ACCIDENT SEVERITY PREDICTION

Amit Kumar Mukul

24th Aug 2020

1. Introduction

1.1. Background

Seattle is a seaport city on the Ist Coast of the United States. Average daily traffic in this town is around 1 Million [1]. Every year approx. 200 fatal and serious injuries are caused on the road of Seattle. Including in these figures are fatal accident which has been in double digit since long time [2]. The goal of Seattle Department of Transportation is to reach zero accident each year. Accident on these roads are caused by various factors like weather, road condition, influence of alcohol etc. Accident on the roads also leads to long traffic jams and hence kill important time of travellers.

Interest group for this project can be Seattle Department of Transportation and anyone who is planning to pass by this town using its highway. Problem statement of this project is to predict the accident severity so that target audience would drive more carefully or even change his/ her travel plan for given conditions.

1. https://www.seattle.gov/Documents/Departments/SDOT/About/DocumentLibrary/Reports/2018_Traffic_Report.pdf
2. https://www.seattletimes.com/seattle-news/transportation/seattle-traffic-deaths-and-injuries-down-slightly-last-year-most-of-the-fatalities-Ire-pedestrians/

1.2. Data Description

For this project I will take data “Collisions-All Years” from Seattle Department of Traffic (SDOT). This data set contains all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe for this dataset is from 2004. The dataset is downloaded from below link:
https://s3.us.cloud-object-storage.appdomain.cloud/cf-cmyses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Important attributes of dataset that will be used to make the Machine Learning model can be understood as below:

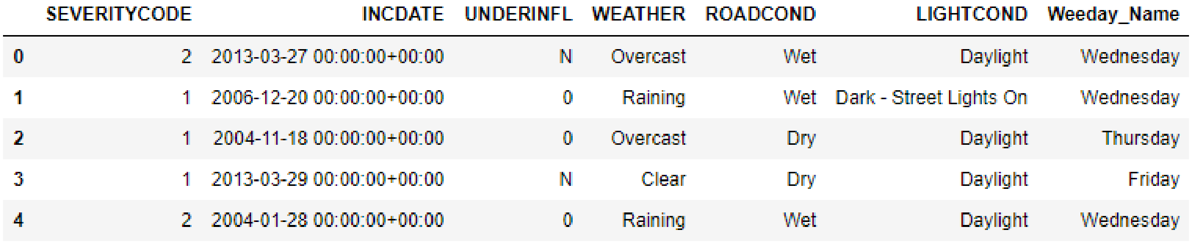
1. SEVERITYCODE: A code that corresponds to the severity of the collision: 1 is for Property Damage & 2 is for Injury.
2. UNDERINFL: Whether a driver involved was under the influence of drugs or alcohol. It is categorical variable and states Y & N for being under an influence or not respectively.
3. WEATHER: A description of the weather conditions during the time of the collision. It is also a categorical variable with different weather conditions like
4. ROADCOND: The condition of the road during the collision. It is also a categorical variable with different road conditions like Dry, Oily, Sand/ Mud/ Dirt, Snow, Wet, Standing Water.
5. LIGHTCOND: The light conditions during the collision. It is also a categorical variable with different light conditions like Dark-No Street Light, Dar-Street Light Off, Dark-Street Light On, Dawn, Daylight, Dusk etc.

There are other attributes as well which majorly defines post-accident data points like place where accident took place, data & time, number of persons got injured etc. which were not relevant for making the model.

2. Methodology

2.1. Data Cleaning

The dataset consisted of many attributes which Ire not relevant to making this predictive modelling, so such columns from the dataset Ire removed. The attributes list finally drilled down 7 columns as shown below including the Iekday name column added to check is any Iekday has useful impact on the severity of accident:



	SEVERITYCODE	INCDATE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	Weeday_Name
0	2	2013-03-27 00:00:00+00:00	N	Overcast	Wet	Daylight	Wednesday
1	1	2006-12-20 00:00:00+00:00	0	Raining	Wet	Dark - Street Lights On	Wednesday
2	1	2004-11-18 00:00:00+00:00	0	Overcast	Dry	Daylight	Thursday
3	1	2013-03-29 00:00:00+00:00	N	Clear	Dry	Daylight	Friday
4	2	2004-01-28 00:00:00+00:00	0	Raining	Wet	Daylight	Wednesday

Figure 1: First five rows of dataset with important attributes

After this step, row with respect to empty and unknown parameters against each attribute was dropped. Finally, my dataset became 170,510 rows and 7 columns.

