Predicting Car Accidents Severity

Tal Shapira

github.com/shapital/Coursera_Capstone

Introduction

- Road traffic crashes result in the deaths of approximately 1.35 million people around the world each year and leave between 20 and 50 million people with non-fatal injuries. More than half of all road traffic deaths and injuries involve vulnerable road users, such as pedestrians, cyclists and motorcyclists and their passengers.
- The ability to predict the severity of car accidents, which describes their fatality, can help save lives, being able to determine the suitable medical staff to send to the car accident location.
- The audience for this work is governments and local authorities, which are responsible for road safety and treating car collision victims.

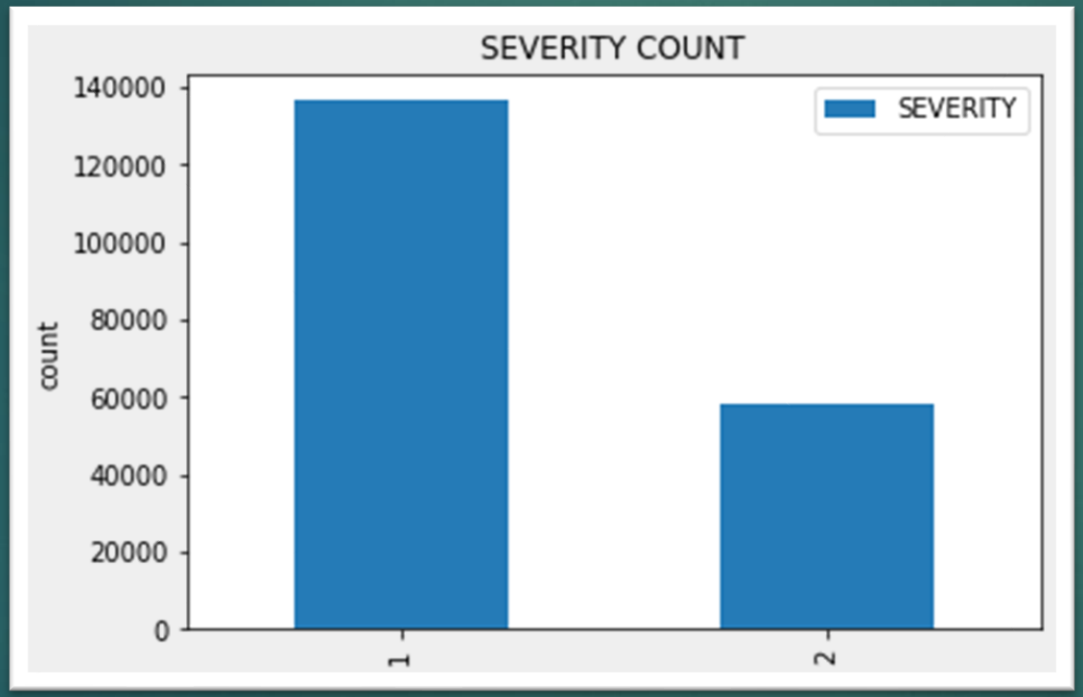
Data

- The data I will be using is car collision in Seattle city, which contains 194,673 cases from 2004 to 2020, each containing information regarding the location and address type, date, collision type, person/pedestrian/bicycle/vehicle count, whether the collision was due to inattention or not, whether a driver was under an influence of drugs and alcohol or not, the road/weather/light conditions, and the severity of the collision (1 - property damage or 2 - injury).

- For example, a collision that occurred at the intersection of 5TH AVE NE AND NE 103RD ST on March 27th 2013, in which a motor vehicle struck a motor vehicle, front end at angle, the light condition was daylight, the weather was overcast and the road was wet, no driver was under the influence of drugs or alcohol, the collision was not due to inattention, there were 2 persons involved and 2 vehicles (no pedestrians or bicycles), and the severity code was 2 - injury.

