IBM Data Science Capstone Project – Seattle Accident Data Analysis

Sarveswara Rao Basa

Introduction/Business Problem

City/Town planners always have difficulty to determine where to put stop sign, stop light, extra precautions such as pedestrian crossings with extra indications(flashing lights, road signs or signals) for automotives /cyclist/pedestrians etc. Since Police collect detailed data of the accident locations and other various accident data information, can we use that data to identify the locations of severe accidents. Additionally by predicting weather is a factor in such accidents and then give a predictive tool for road safety planning. This project is to identify such locations from the accident data and independently provide predictive model to 911 dispatchers to identify how much a weather is a contributing factor in the severity of the accident to mobilize additional resources.

1. Can we predict the severity of the accident based on weather condition?
2. Is there any specific region in Seattle where more accidents occur in last two years?

Severity prediction is a useful for 911 operators to mobilize and dispatch right resources ahead of time based on this predicted model.

Visual analysis is useful to town planners for additional infrastructure selection, to reduce number of incidents at high accident zones.

Data Source

We will use the sample Seattle data set provided in the course for this problem solving. Some key information is necessary for this kind of problem prediction is to have latitude/longitude information or accident location information, weather conditions etc. are important. The provided data set has the required basic data for solving the problem.

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The data set has the following features for data analysis, train and test the model.

Feature	Data Type	Comment	
SEVERITYCODE	Int	Target	
(X)Longitude	Float	Keep	For Accident location clusters
(Y)Latitude	Float	Keep	For Accident location clusters
OBJECTID	Int	Unique ID	None
INCKEY	Int	Drop	
COLDETKEY	Int	Drop	
REPORTNO	Int	Drop	

IBM Data Science Capstone Project – Seattle Accident Data Analysis

Sarveswara Rao Basa

STATUS	Text	Drop	
ADDRTYPE	Text	Keep	For Info and accident cluster
INTKEY	Int	Drop – Only reference number	
LOCATION	Text	Keep, Evaluate	For Info and accident cluster
EXCEPTRSNCODE	Text	Drop	
EXCEPTRSNDESC	Text	Drop	
SEVERITYDESC	Text	Keep, evaluate	
COLLISIONTYPE	Text	Keep, evaluate	
PERSONCOUNT	Int	Keep, evaluate	
PEDCOUNT	Int	Keep, evaluate	
PEDCYLCOUNT	Int	Keep, Evaluate	
VEHCOUNT	Int	Keep, evaluate	
INCDATE	Text	Drop as the same is available in 'INCDTTM'	
INCDTTM	Text	Keep, & Transform to DATE , Day and Time Ranges	For clustering on time/weekday/Weekend
JUNCTIONTYPE	text	Drop	Not using
SDOT_COLCODE	Int	Drop	
SDOT_COLDESC	text	Drop	
INATTENTIONIND	Text	Drop – Only one value	
UNDERINFL	Text	Drop	
WEATHER	Text	Keep	Vs Severity Code
ROADCOND	Text	Drop	Not analyzing
LIGHTCOND	Text	Drop	Not analyzing
PEDROWNOTGRNT	Text	Drop – only one value	
SDOTCOLNUM	Int	Drop	
SPEEDING	Text	Drop – Only one value	
ST_COLCODE	Int	Drop	
ST_COLDESC	Int	Drop	
SEGLANEKEY	Int	Drop	
CROSSWALKKEY	Int	Drop	
HITPARKEDCAR	Text	Drop	

Since our intention is to identify the severity of the collision, we will use **"severitycode"** as our target feature.

Dropping Features:

IBM Data Science Capstone Project – Seattle Accident Data Analysis

Sarveswara Rao Basa

Based on the Metadata given for the data set, some features can be dropped as they are mostly informational for the authorities for their tracking. Also the features which will not be used are dropped. These were indicated next to feature as **Drop** in the table.

Dropping the records:

Since this is a large data set, we will use only the most recent 2 years data for the analysis. That means we will use only 2020 & 2019 data, which gives sufficiently large data to work with.

