Predicting the Severity of Road Accidents

Coursera and IBM Data Science Professional Certificate Capstone Project

By Liis Monson| Sept 2020

The immediate aftermath and the emergency services’ reaction to an accident are crucial elements to deal with the accidents the most efficiently, and therefore reduce the amount of injuries, complications, impact, cost, and save lives.

Can we predict the accident severity rate with relatively high probability using environmental features of an accident and personal characteristic of the travellers?

The Data

Original Data

- UK Government Data for Leeds traffic accidents in 2018
- 1995 rows, 21 features
- Includes environmental and personal features in numeric format
- Has more years available for out-of-sample testing and improving the modelling data set size

Data Pre-Processing & Exploration

Feature Exploring

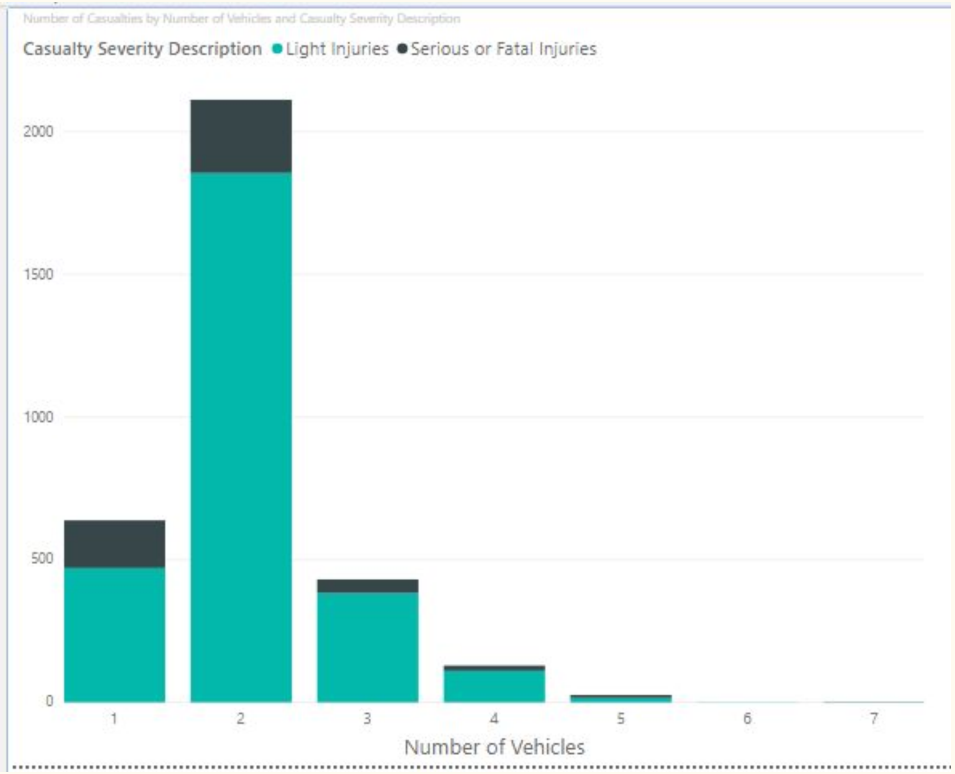
- All features were divided into 3 groups: Severity Assessment, Personal Features, Environmental Features
- Each feature was processed for clarity, formatted into numeric data & analysed in context of Casualty Severity

Casualty Data

- A total Casualty Severity (CS) and Number of Vehicles (NV) in accident were used to calculate a Severity Score (SS) for all accidents
- The formula used was:

