Capstone Project – Car Accident Severity

Ioannis Dimopoulos

August 2020

1. Introduction

Imagine the scenario where you are travelling by car in order to visit friends or family in another city. It is windy and rainy and on the other side of the highway there is a terrible traffic jam. As your journey goes on, you see police cars shouting down your side of the high way as well. You realise that it is a serious accident as you see those involved in the accident being taken to a helicopter, possibly to be transported to the nearest hospital.

Wouldn't it be great if it was possible to have a warning about the severity of a possible car accident, based on external conditions like, weather, light and street conditions? This is aim of this project, using data provided by the city of Seattle.

2. Data

Data source: https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data

The dataset consisted of 220,937 records and 40 columns. The initial available variables were the following:

1. X
2. Y
3. OBJECTID
4. INCKEY
5. COLDETKEY
6. REPORTNO
7. STATUS

8. ADDRTYPE
9. INTKEY
10. LOCATION
11. EXCEPTRSNCODE
12. EXCEPTRSNDESC
13. SEVERITYCODE
14. SEVERITYDESC
15. COLLISIONTYPE
16. PERSONCOUNT
17. PEDCOUNT
18. PEDCYLCOUNT
19. VEHCOUNT
20. INJURIES
21. SERIOUSINJURIES
22. FATALITIES
23. INCDATE
24. INCDTTM
25. JUNCTIONTYPE
26. SDOT_COLCODE
27. SDOT_COLDESC
28. INATTENTIONIND
29. UNDERINFL
30. WEATHER
31. ROADCOND
32. LIGHTCOND
33. PEDROWNOTGRNT
34. SDOTCOLNUM
35. SPEEDING
36. ST_COLCODE
37. ST_COLDESC
38. SEGLANEKEY
39. CROSSWALKKEY

40. HITPARKEDCAR

The description of these variables can be find in the following link: https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

The dataset had a lot of missing values and the machine learning algorithms that were deployed for this project required a full dataset, with no missing values. First, I checked the number of missing values of each variable and I rejected the variables were over half of the records were had missing values. The variables rejected with this process were INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, INATTENTIONIND, PEDROWNOTGRNT and SPEEDING.

Next, I checked the remaining variables and selected only those that seemed to be relevant to the question of the project. First of all, I kept SEVERITYCODE which is the target variable of the project, as this is the variable we want to be ale to predict, based on the machine learning models developed during this project. The other variable I kept in the dataset to be used as feature variables were WEATHER, ROADCOND and LIGHTCOND.

Afterwards, I removed any remaining record with missing values, bringing the dataset to 194,209 records. Next, I sorted the remaining records based on each of the remaining variables to check for any records that didn't make sense or added any useful information to the analysis. Based on this, I removed a record with had SEVERITYCODE "0" which stood for "unknown severity", as well as any records where WEATHER, ROADCOND or LIGHTCOND was equal to "Unknown". This brought the final dataset to 175,292 records and 4 columns.

3. Methodology

3.1 Feature Evaluation

Before developing the machine learning models, I had to evaluate the association between the target variable (SEVERITYCODE) with each of the candidate feature variables. Because all the variables in the final dataset were categorical, bar plots were the most appropriate way

of visual presentation of the association between the variables and chi-squared test was the most appropriate statistical test to evaluate these associations.

3.2. Machine Learning Algorithms

The categorical nature of the variables included in the final dataset dictates the use of classification machine learning algorithms in order to pursue the aim of this project. I used three different algorithms and I checked evaluation metrics to decide which algorithm makes the best predictions. I used K Nearest Neighbours (KNN), Decision Tree and Logistic Regression. For KNN I check which number of K produced the best results; I check for Ks ranging from 1 to 50. For KNN and Decision Tree I the jaccard and F-1 metrics, while for the Logistic Regression algorithm I also checked LogLoss.

