Coursera

IBM Professional Data Science Certificate Capstone Project Report

Factors Effecting Accident Severity in the City of Seattle, WA

Chad Hackert

Oct 2020

https://github.com/CHackert/Coursera_Capstone

Introduction / Business Problem

The US National Safety Council (NSC) has estimated that in the first six months of 2020 the total number of US motor-vehicle fatalities reached 18,300, motor-vehicle related injuries reached 2,086,000 and property damage was over $206 billion USD (ref1). Helping drivers and pedestrians be more informed of the most likely factors leading to an accident before travelling should substantially reduce these numbers and provide for safer and less stressful motor-vehicle travel experience. In addition to the travelling public City Government can use the results of this project to identify and correct high accident controllable infrastructure factors and impose laws directly related to traffic safety. Using data collected by the Seattle Police Department (ref2), this project aims to provide useful information regarding varying conditions and their effects on accident severity in the City of Seattle.

Data

Data Overview

The traffic accident data to be used for this project was collected by the Seattle Police Department (SDP) during routine accident reports and made publicly available by the Department of Transport (DOT) (ref2). This data set covers accidents in the City of Seattle from January, 2004 to May, 2020 and includes 194673 individual reports. The data collected by the SPD falls into 38 features (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, SDOTCOLNUN, SPEEDING, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR) which have either administrative or accident related details. Using the accident detail data (ie. Severity, Location, Weather, more) this project will attempt to predict the most common factors effecting the severity of collisions.

Data Cleaning and Preparation

While the data set is large and contains volumes of useful features not all of these features are causal to the accidents or relate to the ultimate severity. In order to examine the factors leading to severe accidents features that are administrative in nature or redundant were removed. Of the 190000+ reports many were missing data related to the severity of the accident reported on. Driver inattention, pedestrians not granted right of way and speeding data were incomplete in that the reports only indicated a positive result when completed. The negative results were added to the dataset using a replace function and assigning a 1 or 0 value. In addition, one column (“INTKEY”) which was a numerical value assigned to specific intersections in the city. This column was removed as over 2/3rds of the records did not have information included. After this, any rows in the spreadsheet which were missing data that could not be recovered were dropped. This resulted in 10527 records being removed from the file constituting a 5.4% loss of data.

Methodology

Data Exploration

An initial exploration of the data was completed. First, using map data to overview and visualize the occurrences of all accidents in the City of Seattle. Second, histograms, to determine the number of accidents which occurred given a specific feature which were grouped by the severity of the accident (injury accident vs property damage only accident).

