# **CAN WE PREDICT THE SEVERITY OF THE COLLISION?**

# **INTRODUCTION**

# **BACKGROUND**: In the modern world road traffic accidents are very common in each and every part of the world. Seattle is one the busiest cities and as of 2017, there have been more than 11,000 motor vehicle-involved **collisions** per year. These collisions not only lead to high property damage but may result in injuries and in worst case scenarios even death. In 2017, a total **of** 187 fatal and serious **injury collisions** were reported on **Seattle** streets. The causes behind the collisions may range from physical factors like road conditions, lighting, weather, time of day etc to human factors like inattention, speeding, under influence. Various efforts and steps can be taken in order to minimize these collisions. Every city has devised certain traffic rules and regulations to help in this regard. Even, WHO has given certain recommendation to be followed by the government agencies to have an impact on the rate collisions and more importantly to reduce the cost of property damage as well as of life. Even small regulations sometimes have much greater impact.

# **PROBLEM:** To predict the severity of collision based on the various physical and human factors. And make recommendations and take necessary actions based on those predictions like improving the road conditions, lighting, regulations on speed in various area etc.

# **STAKEHOLDERS:**

# **General Public**- vehicle drivers and pedestrians to take necessary precautions in certain circumstances.

# **Seattle Traffic department**- to prepare plans regarding improving road conditions, lightning, speeding limits, etc

# **Car manufacturing industry**- To increase car re-enforcements at the different sites of cars which are frequently involved and especially responsible for the most serious injuries and fatalities.

# **Weather department**- to make necessary updates and caution messages for different locations depending upon weather.

# **Health and paramedical departments**- to take necessary steps to provide faster emergencies services in the areas more prone to accidents to reduce the loss of life.

# **DATA.**

# Data utilized for this analysis was downloaded from Kaggle Seattle collision dataset as csv file.

# <http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv>

# **DATA UNDERSTANDING AND CLEANING**

# The dataset includes data regarding the severity of collisions and various parameters associated with it in the city of Seattle from year 2004 till now.

**PARAMETERS NOT USEFUL.** After exploring the data, we found it to have 40 parameters, among those some were found to be least relevant.

|  |
| --- |
| 'OBJECTID', |
| 'INCKEY', |
| 'COLDETKEY', |
| 'REPORTNO', |
| 'STATUS', |
| 'EXCEPTRSNCODE' |
| 'EXCEPTRSNDESC', |
| 'SDOT\_COLCODE', |
| 'PEDROWNOTGRNT |
| 'ST\_COLCODE', |
| 'SDOTCOLNUM', |
| 'SEGLANEKEY', |
| 'CROSSWALKKEY', |
| 'HITPARKEDCAR’ |

# Parameters like 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'EXCEPTRSNCODE','SDOT\_COLCODE','ST\_COLCODE','SDOTCOLNUM','SEGLANEKEY', 'CROSSWALKKEY' mostly consisted of some unique number or code assigned to each case which did not add any more useful information to the analysis.

# Parameters “INCDTTM” and ‘INCDATE” were utilised to form columns like “Year”, “Month”, “Weekday” “Hour” and “TIMEOFDAY”. TIMEOFDAY column was created to see relation ship between the severity of accidents and the time of the day like early morning, morning, noon, evening, night and late night. Columns “INCDTTM” AND “INCDATE” were also eventually dropped. 'UNDERINFL’ column had values of “N”,” Y”,”1”,”0”. So “1” and ”0” were converted to “Y” and “N” respectively. ‘INATTENTION’ and ‘SPEEDING’ columns were found to have only “Y” values. So, it was safe to consider that all NAN values would be “N”.

# Since SEVERITY DESC AND SEVERITYCODE have the same information. SEVERITYCODE was dropped. Also, the unknown values of SEVERITYDESC were replaced by NAN and dropped. SEVERITYDESC values were also put into two categories

# PROPERTY DAMAGE ONLY=1

# INJURIES, SERIOUS INJURIES AND FATALITIES=2

# **DEALING WITH MISSING DATA**

# We found that in our data there were lot of parameters with very high number of missing values. Dropping all those values would result in loss of lot of information but including all of them would result in increased noise and give biased analysis. So, we decided to make two different datasets.