4. Results

4.1 Feature Evaluation

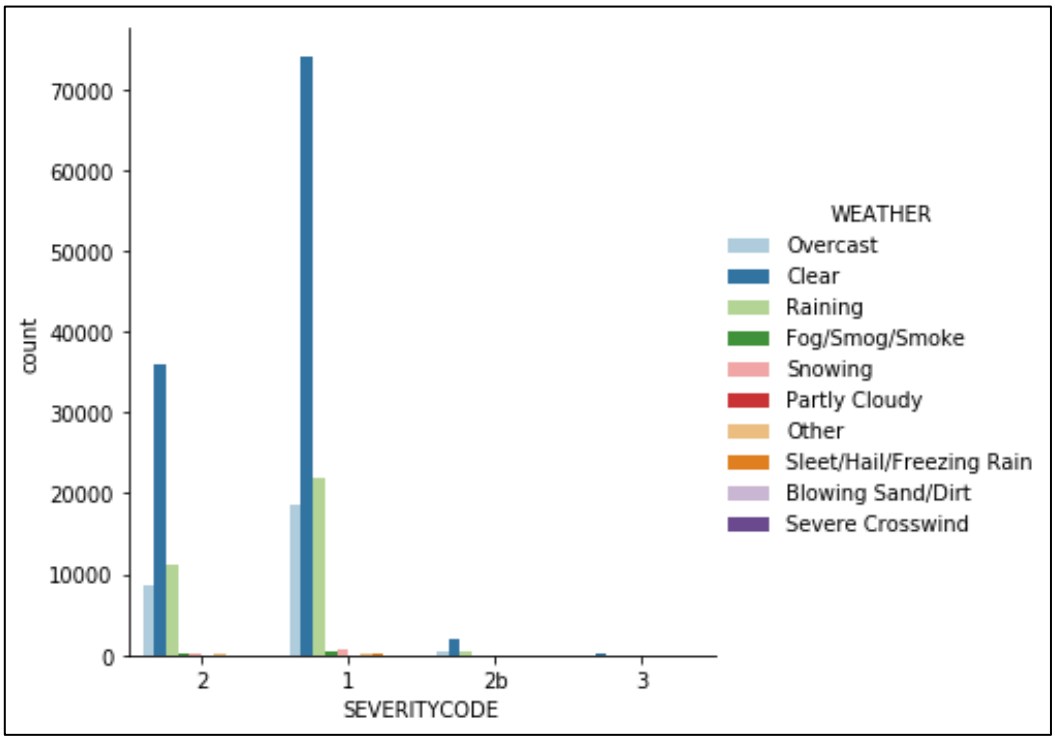

Figure 1. Bar plot of the distribution of accidents of different level of severity across different weather conditions.

