# Predicting the Severity of the Seattle Road Accidents

Nikhil P

September 6th, 2020

## 1. Introduction:

### 1.1 Background:

The seaport city of Seattle is the largest city in the state of Washington, as well as the largest in the Pacific Northwest. As of the latest census, there are approx. 740K people living in Seattle. Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it's no surprise that Seattle sees car accidents every day.

In process of providing the people of Seattle a safer transportation system, SDOT is implementing signals that give pedestrians at crosswalks a three- to seven-second head start before drivers get a green light to make turns. The system, called leading pedestrian intervals, makes pedestrians in the crosswalk more visible to drivers making turns.

So SDOT, SPD & Seattle police are taking all the preventive steps to ensure the Seattle's people a safe journey, and they are the first responders in the event of any unpredicted accident. It would be ideal if they have any pre estimation tool to help them to quickly respond based on the estimated severity of the accident using their data they gathered all these years.

### 1.2 Problem:

The data we have is about accidents in Seattle city. With the data we have, we can predict the severity of the accident. So my attempt here or problem I choose here is to derive the severity of the accident. Generally in accidents, the primary things we would verify are road condition and visibility on the road and sometimes the weather plays a major role especially during rains and winter. Besides these speeding is the key indicator to judge the accident's cause. As all this information is there in that Seattle accident's data.

### 1.3 Audience/Interests:

As we are predicting the severity of the accident here, the primary audience will be the traffic police to do the primary analysis at offsite and take further steps based on the outcome. Besides them, insurance stakeholders can use this to predict the likely outcome if their policyholders had involved in this as an initial investigation report offsite. And a small set of the group would be the people who can use to know the likelihood of an unexpected journey based on factors like weather, road & lighting condition.

## 2. Data acquisition and cleaning:

### 2.1 Data sources:

The data in this data source is provided by SPD and recorded by Traffic Records. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.

### 2.2 Data Cleaning:

We had circa 193K+ records for accidents during the time span starting from 2004 to the recent past with 37 Independent variables and outcome or labelled data will severity of the accident. Severity has been categorized into 5 categories by SDOT.

A code that corresponds to the severity of the collision:

3—fatality

2b—serious injury

2—injury

1—prop damage

0—unknown

Out of all records, only circa 30% is having the predicted severity '1' records and the rest were severity 2. So have balanced the number of records for both the outcomes to prevent the bias towards one outcome as majority records can pull off many variations in prediction.

There is a lot of unusable data for our purpose like INCKEY, COLDETKEY, EXCEPTRSNCODE, EXCEPTRSNDESC etc..,. So have selected all the key factors present in the data for prediction which are

1. Weather condition.

2. Road condition.

3. Lighting condition.

4. Collision Address type

5. Speeding flag

After selecting the data, I have seen some missing values and have updated them as follows.

For weather, road condition, light condition & Address type data, have replaced the missing values with the unknown as already a certain number of records were having that values. A majority number of records are not having any data in the speeding flag, so assumed speeding was not there and replaced with no value as speeding not recorded.

After doing this sanitation, I have seen data in good form for normalizing and looked perfect as categorical values.

## 2.3 Feature Selection:

After cleaning the data & balancing the variety in outcomes number, we have 116376 records. So now the feature selection for this data goes like this.

**Data Features selection:**

| Kept features | Dropped features | Reason for dropping features |
|---|---|---|
| ADDRTYPE | OBJECTID, INCKEY, COLDETKEY, LOCATION | Location is not creating any difference here as no location has been the primary reason for accident. |
| NA | EXCEPTRSNCODE, EXCEPTRSNDESC, COLLISIONTYPE | All these fields will be outcomes from the onsite investigation so should not use as prediction. |

