COURSERA IBM APPLIED DATA SCIENCE CAPSTONE PROJECT: CAR ACCIDENT SEVERITY IN SEATTLE

By A. Mohan

1. Introduction

The city of Seattle collects data on motor vehicle collisions, and the data sets are available for the last several years. It is the objective of this project to analyze the data and use the insights gained from the data to predict the patterns in conditions that lead to accidents, and the severity of the accidents.

These insights can be made available to the city and the public, who are the audience/ stakeholders, and can be used to enhance public safety and minimize accidents and their severity. For example, if accidents are found to occur at a higher rate under some weather or road conditions and in some locations, the members of the public may try to avoid driving under those weather conditions in those locations in future, whenever possible. City officials may also take measures to improve public safety. For example, if many accidents occur due to distracted driving, the city officials may try to educate the public to be more mindful of the risks, or require defensive driving training.

In this report, the data set provided will be analyzed and recommendations will be made based on the conclusions from the data analysis and machine learning techniques applied to the data.

2. Data

The city of Seattle has provided the data set on car collisions in Seattle from the year 2004 to present. The data set provided contains the following 38 columns:

```
['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',
 'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
 'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',
 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',
 'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC',
 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',
 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC',
 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR']
```

The dependent variable to be predicted is the severity code (labeled SEVERITYCODE), which describes the severity of the collision, and has the values:

- 1: property damage
- 2: injury

The analysis of the data found the following features to be important:

- X, Y: which specify the longitude and latitude of the accident
- ADDRTYPE: intersection, alley or block
- COLLISIONTYPE, which describes the type of collision, e.g., rear-ended, angled, etc.
- PERSONCOUNT, which describes the number of people involved in the accident
- PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, which describe the number of pedestrians, bikes or vehicles involved in the accident, respectively
- JUNCTIONTYPE, which describes the type of junction, and contains similar information to ADDRTYPE

- ST_COLDESC and SDOT_COLDESC, which contain a description of the collision. In the predictive modeling, COLLISIONTYPE, which captures the relevant information, will be used instead
- INATTENTIONIND, which describes whether the driver was inattentive
- UNDERINFL, which describes whether the driver was driving under the influence of alcohol
- WEATHER, ROADCOND, LIGHTCOND, which describe the weather, road and light conditions
- SPEEDING, which describes whether the driver was speeding
- PEDROWNOTGRNT, which describes whether the pedestrian right of way was not granted
- In addition, some features were extracted from the data. These are:
  - MONTH: The month in which the accident occurred
  - DAYOFWEEK: The day of the week on which the accident occurred
  - WEEKEND: Whether it is the end of the week. However, on analyzing the data, the day of the week was found to be a more important feature
  - SEASON: The season of the accident. However, on analyzing the data, the month was found to be a more important feature
  - YEAR: The year in which the accident occurred

Several columns were found not to be important and were dropped. These are:

- OBJECTID, INCKEY, COLDETKEY, REPORTNO, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SDOT_COLCODE, SDOTCOLNUM, and ST_COLCODE, which are keys and indices for the purposes of record keeping, and will not be used in data analysis
- LOCATION, which is already captured by X and Y
- SEVERITYCODE.1, which is a repeat of SEVERITYCODE
- SEVERITYDESC, which is a string description of the severity code
- STATUS, which describes whether the accident is matched or unmatched
- SEGLANEKEY and CROSSWALKKEY, which describe the lane and crosswalk. These values are mostly 0, and are not descriptive of the accident
- HITPARKEDCAR, which describes whether a parked car was hit. This information is already contained in other variables describing the collision type, for example, the COLLISIONTYPE column
- In addition, INCDATE and INCDTTM are not used further after extracting the month, day of the week, season, year, and whether it is a weekend. The time of the day is already captured by LIGHTCOND, describing the light conditions.

During data cleaning, NaN values were dropped. For the features UNDERINFL, SPEEDING, PEDROWNOTGRNT and INATTENTIONIND, empty values were interpreted as False, and were updated to 0. (Although this may not be accurate, more reliable data are not available at this time.) For the remaining features, empty values were dropped. For ROADCOND, LIGHTCOND, WEATHER AND COLLISIONTYPE, values that are specified as “Other” or “Unknown” are not useful for predictive data modeling, and were dropped before finally implementing predictive modeling.

