Accident Severity Rating (ASR)

**Maris Sekar**

**September 26, 2020**

# **1 Introduction**

## 1.1 Background

Road safety is a major problem that is often taken lightly. The World Health Organization (WHO) reports that “Approximately 1.35 million people die each year as a result of road traffic crashes”. Many people incur a disability as a result of their injury. Many more lives change either directly or indirectly due to road accidents. The personal injury, emotional and mental health impacts are one side of these unfortunate incidents. But there are economic impacts to the individual, their family and their community as well depending on the nature of the accidents. These impacts are often unforeseen and the relationships associated between these impacts are often forgotten as an aftermath to an unfortunate outcome. It is important to be able to assess the accident severity and the accident severity level to incorporate the human fatality, traffic delay, property damage, or any other type of accident parameters that are involved so that the right level of support can be provided to the affected individuals. For example, this can be a standardized and independent rating – Accident Severity Rating (ASR) that can be shared with insurance providers for them to process insurance claims using a logical and structured approach to assess the claim. This can be used in mandatory driver training programs to focus and share awareness of parameters that contribute to improving road safety (for example, we can determine how much more likely someone is to get into an accident if they are speeding) and further reduce road accidents.

## 1.2 Problem

Data such as location, collision type, number of people involved in the accident, number of vehicles involved, incident date, road conditions, weather and whether speeding was a factor may provide great insights when determining the severity level. The objective of this project is to predict the severity level of accidents and determine the major contributing factors to this severity level.

## 1.3 Interest

This can be of interest to the road transport safety communities within the government to better understand the contributing factors and the relationships between them in order to introduce targeted awareness campaigns and programs to reduce road safety incidents. This can be of great interest to insurance companies so that they are able to logically assess a claim based on the logically derived severity rating.

# **2 Data Acquisition & Cleaning**

## 2.1 Data Sources

I used the provided data set named Data-Collisions.csv shows data provided for this project from the Seattle Department of Transportation for the period between 1st January 2004 and 20th May 2020. There was a total of 194,673 collision incidents with 37 attributes. However, there were a lot of missing columns (FATALITIES, INJURIES and SERIOUSINJURIES) when compared to the provided metadata as well as additional columns (REPORTNO and STATUS) that were not present in the metadata. The missing columns are very important in determining the severity. Another major deficiency with this data was SEVERITYCODE column only had two values out of a total of five possible values according to metadata. This made me more uncomfortable to choose this data so I tried to find an alternate dataset.

Fortunately, Kaggle had a better version of this data set provided also by the Seattle Department of Transportation. There was a total of 221,144 collision incidents with 40 attributes. This had more consistent columns including FATALITIES, INJURIES and SERIOUSINJURIES. This data set also had all possible values for SEVERITYCODE which is important because this is the target variable to predict.

## 2.2 Data Wrangling

The data was analyzed and cleansed based on the requirements for this project. The irrelevant columns were first dropped. The remaining columns were renamed with more sensible headers. Next, I checked for duplicate rows. There are no duplicate rows found.

Missing or null values were then checked and corrected by either replacing with Other, 0 or removing the rows if I was unsure of what values need to go there. StateCode was the only variable that required cleansing since no values entered for this field means data was not entered for these lines. I removed the lines that had empty StateCodes. A major decision was to replace Severity ratings “2b” with 4. Since the idea is to convert the dependent target variable into a number.

The string fields were then encoded into number format using the LabelEncoder () function. Data types were originally a mix of integer, float and string. By the end of the data type conversion all of them were converted to integer data type.

Some fields need to be cleaned up so that Y, N become 0s and 1s – one hot encoding. For example, INATTENTIONIND, UNDERINFL and JUNCTIONTYP fields were cleaned up this way.

I removed the following columns, reasons are given on the same line below:

* + EXCEPTRSNCODE & EXCEPTRSNDESC appears to describe exceptions and there are 5638 lines with not enough information or insufficient location information so remove this column
  + PEDROWNOTGRNT – Most are “N” so we are not losing much information by removing this.
  + SDOTCOLNUM – an arbitrary number that is assigned by SDOT.
  + SEGLANEKEY AND CROSSWALKKEY– Mostly has 0. So, we will remove these columns
  + HITPARKEDCAR – Most have “No”.
  + X, Y – don’t need location information for our models.
  + LOCATION (string) – don’t need location information.
  + INCDATE (INCDTTM has date and time) – Don’t need date info.
  + INCDTTM - Don’t need date info. Good for analysis though. Like to look at what time and days most incidents happen.
  + SDOT\_COLCODE – We are already using ST\_COLCODE which seems to contain similar information.
  + SDOT\_COLDESC (Same reason as above)
  + ST\_COLDESC – don’t need description for features.
  + INJURIES – this is heavily correlated with SEVERITYCODE.
  + SERIOUSINJURIES – has a lot of values in 0 bucket.
  + FATALITIES – has a lot of values in 0 bucket.

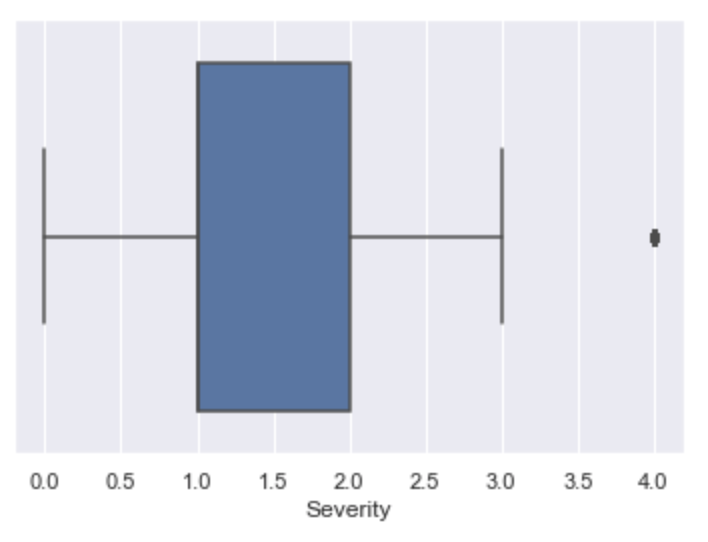
## 2.3 Feature Selection

The following columns were chosen as features:

