida for reference and plot the rest in the appendix.

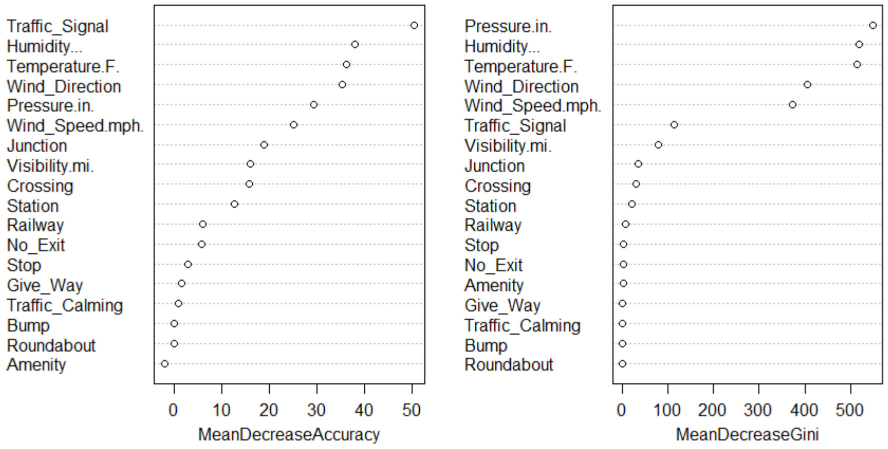

Figure 11: Variable Importance plot: Severity classification for the state of Florida

We obtained fairly low predictive accuracies across models and all the states, which is more than comprehensive for an initial study of this scale to establish that such a relationship between these predictors and the target is weak at best. As a matter of fact, it is barely more accurate than the toss of a coin. Our concluding inference from this study remains that a combination of weather related stimuli and road POIs does not add predictive value in classifying accident severity, and hence, establishes a very weak relationship between the two.

8 Future Research Direction

An unquantifiable number of factors are involved in an accident. Often driver issues (negligence,lack of awareness, intoxication), the actual condition of the road (cracks, potholes, and poor line markings), lack of education on road rules, psychological characteristics of the driver (temper causing road rage) and a lot more uncertainties are involved in crashes. A lot of these are subjective and measuring such factors is a large scale task beyond the scope of this study. A better understanding of the isolated impact of road POIs or just weather stimuli can be better performed when autonomous vehicle crashes are judged, since the human factor involved in these cases is negligible. But autonomous vehicles are not at such common levels of usage or sophistication yet. Similarly, topic modeling of insurance statements from such crash incidents can shed better light on what the humans involved in the crash thought the role of weather, or road POIs or any other such predictors is. At a simpler level, yet higher computational effort, studying all such crashes across all the states of the US instead of isolating four states would be the right first step post this study to attempt to obtain a better predictive accuracy of accident severity to establish a relationship between these covariates.

9 Appendix

9.1 Data Description

The descriptions of each of the columns in the dataset (as defined by the compilers of the data) used above is presented below. [22]

- ID: This is a unique identifier of the accident record.
- Source: Indicates source of the accident report (i.e. the API which reported the accident.)
- TMC: A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.
- 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).
- Start Time: Shows start time of the accident in local time zone.
- End Time: Shows end time of the accident in local time zone.
- Start Lat: Shows latitude in GPS coordinate of the start point.
- Start Lng: Shows longitude in GPS coordinate of the start point.
- End Lat: Shows latitude in GPS coordinate of the end point.
- End Lng: Shows longitude in GPS coordinate of the end point.
- Distance(mi):The length of the road extent affected by the accident.
- Description: Shows natural language description of the accident.
- Number: Shows the street number in address field.
- Street: Shows the street name in address field.
- Side: Shows the relative side of the street (Right/Left) in address field.
- City: Shows the city in address field. Yes
- County: Shows the county in address field.
- State: Shows the state in address field.
- Zipcode: Shows the zipcode in address field.
- Country: Shows the country in address field.
- Timezone: Shows timezone based on the location of the accident (eastern, central, etc.).

