## 1. Introduction

Car accidents happen every day around, regardless of the weather. However, with certain season approaching, the concern of car accident due to bad weather increases. Traffic is difficult enough to deal with in perfect conditions. When you add fog, rain, sleet, snow, or freezing rain, you could create conditions that sounds perfect for a car accident.

The problem we are trying to answer in this project is that given the weather, the road conditions and some other variables, could we predict the possibility of a driver getting into a car accident and estimate how severe it would be.

This could provide some warnings to the drivers so that they would drive more carefully. These results could aid issuance companies' risk estimation and stop them from using bad weather as an excuse to deny valid car accident claims.

## 2. About the data

The data we used were downloaded directly from the Coursera course website and contains the data of 194673 cases of car accidents happened in Seattle, starting from 2004 to present. The dataset is generated by the SDOT Traffic Management Division and the Traffic Records Group. The dataset contains 37 descriptive attributes including weather and road condition of the car crash, and the dataset provided the estimate severity of each car accident.

After some exploratory data analysis, we found that some attributes were either dominated by one category or simply have too many missing entries. These columns were left behind with the attributes that are not related to our problems, like the angle of collision. Eventually, we selected 5 attributes to train our machine learning models. They are the location type, the weather condition, the road condition, the light condition of the crash, and whether the driver was driving under influence. Our target field is the severe code, which is what we want to estimate: the level of severeness.

## 3. Methodology

### 3.1 Data cleaning and selection

The first step we did is inspect all 37 attributes' data type and description. We removed the attributes that are not helpful to solve our problem, such as the angle of collision and the case code given to the collision by SDOT. Like mentioned in the previous section, we also found that some attributes were either dominated by one category or simply have too many missing entries. These attributes include whether the crash involves hitting a parked car. All these columns were out of consideration for our exploratory data analysis.

Eventually, we were left with six independent variables, the location type, the weather condition, the road condition, the light condition of the crash, and whether the driver was driving under influence. Also, to aid our calculation, we numerated the categorial variables into binary codes.

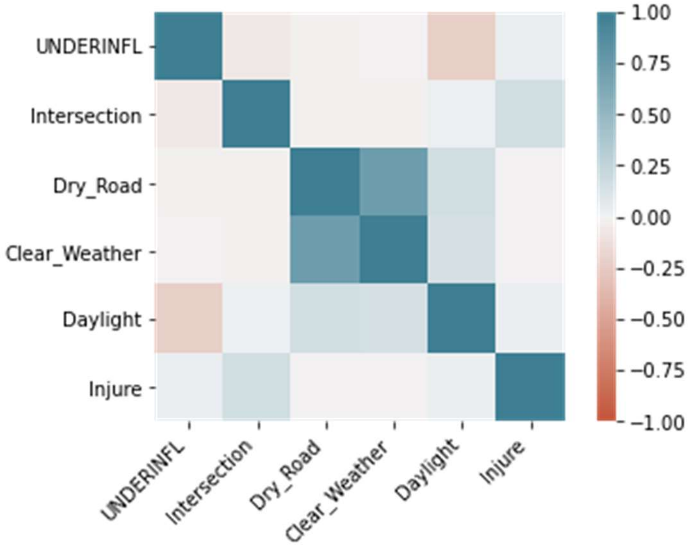
To aid our calculation, we further generalized the categorical variables into binary variables, so that we can approach the problem by building a binary classification model.

Noticed that the data we were given only contains two types of severity code, 1 for property damage, 2 for injury. We modified the attributes by subtracting 1 from all entries. Now we have 0 for no injury/property damage, 1 for injury.

Also, majority of the weather column is occupied by 'clear', 'raining', and 'overcast'. We decided to modify the weather column to have only two variables, 1 for clear sky, and 0 for bad weather. Similar actions were taken on other attributes as well. See the code for details.

### 3.2 Exploratory Data Analysis

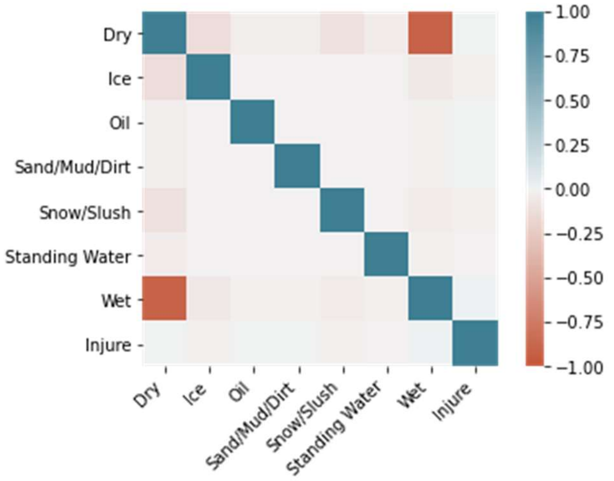
Now we look deeper into our filtered attributes. We further calculated the correlation between each attribute with our target field, and the result is visualized in the following figure.

