Coursera Capstone Project - Car Accident Severity

Ioannis Dimopoulos

August 2020

Introduction & Data Source

- The aim of the project is to develop a model able to predict the possibility of a car accident and its severity based on external conditions.

- The data used for this project were provided by the city of Seattle:
  - Source: https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data
  - Description: https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

Data Cleaning

- Original dataset: 40 variables, 220,937 records.
- First, I removed the variables, where over half of the records had missing (NaN) values.
- Next, I kept only variables relevant to the project. The were:
  - Target variable: SEVERITYCODE
  - Candidate feature variables: WEATHER, ROADCOND, LIGHTCOND
- Afterwards, I removed any remaining records with NaN and the records where SEVERITYCODE was “0” and WEATHER, ROADCOND or LIGHTCOND were “Unknown”
- Final dataset: 4 variables, 175,292 recotds.

Methodology

- Feature Evaluation:
  - Visual Evaluation: Bar plots
  - Statistical Evaluation: Chi-squared test
- Machine Learning Algorithms:
  - Three Classification Machine Learning Algorithms:
    - K Nearest Neighbours (KNN) (tested Ks from 1 to 50 to find the best K)
    - Decision Tree
    - Logistic Regression

Results – Bar Plots

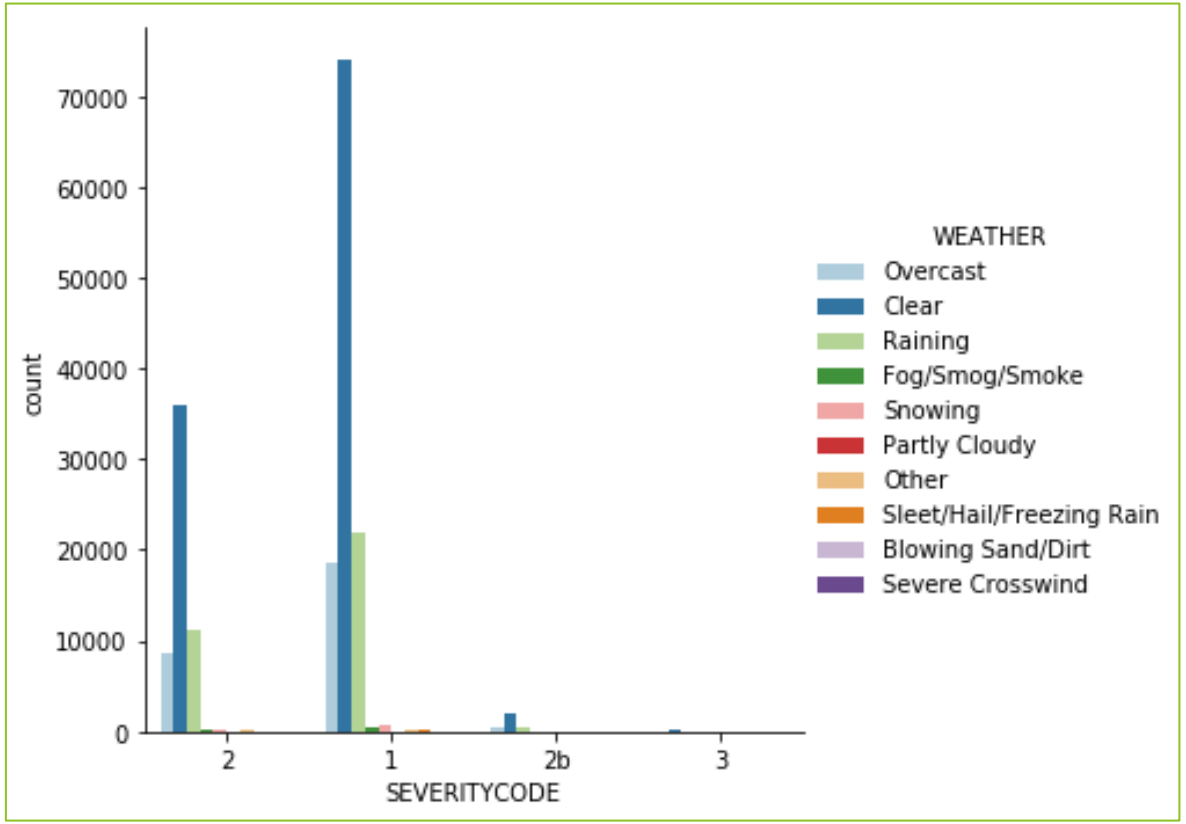

Figure 1. Distribution of accidents of different level of severity across different weather conditions