| | | |
|---|---|---|
| NA | PERSONCOUNT<br>PEDCOUNT<br>PEDCYLCOUNT<br>VEHCOUNT<br>INCDATE<br>INCDTTM | All these fields are just incident reporting number for monitoring the people involved in the accident. |
| NA | JUNCTIONTYPE<br>SDOT_COLCODE<br>SDOT_COLDESC<br>INATTENTIONIND<br>UNDERINFL | All these fields will be outcomes from the onsite investigation so should not use as prediction. |
| WEATHER<br>ROADCOND<br>LIGHTCOND | SEVERITYCODE,<br>SEVERITYDESC | As severity code is labelled data, it should be dropped |
| NA | PEDROWNOTGRNT<br>SDOTCOLNUM | Not a useful data to indicate the Severity code |
| SPEEDING | ST_COLCODE<br>ST_COLDESC<br>SEGLANEKEY<br>CROSSWALKKEY<br>HITPARKEDCAR | All these fields will be outcomes from the onsite investigation so should not use as prediction. |

So, we have selected these 5 Features to predict the label 'SEVERITYCODE' of the accident and to train the model using these 5 dependent variables.

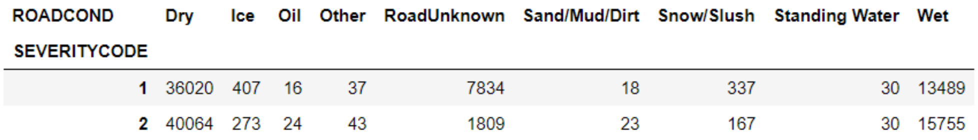
# 3. Exploratory Data Analysis:

## 3.1 Calculation of Target Label:

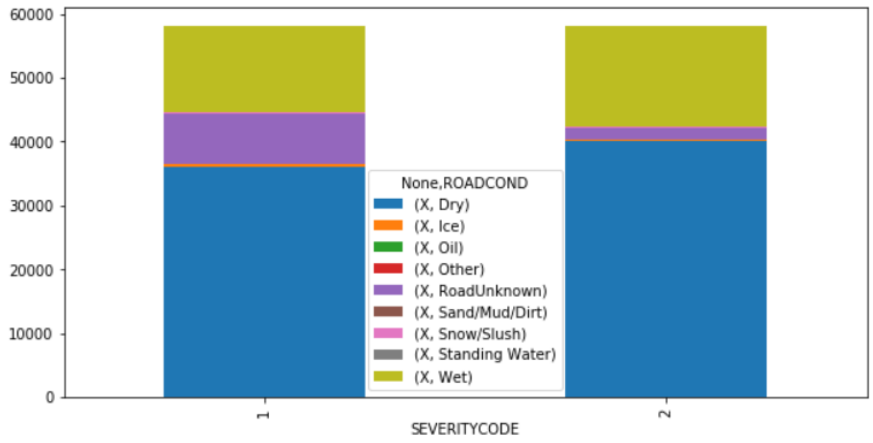
The outcome from the prediction model should be any one of the {'0', '1', '2', '2b', '3'}. As our data is having only 2 outcomes in the data records so prediction from our model will be '1' or '2'.

## 3.2 Correlation between the Severity Code and the Road condition:

From the data, we can see that the most accidents happened when road is dry, 60% of the events occurred in dry road from all the readings we have in this data. And Severity 2 type accidents are more than severity 1 on dry roads.

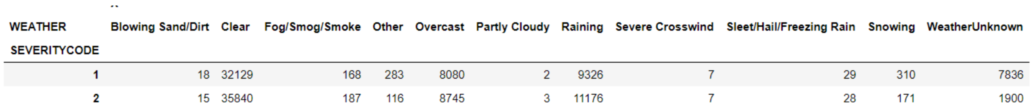


| ROADCOND<br>SEVERITYCODE | Dry | Ice | Oil | Other | RoadUnknown | Sand/Mud/Dirt | Snow/Slush | Standing Water | Wet |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 36020 | 407 | 16 | 37 | 7834 | 18 | 337 | 30 | 13489 |
| 2 | 40064 | 273 | 24 | 43 | 1809 | 23 | 167 | 30 | 15755 |