2.2. Exploratory Analysis

For exploratory analysis, I plotted grouped bar chart to determine the impact of each attributes on SEVERITYCODE. In each plot Green Bar is with respect to code 1 which indicated Property Damage and Blue Bar is with respect to code 2 which indicates Injury in the accident. First, I plotted SEVERITYCODE versus Weather to check Count of Property Damage & Injury with respect to different weather condition.

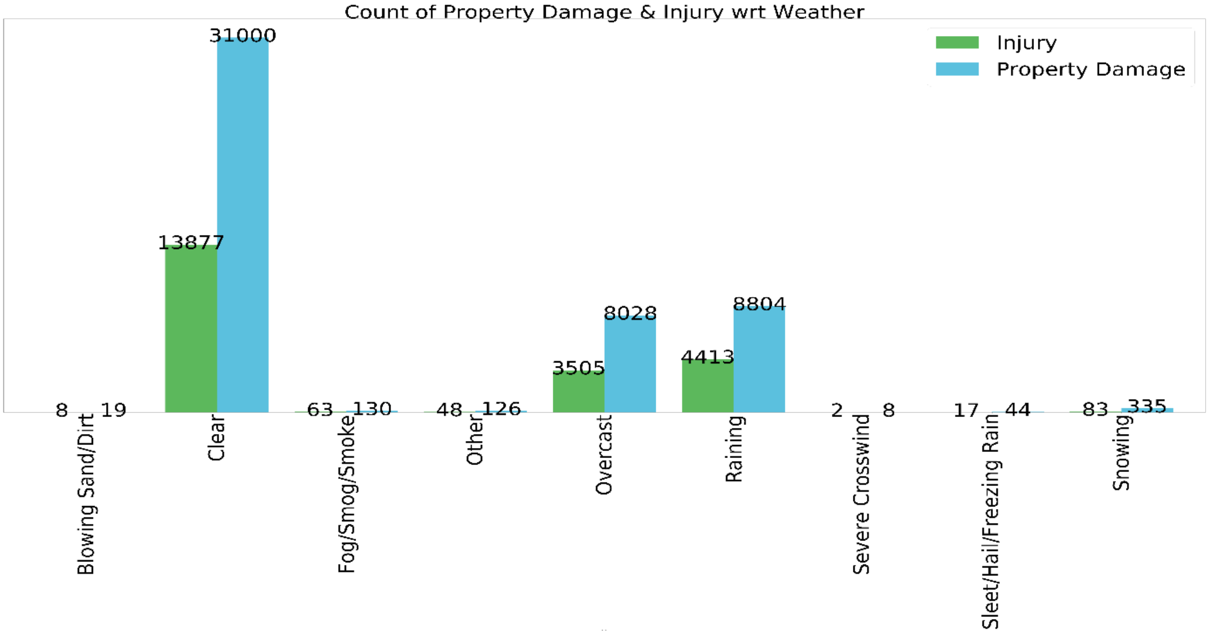


Figure 2: **Count of Property Damage & Injury with respect to Weather**

The plot implies that Clear weather accounts for majority of accident and all categories show varied proportion for Injury with respect to Property Damage. Snowing predicts lower chance of Injury. It concludes that this is an important attribute and will be part of training my predictive model.

Similarly, I plotted for SEVERITYCODE versus ROADCOND to check Count of Property Damage & Injury with respect to road condition.

