Car Accident Severity Prediction

Ruth Hashkes
September 25, 2020

1. Introduction

1.1 Background

Thousands of car collisions happen yearly in Seattle, some of them of high severity and even fatal. Many factors contribute to determining the severity of a car accident, including weather, type of vehicle involved, driving under the influence and so on. Seattle Police Department has released a dataset with many details of car collisions in an attempt to predict them and reduce the rate and severity of them.

1.2 Business Problem

Prediction of severity of a car collision would be beneficial to many entities, starting with the drivers themselves who could drive more carefully, or change route altogether, or insurance companies who could determine insurance policies based on location data with regard to collision severities, and of course to decision makers and local infrastructure supervisors who could change these predictions by directly addressing road quality factors that might be contributing to this issue.

2. Data

2.1 Data Source

The Seattle Collisions Dataset consists of 194,673 car collisions from 2004-2020 with 37 attributes. This data was collected and shared by the Seattle Police Department and Accident Traffic Records Department.

2.2 Data cleaning and Feature selection

The label of our dataset (the dependent variable) is Severity, containing two levels: high (Injury Collision) and low (Property Damage Only Collision).

I removed column containing unique identifiers, columns filled with one value, redundant columns and columns containing many missing values ('OBJECTID', 'INCKEY', 'COLDETKEY','REPORTNO', 'STATUS', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'INCDATE', 'INTKEY', 'SEVERITYDESC', 'SDOTCOLNUM', 'ST_COLCODE', ' ST_COLDESC').

Values of "unknown" or "Other" were replaced with NaN. Data coding for columns "Inattention", "Pedestrian right of way not granted" and "Speeding" was fixed with 0 for No and 1 for Yes.

New date columns were created: year, month, and weekday, to see if there is a change over time or a seasonality effect over collisions.

After cleaning and selecting features the dataset consists of 143,747 car collisions and 28 variables.

3. Methodology

3.1 Exploratory Data Analysis

First, I examined the categorical variables and their connection to the dependent variable, severity. There were 94,821 low severity collisions, and 48,926 high severity collisions, meaning the data is mostly balanced.

It is clear there is a connection between location of the collision to the severity. Collisions at intersections tend to be more severe (Figure 1, since there are twice as much low severity collisions, we would expect to see that 2:1 proportion in both locations).

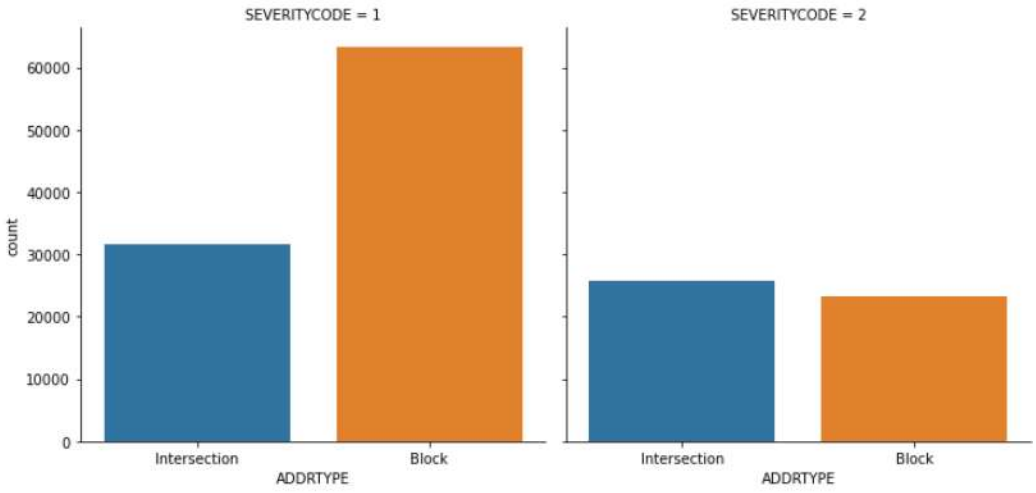

Figure 1: Collision Severity by Address Type

There is also a clear connection between collision types and severity (Figure 2). Sideswipes and collisions with parked cars seem to be more connected to low severity, while collisions involving pedestrian and cyclists seem to result in a high severity collision.