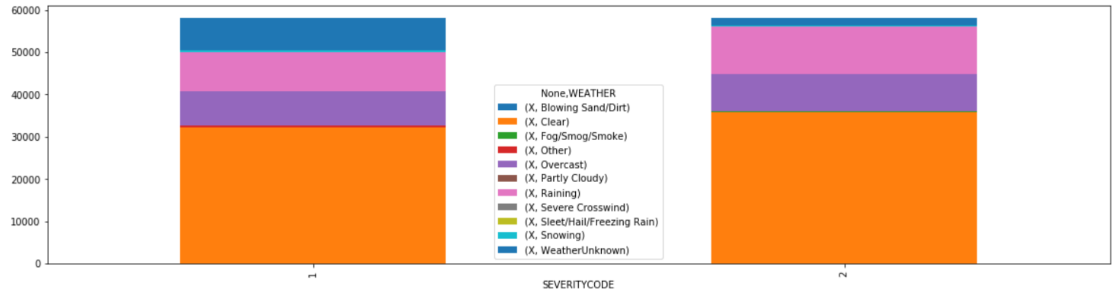


## 3.3 Correlation between the Severity Code and the Weather condition:

From the data, we can see that the most accidents happened when weather is very clear, approx... 60% of the events occurred when sky is clear from all the readings we have in this data. And Severity 2 type accidents are more than severity 1 on clear sky days.

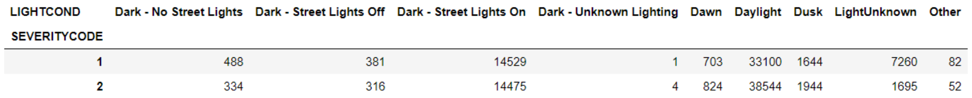


| WEATHER<br>SEVERITYCODE | Blowing Sand/Dirt | Clear | Fog/Smog/Smoke | Other | Overcast | Partly Cloudy | Raining | Severe Crosswind | Sleet/Hail/Freezing Rain | Snowing | WeatherUnknown |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 18 | 32129 | 168 | 283 | 8080 | 2 | 9326 | 7 | 29 | 310 | 7836 |
| 2 | 15 | 35840 | 187 | 116 | 8745 | 3 | 11176 | 7 | 28 | 171 | 1900 |

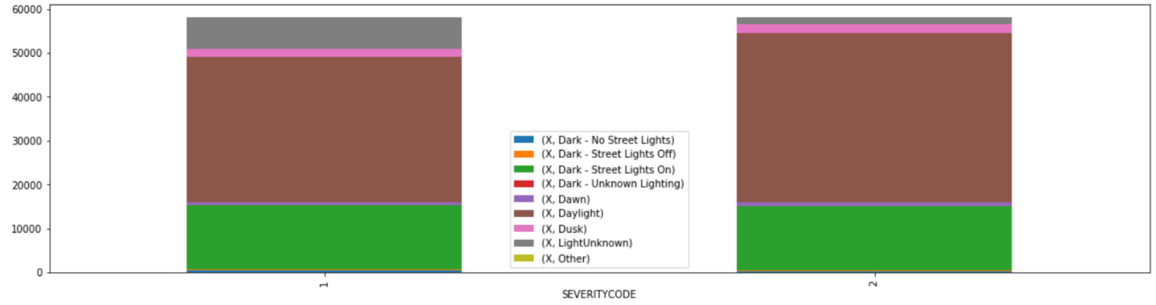


## 3.4 Correlation between the Severity Code and the Light condition:

From the data, we can see that the most accidents happened during day time, approx... 50% of the events occurred when light is clear all the readings we have in this data. And Severity 2 type accidents are more than severity 1 during the day times.