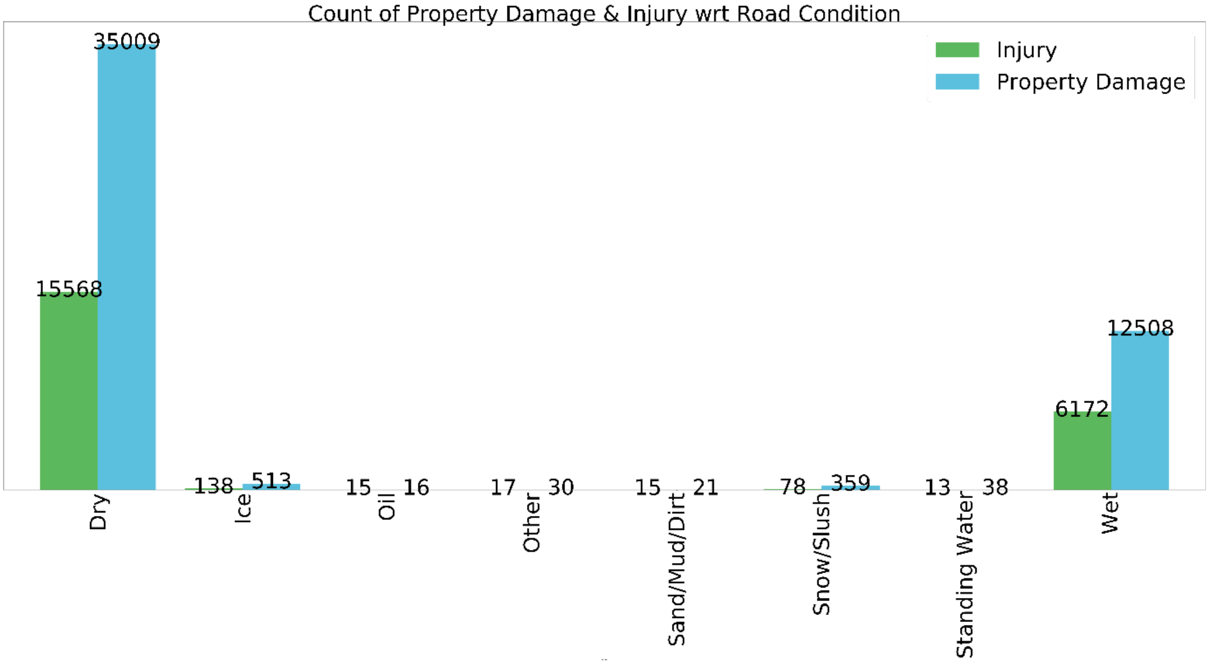


Figure 3: **Count of Property Damage & Injury with respect to Road Condition**

The plot implies that the proportion of road accident with respect to Oil and Other category have higher chance of getting injured in an accident compared to other road conditions. So, I have to consider ROADCOND attribute as necessary for my model.

Similarly, I plotted for SEVERITYCODE versus UNDERINFL to check Count of Property Damage & Injury with respect to influence of drugs and alcohol.

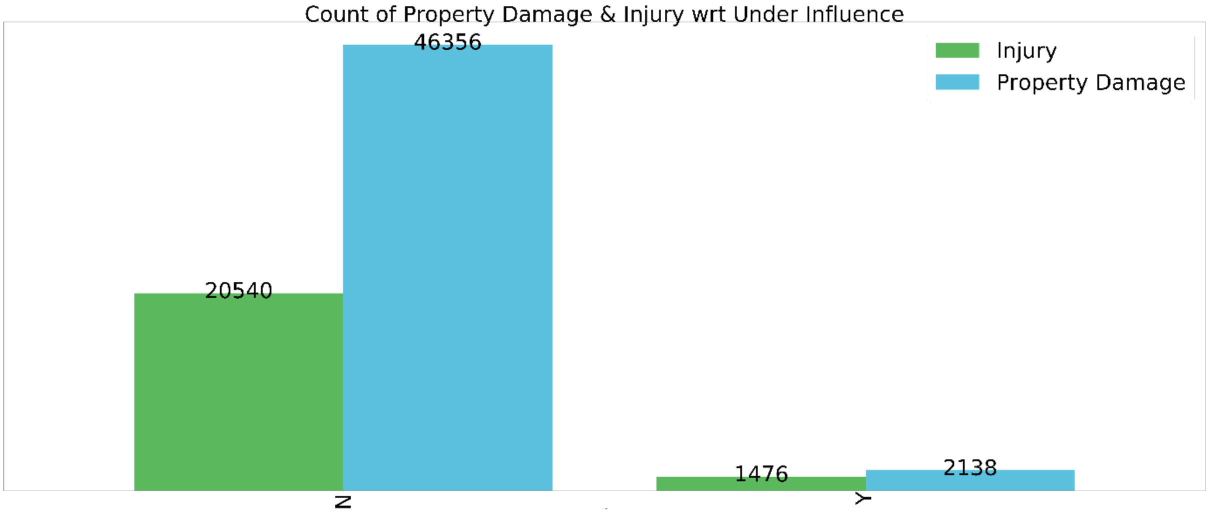


Figure 4: Count of Property Damage & Injury with respect to Under Influence

The plot implies that the proportion of people under influence of Drug or Alcohol have higher change of getting injured in an accident. So, I have to consider UNDERINFL attribute as necessary for my model.

Similarly, I plotted for SEVERITYCODE versus LIGHTCOND to check Count of Property Damage & Injury with respect to light condition.