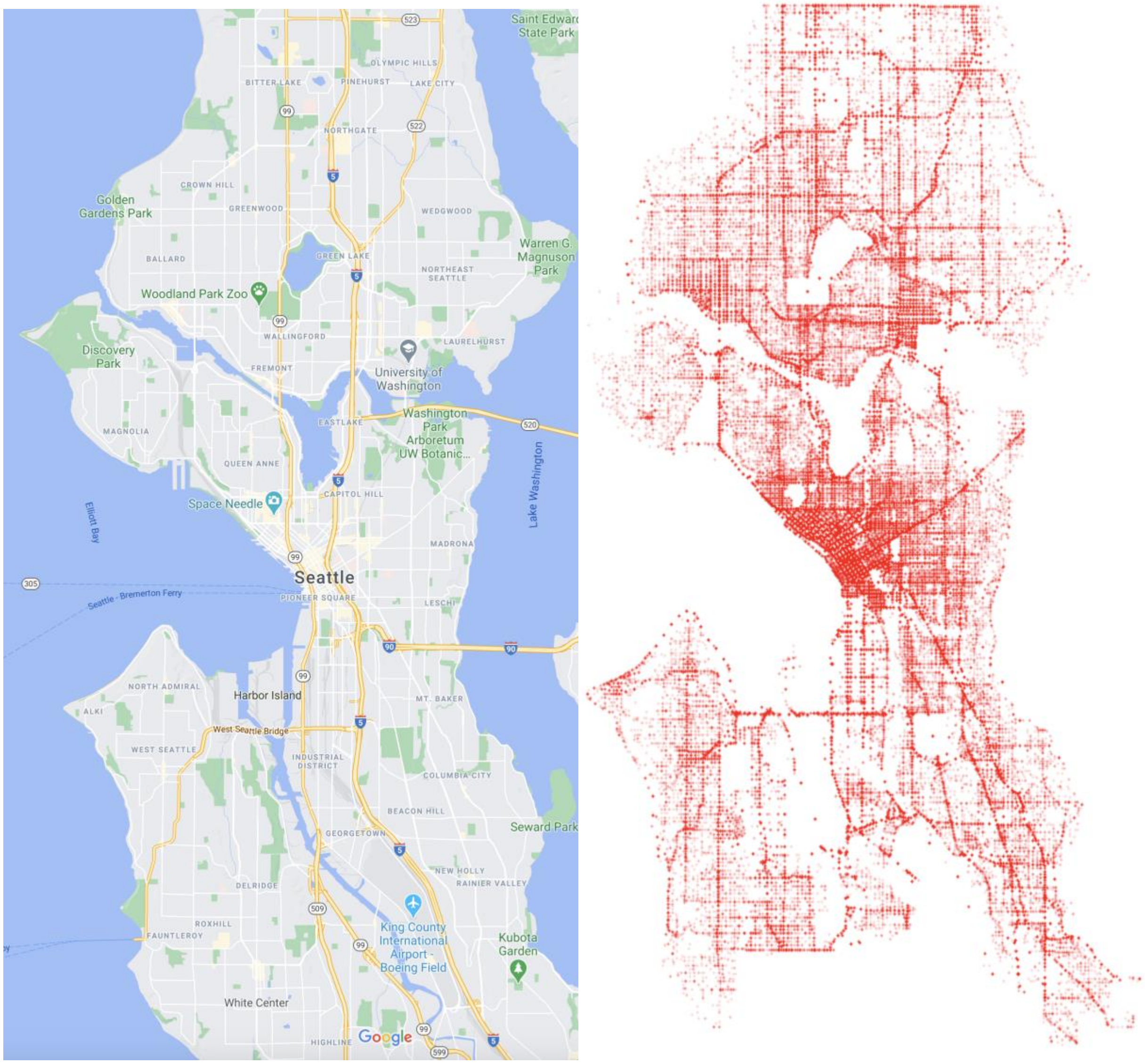


Fig 1. Comparison of the roadmap of the City of Seattle from Google Maps and an accident only location map generated from the project data.

The results from the visual comparison of the road map and accident location map seem to align with a reasonable hypothesis that the majority of the accidents would occur where the majority of cars congregate. The location of accidents fits extremely well with the city centre and major roadways and inversely to the City's parks, waterways and low traffic features such as King's County International Airport.

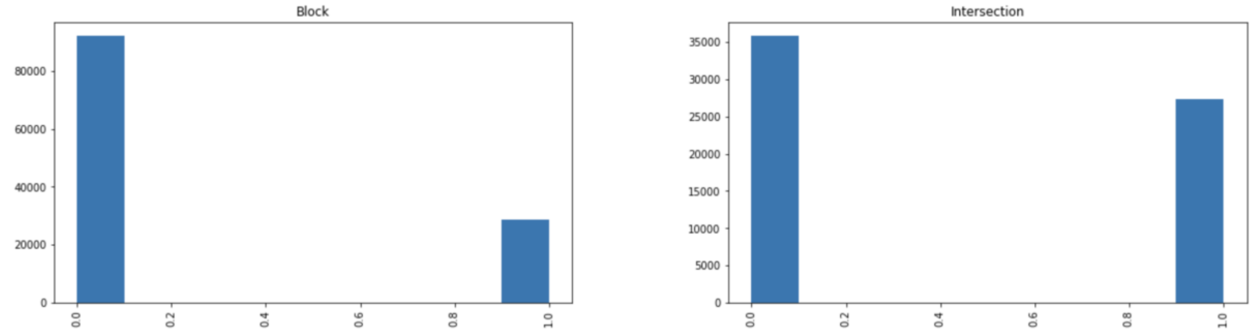


Fig 2. Histograms of the number of injury accidents vs the number of property damage only accidents. Example ('SEVERITYCODE') grouped by their location ('ADDRTYPE').