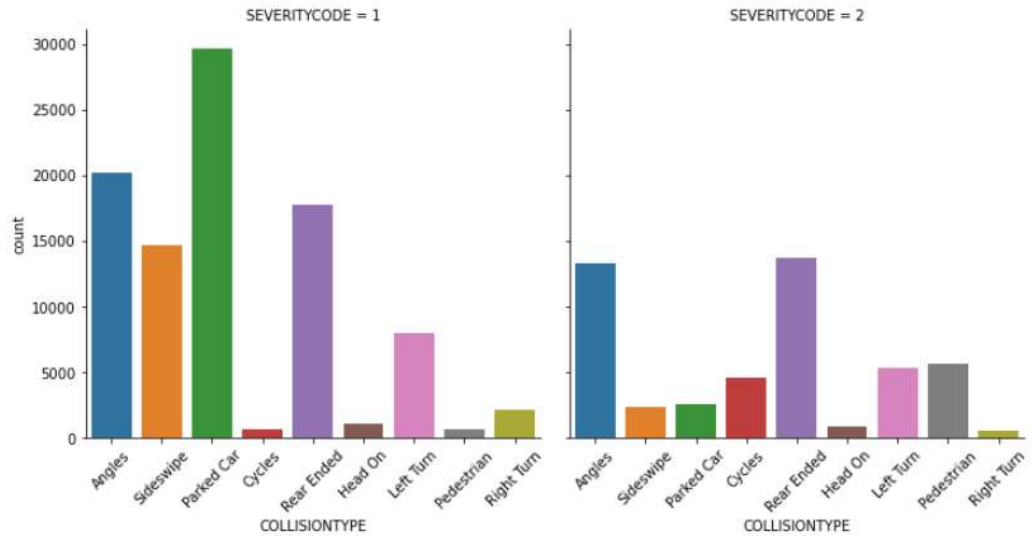

Figure 2: Collision Severity by Collision Type

We Can see a decrease in low severity collisions over the years from 8000 to 5000, but high severity collisions stayed quite the same around 3000 a year (Figure 3).

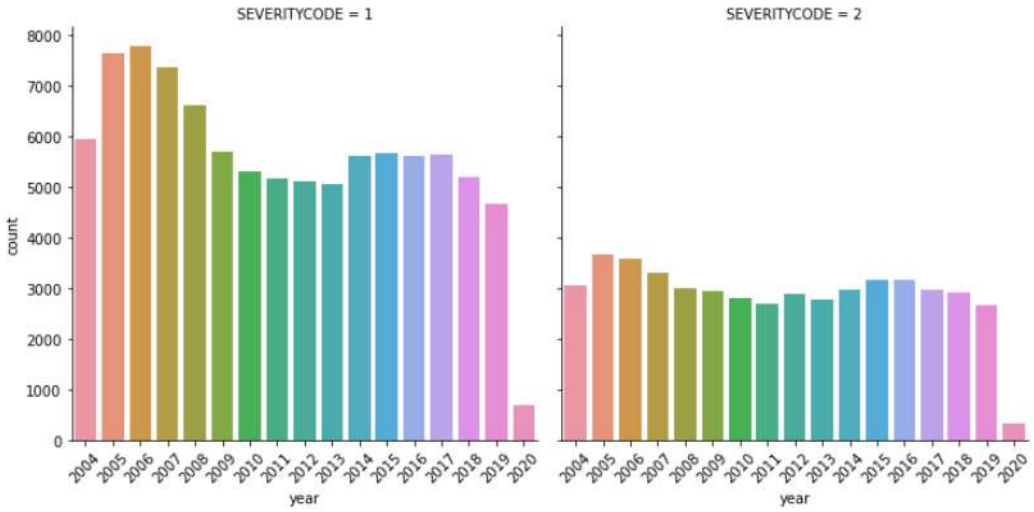

Figure 3: Collision Severity by Year

Collisions under the influence of drugs or alcohol are more common on Saturdays and Sundays as opposed to those not under the influence which are most common on Fridays (Figure 4). This plot also demonstrates the effects of COVID-19 on year 2020 which has data for 5 months but has much lower rates of collisions due to lockdowns and social distancing measures.