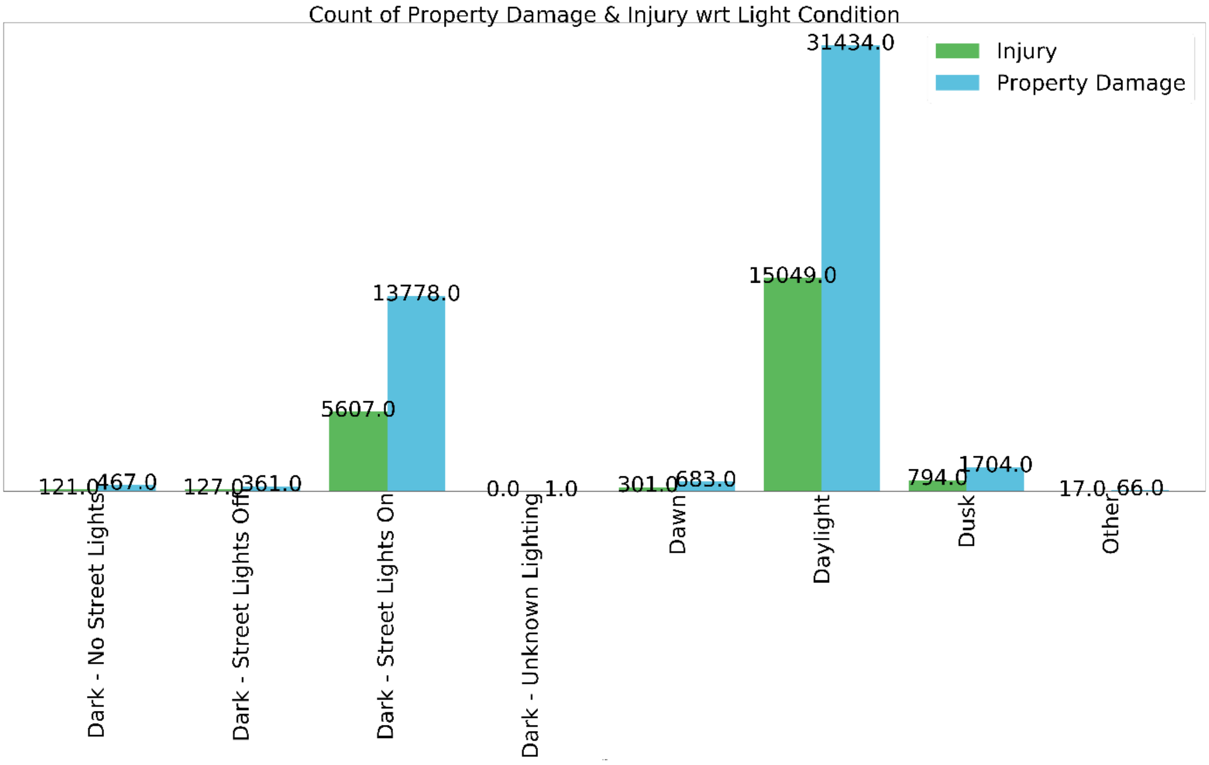


Figure 5: **Count of Property Damage & Injury with respect to Light Condition**

The plot implies that the proportion of road accident with respect to Dark – No Street Lights, Dark Street Lights Off and others have lower chance of getting injured in an accident compared to other light conditions. So, I must consider LIGHTCOND attribute as necessary for my model.

Similarly, I plotted for SEVERITYCODE versus weekday to check Count of Property Damage & Injury with respect to weekdays.