# **1****.** **DATA INCLUDING UNKNOWN, OTHERS, NOT MENTIONED VALUES**

# We tried to include as much data as possible for the analysis.

# **2****. DATA WHERE ALL NAN, UNKNOWN, OTHER AND NOT MENTIONED VALUES WERE REMOVED.**

# **DATA INCLUDING UNKNOWN, OTHERS, NOT MENTIONED VALUES**

# For our analysis we first include most of the data even with NAN values. All the NAN values for parameters '' ADDRTYPE, 'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'SDOT\_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'SPEEDING','ST\_COLDESC were converted to either “Unknown” or “not mentioned” or “not stated” value.

# As we had missing values in the parameters X, Y and location. We decided to drop X, Y, LOCATION. We had the following parameters:

|  |  |
| --- | --- |
| **PARAMETERS** | **Description.** |
| COLLISIONTYPE | Type of collision rear, side swipe etc |
| ADDRTYPE | Whether collision occurred at block or intersection |
| SEVERITYCODE | Code to describe severity of collision |
| SEVERITYDESC | Description of severity, property damage, injuries, serious injuries, fatalities etc. |
| JUNCTIONTYPE | Type of junction, intersection, ramp, etc |
| SDOT\_COLDESC | Type of collision with a car, pedestrian, sidehit etc. |
| INATTENTIONIND | Whether the collision occurred due to inattention |
| UNDERINFL | Whether the collision occurred to the person being under influence |
| WEATHER | Condition of weather at the time of accident |
| ROADCOND | Condition of road at the time of accident |
| LIGHTCOND | Status of lighting at the time of accident |
| SPEEDING | Whether the collision occurred when the vehicle was speeding. |
| Weekday | Weekday the collision occurred |
| Hour | The hour the collision occurred |
| Year | The year the collision occurred |
| Month | The month in which the collision occurred |
| TIMEOFDAY | Time of the day at the time of accident , morning, noon, evening, night, late night etc |
| PERSONCOUNT | Number of persons involved in accident |
| PEDCOUNT | Number of pedestrians involved in accident |
| PEDCYCLIST COUNT | Number of cyclists involved in accident |
| VEHICLE COUNT | Number of vehicles involved in accident |

# **DATA WHERE ALL NAN, ‘UNKNOWN’, ‘OTHER’ AND ‘NOT MENTIONED’ VALUE WERE REMOVED IN THIS WE INCLUDED PARAMETERS X and Y.**

# In this dataset we included above parameters and also the X and Y parameters. We dropped all the NAN, ‘unknown’ and ‘others’ values.