* ADDRTYPE
* INTKEY
* COLLISIONTYPE
* PERSONCOUNT
* PEDCOUNT
* VEHCOUNT
* PEDCYLCOUNT (Not included because pedestrian/person similar)
* JUNCTIONTYPE
* WEATHER
* ROADCOND
* LIGHTCOND
* SPEEDING
* INATTENTIONIND
* UNDERINFL
* ST\_COLCODE

Target is the SEVERITYCODE field

## 2.4 Outliers

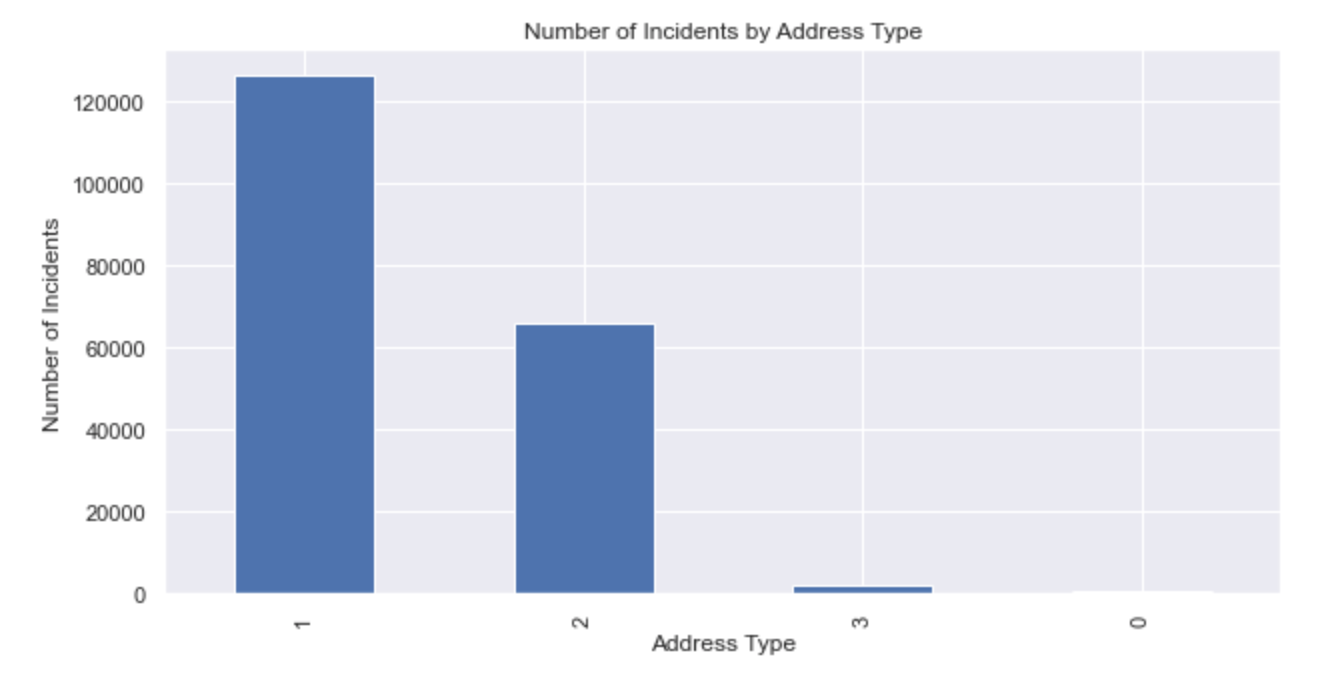


The SEVERITYCODE was checked for outliers. There were a bunch of 4s that were outside the box plot. But I left this in the data set to ensure 4s get picked up (even though the chance of picking it up is slow). This helped one of the algorithms (Decision Tree) to be able to predict the 4s.

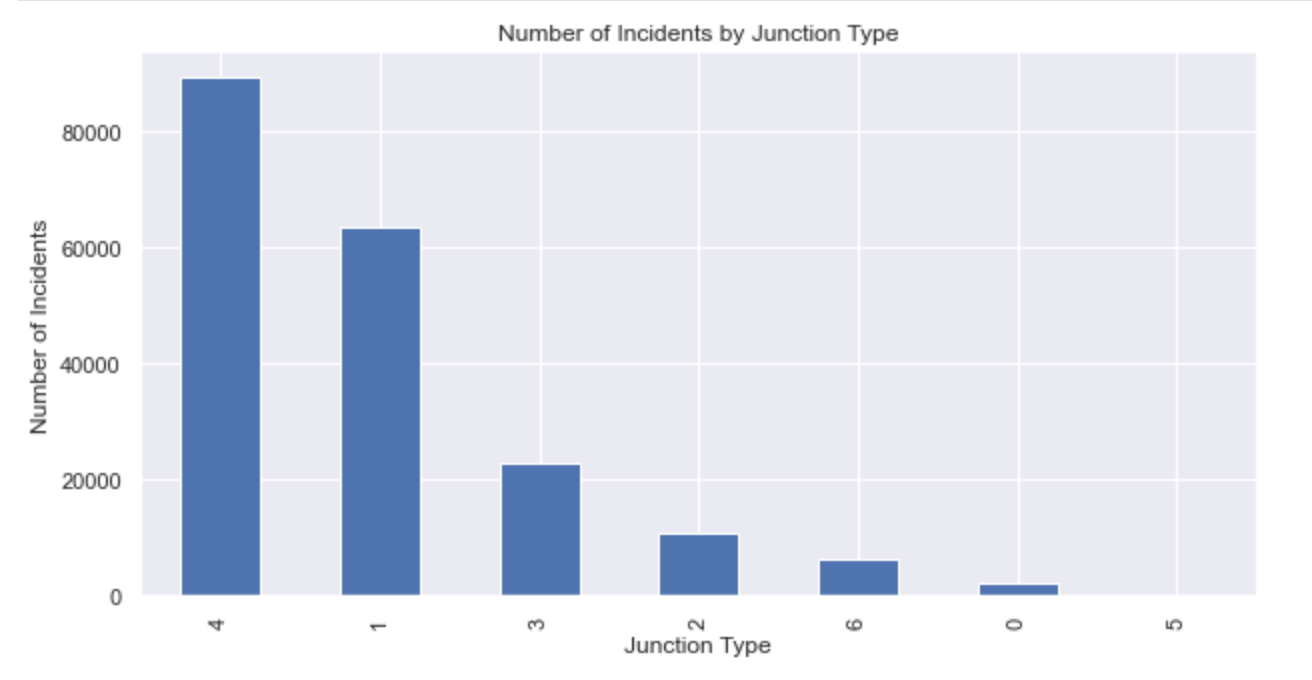
Another outlier was found for the INJURIES columns which was originally used as a feature but later ignored. There was one incident that had greater than 70 injuries recorded. This seemed odd and was excluded from the training and test dataset.