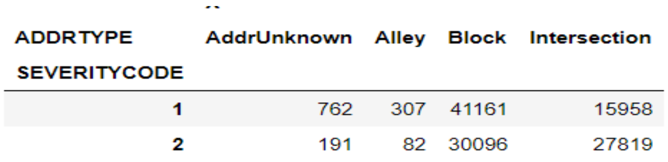


| LIGHTCOND<br>SEVERITYCODE | Dark - No Street Lights | Dark - Street Lights Off | Dark - Street Lights On | Dark - Unknown Lighting | Dawn | Daylight | Dusk | LightUnknown | Other |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 488 | 381 | 14529 | 1 | 703 | 33100 | 1644 | 7260 | 82 |
| 2 | 334 | 316 | 14475 | 4 | 824 | 38544 | 1944 | 1695 | 52 |

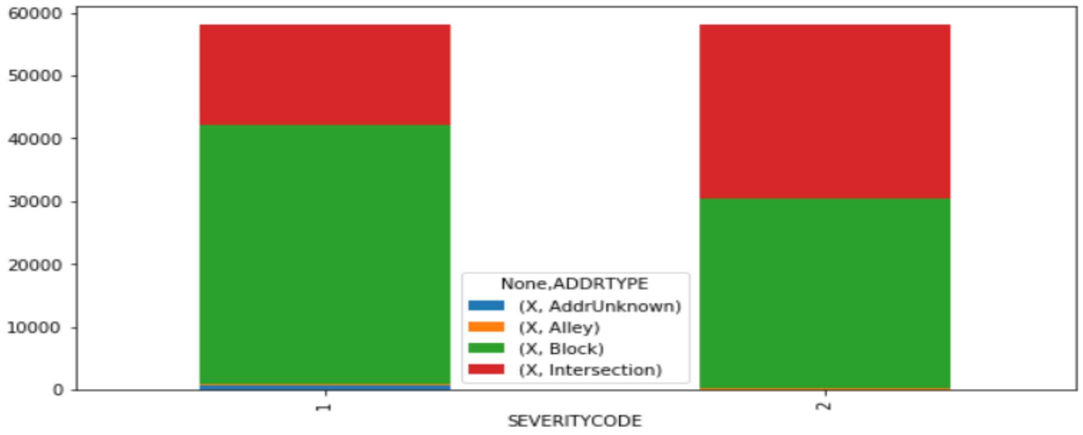


## 3.5 Correlation between the Severity Code and the Collision address Type:

From the data, we can see that the most accidents happened across blocks & intersections, approx... 90% of the events occurred. And Severity 1 type accidents are more than severity 2 across blocks.

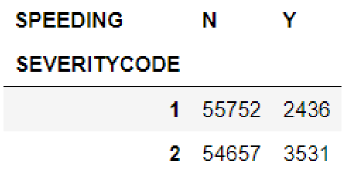


| ADDRTYPE<br>SEVERITYCODE | AddrUnknown | Alley | Block | Intersection |
|---|---|---|---|---|
| 1 | 762 | 307 | 41161 | 15958 |
| 2 | 191 | 82 | 30096 | 27819 |

