Predicting Road Accidents : An Analysis of the impact of Road Condition and Weather to Car Accidents in Seattle USA

Marc Jerrone Castro

August 23, 2020

1. Introduction

1.1 Background

Every year, approximately 6 million car crashes occur in the U.S. where 72% of these crashes result in property damage, 27 % resulting in injuries ranging from bruises to life-changing disabilities, and 6% resulting in death. Although most of these crashes are caused by alcohol, speeding, and general recklessness - there is the underlying factor of weather and road conditions that may have escalated these avoidable accidents. Factors such as road traction, rain, lighting, and even the type of traffic in the area may play a role in predicting these accidents and could potentially serve as a tool to avoid them.

1.2 Problem

Collision data which captures road conditions and general weather might assist in identifying key or significant factors in crashes within the Seattle Area. This project aims to identify correlations between these factors and road crash severity and predict the probability of a crash occurring and it's severity with respect to these factors.

1.3 Interest

Not only the Seattle City Government, but city governments of areas with similar climates and conditions may apply a predictive model to warn drivers of the potential harms they may encounter on the road. Local news stations may also give out warnings about these factors and relay potential car crash severities such that local drivers would be more attentive given these road conditions. Finally, the primary targets of this study are the drivers whereas the resulting model from this project may further be developed as a real-time application that calculates risk and potential dangers given information such as weather, time, and road condition.

2. Data Acquisition & Learning

2.1 Data Source

Seattle **c**ar accident data such as car crash severity, type of street, weather, and other information from 2004 up to this year can be found from the following link:

https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d

The data contains 220,937 entries with 40 columns containing information such as the severity of the accident, address type, the weather, road lighting and other information such as number of people involved and exact date and time of accident.

2.2 Data Cleaning & Feature Selection

Although the data contains a lot of useful information, we limited our scope to some portion of the data. For this project, we limited our study to the following features:

Table 1. Selected Features from the Seattle Collision Dataset.

FEATURE	**DESCRIPTION**	**DATA TYPE**
VEHCOUNT	The number of vehicles involved in the collision.	INTEGER
INJURIES	The number of total injuries in the collision	INTEGER
SERIOUSINJURIES	The number of total serious injuries in the collision	INTEGER
FATALITIES	The number of total fatalities in the collision	INTEGER
'WEATHER'	A description of the weather conditions during the time of the collision.	CATEGORICAL STRING
'ROADCOND'	The condition of the road during the collision.	CATEGORICAL STRING
'LIGHTCOND'	The light conditions during the collision.	CATEGORICAL STRING

One issue we had to address with the data were the categorical variables and our newly binned features. First we addressed the issue of missing values in our categorical values. For the purpose of consistency, all missing values were removed from the dataset entirely as converting them to their respective highest modes would only further contribute to the class imbalance that is inherent within the dataset. Afterwards, we identified that vague values existed within some of the categorical variables - whereas we opted to remove all data points with a value of ‘Others’ while opting to replace all ‘Unknowns’ with the most common value. As these features are represented by a string, we first address these variables by generating one-hot encodings for each data point and removing an extra column to accommodate the errors caused by the dummy variable trap.

Upon close inspection of the dataset, we also observed that some categorical features had similar types and are thus redundant. We opted to group similar features as to limit the impact of class imbalance within our dataset. Take for example how we grouped the Weather features where we classified special types of weather as Extreme Weather and classify Overcast and Partly Cloudy as Cloudy Weather as these usually follow each other and are highly similar. As for the road conditions, we opted to combine oil and ice as they are all road conditions relating to a lower level traction on the road. We also opted to simply absorb sand/mud/dirt into Dry as they are mostly similar. We also decided to combine wet and standing water into one class as well as they involve the presence of water on the road. We performed similar grouping on the lighting conditions by classifying them based on how well lit the area is.

Lastly, we generated a new measure for determining how severe a crash was as the SEVERITYCODE within the dataset was unable to capture the total number of people affected and was solely dependent on indicators such as *Did someone get minor injuries? Did someone die?*. Our scoring measure, LIFECOST, was factored based on the number of individual indicators and scaled based on their general impact to life. Whereas, vehicular damage was scaled by 1, minor injuries were scaled by 10, major injuries by 100, and deaths by 1000. These were then classified based on how large of an impact the crash resulted in.