A supervised learning approach was used to fit and predict the severity of collisions based on the above attributes. The Pearson correlation coefficients were calculated to check for correlations between the features, and the features selected were found to be uncorrelated or at most weakly correlated. The

supervised learning methods used were K Nearest Neighbors (KNN), Logistic Regression, Decision Tree and Support Vector Classification methods.

The fits and predictions will be used to make recommendations that may help reduce collisions. The fits are evaluated using metrics including accuracy score, log loss, precision, recall, F1-score, and ROC curves. The results also identify further important factors not present in the data set, which may be collected to improve the predictions in future.

3. Methodology

To understand the locations of the collisions in Seattle and identify areas with large number of collisions, the collisions were plotted using the longitude and latitude provided in the data set. Figure 1 shows the map with the collision data. It is seen from Figure 1 that the collisions are denser near the city center. This is even more evident in Figure 2, showing collisions of severity code 2 (resulting in injuries). The collisions are concentrated near the city center.

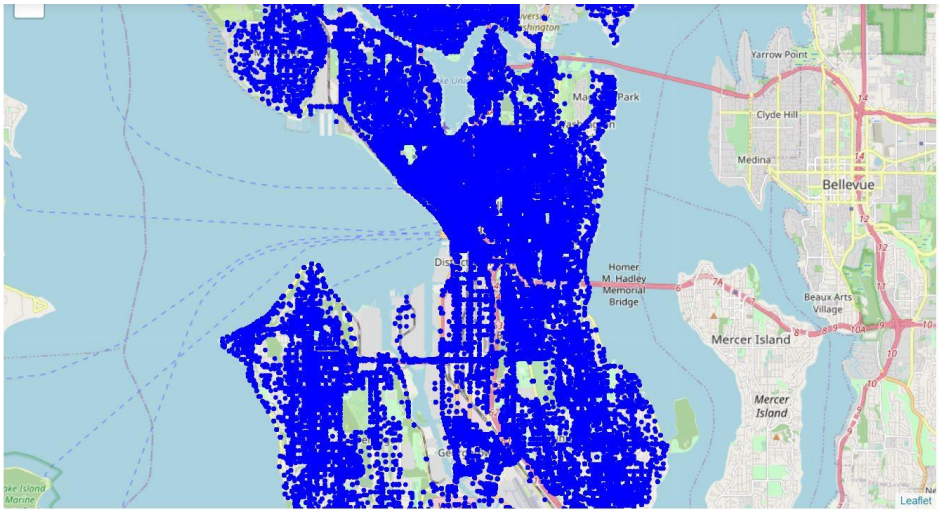

Figure 1: Map of Seattle showing collisions

Figure 3 shows the number of collisions plotted against the month. October has the largest number of collisions. This could be because rainfall starts to increase in October at the end of the fall season. It is also seen that there is a rise in the number of collisions at the transition between seasons, for example, in March, May and October. This could be due to distracted driving and drivers not paying enough attention to the changing conditions. Conversely, the fewest accidents are in February, which could be because winter is coming to an end, but drivers are still cautious from their exposure to more difficult weather conditions in January.