Dropping NaN records

Since we do not want to make any assumptions on the blank values of a feature on **' SEVERITYCODE'** or any other features, any records have Blanks are dropped.

Data Preprocessing and formatting

INCDTTM is combined string and same is separated into Date, Time and 'Day of the week' and those additional features are added with intention to find any patterns for Severity prediction.

Since WEATHER is our main independent variable and it has list values, using **one hot encoding** method, the following additional features were created and added to the dataset.

Blowing Sand/Dirt

Clear

Fog/Smog/Smoke

Other

Overcast

Partly Cloudy

Raining

Severe Crosswind

Sleet/Hail/Freezing Rain

Snowing

Feature Selection

Target Feature is SEVERITYCODE (Dependent)

Weather Features (Independent)

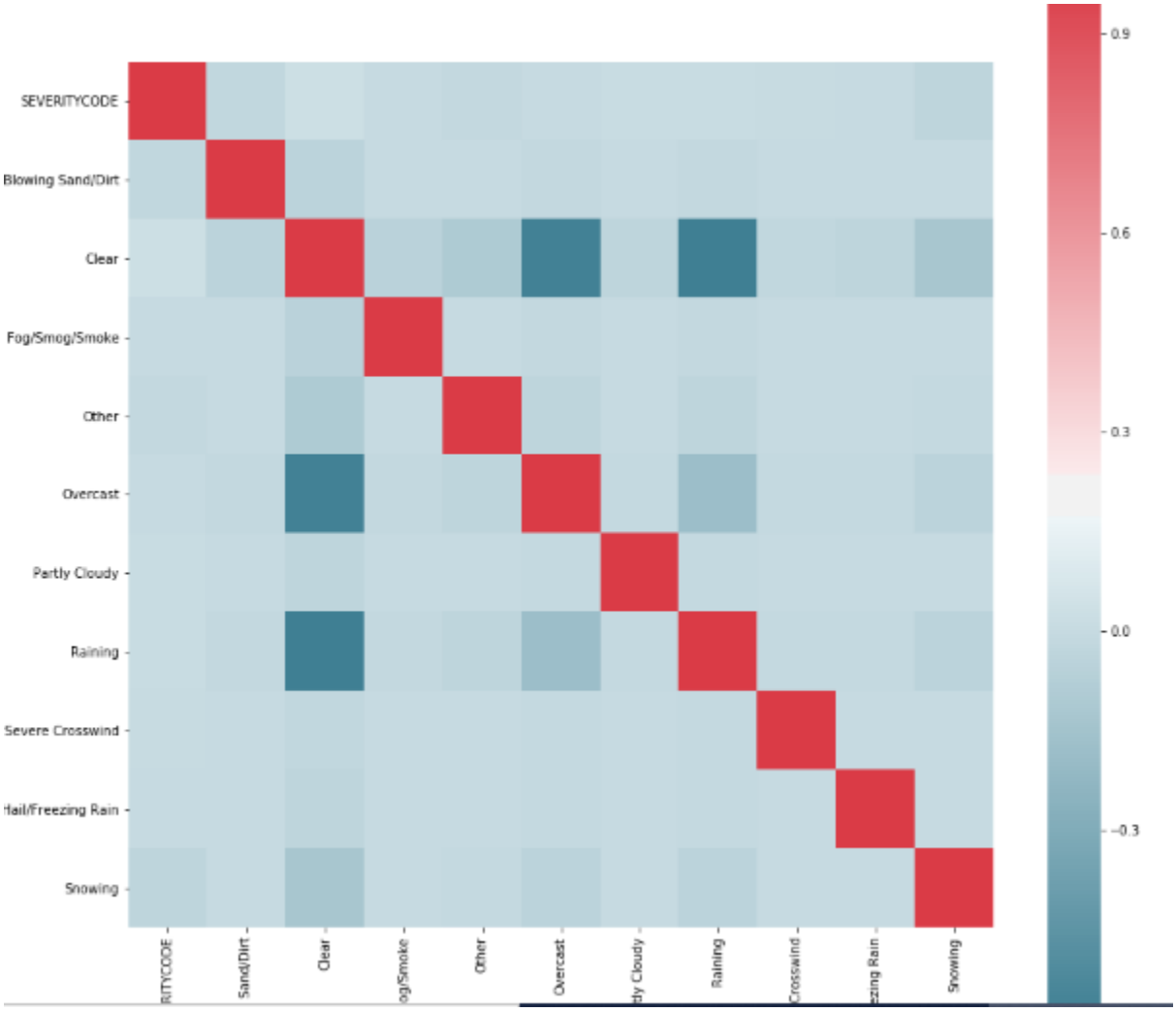
Since only two codes are available in the data sets (**1-Property Damage only**, **2-Injury collision**), we will use **classifier methodology algorithms** for modeling.