Table 2. Classification of Life Cost Scoring

Class	**Conditions** *x = LifeCostScore	**DESCRIPTION**
0	x<= 9	Indicates mostly Property Damage
1	(x>9) and (x<=50)	Indicates Minor Crash involving some minor injuries

2	(x>50) and (x<=100)	Indicates a Severe Crash with multiple people having minor injuries
3	(x>100) and (x<=500)	Indicates a Major Crash with major injuries
4	(x>500) and (x<=2000)	#Indicates a Major Crash with major injuries and death
5	x>2000	#Indicates a Massive Crash with multiple deaths

3. Exploratory Data Analysis (Methodology)

3.1 Relationship of Weather and Road Conditions to Road Crash Severity

Our hypothesis was severe crashes generally occur more frequently under non-ideal weather and road conditions. Based on Figure 1 and 2, we can observe that the data supported our case. We observed that although most crashes had similar severities between non-ideal and ideal conditions, crashes that had a LIFECOST class of 4 or 5 were generally more prevalent under non-ideal conditions. This suggests that a certain level of correlation was present between LIFECOST and Weather and Road Conditions for predicting the likelihood of crashes with a LIFECOST class of 4 or 5.

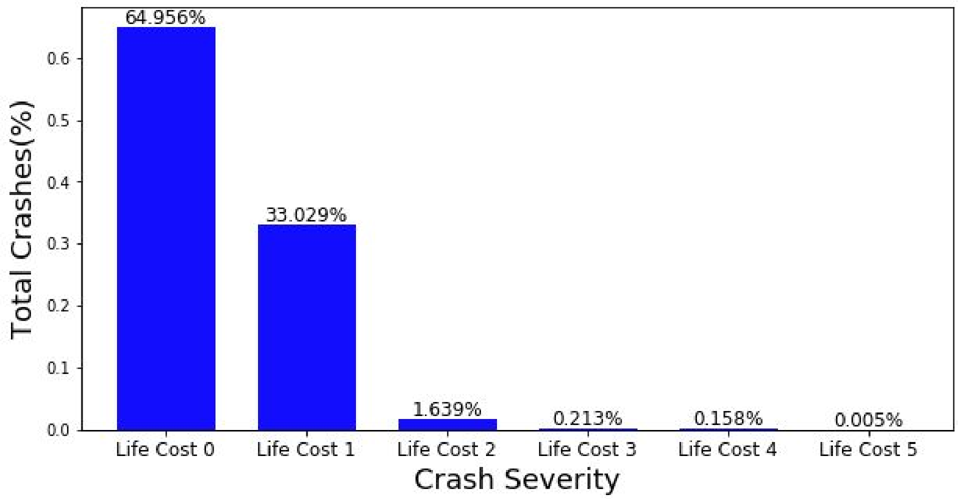


Figure 1. Distribution of Crash Severity based on Life Cost under ideal conditions (Clear weather, Dry Road, Well Lit Road).

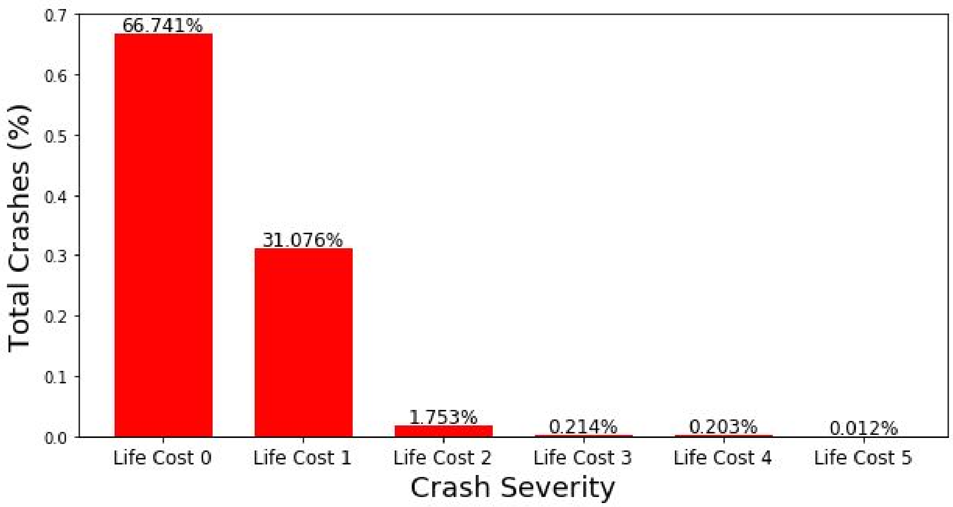


Figure 2. Distribution of Crash Severity based on Life Cost under non-ideal conditions

As we had observed that crashes with under ideal conditions and non-ideal conditions had similar chances of occurring for LIFECOST class of 1, 2, or 3, we have decided to focus our study on the potential for death and injuries to occur under non-ideal conditions.