# **PARAMETER SELECTION.**

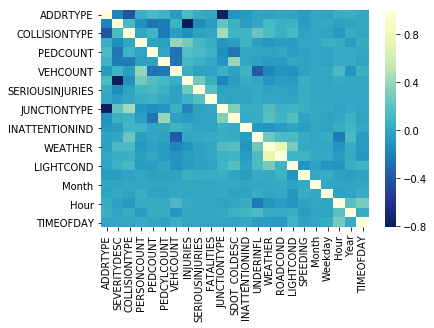
|  |  |  |  |
| --- | --- | --- | --- |
|  | **DATA INCLUDING ‘UNKNOWN’,‘OTHERS’, ‘NOT MENTIONED’ VALUES** | | |
|  |  | | REASON |
| **PARAMETERS** **KEPT** | COLLISIONTYPE ADDRTYPE SEVERITYDESC JUNCTIONTYPE SDOT\_COLDESC INATTENTIONIND UNDERINFL WEATHER ROADCOND LIGHTCOND SPEEDING Weekday Hour Year Month TIMEOFDAY PERSONCOUNT PEDCOUNT PEDCYCLIST COUNT VEHICLE COUNT | | All these parameters were considered to may have an impact on the type of collision |
| **PARAMETERS****DROPPED** | 'OBJECTID','INCKEY','COLDETKEY','REPORTNO','STATUS','EXCEPTRSNCODE''EXCEPTRSNDESC','SDOT\_COLCODE','PEDROWNOTGRNT'ST\_COLCODE','SEGLANEKEY','CROSSWALKKEY','HITPARKEDCAR’‘SEVERITYCODE’LOCATIONXYINCDTTM‘INCDATE | | All these parameters contain codes or indexes which don’t give any more significant information.Duplicate information with SDOT\_COLCOSEVERITYDESC gives the same Information Had many NAN values deleting which would have resulted in loss of information for other parameters as well and including them as a separate category’ UNKNOWN would have proved to be skewed the dataNew parameters like day, month, year and time of day were created to successfully get the useful information from these parameters. |
| **DATA WHERE ALL NAN, ‘UNKNOWN’, ‘OTHER’ AND ‘NOT MENTIONED’ VALUE WERE REMOVED** | | | |
|  | | REASON | |
| COLLISIONTYPE ADDRTYPESEVERITYDESC JUNCTIONTYPESDOT\_COLDESC INATTENTIONINDUNDERINFL WEATHERROADCOND LIGHTCONDSPEEDING WeekdayHour YearMonth TIMEOFDAYPERSONCOUNT PEDCOUNTPEDCYCLIST COUNT VEHICLE COUNTX Y | | All these parameters were considered to may have an impact on the type of collision.We included parameters X and Y also.All the missing values, others, not mentioned values were dropped in order to reduce the noise. Along with all the parameters | |

# **METHODOLOGY:**

# We utilized matplotlib and seaborn to see the distribution of various parameters across the data. We utilized bar graph mostly to visualize data. We determined the frequencies of various values of parameters by value\_counts() and plotted them as seaborn bargraphs.

# 

We used corr() and heatmap to find the correlation between various parameters.

****

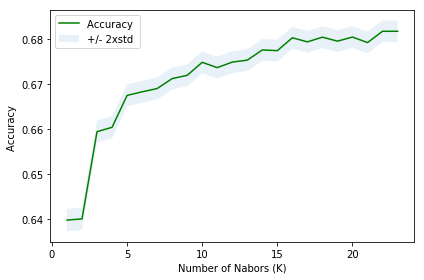
We are found most of our data to be categorical .As for machine learning we require the numeric data. We converted the ‘string’ values to numeric value by using labelencorder.We also found our data to be unbalanced as the property damage only was exceedingly more than injuries. In order to balance the data, we utilized the undersampling of the majority value method. After transforming and balancing the data, we used StandardScaler.preprocessing.fit to standandardize and transform the data. We then split the data into train and test in the ratio of 0.70:0.30.

# **PREDICTIVE MODELS:**

# As the dependant variable in this situation is SEVERITY DESCRIPTION which is a categorical variable, we used classification prediction model.

# **DATA INCLUDING UNKNOWN, OTHERS, NOT MENTIONED VALUES** We utilized various models KNN, SVM, Decision Tree, Logistic Regression. Since our data was unbalanced with more property damage than injuries, we balanced data by under-sampling the majority value. We also utilized Random Forest Classifier on the imbalanced data. We utilized accuracy metrics Jaccard similarity score, f1 score and log loss for validation. We also utilized kfold cross validation.

# KNN: K NEAREST NEIGHBOUR: We randomly selected a value for k as 24. We, then determined the best K value as 22.



# SVM: For SVM we used all rbf, linear, polynomial and sigmoid to find the best predictive model

# DECISION TREE: For Decision Tree we utilized the entropy criterion .

# Logistic Regression:

**Performance of models.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ALGORITHM | JACCARD | F1 SCORE | LOG LOSS |
| 0 | KNN | 0.681780 | 0.681075 | NA |
| 1 | SVM-rbf | 0.696092 | 0.695977 | NA |
| 2 | SVM-linear | 0.666560 | 0.665305 | NA |
| 3 | SVM-Polynomial | 0.683813 | 0.683611 | NA |
| 4 | SVM-sigmoid | 0.562582 | 0.562575 | NA |
| 5 | Logistic Regression | 0.661718 | 0.660396 | 0.593739 |
| 6 | Decision Tree | 0.701843 | 0.699923 | NA |
| **7** | RANDOM FOREST | 0.717438 | 0.699599 | NA |

**RANDOM TREE CLASSIFIER ON THE IMBALANCED DATA**. We utilized random tree classifier on the imbalanced data to make predictive model. The accuracy score for Random Forest Classifier is 0.7561849229487714.