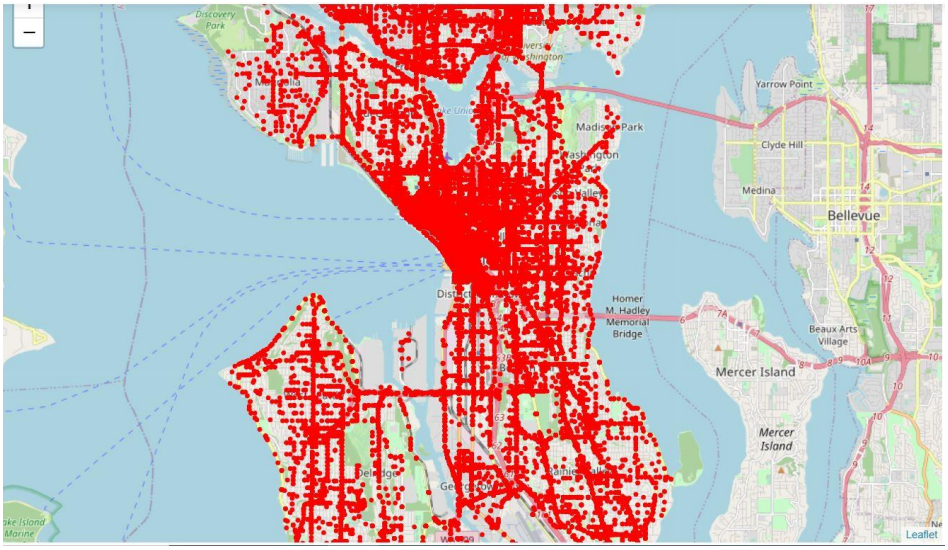

Figure 2: Map of Seattle showing collisions of severity code 2 (resulting in injuries)

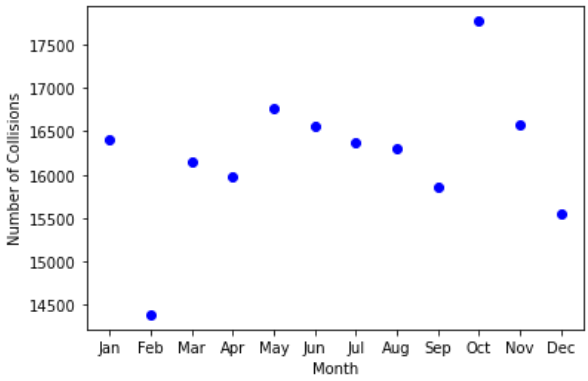

Figure 3: Number of collisions vs month

Figure 4 illustrates the number of collisions for each day of the week. A spike is observed on Fridays, which may be due to driver fatigue at the end of the work week.

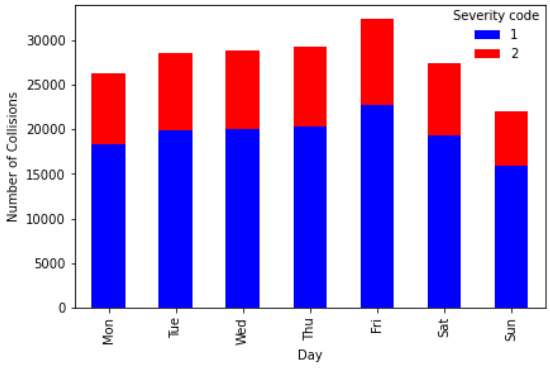

Figure 4: Number of collisions vs day of the week

Figure 5 illustrates the number of collisions for various weather conditions. It is evident that the largest number of collisions occurs on clear days, and far exceeds collisions under rainy conditions. Seattle, which has rainfall for approximately 156 days of the year, has an almost equal number of rainy days and clear days. The large number of collisions occurring on clear days again points to driver fatigue and distracted driving as a major factor. Similarly, Figure 6 shows that most collisions occur when road conditions are dry, and Figure 7 shows that most collisions occur during daylight, or where street lighting is on at night. All of the above observations point to driver fatigue or distraction as a greater factor in collisions than weather, road or light conditions.