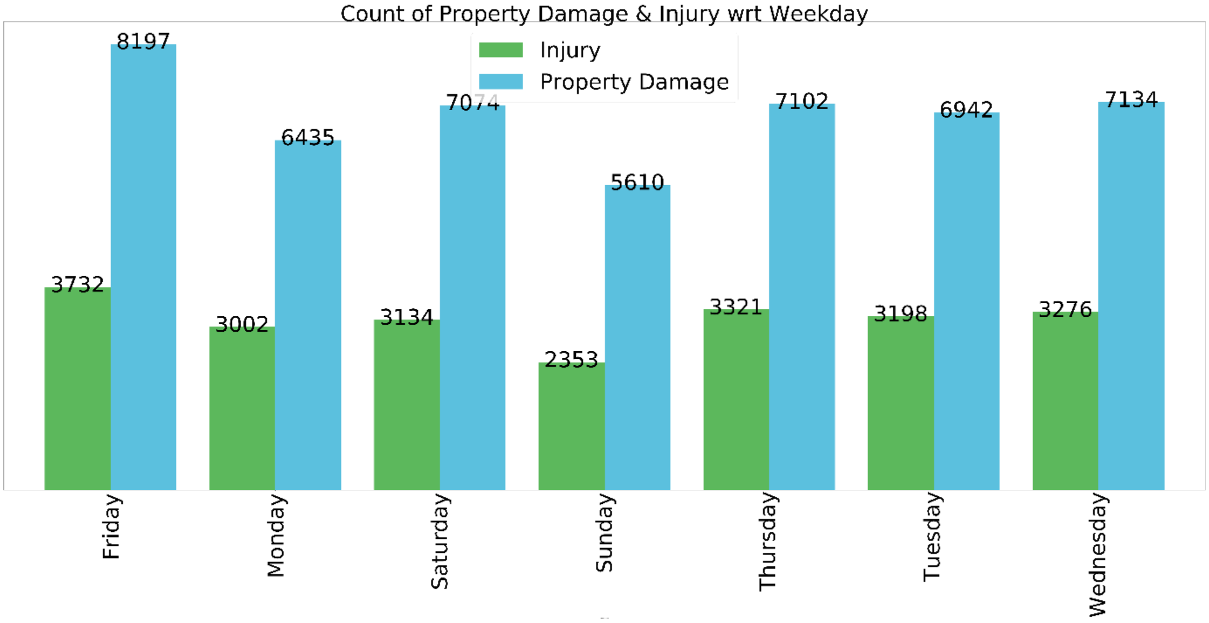
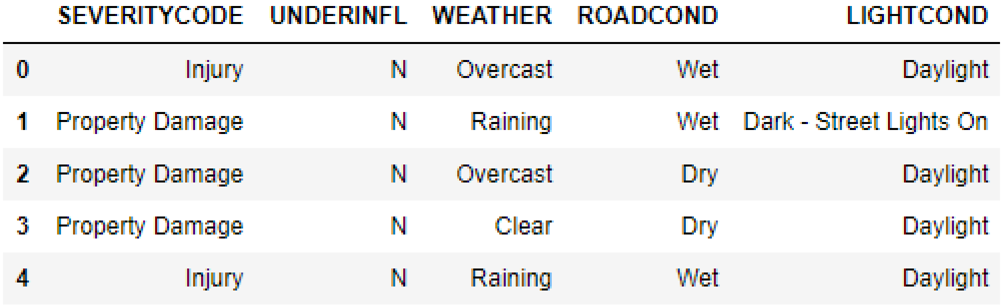


Figure 6: **Count of Property Damage & Injury with respect to Light Condition**

The plot implies that the proportion of road accident with respect to all the weekdays is similar so I can conclude that this attribute is not necessary for my model development. So, I dropped this attribute from my column.

I are left with below dataset on which I will run different Machine Learning Model to calculate the accuracy of my model.



	SEVERITYCODE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND
0	Injury	N	Overcast	Wet	Daylight
1	Property Damage	N	Raining	Wet	Dark - Street Lights On
2	Property Damage	N	Overcast	Dry	Daylight
3	Property Damage	N	Clear	Dry	Daylight
4	Injury	N	Raining	Wet	Daylight

Figure 7: Final Dataset (df_accd)

2.3. Classification Modelling and Evaluation

INCDATE will be dropped from my dataset and as I have showed it has no relevance**.** Since I have the categorical dataset, I will convert each attribute to its dummy variables.

I have defined Feature as my Independent variables which consist of following dummy variable:

X = Feature

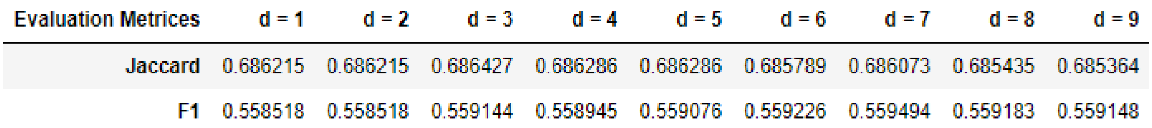
Feature = df_accd[['N', 'Y', 'Blowing Sand/Dirt', 'Clear', 'Fog/Smog/Smoke', 'Other', 'Overcast', 'Partly Cloudy', 'Raining', 'Severe Crosswind', 'Sleet/Hail/Freezing Rain', 'Snowing', 'Dry', 'Ice', 'Oil', 'Other', 'Sand/Mud/Dirt', 'Snow/Slush', 'Standing Water', 'It', 'Dark - No Street Lights', 'Dark - Street Lights Off', 'Dark - Street Lights On', 'Dark - Unknown Lighting', 'Dawn', 'Daylight', 'Dusk', 'Other',]]