Results – Bar Plots

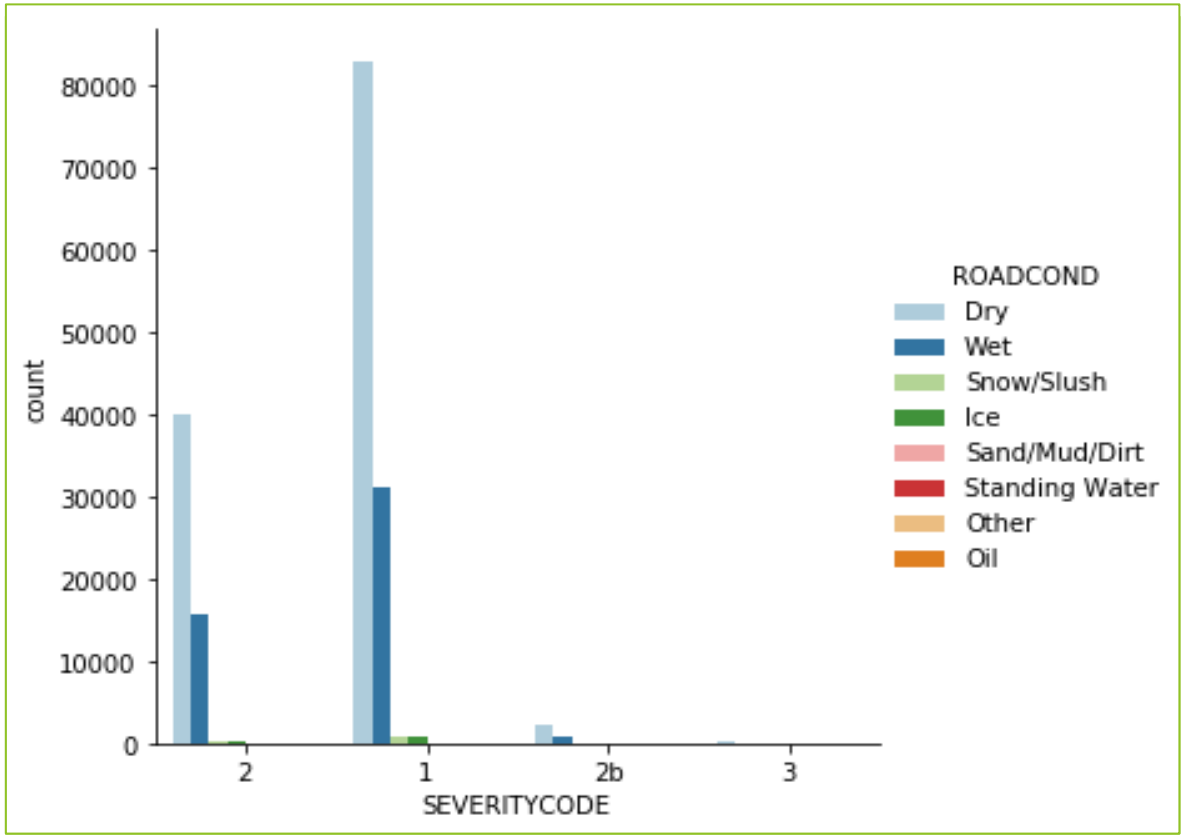

Figure 2. Distribution of accidents of different level of severity across different road conditions.