Methodology – Data Exploration

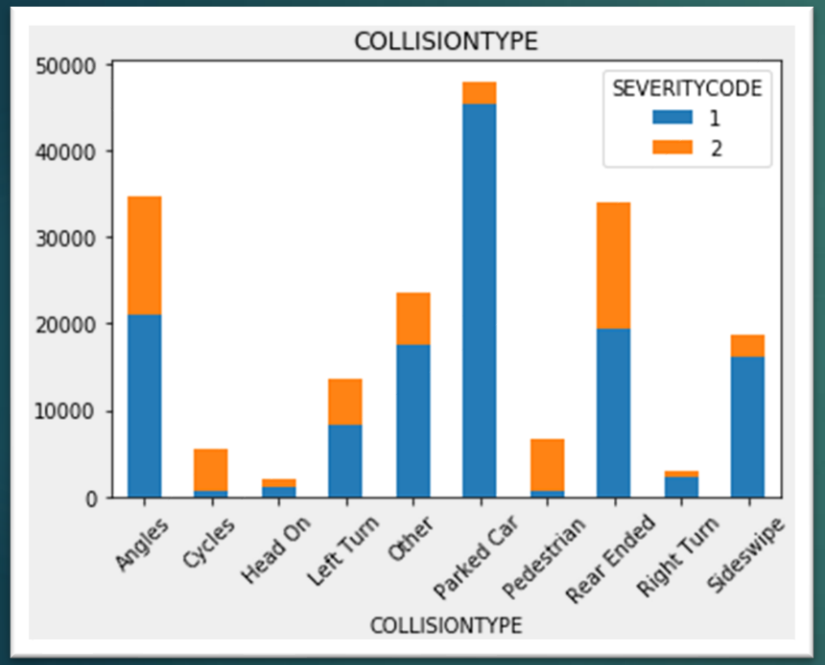
- Let's explore the data. The following graphs depict the distribution of the car accident severity for the entire data set:

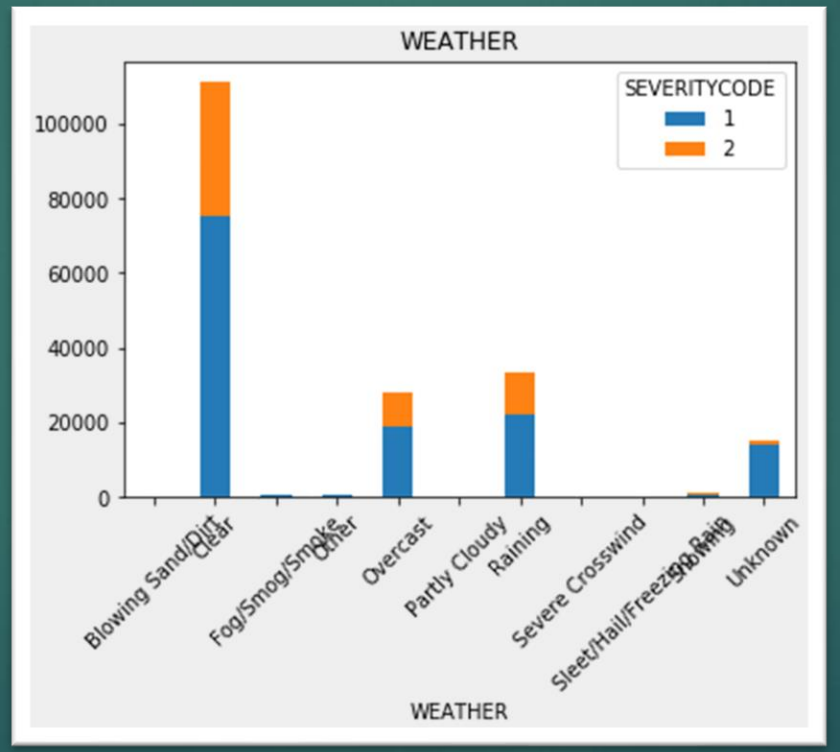


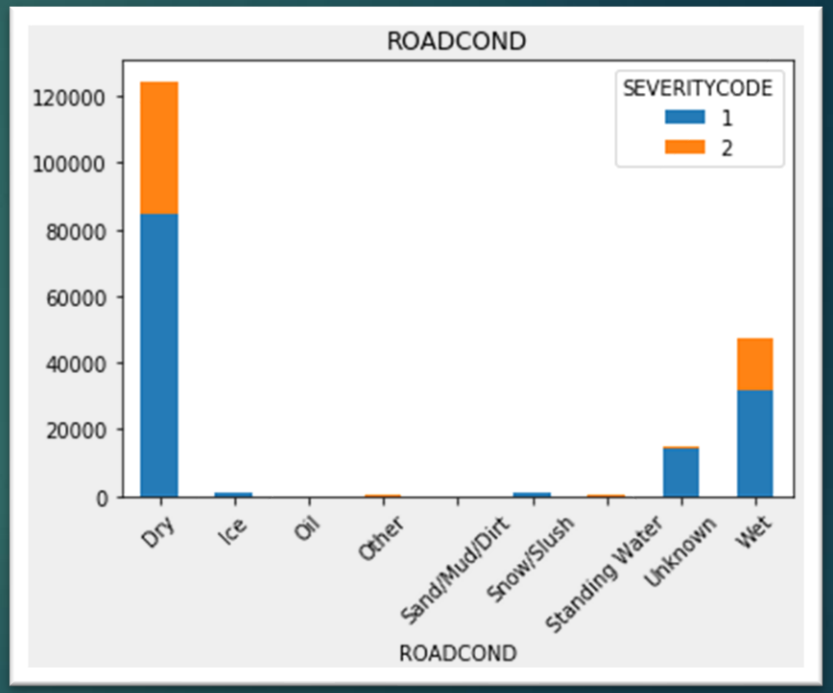
- There a more than twice as many accidents with severity 1 (property damage) than with severity 2 (injuries)

