Predicting Accident Severity in Seattle

1. **Introduction** :

Road Traffic Accidents in Seattle cause Property damage and Injury to people.

Accidents create traffic chaos and require manpower to transport accident victims to the Hospital. This puts additional pressure on the Seattle Police and Traffic department and causes unnecessary delay to the people driving to the work or other city.

1.2 **Problem Description** :

Build the Machine Learning Model (ML) to predict the Accident Severity on the basis of the several parameters such as Weather conditions , Road conditions, Light conditions , Location , junction type and the vehicles over-speeding or not.

This model needs to be build while holding regular discussions with the Seattle Police and Traffic Department

1.3 **Interest** :

Seattle Police and Traffic Department would be very interested in predicting the Accidents to avoid traffic chaos .

People would also be interested so that they can drive more safely or even change their travel plans.

Insurance Companies would also stand benefit so that they can also issue advisory to people if they predict accidents using the model.

2. **Data Understanding** :

2.**1 Data Source** :

We have used the example Data set provided in the Capstone Project.

Data has the details on Accident occurrences in Seattle.

Data has been provided in the form of comma separated version (csv) file “Data-Collisions.csv” :

https://github.com/shashwatdhyani/Coursera_Capstone/blob/master/Data-Collisions.csv

The data file has 38 columns with SEVERITYCODE column repeating hence effectively 37 columns and 65691 rows.

Datasource has the details on the description and Severity of the Accidents happened on a particular day at a given Location and Junction. It has the details on the Weather condition , Light condition and the Road condition when the accident happened. It has record whether the vehicle was over speeding or not. It also has the details how many vehicles and people were involved or whether pedestrians were also involved. Data set also records kind of collision and whether the parked car was hit in the accident.

2.2 **Data Cleaning** :

The value count for SEVERITYCODE shows that data is not balanced and skewed towards Severity 1.

There are 45813 rows for Severity 1 and 18845 rows for Severity 2.

But since the data has been collated by the Seattle Traffic Department we can say the data is not biased.

If we drop the Severity 1 rows to balance the data then we will have to drop too many rows of data which could result in loss of information hence we did not drop any row for Severity 1 code.

There were too many NULL values for SPEEDING , PEDROWNOTGRNT AND INATTENTIONIND.

Metadata information tells that values are Y and N. Since Y was present in the dataset we converted the NULLs into N.

We then removed remaining Null values from the dataset.

We removed the outliers ,like, PERSONCOUNT of value more than 7 has very little occurence hence we dropped the rows having PERSONCOUNT more than 7.

Similarly we dropped the rows where VEHCOUNT was greater than 4.

Grouping the values count for HITPARKEDCAR showed parked car were hit only negligible times.

We found similar observation for PEDCOUNT.

2.3 **Feature Selection** :

We have taken the following columns (features) which would have more impact in predicting the Accident Severity :

'LOCATION','ADDRTYPE','COLLISIONTYPE','INATTENTIONIND','UNDERINFL','WEATHER','SPEEDING','LIGHTCOND','PEDROWNOTGRNT','ROADCOND','JUNCTIONTYPE','PERSONCOUNT','VEHCOUNT'

Dropped the Features like x,y coordinates of the Location ,collision code , description and other codes which are maintained by the Traffic Department for record managements

3. **Predictive Modeling /Methodology** :

This problem falls under the Supervised Classification hence we need to have numerical values for the features.

The majority of the features have categorical values :

'LOCATION','ADDRTYPE','COLLISIONTYPE','INATTENTIONIND','UNDERINFL','WEATHER','SPEEDING','LIGHTCOND','PEDROWNOTGRNT','ROADCOND','JUNCTIONTYPE'

In order to train the model using any of the Classification Algorithm we need to convert the categorical values for these features into the numerical values .

We used Label Encoder method to convert the categorical values of the Features into numerical values

Next we split the data into the Training data set and the Testing data set.

Since it is the Supervised Classification prediction We trained the model with the following Algorithms :

KNN Classifier , Decision Tree Classifier , Support Vector Machine (SVM) and Logistic Regression.

3.1 **KNN Classifier** :

We trained and then tested the accuracy of the Model with several values of n_neighbors and found the best accuracy score of 0.7356 was achieved with n_neighbors=8.

Hence we build a KNN Classifier model with n_neighbors=8 to compare with other Algorithm models.

3.2 **Decision Tree Classifier**:

We trained and then tested the accuracy of the Model with several values of depth and found the best accuracy score of 0.7489079563182527 with depth= 8

Hence we build a Decision Tree Classifier model with depth=8 to compare with other Algorithm models.

3.3 **SVM Model** :

SVM model showed the best accuracy score of 0.75 with kernel=rbf.

Hence we build the SVM model with kernel=rbf

3.4 **Logistic Regression Model** :

Regression model showed the lowest accuracy score of 0.7205

We then evaluated each model based on F1-Score and Jaccard Score.