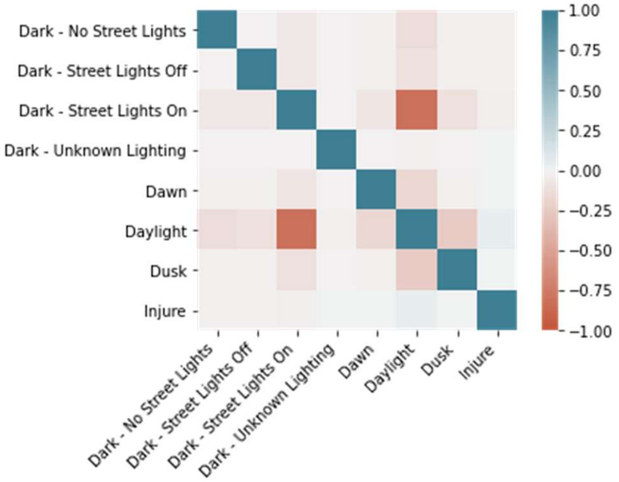


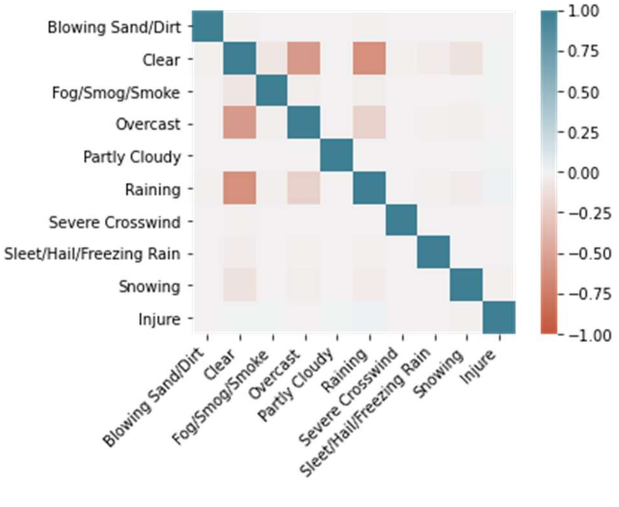
We can see that the severity of the collision is highly correlated with the type of location where is happened, which is at intersections. Besides this, other attributes like weather did not show a significance correlation with the severity of the collision.

Interestingly, we observed a negative correlation between daylight condition and the DUI (Drive under influence) variable. Indicating that people are more likely to be driving under influence when it is getting dark. Also, more than 50% of the collisions involves driving under influences.

At this point, we were wondering if we generalized the categories too early so that we missed some details. We then used one hot key encoding to calculate the correlation between different road, light and weather condition. The results are plotted below:







We can see that none of these conditions is significantly correlated with the severity of injury. Therefore, it is sufficient for us to modify the attributes to binary variables. At this stage, we decide to apply machine learning.

### 3.3 Machine Learning

Since we are using multiple independent variables to predict the severity of injury, which is binary due to our modification mentioned in previous sections, we decided to start solving this binary classification problem with a supervised machine learning method called logistic regression.

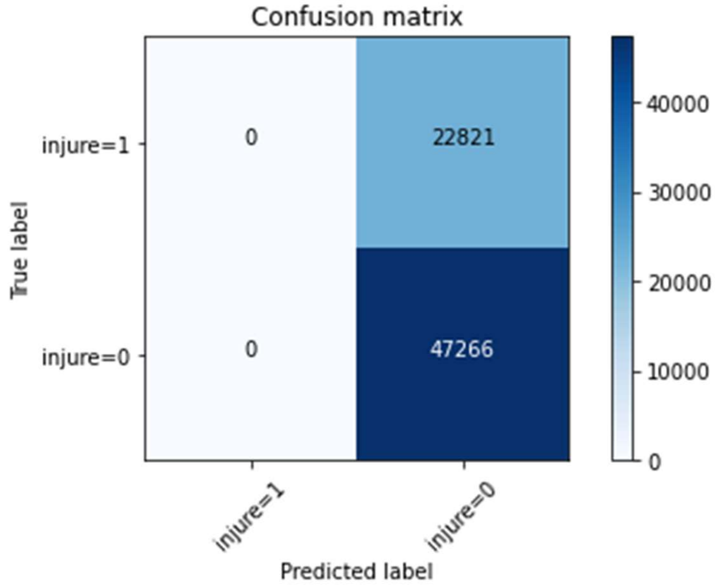
Logistic regression is suitable for this situation because it returns a probability score between zero and one for a given sample of data. In fact, logistic regression predicts the probability of that sample and we map the cases to a discrete class based on that probability.

To solve our problem, we will use the location type (intersection or not), the weather condition (clear sky or not), the road condition (dry road or not), and the light condition (daylight or dark) of the crash as our input variable. We don't consider the DUI attribute because it is controlled by the driver rather than random probability.

We split our data into training set and testing set. The training set contains 60% of our data while the rest 40% is used for testing. Also, to ensure accuracy, we randomly select the sets 5 times.

## 4. Results:

To help evaluate our model visually, we plotted the confusion matrix.



We can see our model predict the collisions recorded in all test cases are property damage. This result means that our data is not balanced. We also used F1-score and Log Loss score to evaluate our model, which have desired values of 1 and 0. Our model returns an average F1-score of 0.67 and Log Loss score of 0.62. This indicates that our model is not very accurate since we have a low F1-score and high Log Loss score.

## 5. Discussion

We also tried other classification models like decision tree and K-Nearest Neighbors with different modeling parameters. Both these methods returned average F1-score around 0.66. However, it is worth to point out that the K-Nearest Neighbors method performs best at K=18 and has significantly higher precision in predicting injuries. The K-Nearest Neighbors method has a 29% accuracy to predict the injury case, while other two methods cannot predict injury at all, which is 0% accuracy.

## 6. Conclusion:

Unfortunately, we were not able to use machine learning model to predict the possibility of a driver getting into a car accident and estimate how severe it would by using the given data in this project. It could be due to the unbalanced we were given, or the lack of sample size of the conditions we want to consider. But the discovery on the K-Nearest Neighbors method shed some light on the correct path. If we have access to a bigger, more diverse and balanced dataset, we may be able to increase our model accuracy. Given the weather, the road conditions and some other variables, it is highly possible to predict the possibility of a driver getting into a car accident and estimate how severe it would be.