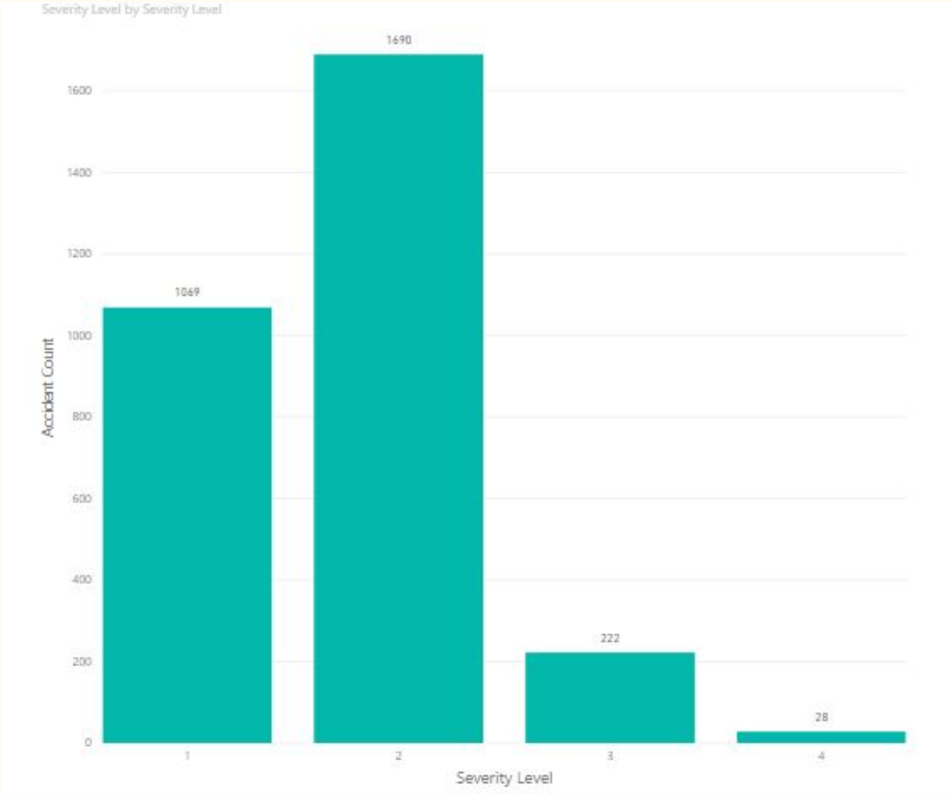
SS = NV + CS1 + CS2 +...+CSn

Where:
CS is a Casualty Severity of each person injured in the accident
n is a total number of casualties