**DATA WHERE ALL NAN, UNKNOWN, OTHER AND NOT MENTIONED VALUES WERE REMOVED**

We utilized data where all the unknown, other, not mentioned values were removed to predict the outcome. In this we found the SEVERITY DESC values were in the ratio1:0.56 and more or less to be balanced. We again utilized KNN, SVM, Decision Tree, Logistic Regression and Random Forest classifier models and the above validation parameters.

LO.

# **Performance of model.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ALGORITHM | JACCARD | F1-SCORE | LOG LOSS |
| 0 | KNN | 0.722473 | 0.696543 | NA |
| 1 | SVM-rbf | 0.730526 | 0.690109 | NA |
| 2 | SVM-linear | 0.721799 | 0.661457 | NA |
| 3 | SVM-Polynomial | 0.728412 | 0.681863 | NA |
| 4 | SVM-sigmoid | 0.616866 | 0.617272 | NA |
| 5 | Logistic Regression | 0.725398 | 0.684980 | 0.550345 |
| 6 | Decision Tree | 0.723126 | 0.697318 | NA |
| - |  |  |  |  |

# We compared the models and the data and the model showing best accuracy metric was chosen

**RESULTS:**

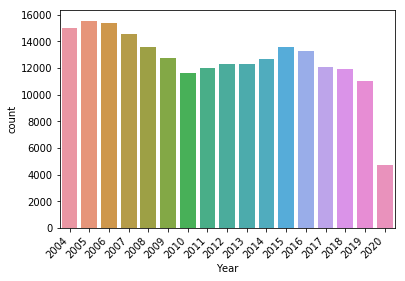
**LOCATION: TOP 15 LOCATIONS WHERE MOST OF THE COLLISIONS TOOK PLACE**

# 

For this information, we utilized as much data as possible to find out the most common locations of collision. We found that BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB AND AURORA AVE N with 298, N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND CORLISS AVE N with 297,BATTERY ST TUNNEL SB BETWEEN AURORA AVE N AND ALASKAN WY VI SB 291 collision.

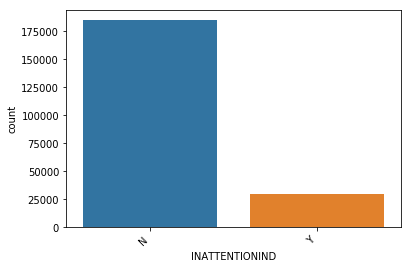
Since Ist and 3rd location seems to be more or less either same or in the similar area, it will the location which is more prone to collision.

**YEARWISE :** Collisions initially decreased from year 2004 to 2010 and then again increased upto 2015. Since then it has shown a gradual decline till now in 2020(which has shown the least number of collisions).