- Airport Code: Denotes an airport-based weather station which is the closest one to location of the accident
- Weather Timestamp: Shows the time-stamp of weather observation record (in local time).
- Temperature(F):Shows the temperature (in Fahrenheit).
- Wind Chill(F):Shows the wind chill (in Fahrenheit).
- Humidity(%):Shows the humidity (in percentage).
- Pressure(in):Shows the air pressure (in inches).
- Visibility(mi):Shows visibility (in miles).
- Wind Direction: Shows wind direction.
- Wind Speed(mph):Shows wind speed (in miles per hour).
- Precipitation(in):Shows precipitation amount in inches, if there is any.
- Weather Condition: Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
- Amenity:A POI annotation which indicates presence of amenity in a nearby location.
- Bump : A POI annotation which indicates presence of speed bump or hump in a nearby location.
- Crossing:A POI annotation which indicates presence of crossing in a nearby location.
- Give Way: A POI annotation which indicates presence of give way in a nearby location.
- Junction: A POI annotation which indicates presence of junction in a nearby location.
- No Exit: A POI annotation which indicates presence of no exit in a nearby location.
- Railway: A POI annotation which indicates presence of railway in a nearby location.
- Roundabout: A POI annotation which indicates presence of roundabout in a nearby location.
- Station:A POI annotation which indicates presence of station in a nearby location.
- Stop:A POI annotation which indicates presence of stop in a nearby location.
- Traffic Calming:A POI annotation which indicates presence of traffic calming in a nearby location.

- Traffic Signal: A POI annotation which indicates presence of traffic signal in a nearby location.
- Turning Loop: A POI annotation which indicates presence of turning loop in a nearby location.
- Sunrise Sunset: Shows the period of day (i.e. day or night) based on sunrise/sunset.
- Civil Twilight: Shows the period of day (i.e. day or night) based on civil twilight.
- Nautical Twilight: Shows the period of day (i.e. day or night) based on nautical twilight.
- Astronomical Twilight: Shows the period of day (i.e. day or night) based on astronomical twilight.

9.2 Classification Trees

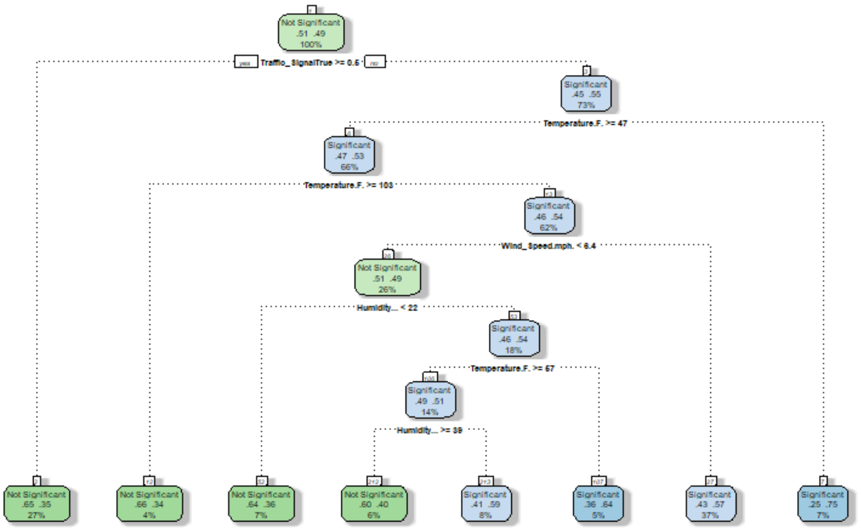

Figure 12: Recursively Partitioned Classification Tree to classify accident Severity for the state of Arizona