3.2 Relationship of Weather and Road Conditions to Injuries

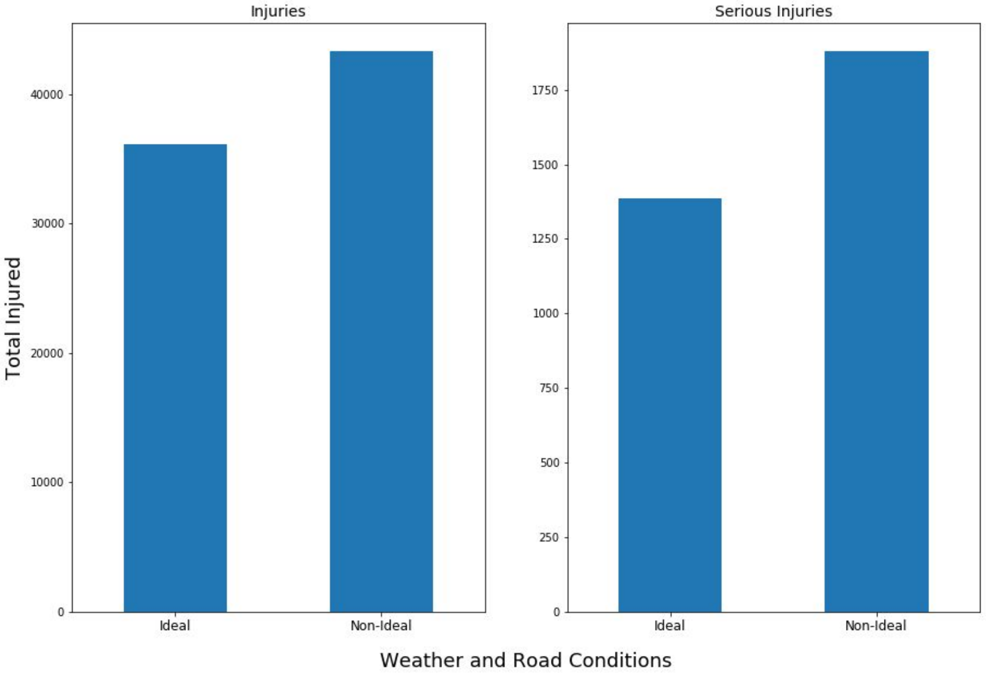


Figure 3. Comparison of Injuries and Serious Injuries count based on Weather and Road Conditions.

Our hypothesis for injuries is that non-ideal conditions should generally have more injuries as there is an increased risk in driving. From Figure 3, we can observe that the data actually agrees with our hypothesis and shows that non-ideal conditions generally result in more injuries (both minor and major). Whereas, there was a 9.0% increase in minor injuries and a 15.1% increase in serious injuries under non-ideal driving conditions in contrast to minor and major injuries under ideal driving conditions.

3.3 Relationship of Weather and Road Conditions to Deaths

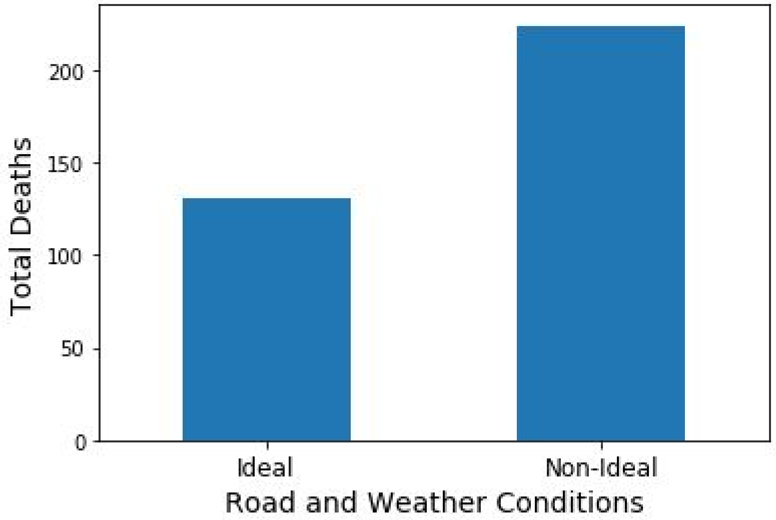


Figure 4. Deaths based on Weather and Road Conditions.