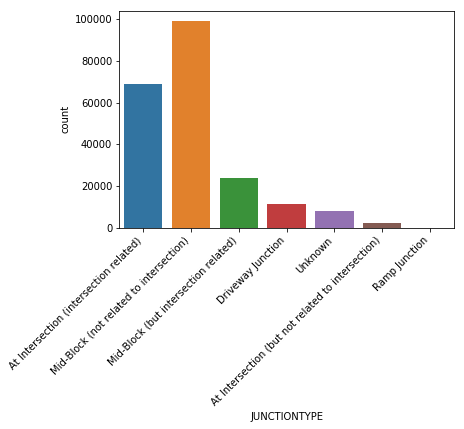


**ADDRESSTYPE:**

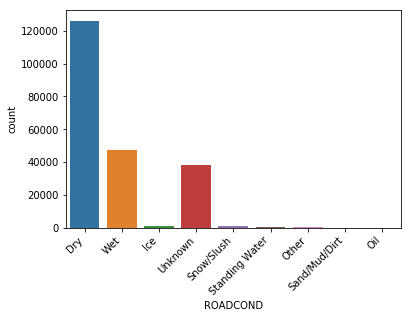
**INATTENTION**



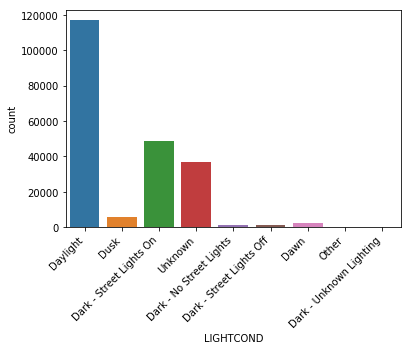
**JUNCTIONTYPE**



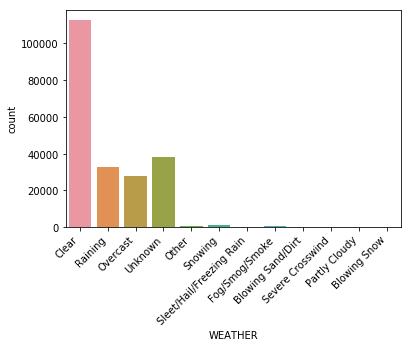
**ROADCONDITION**



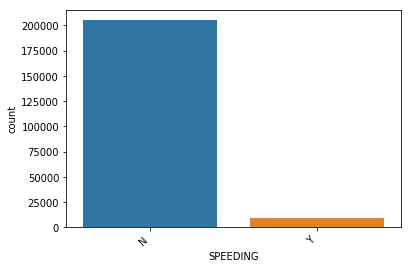
**LIGHT CONDITION**



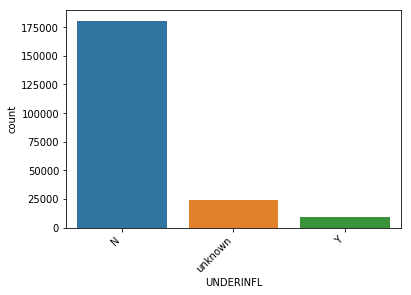
**WEATHER:**



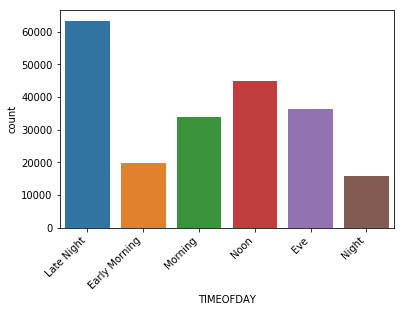
**SPEEDING**



**UNDER-INFLUENCE**



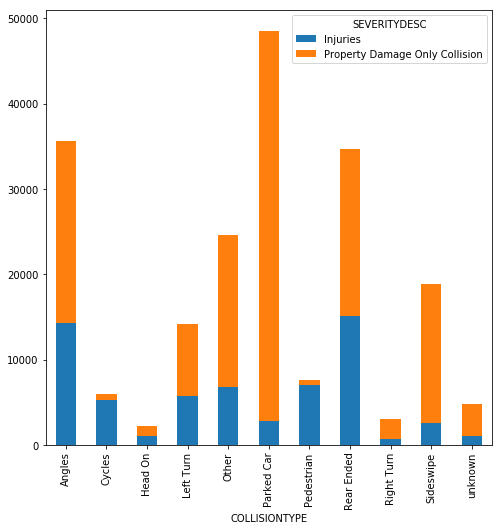
**TIME OF THE DAY**:



**RELATIONSHIP BETWEEN VARIOUS PARAMETERS AND THE SEVERITY DESCRIPTION**

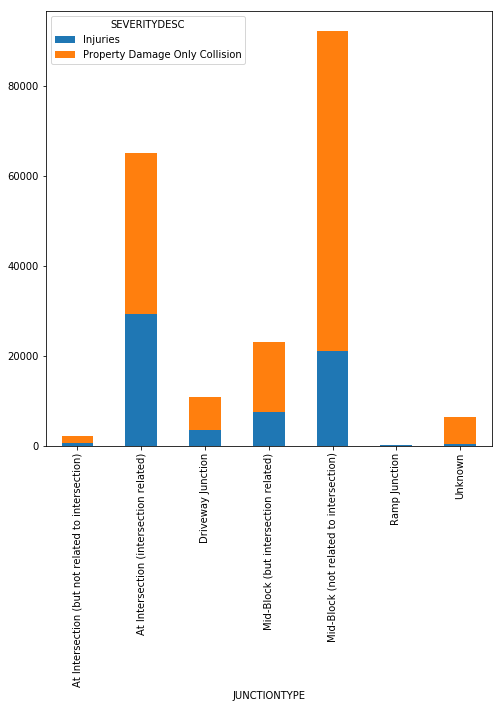
We utilized mostly crosstab and matplotlib stacked bar graphs to analyse the relationship between various variables and the severity description.

**RELATIONSHIP BETWEEN COLLISIONTYPE AND SEVERITY OF COLLISION**



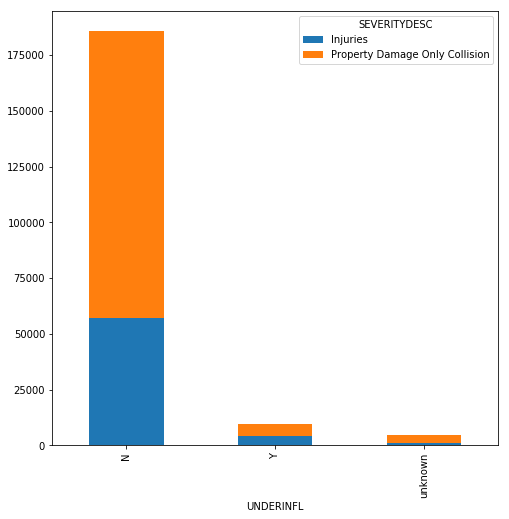
We found that injuries were more common in the pedestrian and cycles group which was quite understandable as the property was either not involved or was least affected by the collision. Among the type of collisions, collisions resulting in injuries were Head on followed Rear Ended and Angles .

**RELATIONSHIP BETWEEN JUNCTIONTYPE AND SEVERITY OF COLLISION**



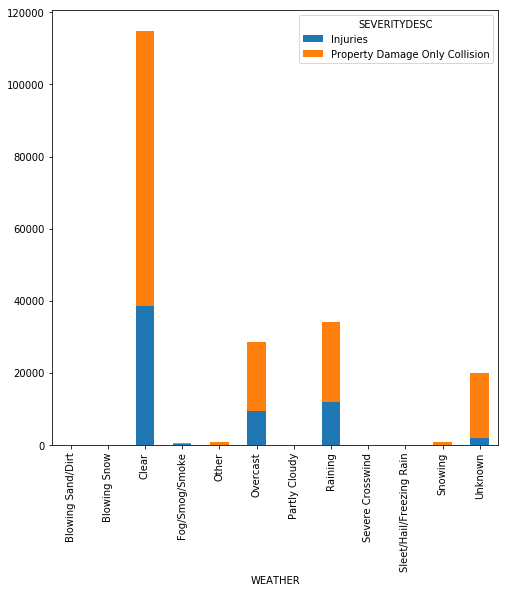
We found that the intersection related collisions were more related to injuries than non-intersection. Also, most of the mid-block collisions led to property damage only.