# **3 Exploratory Data Analysis (EDA)**

## 3.1 Relationships

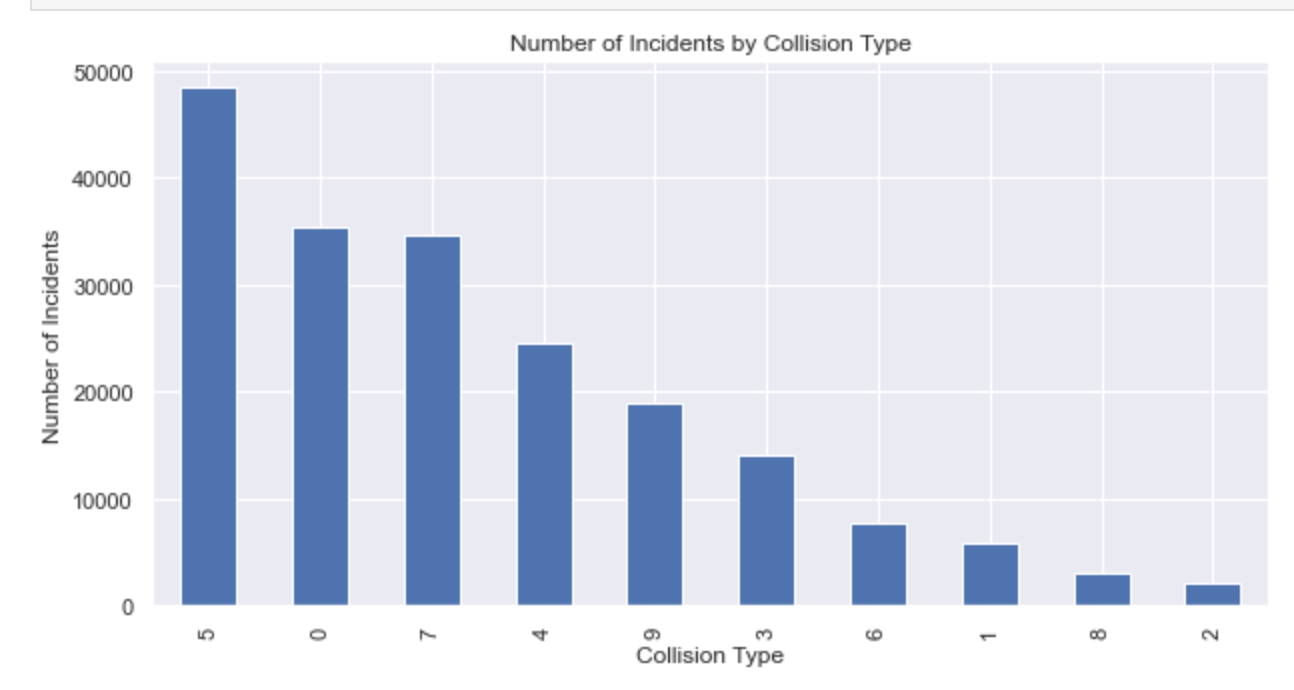


Most accidents occurred near a block as opposed to other types of locations. I would expect most to occur in an intersection but depends on how many intersections are in the city of Seattle.