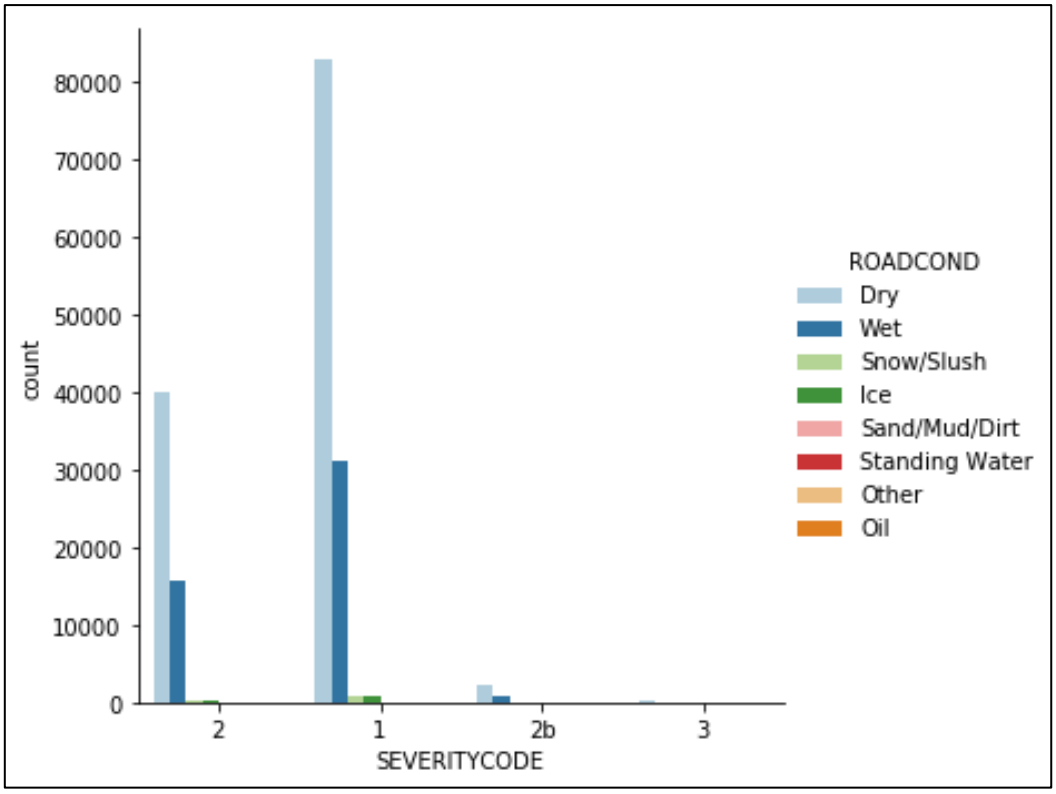

Figure 2. Bar plot of the distribution of accidents of different level of severity across different road conditions.

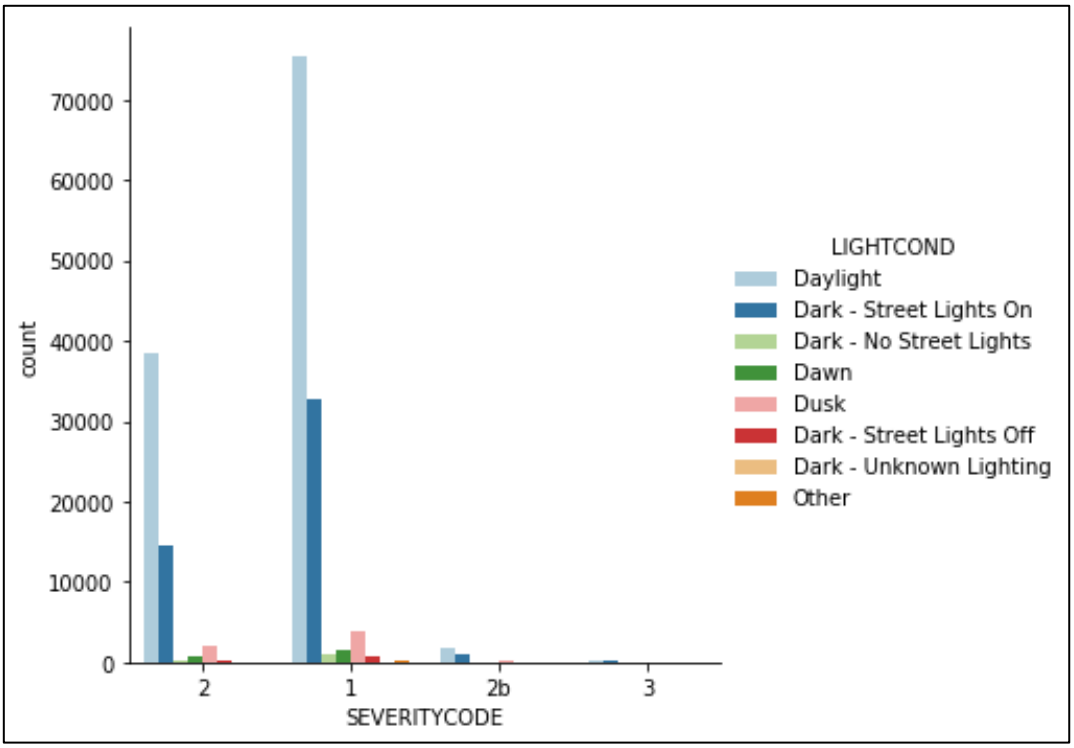

Figure 3. Bar plot of the distribution of accidents of different level of severity across different light conditions.