**RELATIONSHIP BETWEEN UNDERINFLUENCE AND SEVERITY OF COLLISION**



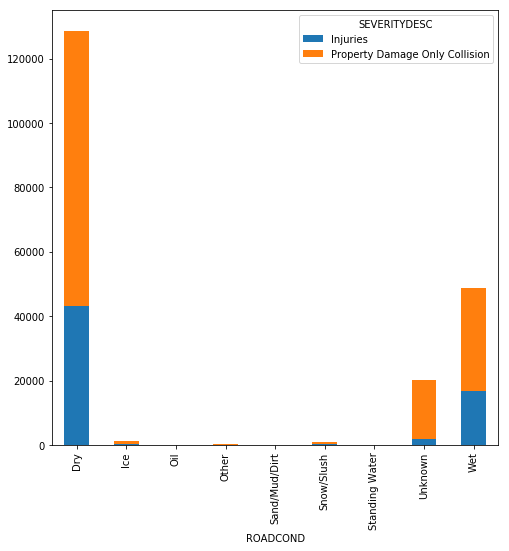
We found although most of the collisions were resulted by people who were not undue influence but almost 50% of the collisions resulted due to people under influence led to the injuries.

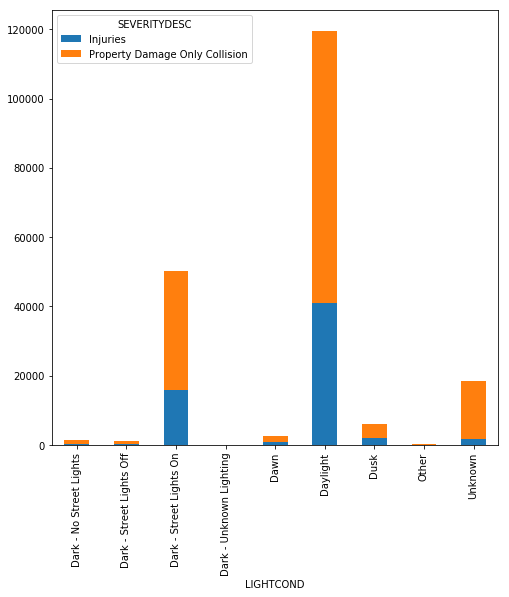
**RELATIONSHIP BETWEEN WEATHER AND SEVERITY OF COLLISION**



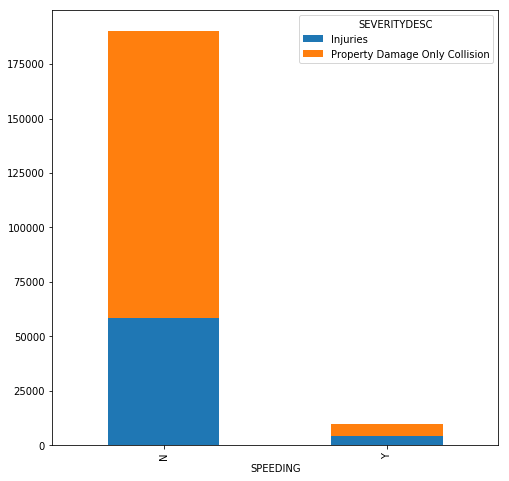
Again, with weather also most of the collisions which resulting in property damage or injuries happened during clear weather.

**RELATIONSHIP BETWEEN ROADCONDITION AND LIGHT CONDITIONAND SEVERITY OF COLLISION**





**RELATIONSHIP BETWEEN SPEEDING AND SEVERITY OF COLLISION**



Speeding again showed that almost in 50% of the cases resulted in the injuries.

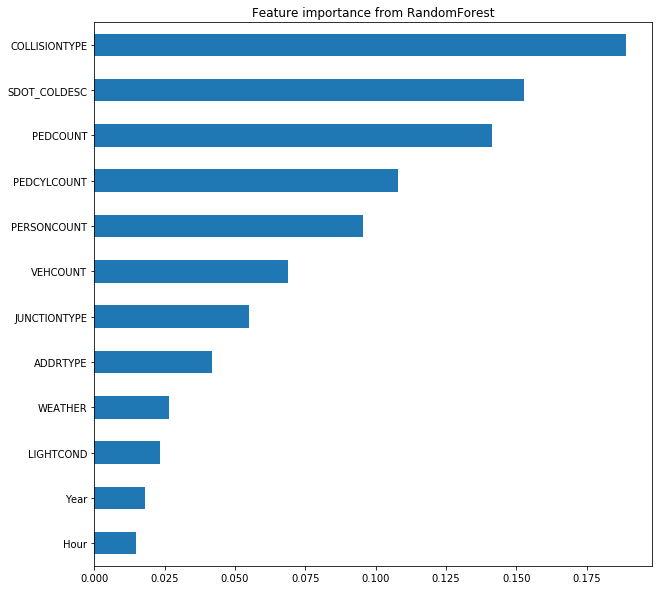
MODEL DEVELOPMENT: We first developed the machine learning classification models with as much data as possible to include. We found the model showing the best accuracy metric was Decission Tree Model with jaccard score of 0.70 in this specific data(after balancing the data).