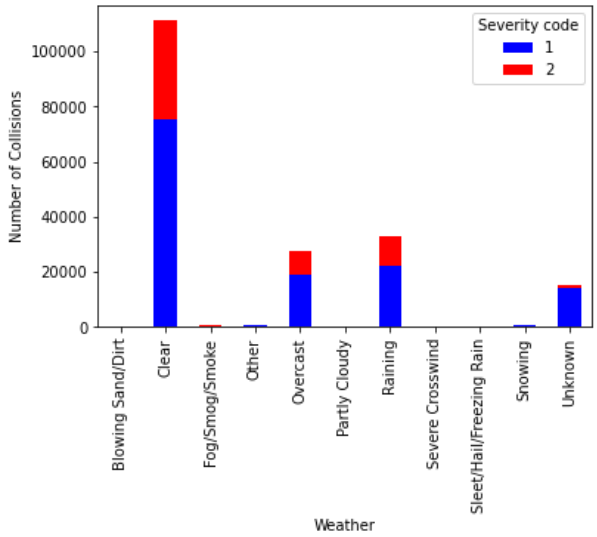

Figure 5: Number of collisions vs day of the week

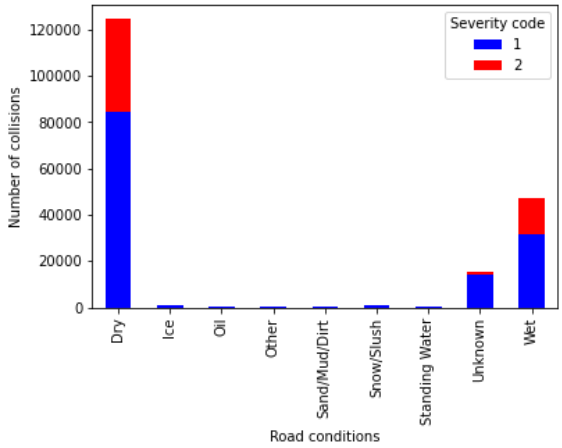

Figure 6: Number of collisions vs road conditions

Figure 8 shows that more accidents occur at mid-block than at intersections. As seen from Figure 9, the largest number of collisions occurs due to a vehicle hitting a parked car, followed by collisions at angles and rear-ending. The largest numbers of severity code 2 collisions, which result in injuries, are due to collisions at angles and rear-ending.

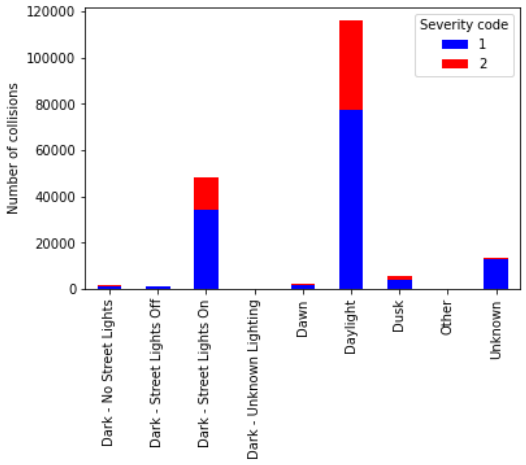

Figure 7: Number of collisions vs light conditions

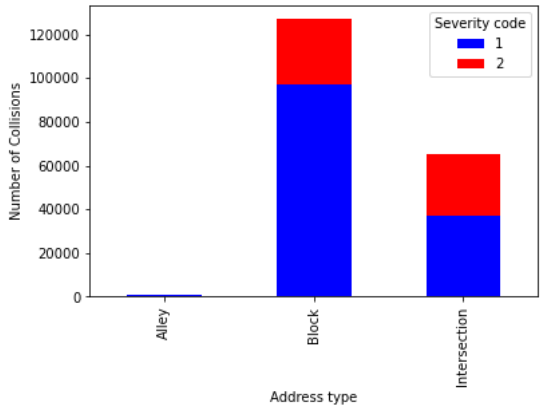

Figure 8: Number of collisions vs address type, i.e., type of junction

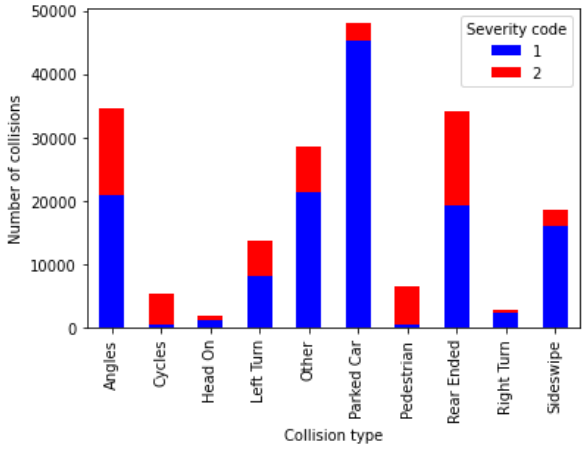

Figure 9: Number of collisions vs collision type

It is difficult to measure whether the driver was speeding or if pedestrian right of way was not granted, after an accident, unless there are witnesses or camera footage. Moreover, it is not likely that many drivers would admit to inattention after an accident, or realize that they were inattentive. The data on