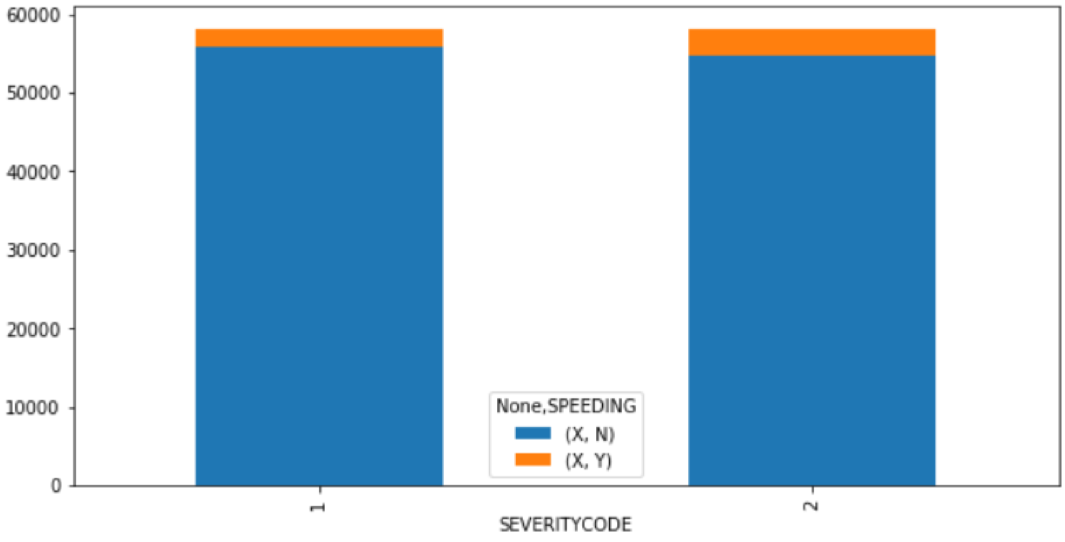


### 3.6 Correlation between the Severity Code and the Speeding flag:

From the data below you can see that only 10% speeding flags were recorded the feature set we have selected to train model. Besides that severity 2 type accidents are more than the severity 1 when unpermitted speeding causes the accident.



| SPEEDING | N | Y |
|---|---|---|
| SEVERITYCODE | | |
| 1 | 55752 | 2436 |
| 2 | 54657 | 3531 |



## 4. Methodology & Predictive Modelling:

### 4.1 Selecting Model:

As the outcome of the model will be a classified class, so we will use the classification models here for building the model. We can use any of these 3 models to build the data.

i. K Nearest Neighbor(KNN)
ii. Decision Tree
iii. Support Vector mission (SVM)

### 4.2 K Nearest Neighbor (KNN)

We should find the best k to build the model with the best accuracy.

“In classification using k-NN, we need to compute the distances between cases based upon their values in the feature (variable) set. The nearest neighbor to a given case has the smallest distances from that case. The distance between two cases can be the Euclidean distance or the city block distance.

The k-NN can be used for categorical or continuous outcome (dependent variable). For categorical dependent variable, k-nearest neighbors (k-NN) can be used to classify cases by classifying to the group which has the most neighbors. The algorithm can be summarized as follows:

1. Specify a positive integer k. (k≥1).
2. Calculate the distance between pairs of cases.
3. If k=1 is chosen, classify the case into the group of its nearest neighbor.
4. If (k≥1), classify the case into the group of the majority k-closest neighbors

In k-NN algorithm, continuous features are optionally coded using adjusted normalization.”

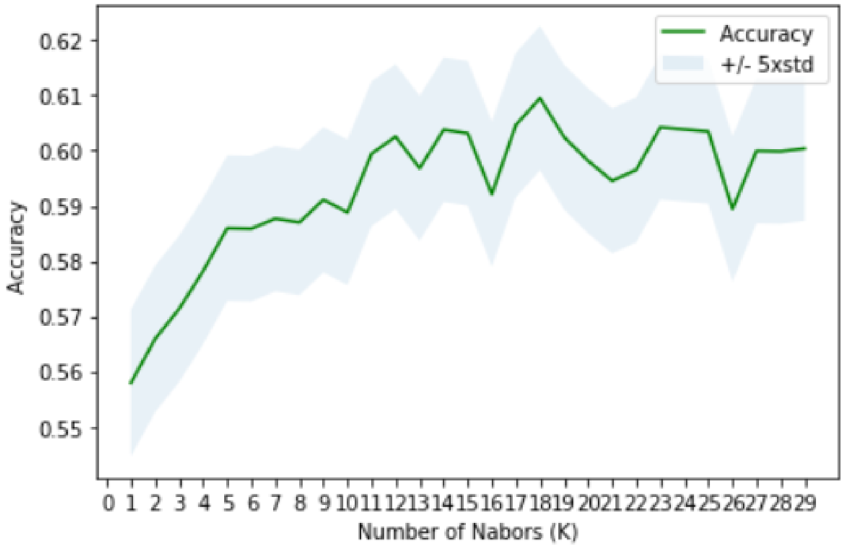
In this application of KNN we choose k=18 and achieved the following scores:

Jaccard_similarity_score for KNN 0.5309197147194454

F1_score for KNN 0.5297590757545433

Recall score for KNN 0.5309197147194454

Precision score for KNN 0.5312112820034307



### 4.3 Decision trees:

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.