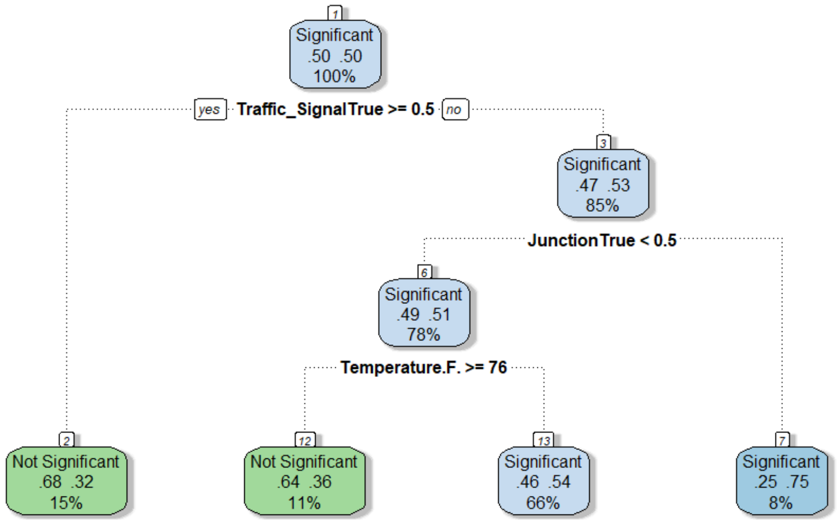

Figure 13: Recursively Partitioned Classification Tree to classify accident Severity for the state of Pennsylvania

9.3 Variable Importance Plots

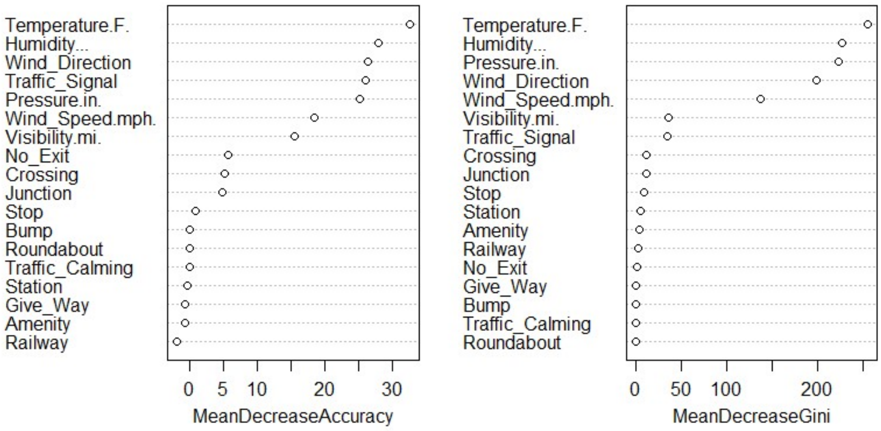

Figure 14: Variable Importance plot: Severity classification for the state of Arizona

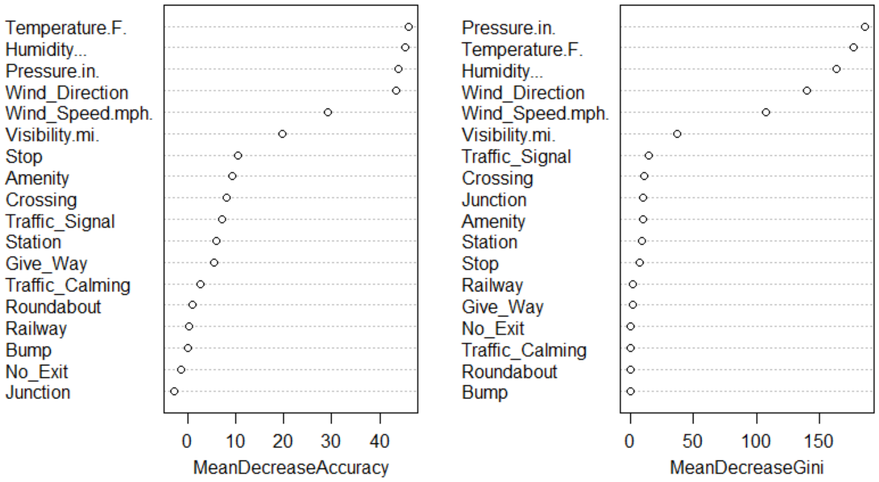

Figure 15: Variable Importance plot: Severity classification for the state of Washington