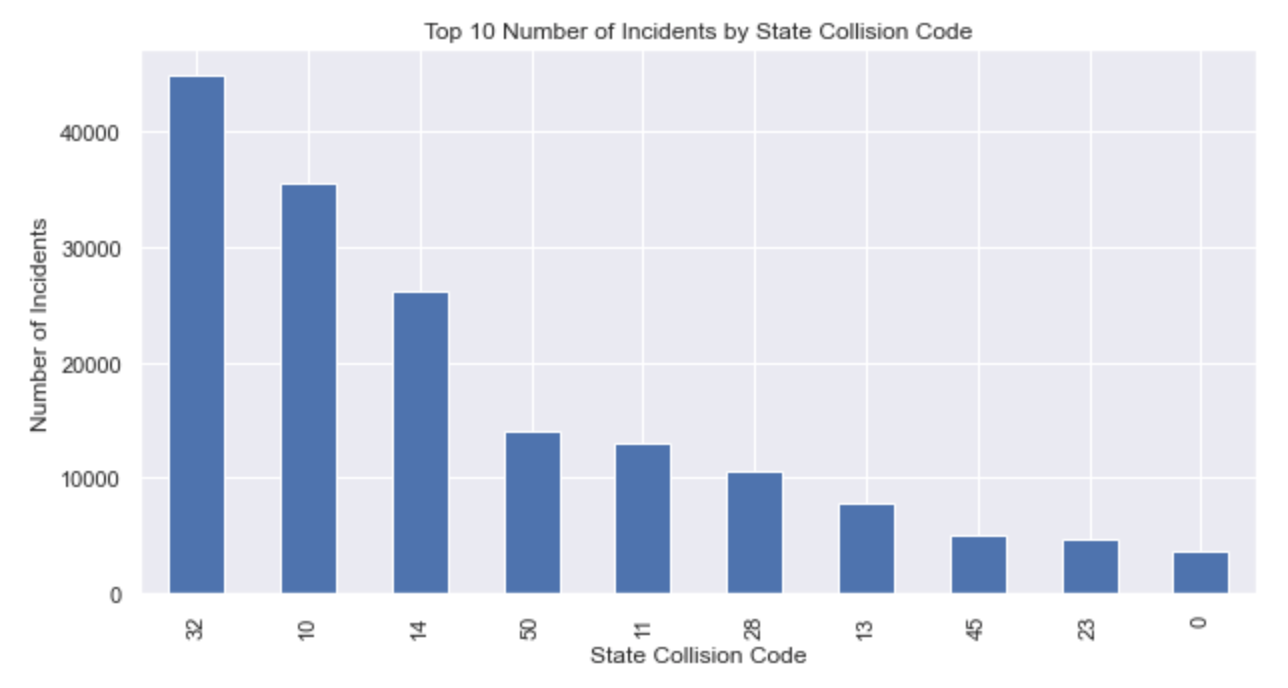


Most incidents occurred Mid-Block and were not related to being near an intersection - this correlates with the bar chart on top of this one that looked at the address types.

"Raining", "Other" and "Overcast" weather occuring the most after this when an incident took place. Most incidents occured during daylight conditions followed by dark conditions. This is counter intuitive but it makes sense that since most people drive during the day the chances of accident greatly improves. The chances of having an incident when the road conditions were dry is higher than any other condition. I believe this is due to the volume of traffic during the dry time but there may be more to this than meets the eye.

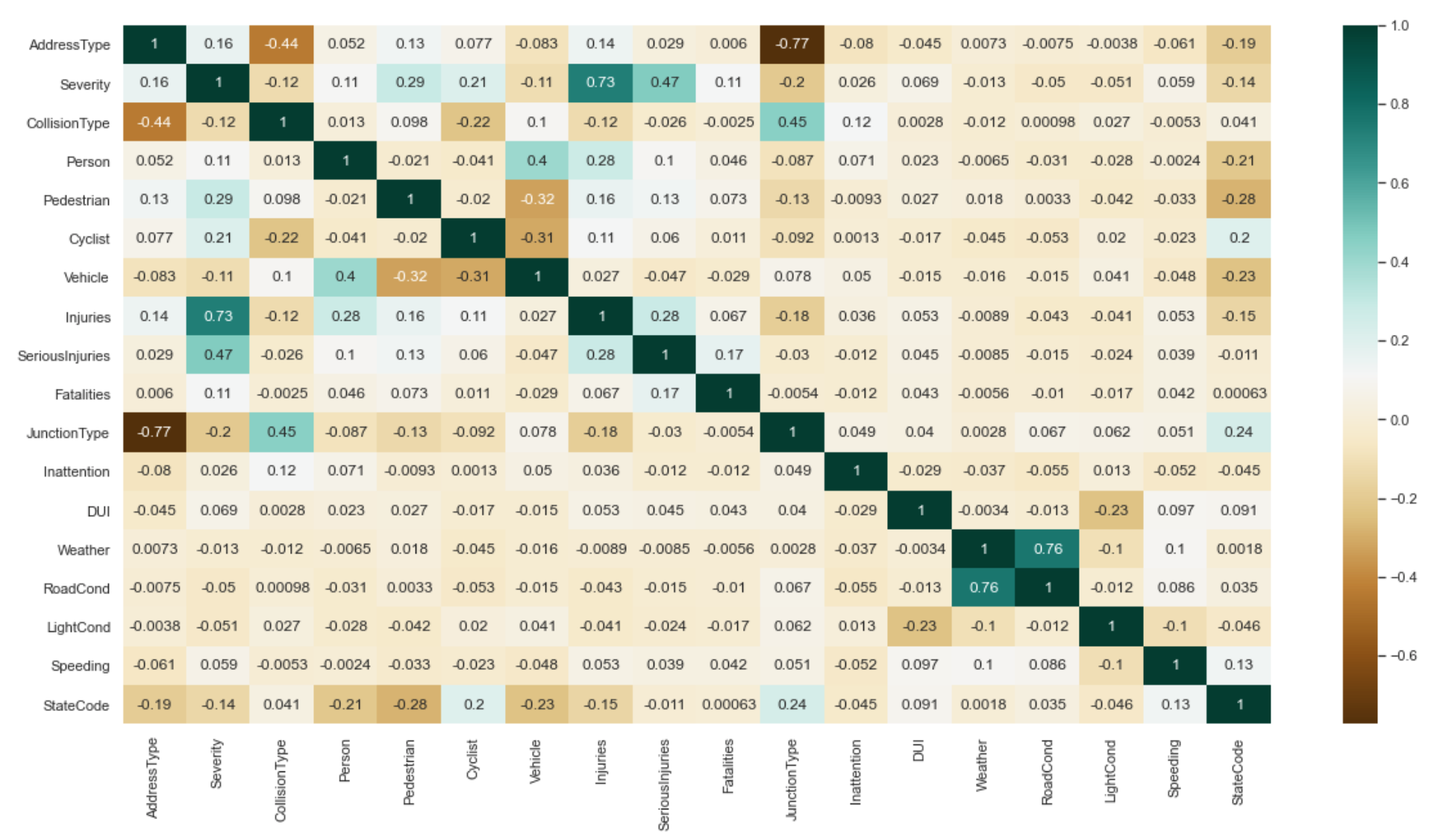


Most collisions occured while parked. Angles and Rear-Ended collisions are almost as high - slightly lower. Head On collisions were a handful which makes sense.



The top three collision codes correspond to "One Parked - One Moving", "Entering At Angle" and "From Same Direction - Both Going Straight - One Stopped - Rear End". These correspond with the incidents by collision type histogram shown above.