The results of the histograms were interpreted with respect to their total count and the relation of injury to property damage only accidents to determine which features were to be selected for modeling. These features were then compared to information published by the US DOT (ref3) regarding most common causes for motor-vehicle accidents and filtered down to eight features. These are Address Type (intersection or on street), Collision Type, Driver Inattention, Impaired Driving, Weather, Road Conditions, Light Conditions and Speeding.

Modeling

As the intended outcomes of modeling are expected to predict the severity of an accident (1 = Injury, 0 = Property Damage) algorithms employed must be suitable to classify the results into a discrete set of classes. As such predictive modeling was completed using K-Nearest Neighbour, Decision Tree and Logistic Regression algorithms. Support Vector Machine was not employed given its row limitation of approximately 1000 rows. The SPD collected data has almost 200000 entries which makes SVM inappropriate for this application.

The K-Nearest Neighbour algorithm was applied using a range of k values from 1 to 9 with varying degrees of accuracy. The best accuracy was found with a K neighbour value of 4 and accuracy of 0.723.

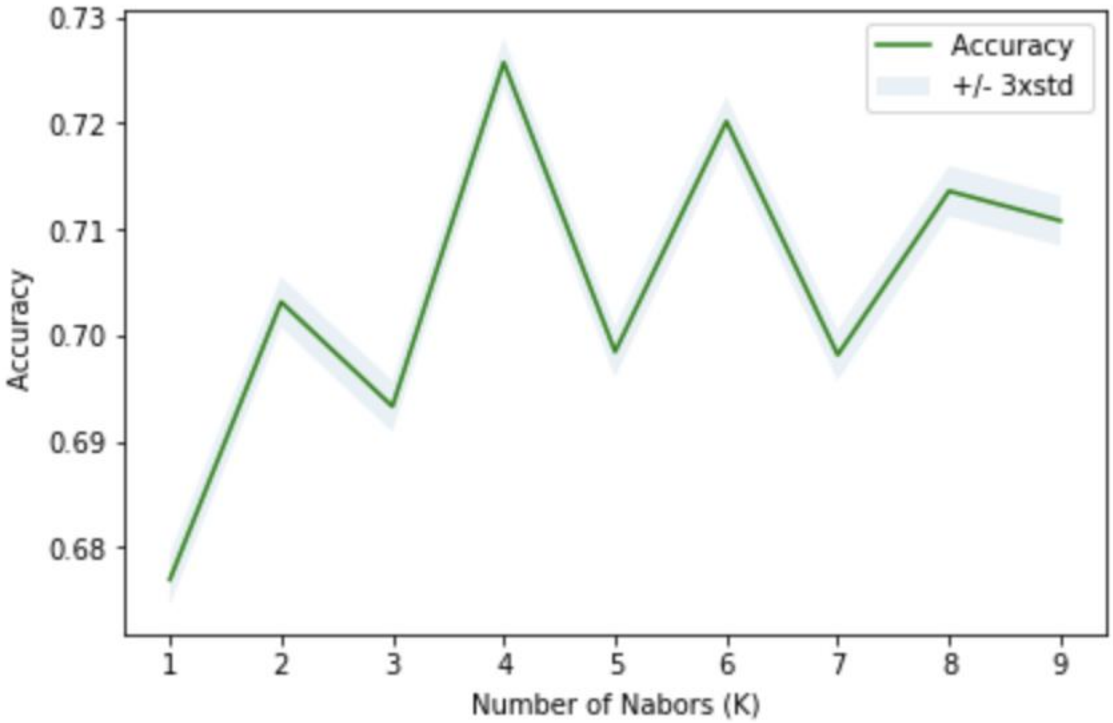
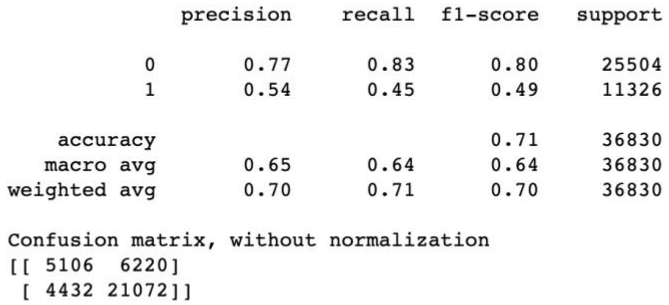
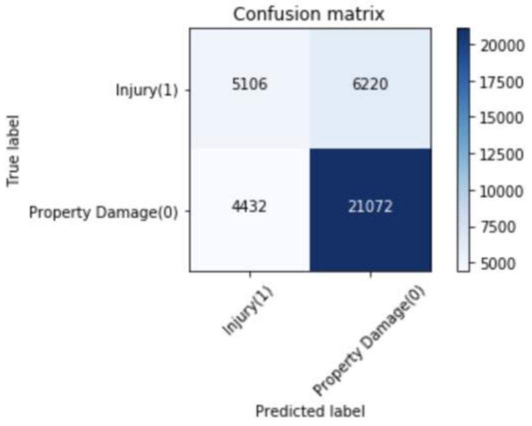