speeding, pedestrian right of way not being granted and inattention are not very reliable for this reason. The data points on whether the driver was under the influence of alcohol or drugs is more reliable as it is generally measured when a traffic incident is investigated. The plots in Figure 10, 11, 12 and 13 show the number of collisions and severity of collisions for the cases of speeding, pedestrian right of way not being granted, inattention and driving under the influence, respectively. When these factors are true, it can be seen that there is a much higher proportion of collisions having severity code 2 and leading to injuries. In particular, most accidents when pedestrian right of way is not granted result in injuries.

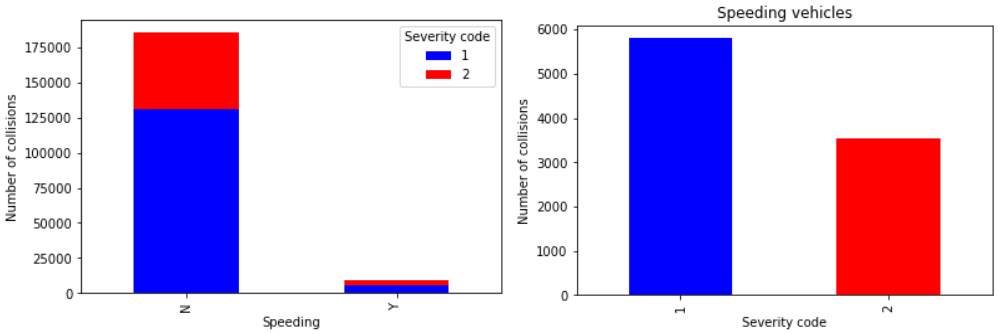

Figure 10: Number and severity of collisions with speeding

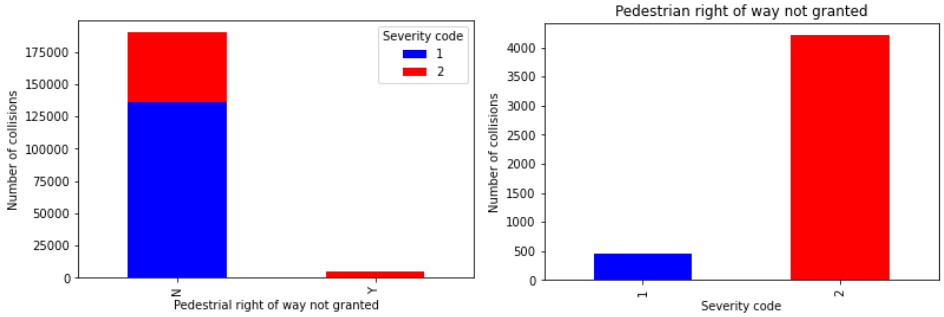

Figure 11: Number and severity of collisions when pedestrian right of way is not granted