Methodology – Data Exploration

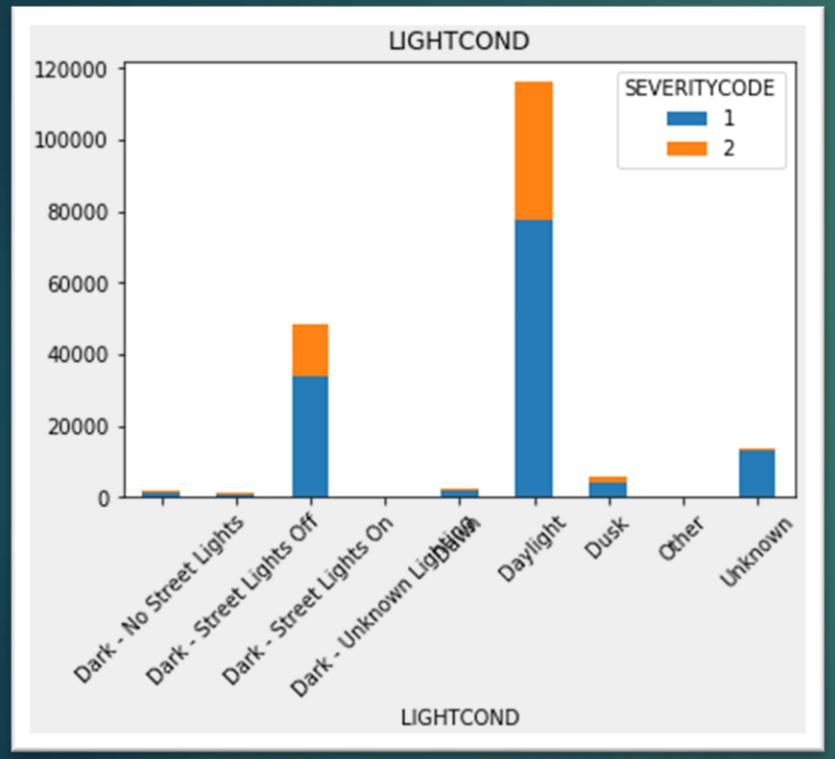
- The following graphs depict the distribution of the car accident severity according to several features:

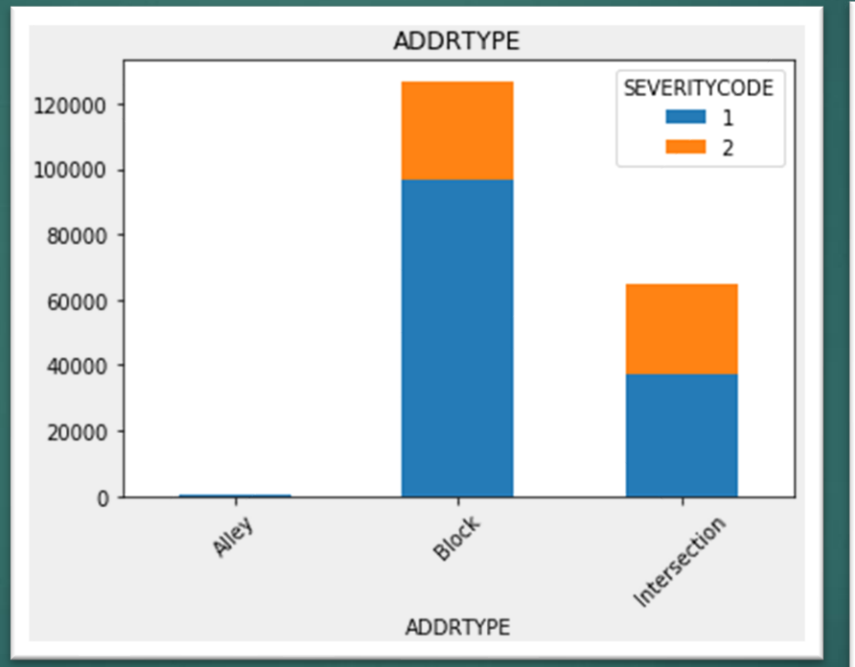


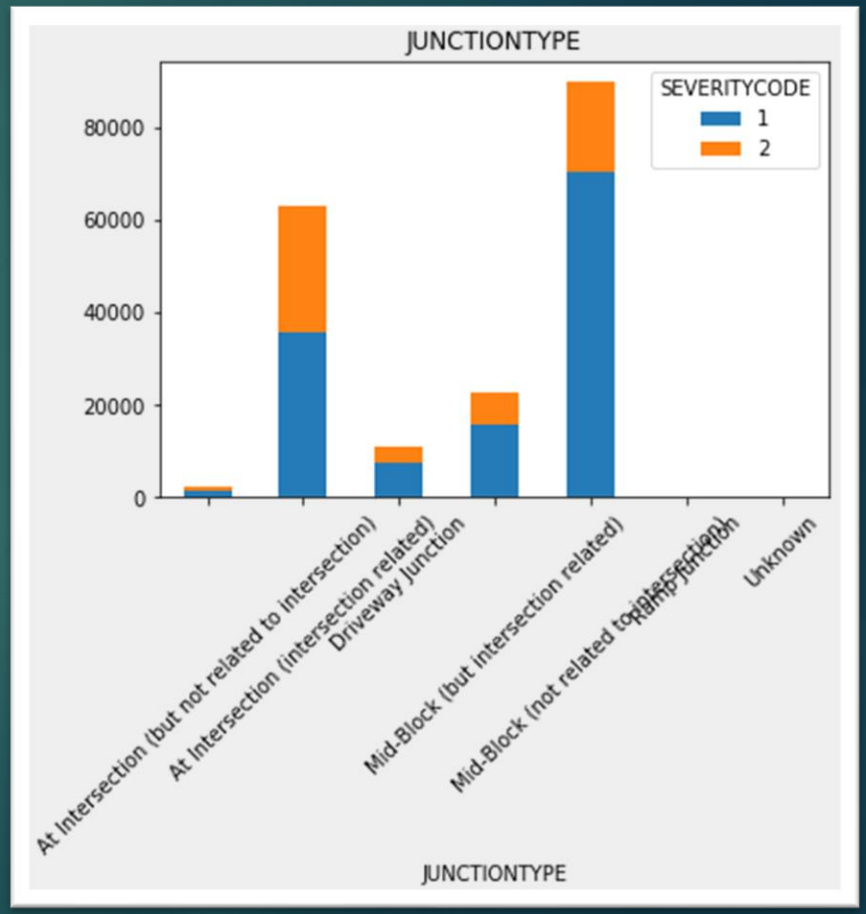




Methodology – Data Exploration







Methodology – Machine Learning Models

- I used 4 different machine learning models to try the predict the car accident severity:
  - K Nearest Neighbor(KNN):
    - the best K was chosen (with the highest accuracy).
  - Decision Tree:
    - criterion = "entropy“
    - max_depth = 10
  - Support Vector Machine:
    - kernel='rbf'
  - Logistic Regression:
    - C=0.01
    - solver='liblinear'

Results

- The following table depicts the accuracy results of the 4 different Machine Learning models according to 3 accuracy metrics:

Algorithm	Jaccard Index	F1-score
KNN (K=8)	0.717105	0.673585
Decision Tree	0.728922	0.679298
SVM	0.727326	0.667747
Logistic Regression	0.726056	0.670279

- All 4 methods have very similar accuracy scores, both according to Jaccard Index and according to F1-score. However, Decision Tree has the best score in both metrics, hence it is the best option for the prediction model

Discussion

- According to the results of the accuracy measures of the different Machine Learning models, the best model is Decision Tree.
- There is of course room for improvement, by addressing features that were dropped, such as COLLISIONTYPE, JUNCTIONTYPE and HITPARKEDCAR.