We also used the Random Tree Classifier on the imbalanced dataset of the same data and found the accuracy score of 0.73.

In our second dataset, in which all of the missing, unknown, not mentioned values were dropped we found that SVM model with rbf showed highest jaccard score of 0.73

We concluded that for the best model for data including all Nan, UNKNOWN etc values and imbalanced Random Forest classifier had the best predictive outcome while for the data where all the NAN values were excluded the SVM (rbf) had . Both had the same accuracy score of 0.73.

Random Forest classifier was also used to find features most contributing to the information gain.

****

Feature importance from RandomForest Classifier Showed the collision type, SDOT COLDesc followed by pedcount, pedcylcount,personcount and vehcount were the most important features affecting the outcome.

**DISCUSSION:**

Based on our analysis we were able to find the top 15 locations having the most number of collisions(>150) . Some of these locations are close to the same area.

Although speeding as a cause was found to be in very small number of collisions but it was found to cause injuries in almost 50 % of cases who were speeding and was with the people under influence.

We also found that among the type of collisions, those resulting more in injuries were Head on followed Rear Ended and Angles.

Again the collisions at the intersections (mostly related to intersections ) also had relatively more injuries as compared to other junction type like block where most of the total collisions took place.

Based on our models we also found the various parameters which had most impact on the outcome. These mainly included the type of collision and the number of people or pedestrian or pedcyclist involved .

Also the among the weekdays most collisions were seen happening on Friday

Although weather, road conditions and light conditions showed most of the collisions resulted on clear weather, dry road and daylight, we couldn’t evaluate the effect of the specific weather like snowny or rainy or wet or icy roads or badlighting or lack of street lights as their numbers were very small. This is completely understandble as during the time 2004-2020 most vehicles move more during the day and mostly during daylight. Also %of days with weather rainy or snowy are less compared to dry days.

RECOMMENDATIONS:

1. The locations with maximum number of collisions may require more emergency services and hospitals in the area, so that the serious injuries and fatalities are kept to minimum.
2. As the number of collisions in the year 2020 shows lowest number so far, one because the year is the not over yet and more importantly can be related to corona virus pandemic which may have led to decrease in traffic. So we may indirectly infer that traffic may have been one of the contributing factors for collisions. Hence steps to decrease the traffic at any moment of time and better traffic control may also affect the rate of collisions.
3. Speeding and under influence was found as a reason only in small number of collisions but in both cases nearly 50% led to injuries. Thus more speeding limits especially at the locations with maximum collisions over the years and more strict regulations for driving underinfluence.
4. We also found the collisions resulting in injuries were found little more in collision at the intersection which may call for different measures at the intersection like speed limits and measures that may provide better visualisation of the other side at the intersection.
5. Head On collisions, angles and rear end collisions resulted in more injuries than any other type of collision. So more reinforcement from the automobile industry at those places in the vehicle may lead to less injuries in a collision.
6. By using our predictive model we may be able to predict the outcome based on the variable input upto 75% accuracy. We were not able to improve the accuracy by including all the data as possib

**CONCLUSION:**

Seattle collision data provides lot of information which we utilized into the prediction model but we strongly believe that there are many other parameters which were not included in data and may have more impact on the type of collision than the others . We may need to include those parameters in future data collection e.g condition of the vehicle, talking on the phone, pedestrian not crossing at zebra crossing etc. These parameters may actually improve accuracy of the model.