Similar to our hypothesis for injuries, we expect that non-ideal conditions would lead to significantly more deaths. As observed, there was a 26.2% increase in deaths under non-ideal conditions. This slight increase is due to the increase in factors that may affect the probability of an accident occurring. Additionally, unlike during ideal driving conditions, there is a reduced effect on stop-gaps for severe accidents during non-ideal driving conditions such as reduced traction that dampens the stopping power of breaks. Mechanical components of vehicles may also be affected during non-ideal conditions such as the freezing of an engine and other similar factors.

3.4 Injuries and Deaths with respect to Non-Ideal Weather and Road Conditions

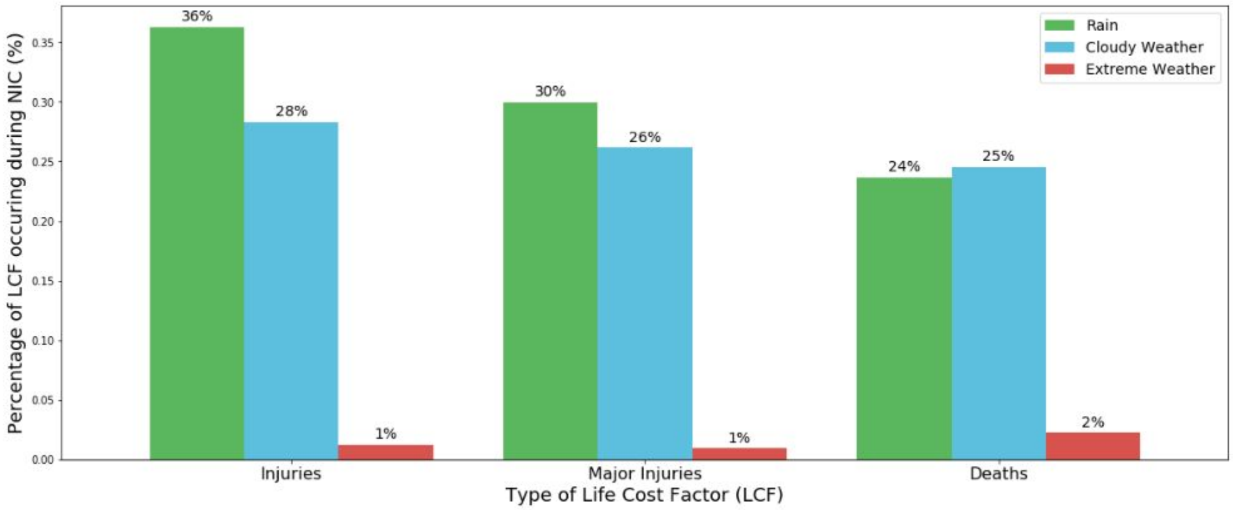


Figure 5. Percent contribution of each non-ideal weather to each LCF during non-ideal driving conditions.

We hypothesize that non-ideal weather conditions would make up the majority of injuries and deaths due to the additional risks during driving. From our initial analysis of the data, we can see that the majority of accidents that resulted in injury or death occurred during rain, cloudy weather, and extreme weather. Note that it was necessary to review these features using their features as there are generally more rainy days in contrast days with extreme weather. From Figure 5, we can see that our data agrees with our results - whereas 65% of accidents that led to injury happened during rainy weather, 57% of accidents that led to major injuries and 51% of deaths happened during non-ideal weather conditions. Although the magnitude of these percentages were smaller than expected, this observation is still relevant as they capture that non-ideal weather conditions are generally more perceptible to accidents that lead to injuries.

