Predicting Car Accidents Severity

Tal Shapira

September 2020

1. Introduction

1.1 Background

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 young are particularly vulnerable on the world's roads and road traffic injuries are the leading cause of death for children and young adults aged 5-29. Young males under 25 years are more likely to be involved in road traffic crashes than females, with 73% of all road traffic deaths occurring among young males in that age. Developing economies record higher rates of road traffic injuries, with 93% of fatalities coming from low- and middle- income countries.

In addition to the human suffering caused by road traffic injuries, they also incur a heavy economic burden on victims and their families, both through treatment costs for the injured and through loss of productivity of those killed or disabled. More broadly, road traffic injuries have a serious impact on national economies, costing countries 3% of their annual gross domestic product.

1.2 The Problem

The severity of a car collision is an important aspect in dealing with traffic accidents. Whether an accident resulted only with property damage or with injured people is a key factor in determining the suitable medical staff to send to the location of future car accidents. Many factors may influence the severity of an accident, such as collision type, weather, road and lighting conditions, number of people / vehicle / pedestrians / cyclists involved, etc.

In this project I intend to build a model that predicts the severity of accident based on their features, based on past collisions.

1.3 Interest

Governments and local authorities that wish to save human lives by determine the suitable medical staff to send to the location of car accident aught to be interested in accurate prediction of the accident's severity. Technological companies that aim to increase road safety would also be interested in this model.

2. Data

2.1 Data Source

The data I will be using is car collision in Seattle city:
[link to csv file]

2.2 Feature selection

The dataset contains 194,673 cases from 2004 to 2020, each containing information such as 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 finally the severity of the collision (1 - property damage or 2 - injury).

The original data file includes 36 columns of features a column for the severity. For simplicity of the model I decided to focus on the following features:

1. PERSONCOUNT – how many people were involved in the accident
2. PEDCOUNT – how many pedestrians were involved in the accident
3. PEDCYLCOUNT – how many bicycles were involved in the accident
4. VEHCOUNT – how many vehicles were involved in the accident
5. INATTENTIONIND – whether the collision was due to inattention or not
6. UNDERINFL – whether a driver involved was under the influence of drugs or alcohol or not
7. PEDROWNOTGRNT – whether the pedestrian right of ways was granted or not
8. SPEEDING – whether speeding was a factor in the collision
9. WEATHER – a description of the weather conditions during the time of the collision
10. ROADCOND – the condition of the road during the collision
11. LIGHTCOND – the light condition during the collision
12. ADDTYPE – collision address type (Alley / Block / Intersection)