Fig 3. Plot of Accuracy from a range of k values from 1 to 9.

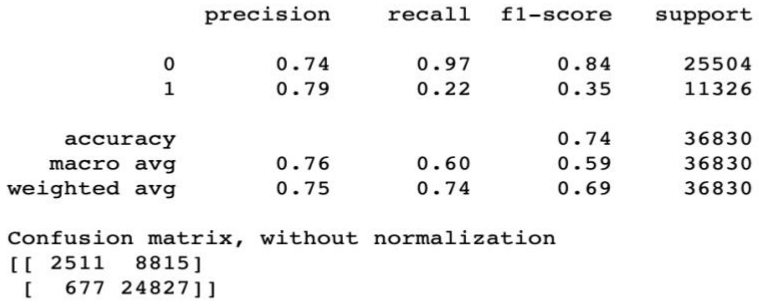
The calculated precision, recall, F1 score and confusion matrix were produced as below:



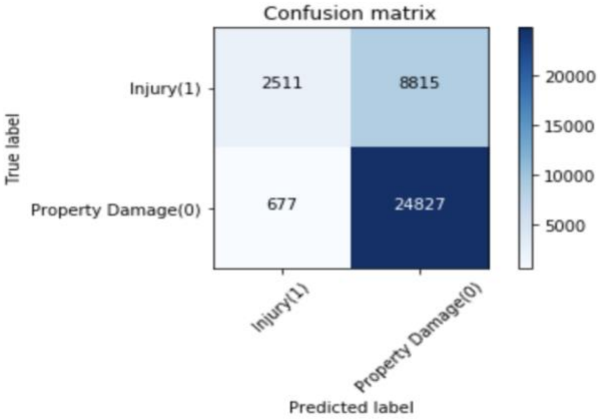
```
              precision    recall  f1-score   support

           0       0.77      0.83      0.80     25504
           1       0.54      0.45      0.49     11326

    accuracy                           0.71     36830
   macro avg       0.65      0.64      0.64     36830
weighted avg       0.70      0.71      0.70     36830

Confusion matrix, without normalization
[[ 5106  6220]
 [ 4432 21072]]
```



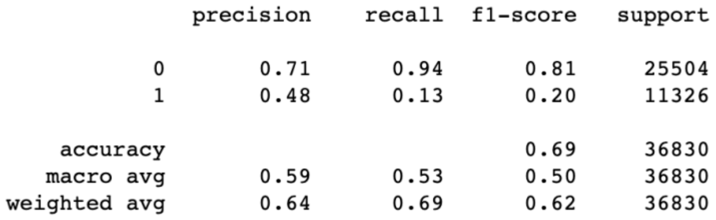
The Decision Tree classification algorithm was also applied to the accident severity data. The calculated precision, recall, F1 score and confusion matrix were produced as below:



```
              precision    recall  f1-score   support

           0       0.74      0.97      0.84     25504
           1       0.79      0.22      0.35     11326

    accuracy                           0.74     36830
   macro avg       0.76      0.60      0.59     36830
weighted avg       0.75      0.74      0.69     36830

Confusion matrix, without normalization
[[ 2511  8815]
 [  677 24827]]
```