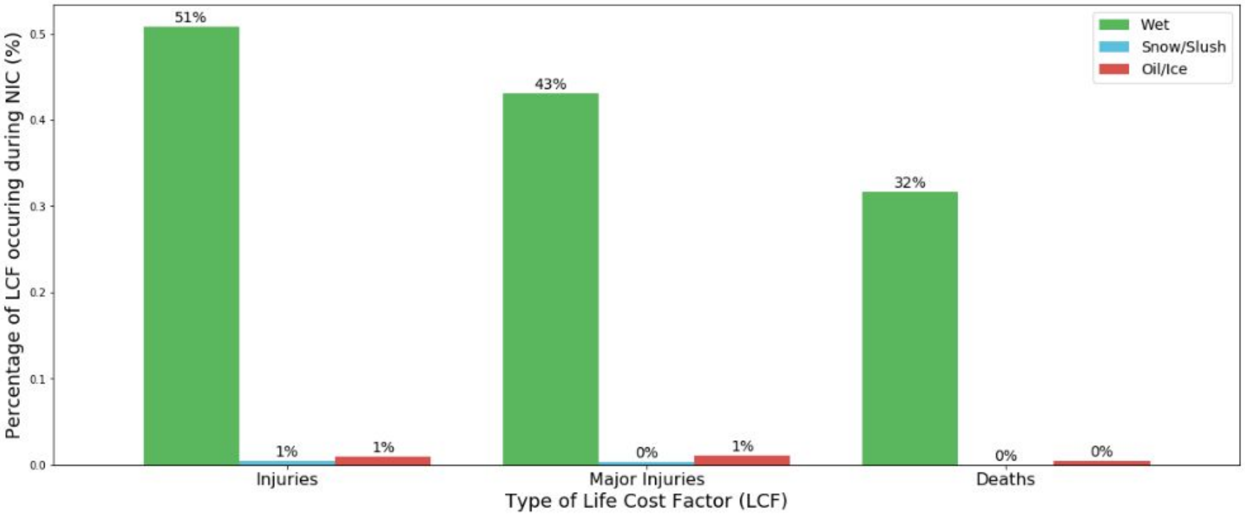


Figure 6. Percent contribution of each non-ideal road conditions to each LCF during non-ideal driving conditions.

As for injuries and deaths with respect to non-ideal conditions, we can see from Figure 6 that most major injuries and deaths actually occurred during ideal road conditions (Dry) - although, injuries occur slightly more frequently during non-ideal conditions. From this, we can conclude that road conditions play a lesser role in the severity of car crashes as injuries and deaths occur more frequently during ideal conditions.

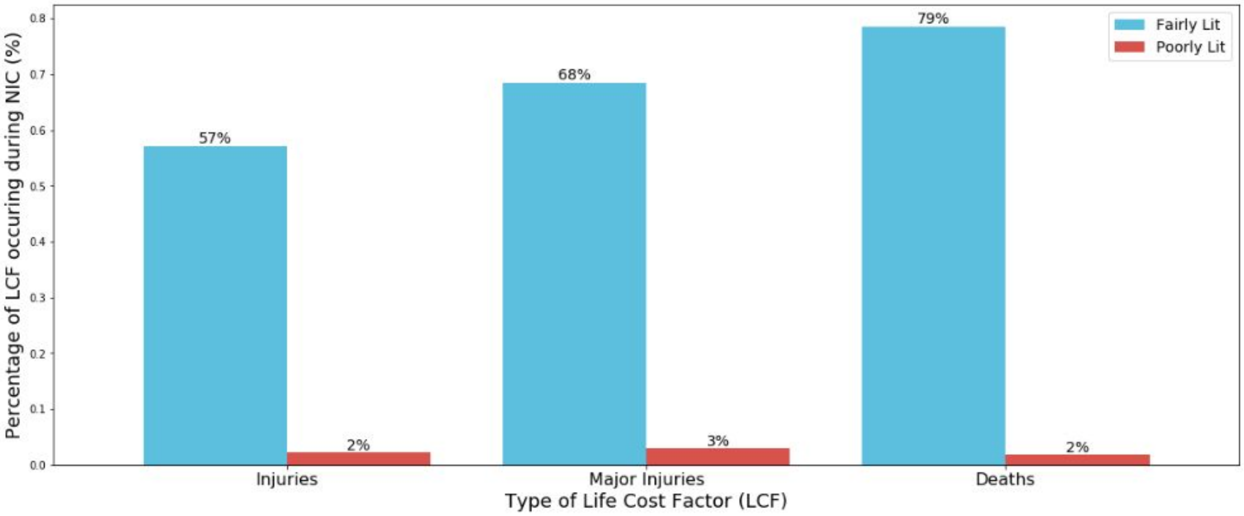


Figure 7. Percent contribution of each non-ideal road lighting to each LCF during non-ideal driving conditions.

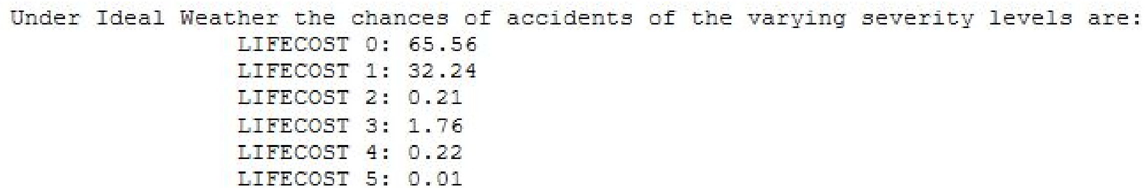
In regards to Light conditions, we hypothesized that under accidents leading to injuries and deaths should occur more frequently under poorly lit conditions. However, based on Figure 7, our data disagrees with our hypothesis as Fairly Lit conditions (e.g. Dawn, Dusk, Dark - Street Light On) composed most of the accidents that occurred in Seattle. We can assume that drivers are more attentive during dark conditions as they have limited visibility and that they become slightly more careless in roads with moderate lighting.