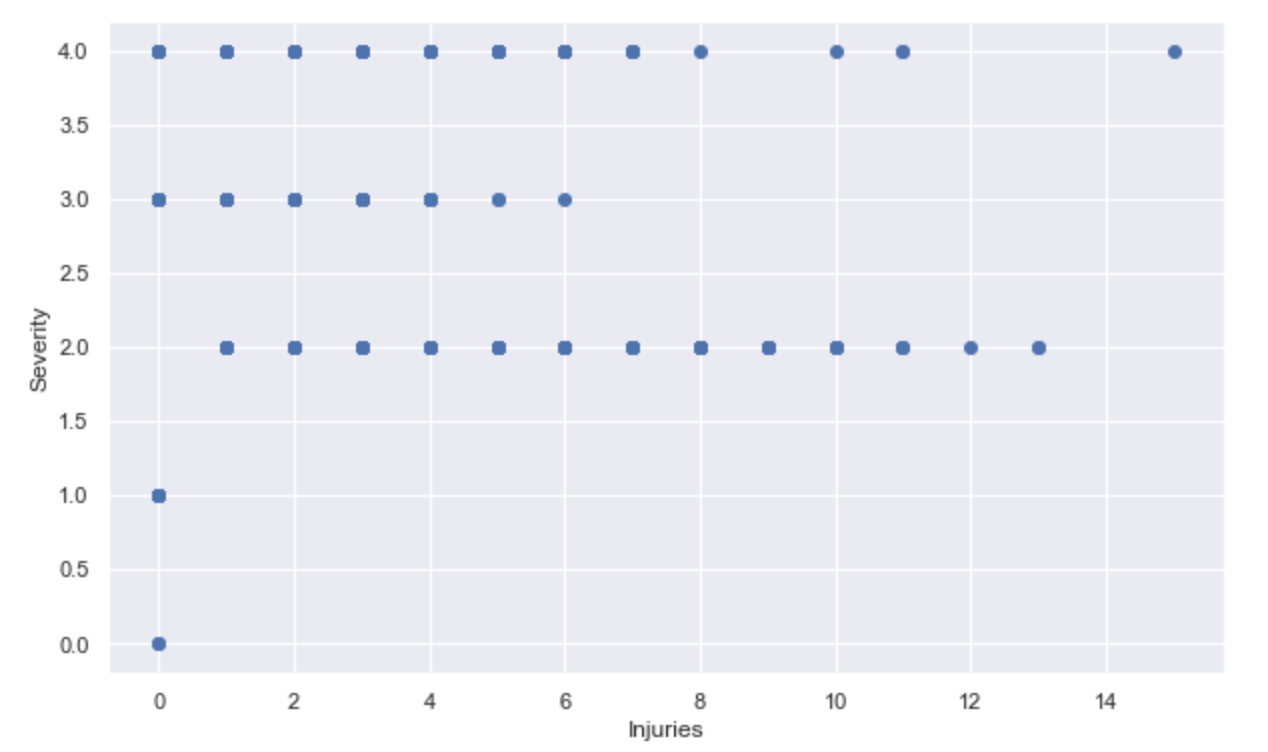
## 3.2 Correlation Matrix



We can see from the above that Severity is the dependent variable that is related to Injuries, but not Serious Injuries and Fatalities. This may not make sense at first because you would expect fatalities and serious injuries to have an effect on Severity. But since the number of fatalities and serious injuries that are > 0 are very low it makes sense that Severity is only heavily influenced by injuries alone. Since injuries (minor injuries) are the most commonly occuring incident.

Severity is also dependent on number of Vehicles, Persons, Pedestrians and Cyclists involved in an incident.

The next best correlation is between number of persons involved in an incident and the no. of vehicles involved. This could be because there is most likely one person per vehicle involved in an incident.