The three bar plot indicate quantitive and qualitative differences of the distribution of accidents of different severity across different weather, road and light conditions.

The three chi-squared tests I ran to evaluate the association between SEVERITYCODE and WEATHER, ROADCOND or LIGHTCOND produced p values smaller than 0.05, therefore all three variables were included as feature variables for the development of the machine learning algorithms.

4.2. Machine Learning Algorithms

First of all, I had to determine the best number of K for the KNN model. Out of the 50 KNN models, the one with K=47 produced the best accuracy which was 0.66% (Figure 4).

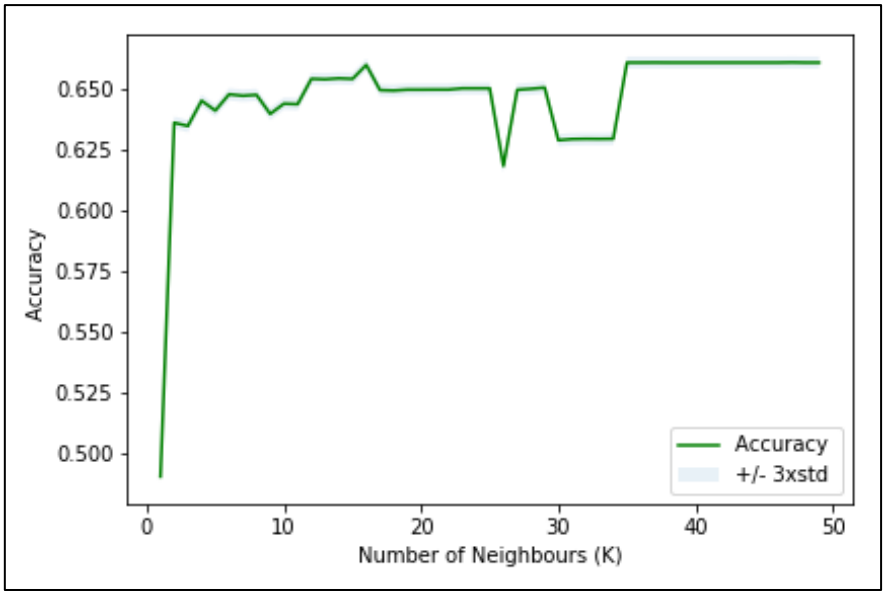

Figure 4. Line plot of the accuracy of the KNN models across different number of K.

The next table summarises the results of the accuracy metrics for the three different machine learning algorithms.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.661342	0.526620	N/A
Decision Tree	0.661086	0.526724	N/A
Logistic Regression	0.661257	0.526421	0.720668

Table 1. Summary of the accuracy metrics for the three different algorithms used in this project. The highest scores are indicted with the use of red font.

KNN produced the highest Jaccard score and while Decision Tree produced the highest F1-score.

5. Discussion

The three algorithms produced very similar metrics; specifically, they were equal up to the third decimal. Although Logistic Regression didn't produce the highest scores, all metric results were very similar. Additionally, Logistic Regression has the unique feature, compared to other algorithms, that apart from predicting the state of the target variable for any unlabelled record, it also produced an estimate of the possibility of that prediction. Based on these two facts, I believe the best model for predicting the possibility of an accident based on weather, light and road conditions is the one produced using the Logistic Regression algorithm.

6. Conclusion

The aim of this project was to create a model to be used for the prediction of the severity of car accidents based on external conditions. I checked the association between severity of the accident with weather, light and road conditions using chi-squared test and it was significant in all three cases. I also developed three machine learning algorithm models for the prediction of the severity of an accident based on these three external conditions, using the KNN,

Decision Tree and Logistic Regression algorithms. All three models produced similar results. However, it is my opinion that the Logistic Regression model is the more appropriate one to use for the aim of the project, as apart from predicting the severity level of future accident, it also provides with the possibility of that prediction.