4. Predictive Modeling

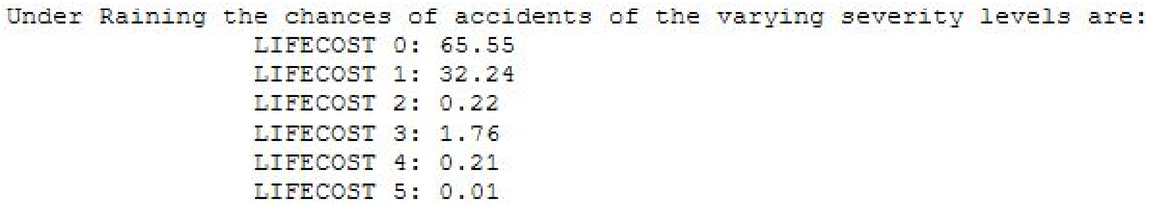
Our main objective for this project is to predict the probability of severe car crashes occurring during non-ideal weather and road conditions. As such we develop a classification model to predict the probability of an accident occurring and attempt to present varying probabilities of a LIFECOST classes.

4.1 Applying Standard Classification Algorithms.and Problems Encountered.

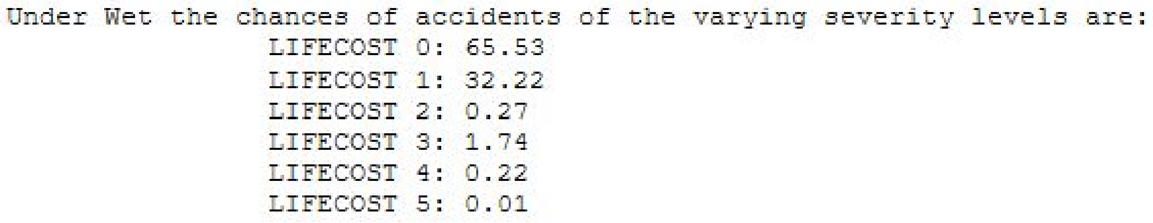
I initially split the dataframe composed of non-ideal weather and road conditions into training and testing sets with a test split ratio of 0.155. Upon training these models (SVM, K-Nearest Neighbors, Decision Tree, and Logistic Regression) and identifying the ideal hyperparameters for each model, I observed that all data points in the testing set were categorized as LIFECOST 0 as they make up the bulk ~67% of the target variable. Note that we opted to utilize the Logistic Regression for our final model as it is easier and faster to train. However, as we are dealing with minimal data and repeating values for our data points with varying labels, it is only expected. Thus we hope to focus our efforts on the prediction probabilities of each classification of LIFECOST. The issue with this approach is that we have yet to resolve the issue of the bulk of target data is dominated by two classes and as such predictive probabilities are also skewed in their favor, whereas given the following scenarios, the probability of an accident occurring are:



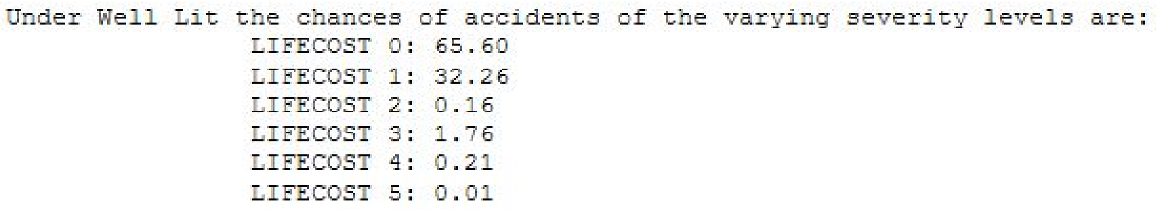
```
Under Ideal Weather the chances of accidents of the varying severity levels are:
                LIFECOST 0: 65.56
                LIFECOST 1: 32.24
                LIFECOST 2: 0.21
                LIFECOST 3: 1.76
                LIFECOST 4: 0.22
                LIFECOST 5: 0.01
```



```
Under Raining the chances of accidents of the varying severity levels are:
                LIFECOST 0: 65.55
                LIFECOST 1: 32.24
                LIFECOST 2: 0.22
                LIFECOST 3: 1.76
                LIFECOST 4: 0.21
                LIFECOST 5: 0.01
```



```
Under Wet the chances of accidents of the varying severity levels are:
                LIFECOST 0: 65.53
                LIFECOST 1: 32.22
                LIFECOST 2: 0.27
                LIFECOST 3: 1.74
                LIFECOST 4: 0.22
                LIFECOST 5: 0.01
```