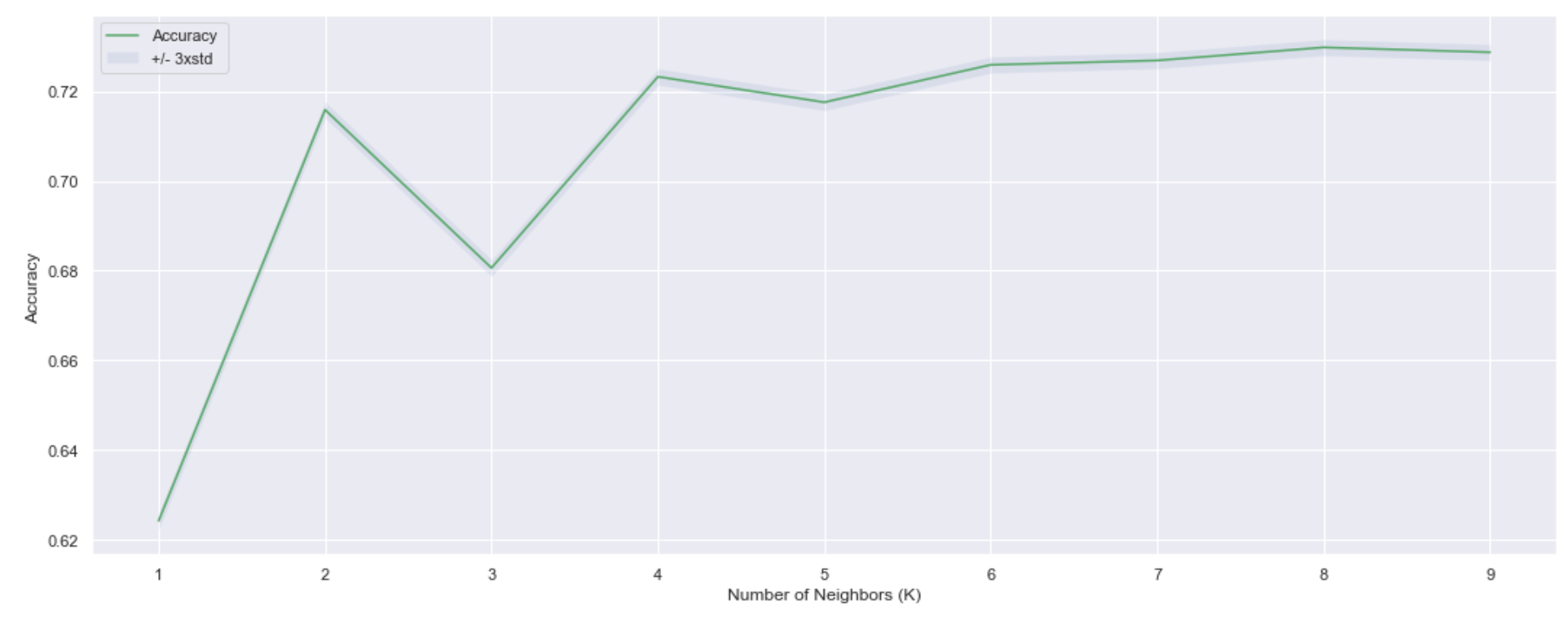


Severity appears to be a step function of the number of Injuries that averages around 2.5 Severity. Could be a linear regression model with y = 2.5 approximately based on the above plot.

# Modeling and Evaluation

I chose the following classification models for this project:

* K-Nearest Neighbors (KNN)
* Decision Tree
* Logistic Regression
* Support Vector Machine (SVM)
* XGBoost



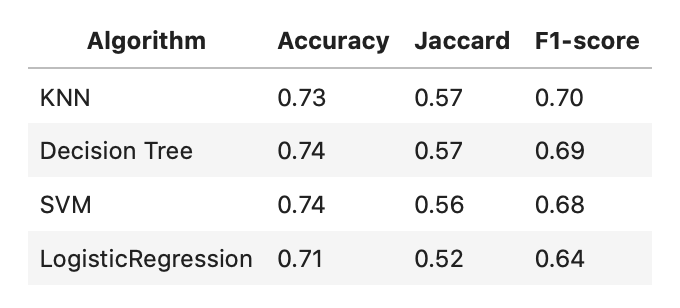
For the KNN algorithm a K of 8 was chosen after an iterative run between K=1 and K=10 since it provided the maximum accuracy as shown above.

I decided to drop the Injuries, SeriousInjuries and Fatalities features in order to predict the target variable Severity. This is because there is a high degree of correlation between Injuries and Severity as seen from the correlation matrix above. I dropped SeriousInjuries and Fatalities as well since most of the incidents are zeroes for these features.

I normalized the training and testing dataset especially since this is a classification problem. For all of the above models except Support Vector Machine (SVM) I used a 70/30 split – 70% training data and 30% testing data to validate the model. For SVM I choose 50,000 records as my training set and 10,000 records as the test set since my computing resources couldn’t handle more than 50,000 records for the SVM model.

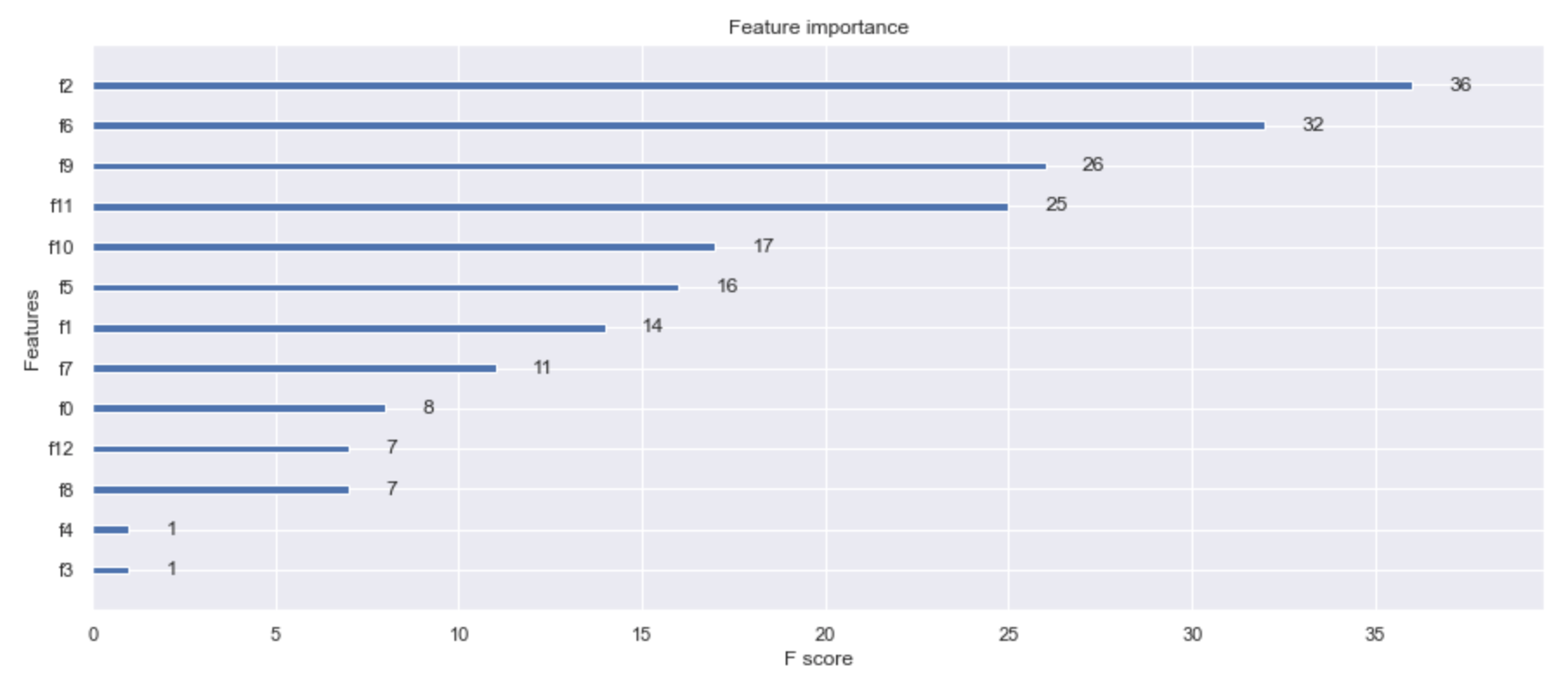
## Evaluation

For every model, the accuracy score, Jaccard index, F1-score as well as the classification report was generated to evaluate the model. The results were tabulated and are given below.