I defined my dependent variable as y which is SEVERITYCODE. i.e.

y = df_accd['SEVERITYCODE'].values

I split my dataset into training and test dataset with test size of 0.2 i.e. 80% of data will be used to train the model and rest 20% of data will be used to test. I will be running following Machine Learning Modelling techniques which is mostly used for predictive classifier modelling:

a. **Decision Tree:** Decision Tree is a supervised learning predictive modelling algorithm that identifies and split a data set based on different conditions. I tried to find an optimal depth in the depth range (d) from 1 to 10 by evaluating my model and getting different Jaccard Score and F1 Score with respect to each depth range. From below table, I selected d=3 which was having maximum Jaccard score and optimal F1 score. Finally, the Decision Tree model was finalised with Jaccard value of 0.686427.



Evaluation Metrices	d = 1	d = 2	d = 3	d = 4	d = 5	d = 6	d = 7	d = 8	d = 9
Jaccard	0.686215	0.686215	0.686427	0.686286	0.686286	0.685789	0.686073	0.685435	0.685364
F1	0.558518	0.558518	0.559144	0.558945	0.559076	0.559226	0.559494	0.559183	0.559148

Table 1: Decision Tree Classifier Evaluation Table

b. **Logistic Regression**: Logistic regression is a machine learning algorithm that utilizes various solvers like Liblinear, saga etc. to classify and works best on binary classification problems, In this modelling I selected different solvers values and regularization value to find optimal Logistic Regression Model. Final Log Model was made from C=0.001 & liblinear

- solvers = ['lbfgs', 'saga', 'liblinear', 'newton-cg', 'sag']
- regularization value (C)= [0.1, 0.01, 0.001]

Following Accuracy Scores were obtained:

	Solvers				
Regularization Value	**lbfgs**	**saga**	**liblinear**	**newton-cg**	**sag**
0.1	0.619	0.618661	0.618660	0.618661	0.618661
0.01	0.618656	0.618656	0.618652	0.618656	0.618656
0.001	0.618627	0.618627	0.618953	0.618627	0.618627

Table 2: Logistic Regression Classifier

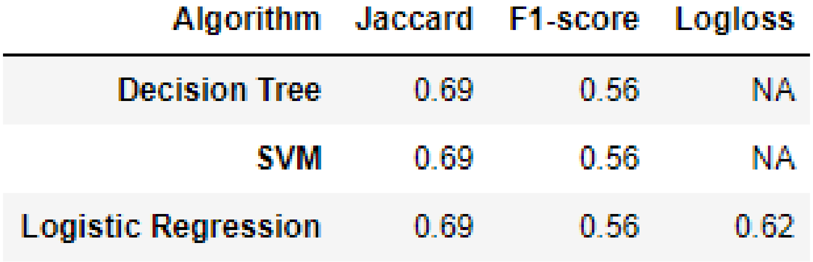
c. **Support Vector Machine**: The algorithm in SVM works by finding a hyper dimensional plane in n-dimensional space where n is number of variables. It is used for binary classification. In this modelling I chose different Kernel Function like sigmoid, poly, rbf and linear and ran the model. Below are the accuracy scores against each Kernel Function out which SVM model with Poly Kernel was finalized:

Sigmoid	0.558518008
Poly	0.558793187
rbf	0.558518008
linear	0.558642697

Table 3: Accuracy against each Kernel function

3. Results

Finally, we got following results, when we tried to calculate the F1 Score, Jaccard Index and LogLoss Score for the Model Built:



Algorithm	Jaccard	F1-score	Logloss
Decision Tree	0.69	0.56	NA
SVM	0.69	0.56	NA
Logistic Regression	0.69	0.56	0.62

4. Discussion

We can use any Classification model from above three models. They have produced somewhat identical Jaccard & F1 Score. As I mentioned in the beginning as well it will solve both the purpose of de risking the possibility of accident and lowering the traffic during accident prone conditions. It will help SDOT official to plan better.

5. Conclusion

In this study, I have analysed the relationship between Accident Severity and Variables that effects the severity of an accident. With data visualization it was really easy to see the proportion of meeting with severe accident causing injury and dependent variables like Road condition, weather condition, light condition. This model can be run by SDOT officials or people who commutes in the city and they can easily identify if given condition can cause accident or not.