Results – Bar Plots

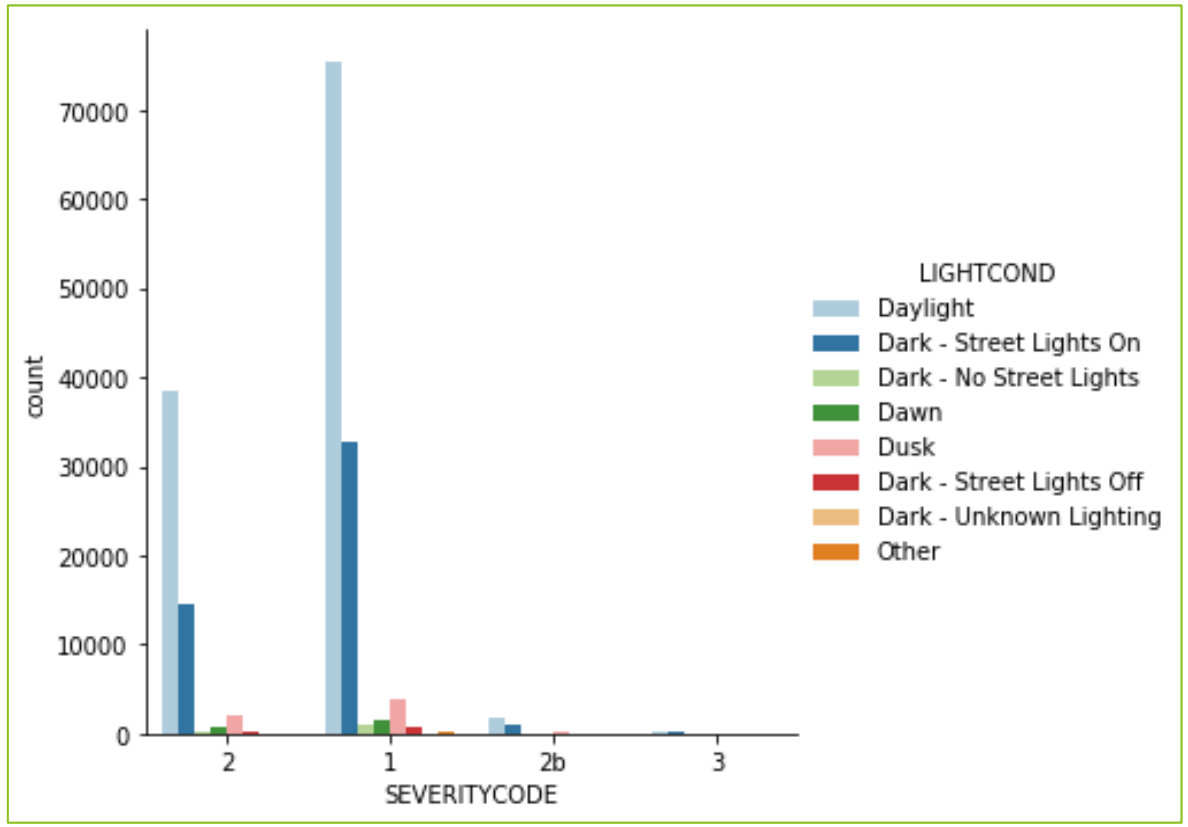

Figure 3. Distribution of accidents of different level of severity across different light conditions.

Results – Chi-squared Test

- The chi-squared test of association between SEVERITYCODE and WEATHER, ROADCOND, LIGHTCOND produced p-values < 0.5
  - → All three associations were significant
  - → All three candidate feature variables were included in the machine learning algorithms

Results – Best K for KNN

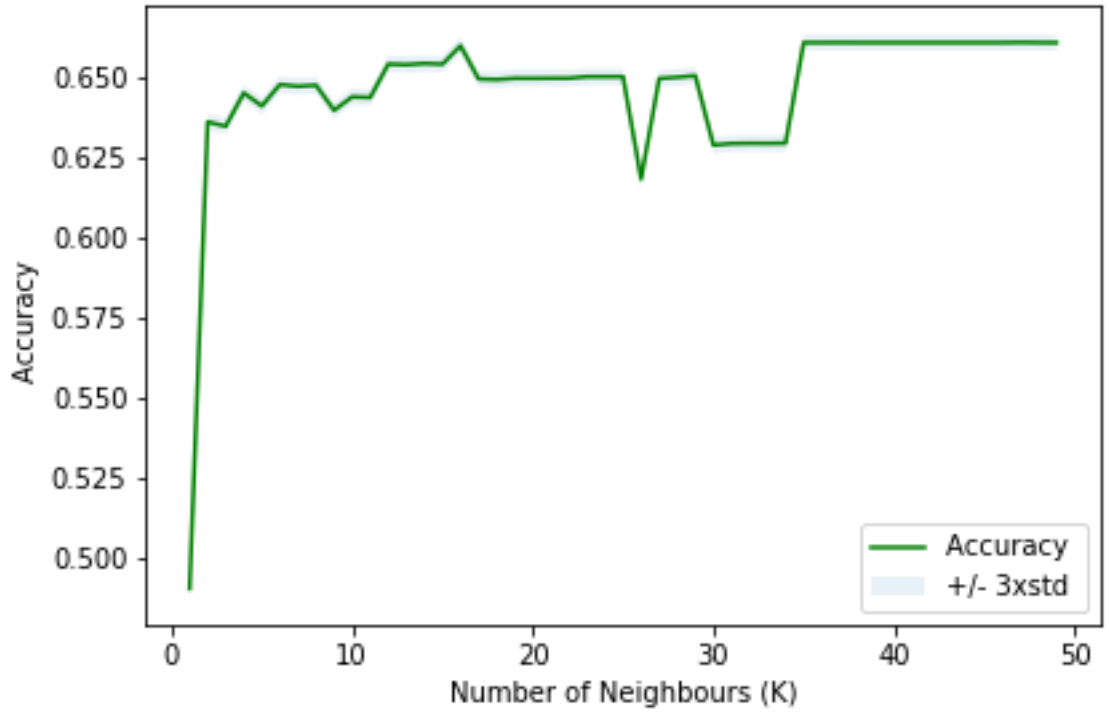

Figure 4. The K that produced the best accuracy (66%) was 47.

Results – Machine Learning Accuracy Metrics

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

Discussion

- The three algorithms produced very similar metrics.
- Logistic Regression produced an estimate of the possibility of the accident prediction.
- Therefore the Logistic Regression model is more informative and more appropriate to meet the aim of this project.