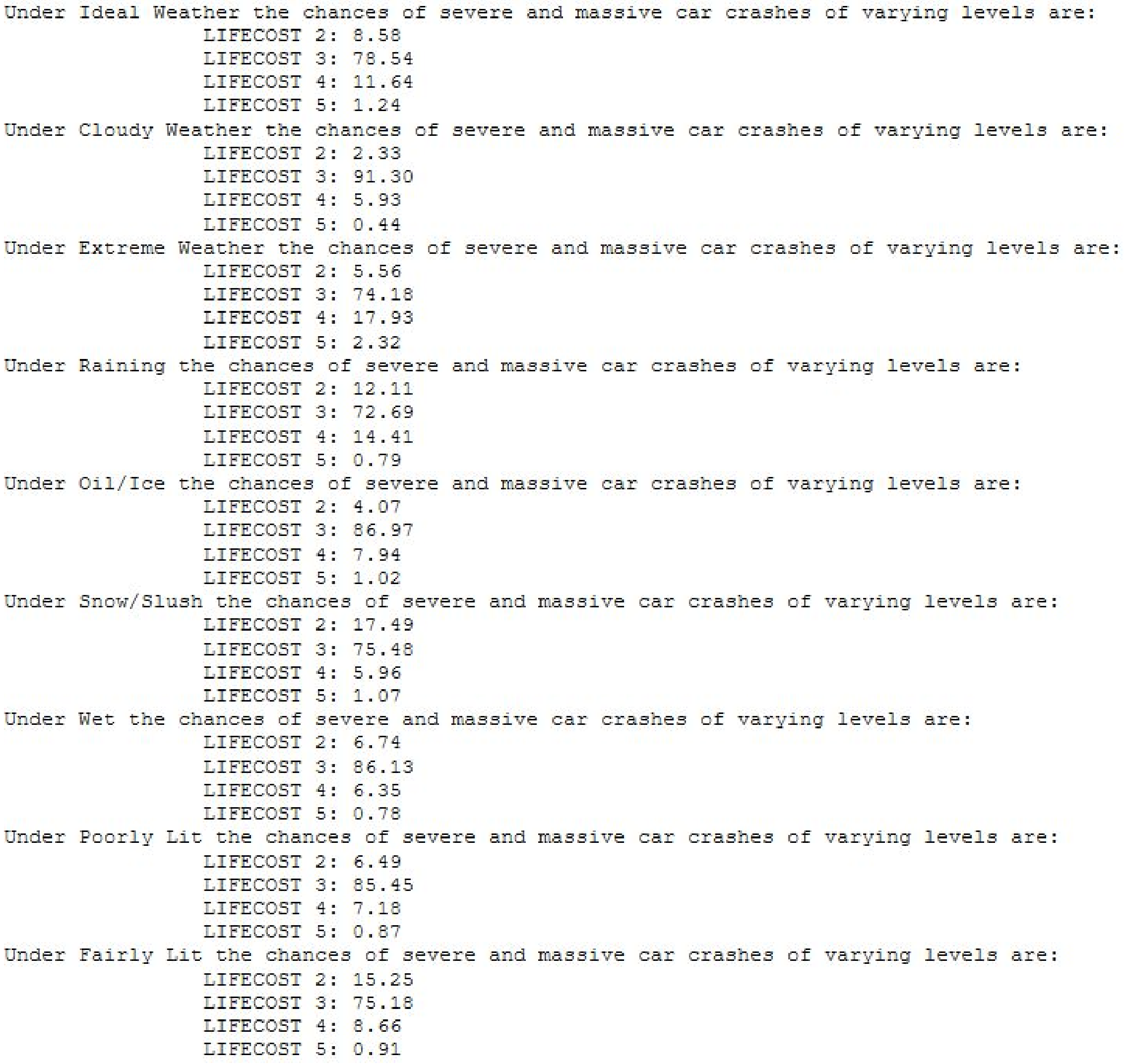
```
Under Well Lit the chances of accidents of the varying severity levels are:
                LIFECOST 0: 65.60
                LIFECOST 1: 32.26
                LIFECOST 2: 0.16
                LIFECOST 3: 1.76
                LIFECOST 4: 0.21
                LIFECOST 5: 0.01
```

As you might observe, although there were varying conditions, there were minimal changes on the probability of each accident occurring with a corresponding

LIFECOST. Due to the large data imbalance we were unable to discern the changes in probability for car crashes involved with major and severe injuries as well as deaths.