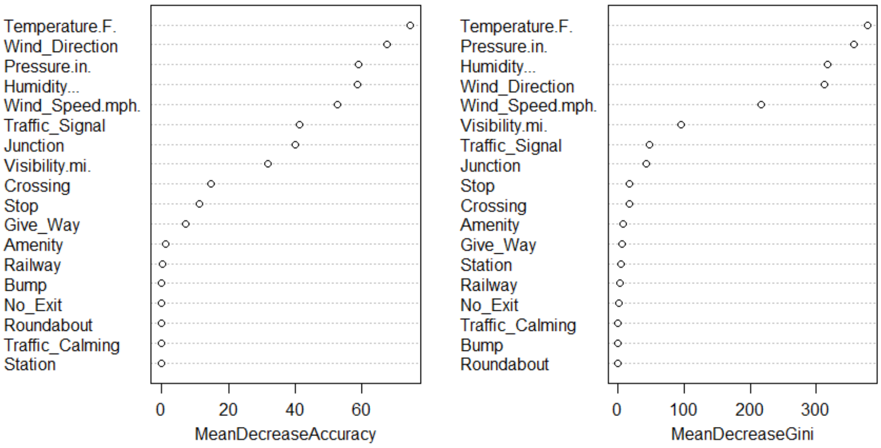

Figure 16: Variable Importance plot: Severity classification for the state of Pennsylvania

9.4 Reference Tables

State	**Model**	**Accuracy**	**Error**	**Hyperparameters**
Florida	Classification Tree	0.5783299	0.4216701	
	Bagging	0.6066253	0.3933747	ntree = 400
	Random Forest	0.6045549	0.3954451	mtry = 10
	AdaBoost	0.6090407	0.3909593	
	MARS	0.6080055	0.3919945	
	Logistic Regression	0.6011042	0.3988958	threshold = 0.5
Pennsylvania	Classification Tree	0.5958873	0.4041127	
	Bagging	0.7027269	0.2972731	ntree = 400
	Random Forest	0.6937863	0.3062137	mtry = 6
	AdaBoost	0.6428252	0.3571748	
	MARS	0.606616	0.393384	
	Logistic Regression	0.606169	0.393831	threshold = 0.45
Washington	Classification Tree	0.5414481	0.4585519	
	Bagging	0.6505771	0.3494229	ntree = 500
	Random Forest	0.6463799	0.3536201	mtry = 8
	AdaBoost	0.5844701	0.4155299	
	MARS	0.5466946	0.4533054	
	Logistic Regression	0.5414481	0.4585519	threshold = 0.5
Arizona	Classification Tree	0.5671168	0.4328832	
	Bagging	0.6536934	0.3463066	ntree = 250
	Random Forest	0.6521048	0.3478952	mtry = 11
	AdaBoost	0.5949166	0.4050834	
	MARS	0.6155679	0.3844321	
	Logistic Regression	0.5782367	0.4217633	threshold = 0.55

Table 4: Out of Sample Model Performance

References

[1] www.nsc.org. (n.d.). Retrieved from https://www.nsc.org/Portals/0/Documents/NewsDocuments/2018/December-2017.pdf.

[2] 2018 Fatal Motor Vehicle Crashes: Overview. (2019, October). Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812826.

[3] Chong, M., Abraham, A., Paprzycki, M. (2004, August). Traffic accident data mining using machine learning paradigms. In Fourth International Con- ference on Intelligent Systems Design and Applications (ISDA'04), Hungary (pp. 415-420)

[4] Qiu, C., Wang, C., Fang, B., & Zuo, X. (2014). A multiobjective particle swarm optimization-based partial classification for accident severity analysis. Applied Artificial Intelligence, 28(6), 555-576.