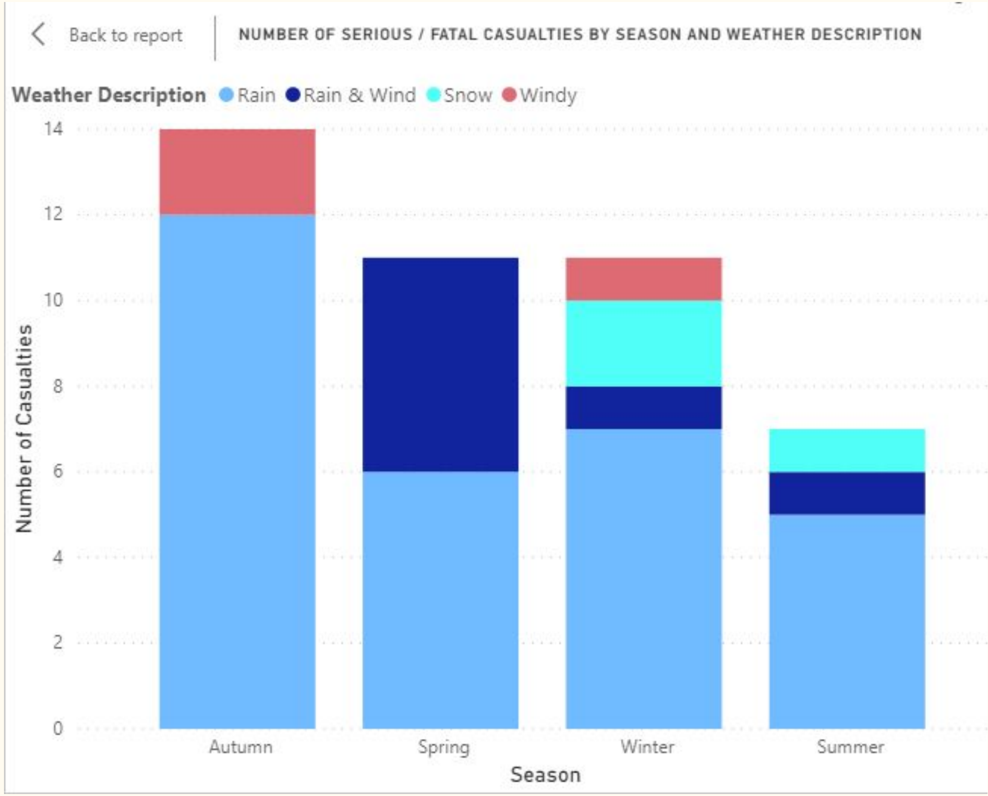
Severity Assessment

- Each accident was assigned Severity Level dependent on Severity Score
- There are 7 levels in total from Level 0 (incident) to Level 6 (potential disaster situation)
- All levels are related to each other exponentially. The data set had min Level 1 and max Level 4



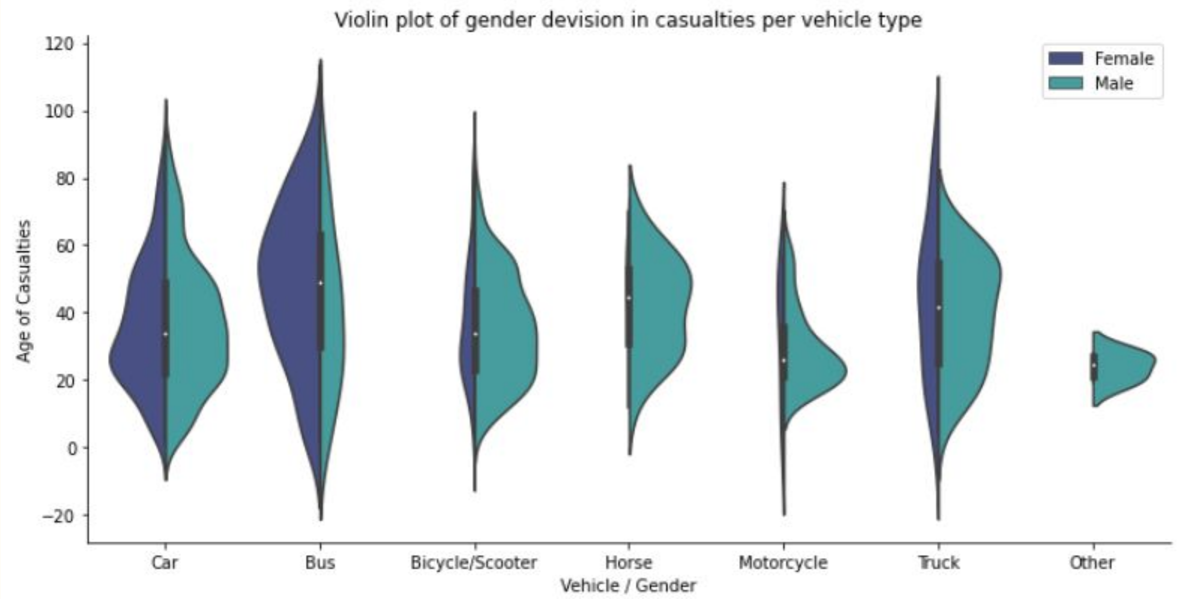
Environmental Features

- The following environmental features were included. These all affected the casualty severity and number of accidents in some way:
    - Seasonality
    - Time of Day
    - Lighting Conditions
    - Weather
    - Road Surface Conditions



Personal Features

- The following travellers' features were taken into account for modelling:
  - Gender
  - Age
  - Choice of vehicle