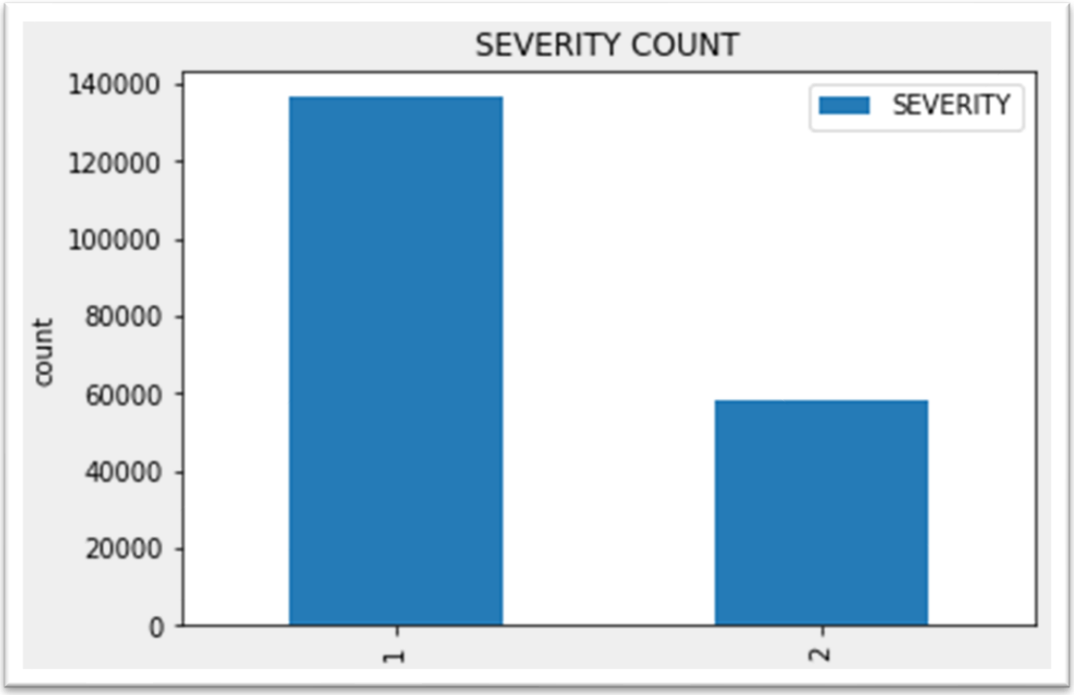
2.3 Data Cleaning

Since the dataset is created manually, it contains many missing values and unknowns. Hence a data cleaning process is required.
For the fields (INATTENTIONIND, PEDROWNOTGRNT, SPEEDING, UNDERINFL) I replaced missing values with 0's.
For the fields (WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE) I dropped rows having "Other" or "Unknown" values, since they do not contribute information. Also, I dropped rows with missing values.

3. Methodology

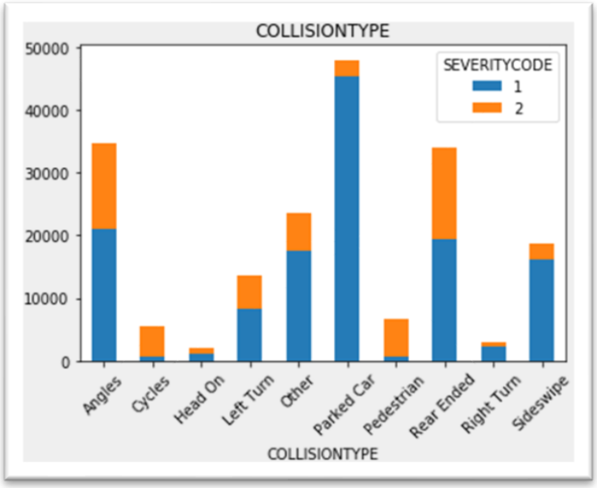
3.1 Exploratory Data Analysis

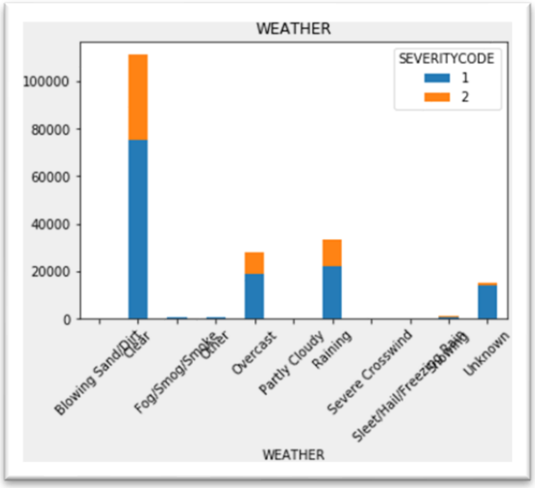
The first step is visualizing the target value distribution, i.e. the severity:

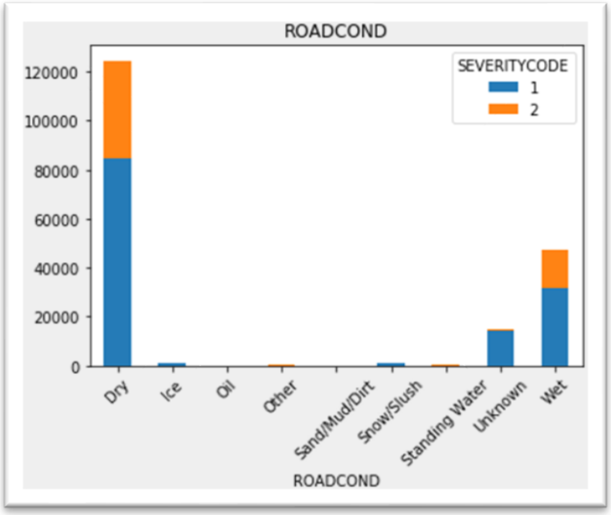


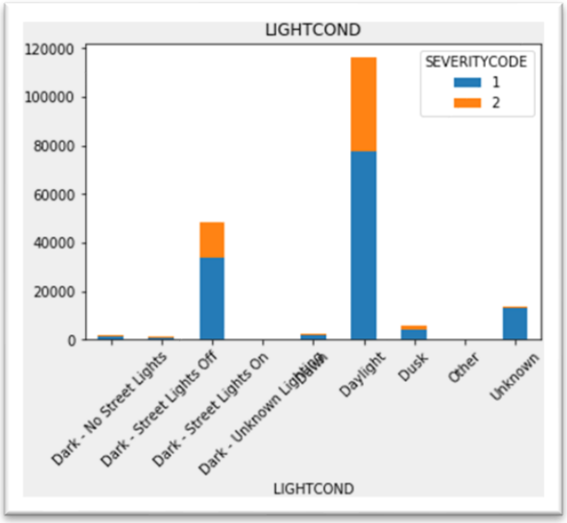
The histogram shows that the dataset is somewhat imbalanced, with more than twice as many cases with severity 1 (property damage) than severity code 2 (injured people). However, this imbalance is tolerable since there are tens of thousands of cases for each category.