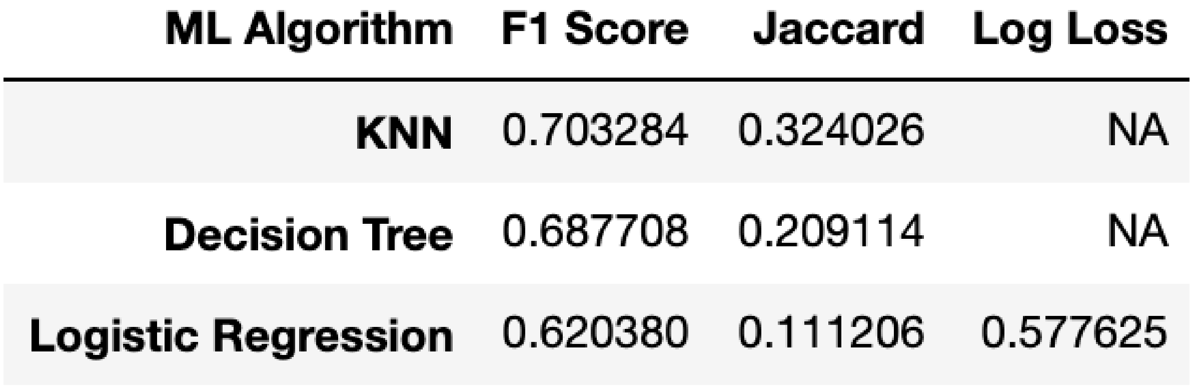
Finally, the Logistic Regression algorithm was applied to the accident severity data. The calculated precision, recall, F1 score were produced as below:



```
              precision    recall  f1-score   support

           0       0.71      0.94      0.81     25504
           1       0.48      0.13      0.20     11326

    accuracy                           0.69     36830
   macro avg       0.59      0.53      0.50     36830
weighted avg       0.64      0.69      0.62     36830
```

Results

Post modeling evaluation was performed on the three classification algorithms finding F1, Jaccard values for the K-Nearest Neighbour, Decision Tree and Logistic Regression models with the addition of a Log Loss evaluation to the Logistic Regression model. The end results are summarized below:



ML Algorithm	F1 Score	Jaccard	Log Loss
KNN	0.703284	0.324026	NA
Decision Tree	0.687708	0.209114	NA
Logistic Regression	0.620380	0.111206	0.577625

The model with the best accuracy was the Decision Tree model with approximately 74%, followed by K-Nearest Neighbour with 71% and finally, Logistic Regression at approximately 69%.

Discussion

After exploring the data and applying three classification prediction models and the result generated therein, three important conclusions regarding accident severity can be drawn.

1) The majority of both Injury and Property Damage motor-vehicle accidents occur in the areas where there is greater vehicle traffic. The City of Seattle core and the surrounding major roadways account for the majority of accidents.
2) The majority of severe (Injury) motor-vehicle accident occur during the daytime under dry road conditions and clear visibility.
3) Predictive modeling based on K-Nearest Neighbour, Decision Tree and Logistic Regression algorithms yielded three useful models.

After considering the above conclusions some recommendations can be made. First further work on this data set should be undertaken to better understand where and why the majority of severe accidents occur in the best of weather and road conditions. The author would suggest that a month, day of week and time of data analysis be undertaken to see if the majority of the accidents occur at specific date/time intervals such as Friday or the rush hour time period.

Conclusion

In an effort to provide better information to the driver and authorities in the City of Seattle this project was undertaken to analyze data collected by the SPD for motor vehicle accidents.

References

ref1: US National Safety Council - https://injuryfacts.nsc.org/motor-vehicle/overview/preliminary-estimates/

ref2: Seattle MV Accident Data - https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

ref3: DOT National Motor Vehicle Crash Causation Survey - https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059