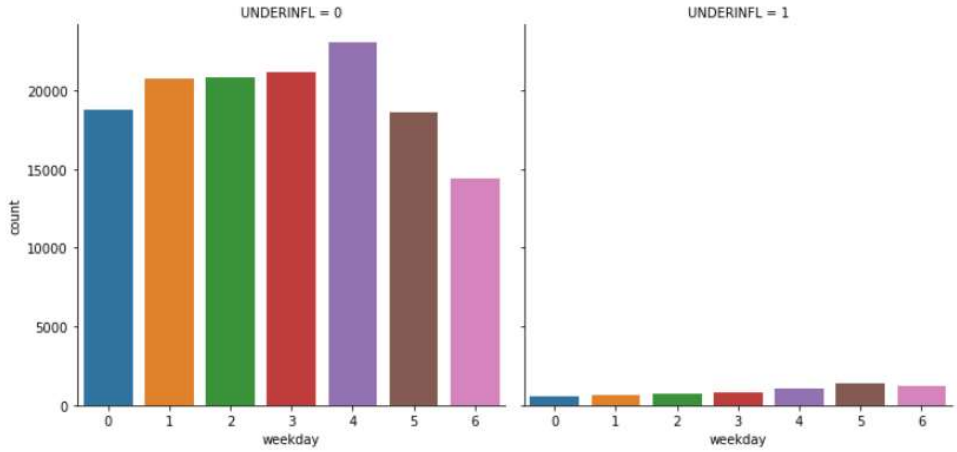

Figure 4: Collisions Under the Influence of Alcohol or Drugs by Day of the Week

Pattern of weather, road and light conditions seem similar between severities. Inattention, driving under the influence of alcohol or drugs, not granting pedestrian right of way, and speeding, all have connections to higher severity of collision, while hitting a parked car does not (not shown).

The numeric variables of number of people involved in the collision, pedestrians, cyclists, and number of vehicles are all connected to collision severity (t-tests, $p<0.0001$ for all). The higher number of people involved the more severe the collision is, and the opposite is true for number of vehicles.

3.3 Predictive Modeling

The problem is a supervised learning classification problem – to predict which severity would a collision have, a low or high one. To predict this label, I used 3 algorithms – K-Nearest Neighbor, Random Forest, and Logistic Regression.

To use these algorithms, I created dummy variables from the categorical variables, split the data into 80% train data and 20% test data, and then normalized it.

4. Results

All three models had similar results.

	Precision	Recall	F1-Score	AUC
K-Nearest Neighbor	0.70	0.71	0.70	0.64
Random Forest	0.74	0.73	0.69	0.63
Logistic Regression	0.73	0.73	0.70	0.64

According to the ROC plot it is clear the logistic regression is the most suitable at balancing true positive and false positive rates (Figure 5).

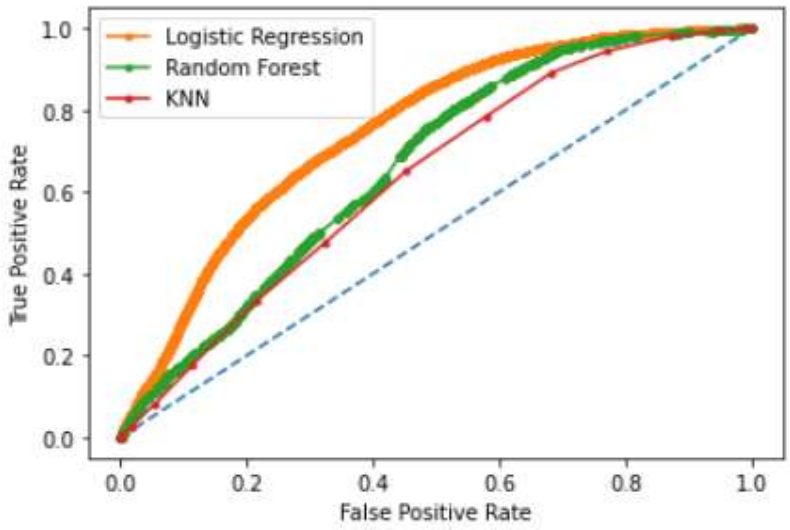

Figure 5: ROC Plot

Feature importance showed the top features for all models are number of people involved in the collision, pedestrians, cyclists, number of vehicles, parked car, block or intersection, month, weekday, driving under the influence of alcohol or drugs, sideswipe or left turn and more.

5. Discussion

Taking in the models and analysis, it is clear it is possible to predict collision severity with a high certainty. There are many factors to consider, which differ between entities.

First and foremost, accidents between vehicles and pedestrians are most prone to be of high severity. The number of pedestrians involved and not giving the pedestrians their right of way are big risk factors.

It seems that weather, road conditions and light conditions have little impact of severity. On the other hand, driving offences such as driving under the influence of alcohol or drugs, inattention, or speeding are connected to the severity.

Thus, my recommendation for drivers would be to abide the transportation laws and pay careful attention to pedestrians. There is no need to be fearful of harsh weather, in my opinion this is because people drive extra careful in these circumstances.

My recommendations to insurance policies would be to check the driver's prior driving history and take a harsher stance with known offendors.

My recommendations to the decision makers and local authorities would be to enhance law enforcement on transportation offenses such as DUIs, speeding and so forth.

6. Conclusion

In this study I analyzed car collision data and the connection of different factors to collision severity. I identified that collisions involving pedestrians are the most fatal, that weather and road and light conditions have little effect, and that driving offenses such as DUIs and speeding are dangerous. These models can help future decision-making for drivers, insurance companies or local authorities to prevent the next fatal collision from taking place.