Evaluation metrics for all the Models¶

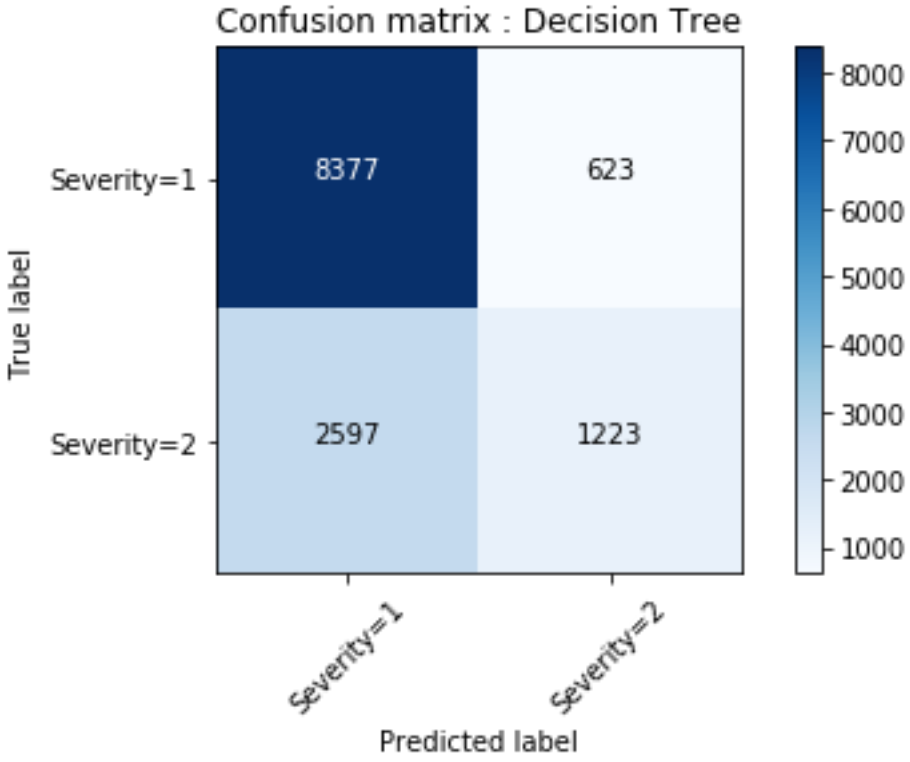
Algorithm	Jaccard	F1-Score	Log Loss
KNN	0.7356	0.7076	NA
Decision Tree	0.7488	0.7174	NA
SVM	0.7465	0.6948	NA
Logistic Regression	0.7205	0.6649	0.5504

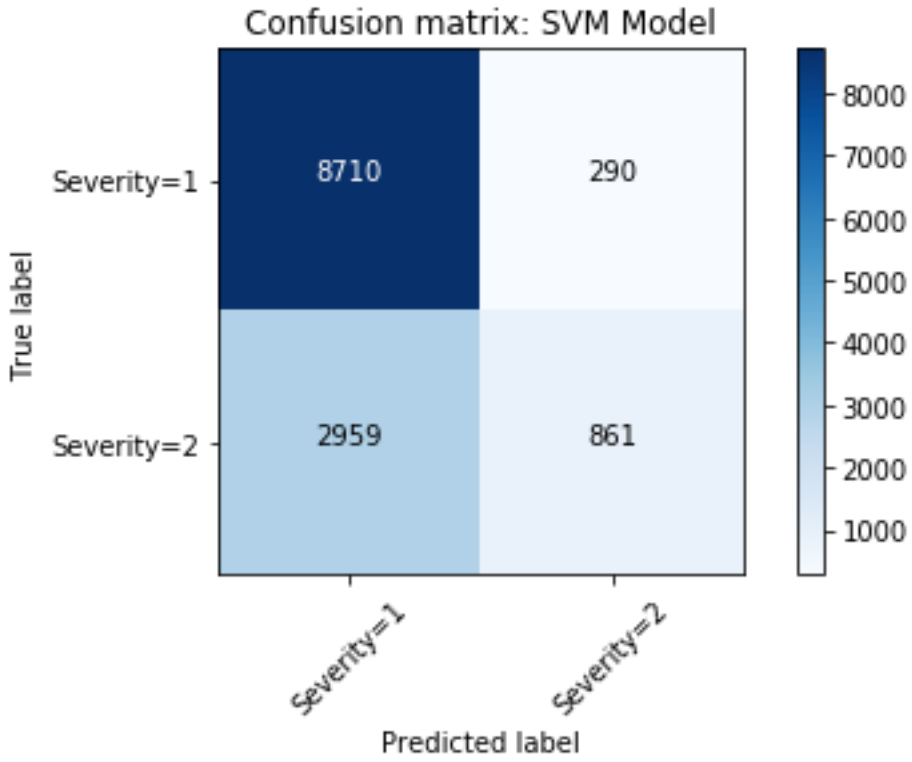
Decision Tree model has the highest Jaccard Score (0.7488) and the highest F1-Score (0.7174) among all the Models.

SVM model has the next highest Jaccard Score of 0.7465 and the F1-Score of 0.6948.

KNN model has the Jacard Score of 0.7356 while Logistic Regression Model has the Jacard score of 0.7205

Since Decision Tree model and SVM model had small difference in Jaccard score we built the Confusion Matrix for Decision Tree model and SVM model.





From the Confusion Matrix we found that Decision Tree model can predict Severity 2 Accidents more accurately (almost by 50%) than SVM Tree model.

There is not much difference in predicting Seveity 1 Accidents between the two models

Results :

From the Confusion Matrix we can see that Decision Tree model can predict Severity 2 Accidents more accurately (almost by 50%) than SVM Tree model.

There is not much difference in predicting Seveity 1 Accidents between the two models.

Since Decision Tree model also has highest Jaccard score and F1 -Score we choose Decision Tree model for predicting the Accident Severity in Seattle.

Conclusion :

The Machine Learning Model built in this Project can be used to predict the Accident Severity in Seattle.

We found most of the Accidents occurred at Block Addrtype.

Alcohol or drug influence did not cause many accidents.

It can be deployed to warn the people of the likelihood of the Accident which could be helpful in avoiding traffic chaos and unnecessary delays.

It will help save on the manpower and the Cost to the local Governing bodies which would have otherwise incurred by transporting more victims to the hospitals.

Insurance companies would also get benefit by issuing advisory to the people.