Modelling

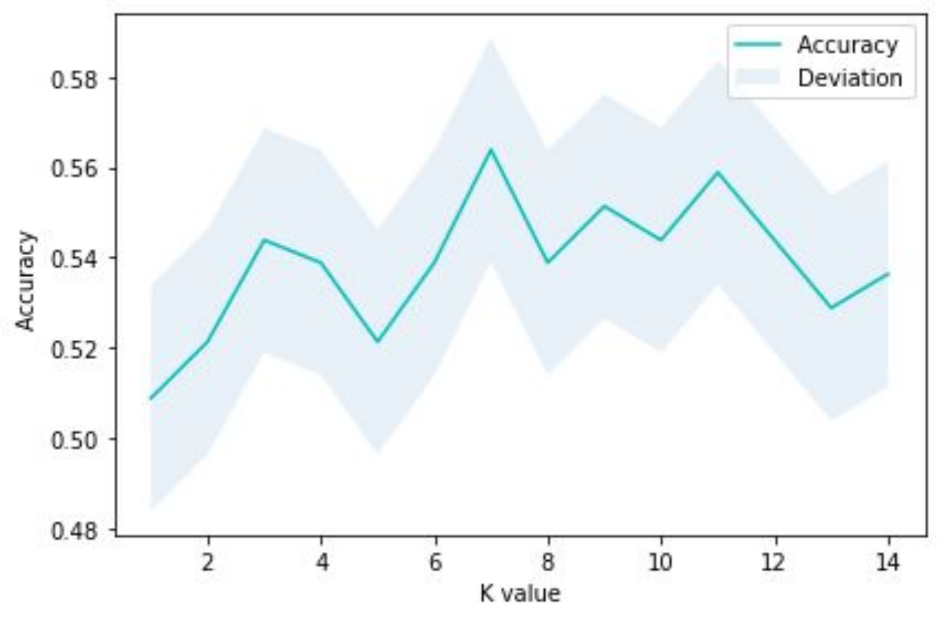
Choice of ML Models

- Classification models
- Models that are designed to predict more than 2 classes
- KNN and Decision Tree selected

K-Nearest Neighbor

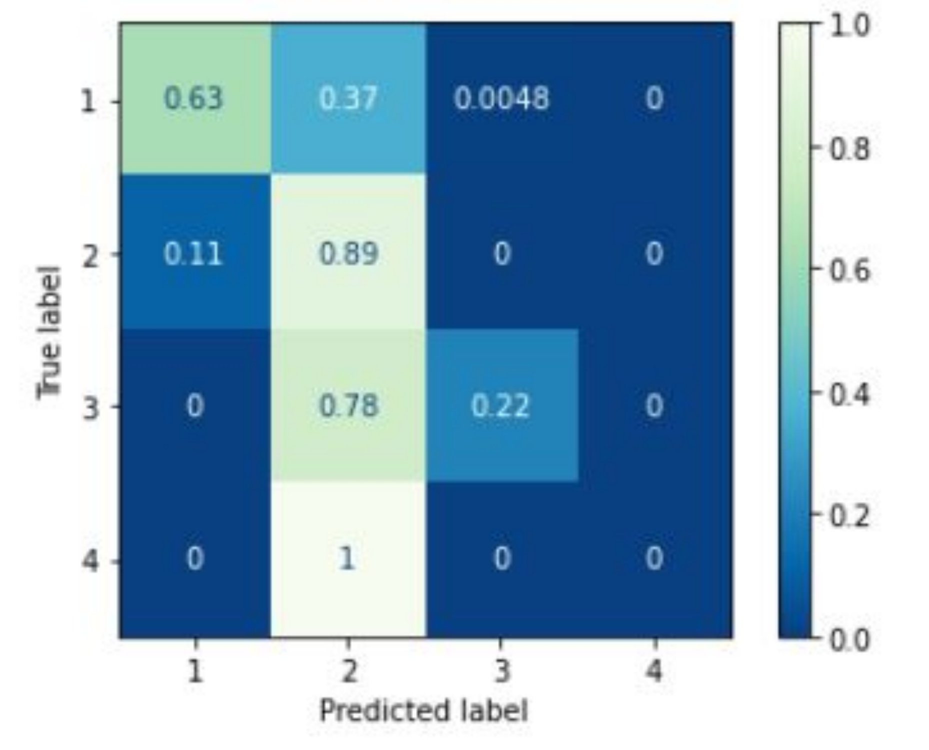
- Training/test 80/20 split was found to work the best
- The best K==7
- Accuracy Score: 57%
- F1 Score: 56%
- Jaccard Score: 57%