In this application of a decision tree, we achieved the following scores:

Jaccard_similarity_score for Decision Tree 0.608913585197491

F1_score for decision Tree 0.6053587959973631

Recall score for decision tree 0.608913585197491

Precision score for decision tree 0.6129466428339209

### 4.4 Support Vector Machine:

Support Vector Machine (SVM) has been reported to be a flexible classifier by data mining community and delivers quite promising results in handling imbalanced dataset. SVMs belong to the general category of kernel methods which are algorithms that depends on the data only through the dot-products. These dot-products can be transformed by a kernel function which computes a dot-product in high dimensional feature space.

As a classifier, the SVM has the ability to generate nonlinear decision boundaries using methods designed for linear classifiers. In addition, the use of kernel functions allows the data scientist, and SVM users to apply a classifier to any datasets which have no obvious representation in terms of patterns. The strengths of SVM have made this classifier popular in scientific and medical studies. SVM has been used in recognition of human movement patterns and experiments on different kernels on SVM have shown that it was successful when applied to imbalanced data. Classification of SVM is achieved by realizing a linear or non-linear separation surface in the input space.

In this application of a SVM, we achieved the following scores:

Jaccard_similarity_score for SVM 0.5309197147194454

F1_score for SVM 0.5297590757545433

Recall score for SVM 0.5309197147194454

Precision score for SVM 0.5312112820034307

# 5. Results

Model evaluation will be carried out by scoring the model using following indexes/scores.

The two scores we will be looking at are the Jaccard Score and the F1 Score, which is based on recall and precision scores:

1. ***Jaccard Score:*** Also called Jaccard index, Intersection over Union and the Jaccard similarity coefficient (originally given the French name coefficient de community by Paul Jaccard), is a statistic used for gauging the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets5

2. ***F1 Score:*** measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive.

3. ***Recall Score:*** Recall is the ratio of correctly predicted positive observations to all observations in a class. The question recall answers for instance is: Of all the accidents labeled as Injury that truly involved an Injury, how many did we label correctly

4. ***Precision Score***: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer for instance is of all accidents labeled as Injury, how many involved Injury? High precision relates to the low false positive rate.

| Metrics | KNN | Decision Tree | SVM |
|---|---|---|---|
| Jaccard Index | 0.530 | 0.608 | 0.530 |
| F1 Score | 0.529 | 0.605 | 0.529 |
| Precision | 0.531 | 0.608 | 0.530 |
| Recall | 0.530 | 0.612 | 0.531 |

After evaluating the scores for all the models we trained, Decision tree looks better than remaining with accuracy rate around 60%, model accuracy can be increased with help of more data as this data is not having all the possible outcomes.

## 6. Conclusions/Discussions:

In this case study, I have analyzed the Seattle accidents data to predict the severity of the accident using the 5 features weather condition, road condition, lighting condition, collision address type & speeding index.

The results highlighted the unexpected insight as most accidents were happening when all the things were perfect like most accidents occurred during day times if we thought that lighting may be reason.