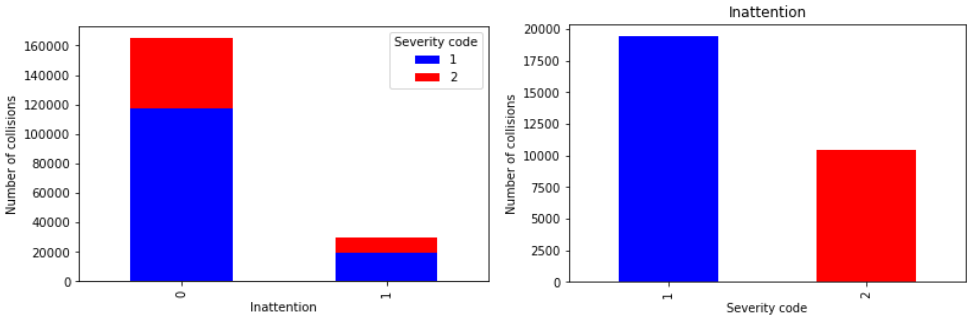

Figure 12: Number and severity of collisions when inattention is recorded

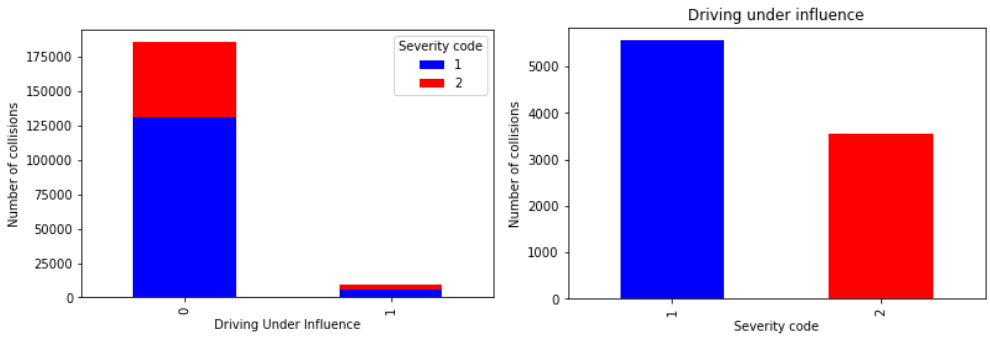

Figure 13: Number and severity of collisions when the driver is under the influence of alcohol

Based on this exploratory data analysis, the key features were identified for predictive modeling. In the next section, classification models are built to predict the severity of accidents.

4. Results

Based on the data analysis of the previous section, the following 15 independent variables were identified as the features for predictive modeling.

- MONTH
- DAYOFWEEK
- INATTENTIONIND
- UNDERINFL
- WEATHER
- ROADCOND
- LIGHTCOND
- SPEEDING
- PEDROWNOTGRNT
- ADDRTYPE
- COLLISIONTYPE
- PERSONCOUNT
- VEHCOUNT
- PEDCOUNT
- PEDCYLCOUNT

As mentioned previously, NaN values were dropped. Any “Unknown” or “Other” values are not useful for predictive modeling, and were dropped. The final data set contained 146,883 rows. The Pearson correlation coefficients were calculated for the features and it was confirmed that they can be treated as independent variables. The dependent variable is the severity code (SEVERITYCODE). The data set was split into a training set and a test set, with 20% of the data being used as the test set. The accuracy score, F1 scores and log loss are based on the results from the test set.

The K Nearest Neighbors (KNN), Logistic Regression, Decision Tree and Support Vector Classification methods were applied. The number of neighbors for the KNN method was varied between 1 and 9, and the value providing maximum accuracy was selected. This is illustrated in Figure 14. (For this binary

classification problem, accuracy and Jaccard similarity scores are the same, and accuracy scores are reported here.) Similarly, the maximum depth of the decision tree and the regularization parameter in logistic regression were varied. In the support vector method, only the linear kernel was used due to computational limitations (as the computation could not be completed with the rbf or sigmoid kernels on a laptop with the Intel Core i5-6300U 2.4 GHz CPU even after several hours).

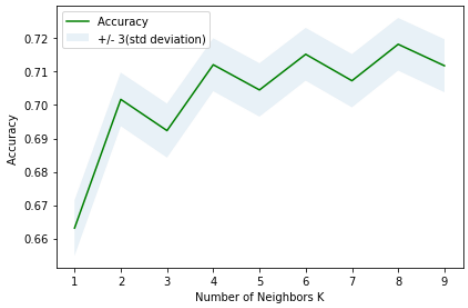

Figure 14: Number of neighbors vs accuracy in KNN model

The performance of the classification models is summarized in Table I. The models are able to predict severity code 1 collisions well, but do not perform well for predicting severity code 2 collisions.

Table I: Performance of classification models

Algorithm	Accuracy	F1-score, code 1	F1-score, code 2	LogLoss
KNN	0.72	0.81	0.47	
Decision Tree	0.74	0.83	0.44	
SVM	0.73	0.83	0.35	
Logistic Regression	0.73	0.82	0.41	0.55

The ROC curve showing the true positive rate vs the false positive rate is illustrated in Figure 15. The Decision Tree Classification model performs better than the remaining models tested. The KNN and Logistic Regression model follow. The SVC does not perform well with the linear kernel. Performance could potentially be improved with a different kernel, but could not be attempted at present due to lack of computational resources to run the computationally intensive model.