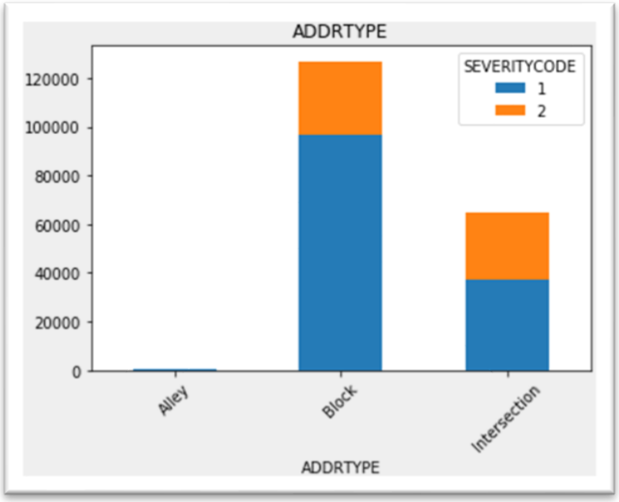
The next step is to visualize the distribution of the car accident severity according to several features:

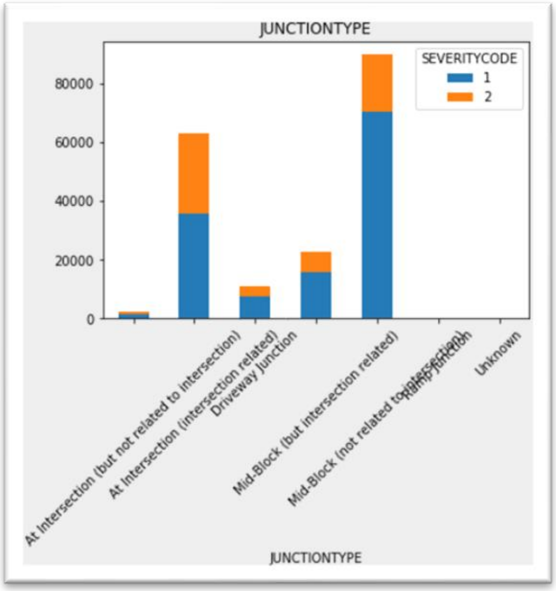












Most of the features show roughly the same ratio between the collision severity types as the overall ratio.
In some cases, the ratio tends more towards severity code 1, such as collision type – Parked Car.
And In some cases, the ratio tends more towards severity code 2, such as collision type – Rear Ended.

3.2 Predictive Modeling

I used different machine learning classification models to try the predict the car accident severity. These are all supervised learning models, which predict the outcome with varying accuracy. The model with the best accuracy will be chosen.
Before fitting the models to the dataset, I used a standard scaled to give each column of the features dataset a mean of 0 and a standard deviation of 1. Then used a train/test split method, which randomly chose 80% of the lines for training the model, and the remaining 20% to test the model.

The following machine learning classification models were used:

1. K Nearest Neighbor(KNN):
   a. the best K was chosen (with the highest accuracy).
2. Decision Tree:
   a. criterion = "entropy"
   b. max_depth = 10
3. Support Vector Machine:
   a. kernel='rbf'
4. Logistic Regression:
   a. C=0.01
   b. solver='liblinear'

4. Results

The following table depicts the accuracy score of each Machine Learning model according to Jaccard Index and F1-score metrics.

Algorithm	Jaccard	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.

5. Conclusion

This report attempted to use data from past car collisions to predict the severity of future car accidents. Many features were included in building the prediction model, among which are weather, road type and lighting condition. After clearing less important properties from the data set, removing samples with missing or useless information, a clean data set was used to train and test 4 different Machine Learning models. The accuracy of the models was tested using 2 different metrics, and the model that proved best in both cases is Decision Tree with max depth of 10.

An application of this model could be determining the suitable medical staff to send to the location of future car accidents, based on the predicted severity of the accident based on its features.

There is of course room for improvement, by addressing features that were dropped, such as COLLISIONTYPE, JUNCTIONTYPE and HITPARKEDCAR.