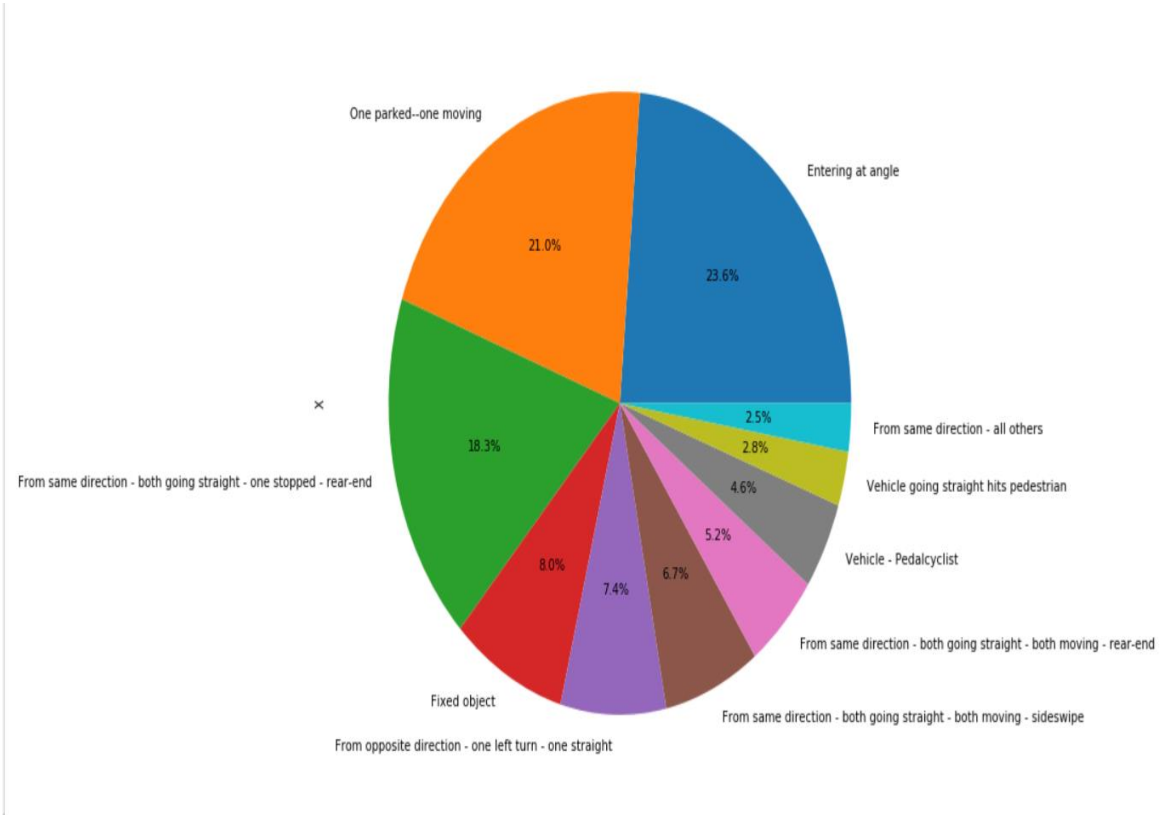
Most accidents occurred during the clear sky if we think rains, fog or snow might cause the accident.

Likewise most accidents happened on the dry roads if we think road's behavior may be the reason.

And in context of speeding, only 10% of the selected data are having speeding Index is Y, so speeding is not the complete reason. Many accidents happened across the intersections & blocks.

But the one important finding I have noticed from the data which is not a feature variable to predict the severity, the field 'ST_COLDESC' which is post investigation outcome by SDOT to categorize the collision.

The following pie chart depicts the share of the most collision types in the data, and the shocking finding is that collisions happened mostly with parked vehicle, with vehicle going in same direction and vehicles entering in angels.



The conclusion from this finding is to manage the vehicles going in same direction to follow proper distancing between the vehicles going in the same direction and following the proper lines to avoid slipping into parking lanes.

Riders should make use of side signals mostly to avoid the accidents between the vehicles going in same direction or with parked cars.

## 7. Future Directions:

1. Most of the data is not fitted into training data due to lack of all possible scenarios, it would be good to include data with all possible outcomes instead of only Severity code 1 & 2.
2. We can avoid training this data into SVM model as it is taking lot of time and data is not very highly dimensional too.

## 8. References:

1. Data - https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
2. https://www.colburnlaw.com/seattle-traffic-accidents/
3. https://www.seattletimes.com/seattle-news/transportation/seattle-traffic-deaths-and-injuries-down-slightly-last-year-most-of-the-fatalities-were-pedestrians/