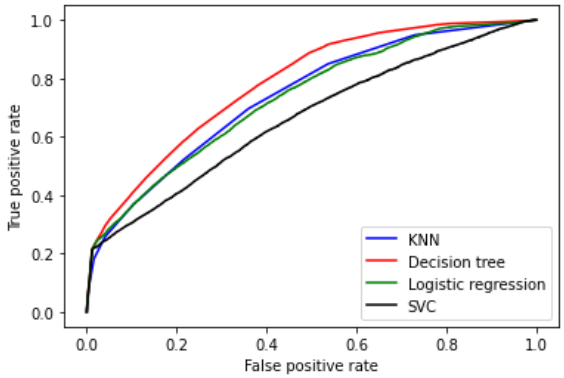

Figure 15: ROC curves for the classification models used

Severity code 2 collisions leading to injuries are not predicted well in this approach. This might be because the feature set does not fully capture all the relevant information. Injuries can be prevented in car accidents by wearing a seat belt, or if car manufacturers improve safety devices like airbags. Newer and high-end cars may also provide features such as blind spot warnings and automatic braking. Since car manufacturers are expected to provide improved safety features over time, the number of collisions is expected to decline over the years. This is illustrated in Figure 16. On the whole, there is a decrease in the number of collisions over the years. There is a spike in 2014-2016, which could have been due to improved recording of accidents, but this could not be confirmed. The year was not included in the predictive modeling, as it is not a predictive feature in itself. For severity code 2 collisions, no clear decrease was observed over the years (as shown in the Python notebook). Wearing a seat belt may be an important feature in preventing severity code 2 collisions resulting in injuries. More data are needed on whether seat belts were in use, and the safety features available in the involved cars.

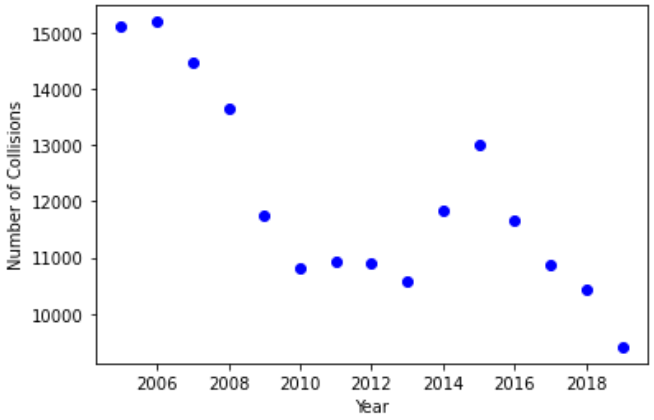

Figure 16: Number of collisions vs year

5. Discussion

In this report, car accidents in Seattle over the last 15 years were analyzed, and predictive models were built to predict the severity of car accidents. The data revealed that factors like driver fatigue and distracted driving play a bigger role than external factors like road conditions, weather or light conditions.

A supervised learning approach based on KNN, Logistic Regression, Decision Tree and Support Vector Classification methods was adopted. Accidents leading to property damage could be predicted well. However, accidents leading to injuries could not be predicted well, and the models showed a low recall, i.e., a large number of false negatives. More advanced machine learning models can be attempted in future with a lower threshold for injury probability, to better predict injuries. It is also possible that factors like driver inattention, speeding, and pedestrian right of way not being granted, were not reliably recorded, making it difficult to predict injuries. Potentially, more data may be needed to build improved models; for example, whether seat belts were in use, and the ranking of safety features available in the involved cars.

6. Conclusions

In this report, traffic collisions in Seattle are analyzed, and predictive models are built based on supervised learning methods. One potential future direction is to build more advanced models based on deep neural networks to predict accident severity. Setting a lower threshold for the probability of injuries may reduce the number of false negatives in injury prediction. However, this may reduce the precision of the model. Another direction is to improve data collection. If factors like driver inattention, speeding, and pedestrian right of way not being granted, are more reliably recorded, predictions may be improved. In addition, collecting data on whether seat belts were in use, and the ranking of safety features available in the involved cars, may also assist in improving the injury prediction rate.