IBM Data Science Capstone Project – Seattle Accident Data Analysis

Sarveswara Rao Basa

Correlation Heatmap:

To eliminate any features highly correlated, correlation heatmap is plotted.



Modeling

The data set is randomly divided into training (75%) and testing (25%)

```
Train set: (7422, 11) (7422,)
Test set: (2474, 11) (2474,)
```

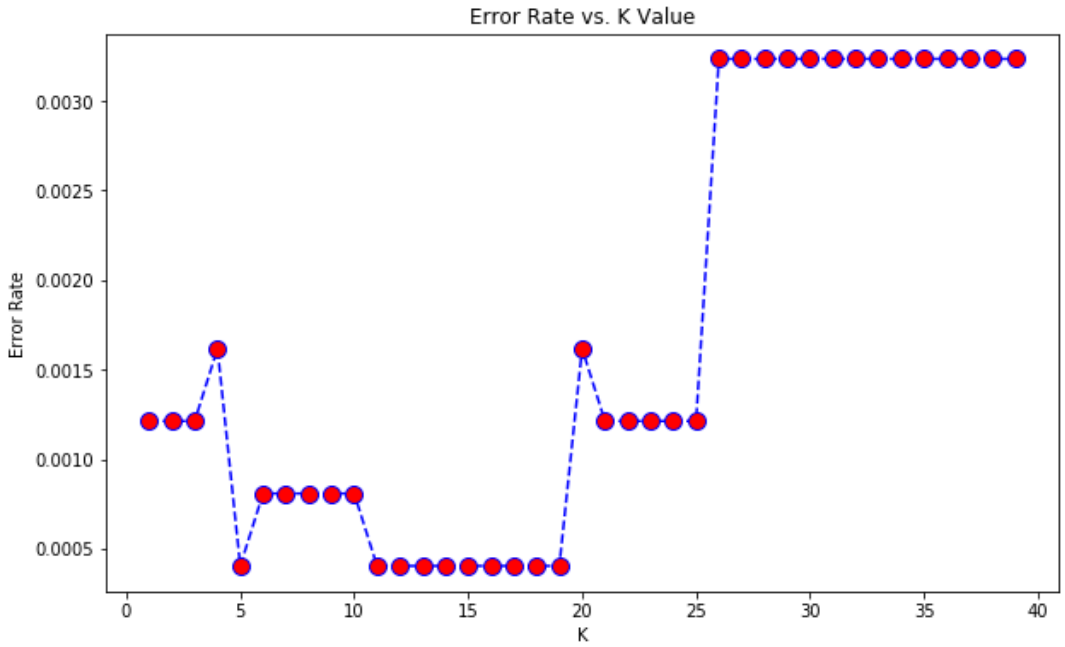
The predictive model building is done using train dataset with following algorithms:

KNN

Found the optimum K as 3

IBM Data Science Capstone Project – Seattle Accident Data Analysis

Sarveswara Rao Basa



Using K = 3,

Build the model using train dataset and then predicted the values for Test Data set and predicted the KNN J-CARD Index & F1-Score for test data set.

```
KNN Jaccard index: 0.99879
KNN F1-score: 0.99879
```

The similar modeling is done with Decision Tree, Support Vector Machine and logistic regression.

Decision Tree

```
DT Jaccard index: 1.00000
DT F1-score: 1.00000
```

Since the weather conditions are mutually exclusive, this is not useful.