4.2 Our Solution to Class Bias

To remove the bias of having move values of LIFECOST 0 and LIFECOST 1, we opted to simply isolate the data points with LIFECOST value greater than or equal to two. By doing this, we are able to give an insight into car crashes that deal with severe crashes involving a number of injuries (mostly major) and potentially fatalities. Whereas, considering the following scenarios, the likelihood of severe to massive crashes occurring are:



```
Under Ideal Weather the chances of severe and massive car crashes of varying levels are:
                LIFECOST 2: 8.58
                LIFECOST 3: 78.54
                LIFECOST 4: 11.64
                LIFECOST 5: 1.24
Under Cloudy Weather the chances of severe and massive car crashes of varying levels are:
                LIFECOST 2: 2.33
                LIFECOST 3: 91.30
                LIFECOST 4: 5.93
                LIFECOST 5: 0.44
Under Extreme Weather the chances of severe and massive car crashes of varying levels are:
                LIFECOST 2: 5.56
                LIFECOST 3: 74.18
                LIFECOST 4: 17.93
                LIFECOST 5: 2.32
Under Raining the chances of severe and massive car crashes of varying levels are:
                LIFECOST 2: 12.11
                LIFECOST 3: 72.69
                LIFECOST 4: 14.41
                LIFECOST 5: 0.79
Under Oil/Ice the chances of severe and massive car crashes of varying levels are:
                LIFECOST 2: 4.07
                LIFECOST 3: 86.97
                LIFECOST 4: 7.94
                LIFECOST 5: 1.02
Under Snow/Slush the chances of severe and massive car crashes of varying levels are:
                LIFECOST 2: 17.49
                LIFECOST 3: 75.48
                LIFECOST 4: 5.96
                LIFECOST 5: 1.07
Under Wet the chances of severe and massive car crashes of varying levels are:
                LIFECOST 2: 6.74
                LIFECOST 3: 86.13
                LIFECOST 4: 6.35
                LIFECOST 5: 0.78
Under Poorly Lit the chances of severe and massive car crashes of varying levels are:
                LIFECOST 2: 6.49
                LIFECOST 3: 85.45
                LIFECOST 4: 7.18
                LIFECOST 5: 0.87
Under Fairly Lit the chances of severe and massive car crashes of varying levels are:
                LIFECOST 2: 15.25
                LIFECOST 3: 75.18
                LIFECOST 4: 8.66
                LIFECOST 5: 0.91
```

As you might observe, the increases and decreases in likelihood of each class of severity from occurring changes depending on the condition of the weather and the road. Such that from these scenarios, we can derive the following conclusions given the assumption that there is an immensely greater chance of getting into an accident involving no major injuries:

- During cloudy weather, there are generally more crashes involved with major injuries and less with death.
- During extreme weather, it is 6.29% more likely to be involved in accidents involving multiple major injuries and death in contrast to ideal driving conditions. While nearly double the likelihood of involvement with accidents that have multiple deaths.
- During rain it is 2.77% more likely to have more major injuries and a higher chance of death in contrast to ideal driving conditions.
- On a wet road, you are more exposed to accidents involving major injuries than to those involving minor injuries.
- With a fairly lit road, there is ~ 80% increase of the possibility that you are involved in a crash with multiple minor injuries and are less likely to be involved in more severe crashes. This is similar for crashes during snowy/slush roads where there was a lower probability for high severity crashes to occur and are more inclined toward crashes with lesser life impact.

5. Conclusions

In this study, I analyzed the relationship between car crash severity under ideal and non-ideal driving conditions with respect to weather, road, and light conditions in Seattle USA from the period of 2004 to present. I was able to identify that there were significant differences in the likelihood of higher severity between ideal and non-ideal driving conditions. Specifically, I noticed that some conditions resulted in higher probabilities of major injuries and deaths such as extreme weather, rain, and wet roads. There were also instances that resulted in a lower probability for accidents involving multiple major injuries such as driving on fairly lit roads (Dusk, Dawn, Dark - with Street Lights On) in contrast to well lit roads(Mid-day light). The resulting model is promising as it informs drivers and city officials of the potential dangers of non-ideal driving conditions such that they are able to address these problems.

6. Future Directions

As this study was mainly concerned with categorical data, one way to improve the model would be to find a dataset that contains specific nominal information on the weather and road conditions during the time of accident. Information such as wind speed, temperature, humidity, friction, and etc. could potentially determine key factors that result in these crashes.