All of the models have similar scores with a F1-score of around 70%. Accuracy is around 75% for out of sample test set.

The RMSE was calculated for the XGBoost model for evaluation and was found to be 0.612. The error is very small so XGBoost algorithm seems to be doing a good job with this data set.

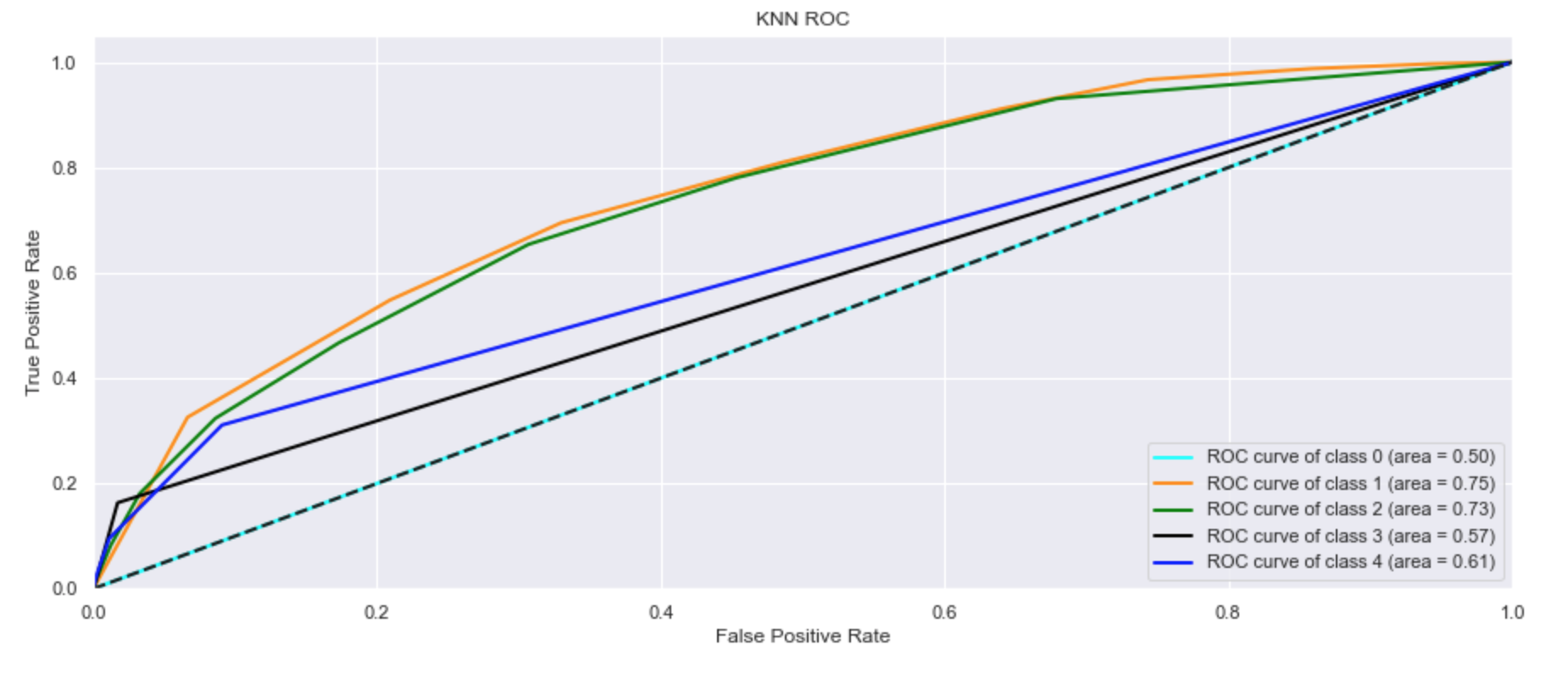


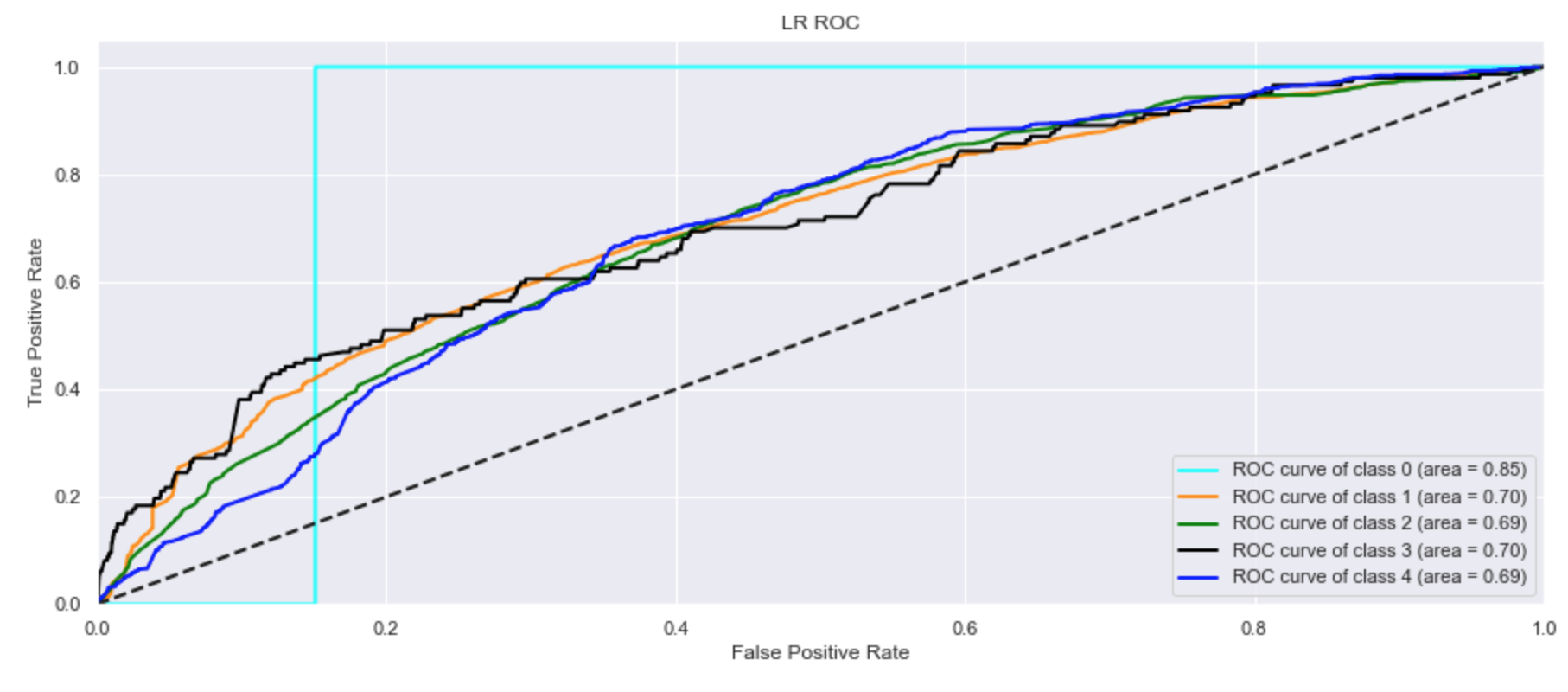
The feature importance was calculated for the XGBoost algorithm and graphed as shown above. As can be seen above the feature with the most important is the Junction Type and the Number of Persons involved in an incident.