[5] Kwon, O. H., Rhee, W., & Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. Accident Analysis & Prevention, 75, 1-15.

[6] Shanthi, S., & Ramani, R. G. (2012, October). Feature relevance analysis and classification of road traffic accident data through data mining techniques. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 1, pp. 24-26).

[7] Farmer, C. M. (2017). A projection of United States traffic fatality counts in 2024.

[8] Chong, M., Abraham, A., & Paprzycki, M. (2004, August). Traffic accident data mining using machine learning paradigms. In Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04), Hungary (pp. 415-420).

[9] Theofilatos, A. (2017). Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. Journal of safety research, 61, 9-21.

[10] Malin, F., Norros, I., & Innamaa, S. (2019). Accident risk of road and weather conditions on different road types. Accident Analysis & Prevention, 122, 181-188.

[11] Sherretz, L. A., & Farhar, B. C. (1978). An analysis of the relationship between rainfall and the occurrence of traffic accidents. Journal of Applied Meteorology, 17(5), 711-715.

[12] Savolainen, P. T., Mannering, F. L., Lord, D., & Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis & Prevention, 43(5), 1666-1676.

[13] Shankar, V., & Mannering, F. (1996). An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. Journal of safety research, 27(3), 183-194.

[14] Shankar, V., Mannering, F., & Barfield, W. (1996). Statistical analysis of accident severity on rural freeways. Accident Analysis & Prevention, 28(3), 391-401.

[15] Montella, A., Aria, M., D'Ambrosio, A., & Mauriello, F. (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis & Prevention, 49, 58-72.

[16] Rovšek, V., Batista, M., & Bogunović, B. (2017). Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree. Transport, 32(3), 272-281.

[17] Badr, W. (2019, April 20). Having an Imbalanced Dataset? Here Is How You Can Fix It. Retrieved from https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb.

[18] Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019, November). 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). ACM.

[19] Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 875-886). Springer, Boston, MA.

[20] Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A Countrywide Traffic Accident Dataset. arXiv preprint arXiv:1906.05409.

[21] Logistic Regression. (n.d.). Retrieved from https://www.stat.cmu.edu/ cshalizi/uADA/12/lectures/ch12.pdf.

[22] James, G., Witten, D., Hastie, T., Tibshirani, R., (2017). Logistic Regression. In An introduction to statistical learning: with applications in R (pp. 130–138). New York: Springer.

[23] Multivariate Adaptive Regression Splines. (n.d.). Retrieved from http://uc-r.github.io/mars.

[24] James, G., Witten, D., Hastie, T., Tibshirani, R.(2017). Regression Splines. In An introduction to statistical learning: with applications in R (pp. 271–276). New York: Springer.

[25] Classification and Regression Trees. (2009, November). Retrieved from https://www.stat.cmu.edu/ cshalizi/350/lectures/22/lecture-22.pdf.

[26] James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). Classification Trees. In An introduction to statistical learning: with applications in R (pp. 311–313). New York: Springer.

[27] Mukherjee, S., & Papazaharias, D. (2019, October). Regression and Classification Trees. Data Analytics and Predictive Modeling, IE 500. Buffalo, NY.

[28] James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). Bagging. In An introduction to statistical learning: with applications in R (pp. 316–318). New York: Springer.

[29] Mukherjee, S., & Papazaharias, D. (2019, October). Bagging, Random Forest. Boosting. Data Analytics and Predictive Modeling, IE 500. Buffalo, NY.

[30] James, G., Witten, D., Hastie, T., Tibshirani, R.(2017). Random Forests. In An introduction to statistical learning: with applications in R (pp. 319–320). New York: Springer.

[31] James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). Boosting. In An introduction to statistical learning: with applications in R (pp. 321–322). New York: Springer.