Verdict: not awesome results



Decision Tree

- Training/test 80/20 split was found to work the best
- Accuracy Score: 74%
- F1 Score: 73%
- Jaccard Score: 74%
- Confusion Matrix suggests that higher level classes are classified lower because the small amount of data in higher levels (model considers them outliers)



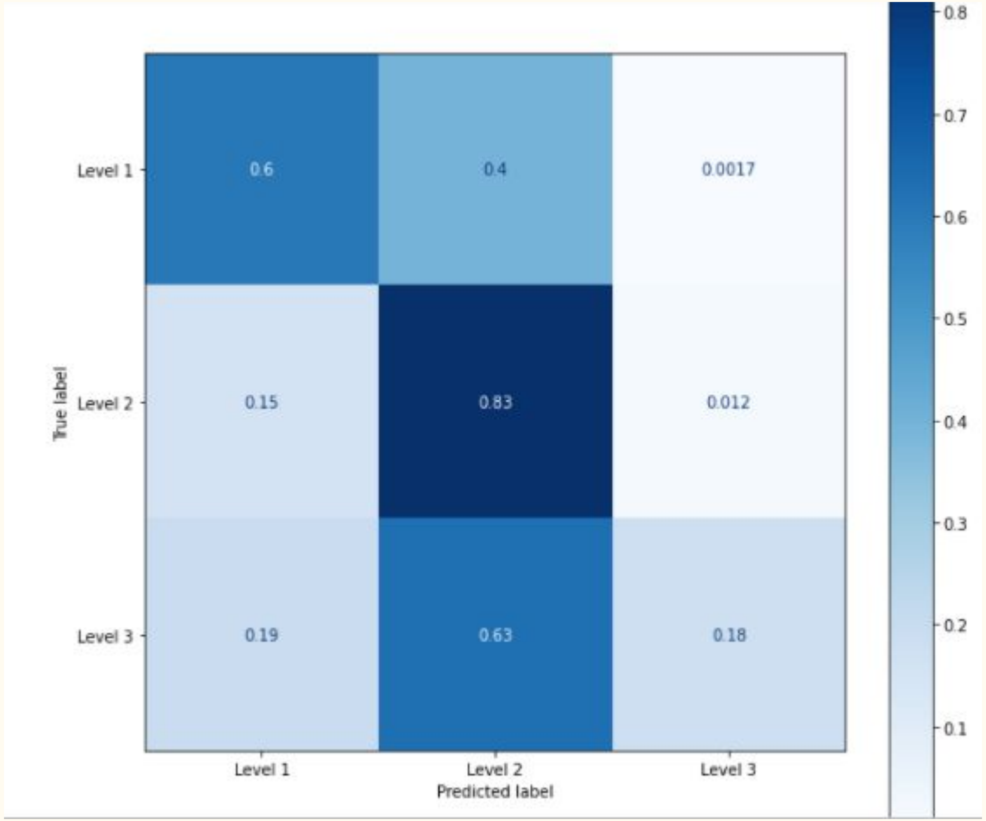
Model Evaluation

- Out-of-sample data set from 2017 was used

- It was formatted to be exactly the same as 2018 data set which was used to develop the model

- F1 Score: 67.8%
- Jaccard Score: 68.3%

- Model is overestimating the severity levels for Level 1 accidents, and is pretty lost with accidents with severity levels >=3



Conclusion

- The severity of an accident can be predicted with a reasonable accuracy with analysing features like accident time, date, weather, lighting and road conditions, as well as number of vehicles and people involved, drivers’ age, gender and the choice of vehicle.
- The presented Decision Tree model predictions could be used by emergency services with caution
- More data needs to be added in order to improve the accuracy of the model, especially for accidents with relatively higher Severity Levels. This could be achieved by either adding more older data from Leeds, or adding data from other similar towns