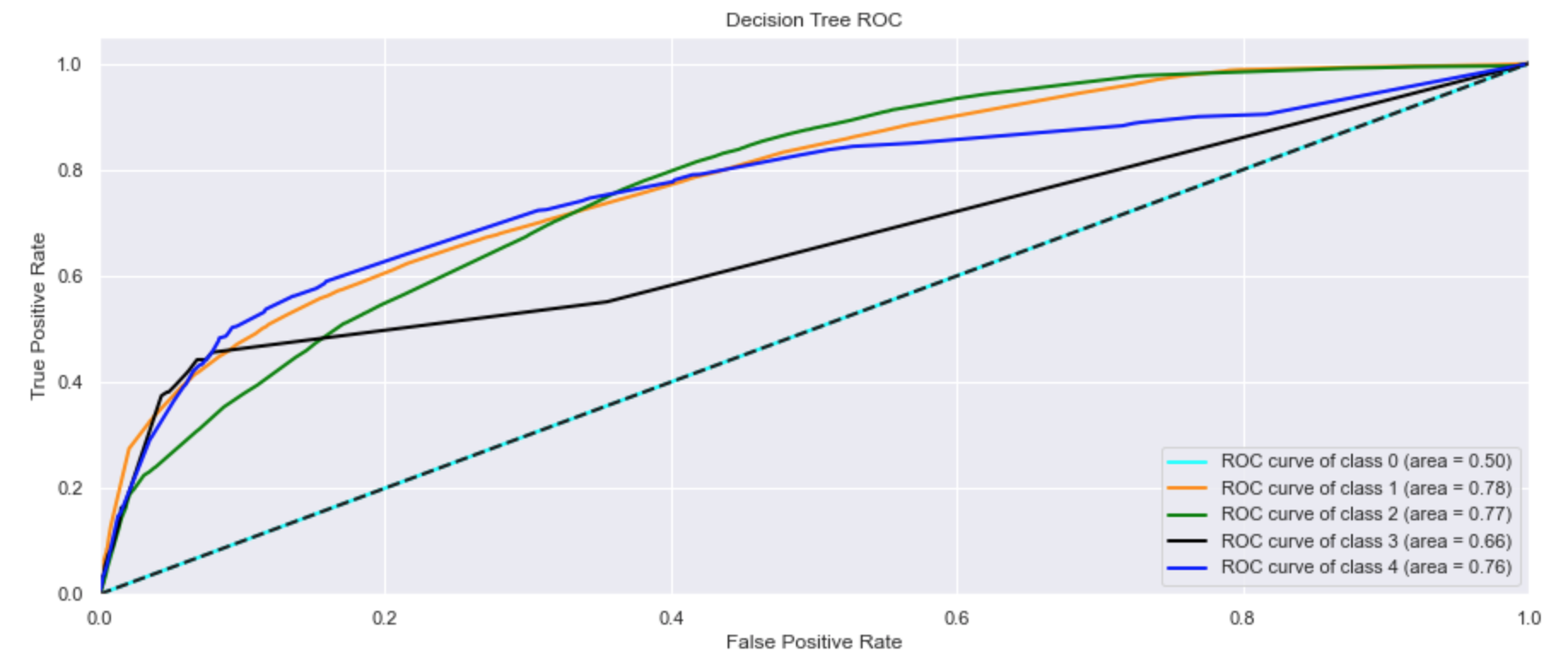
The confusion matrix for the Logistic Regression algorithm is given below. Although most 1s were correctly classified, 2s were wrongly classified as 1 quite a number of times.



The ROC of the KNN, Logistic Regression and Decision Tree algorithms was also plotted as shown below:







From the above ROC plots it can be seen clearly that the Decision Tree ROC for all the classes look very attractive - classes 0, 1, 2, 3 and 4 had more True Positives compared to the False Positives which makes its curves steeper compared to the other two ROCs. One thing to note is the best case is with the Logistics Regression for just class 0 which reached the maximum number of True Positives the fastest.

# Conclusion

The best model for this problem appears to be the Decision Tree model because it is able to predict almost all classes of Severity ratings compared to the rest of the models. This is important especially due to the enormous emphasis on the "1" rating for Severity. Especially for safety applications I would rather be able to predict more classes than predict one class reliably. Further research and testing needs to be done with Decision Tree models to see the feasibility of an application using these algorithms.