Support Vector Machine

```
SVM Jaccard index: 0.99879
SVM F1-score: 0.99879
```

Logistic Regression

```
LR Jaccard index: 1.00
LR F1-score: 1.00000
LR LogLoss: 0.04323
```

IBM Data Science Capstone Project – Seattle Accident Data Analysis

Sarveswara Rao Basa

Summary:

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.99838	0.99838	NA
Decision Tree	1.00000	1.00000	NA
SVM	0.99797	0.99798	NA
LogisticRegression	1.00000	1.00000	0.04323

Conclusion:

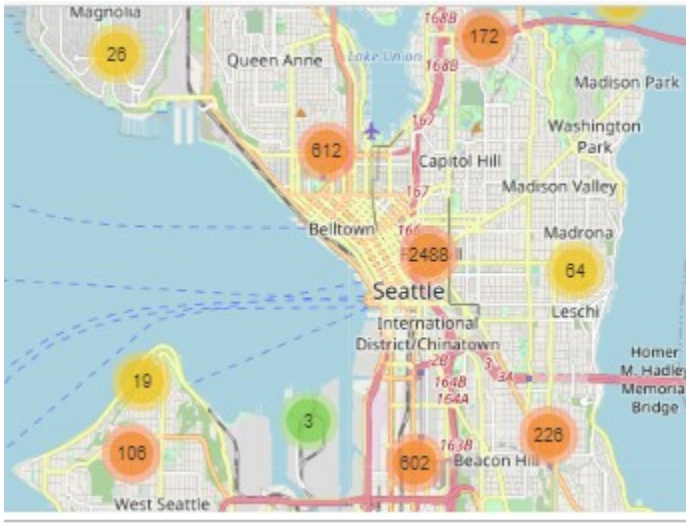
We can use KNN, SVM and Logistic regression for the prediction.

The most significant weather conditions which are contributing to Accidents are: **Overcast**, **Raining** and **Snow**

Identifying the high amount of accidents location

For this analysis, Folium library is used and mapped the accidents on Seattle geography map.

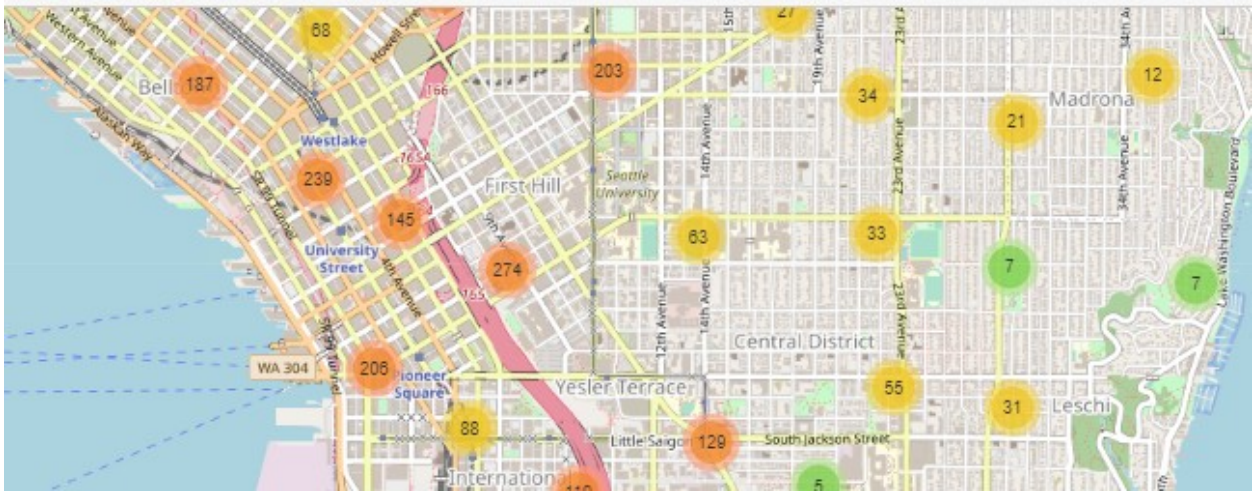
Using Marker Cluster plugin, the map is made more interactive.



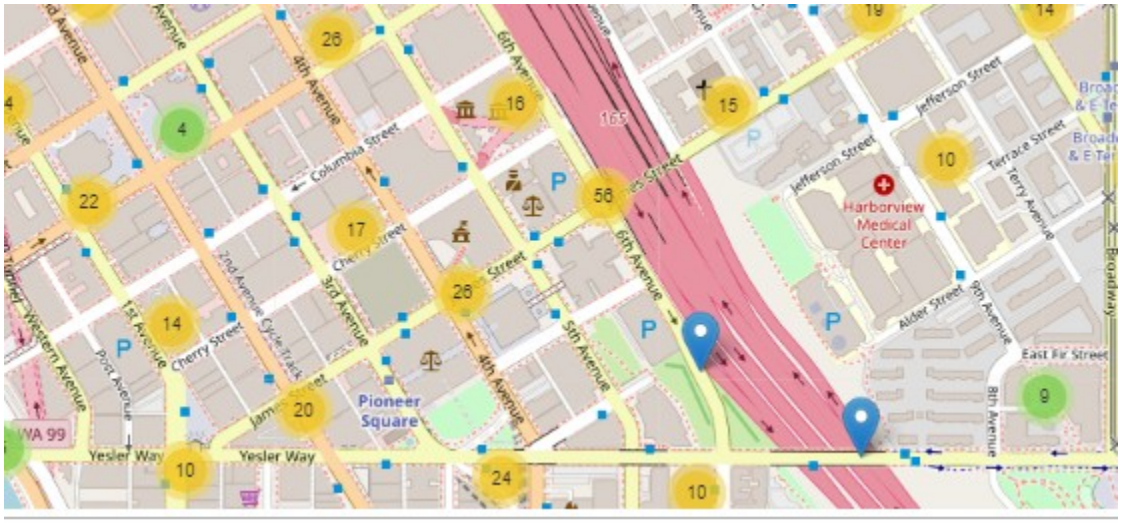
On zooming 2488 count cluster:

IBM Data Science Capstone Project – Seattle Accident Data Analysis

Sarveswara Rao Basa



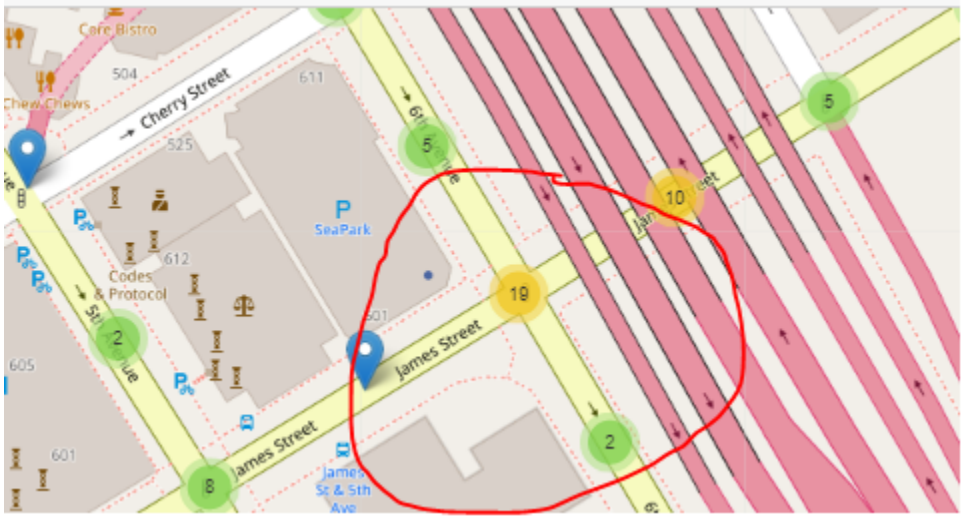
On further zooming on 274 cluster:



On zooming further on 56 Cluster:

IBM Data Science Capstone Project – Seattle Accident Data Analysis

Sarveswara Rao Basa



Conclusion:

One of the high accident incidents intersection is: James Street/6th Avenue.

Now a town planner can do further analysis to find what are the causes for the high incidents and make corrective measures.