Capstone Project – Predicting Car Accident Severity Based on Collision Data

Introduction

Car accidents are one of the most common accidents in the world and are the 13th leading cause of death overall among all causes.

By analyzing collision data from 2004 to 2020, the business interest is to find correlations between the severity of accidents and a wide range of potential factors including the accident location, weather condition, lighting condition, accident time, etc.

We believe that by trying to understand the relationship between those factors and the severity of the accident, it will be helpful to guide people when driving in similar conditions, or even predict the outcome of accidents based on the accident conditions such that the first responders will be more prepared when they are reaching to the scene.

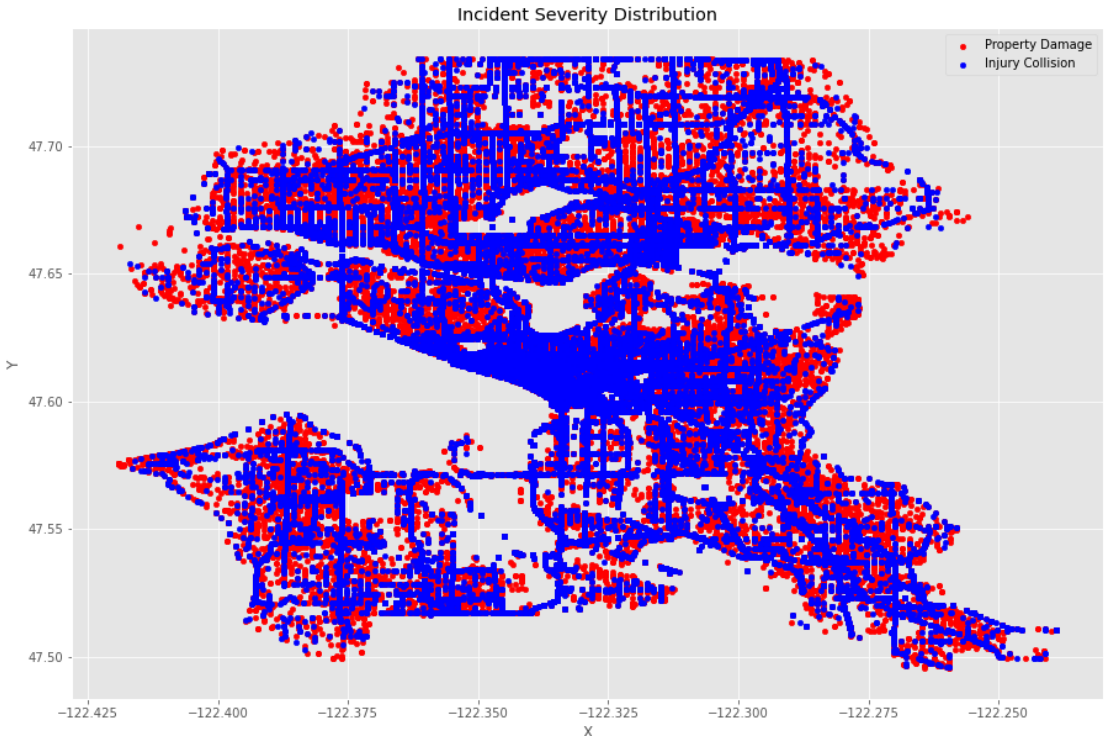
Data

The data is from Traffic Records, Seattle Police Department and can be accessed through this link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

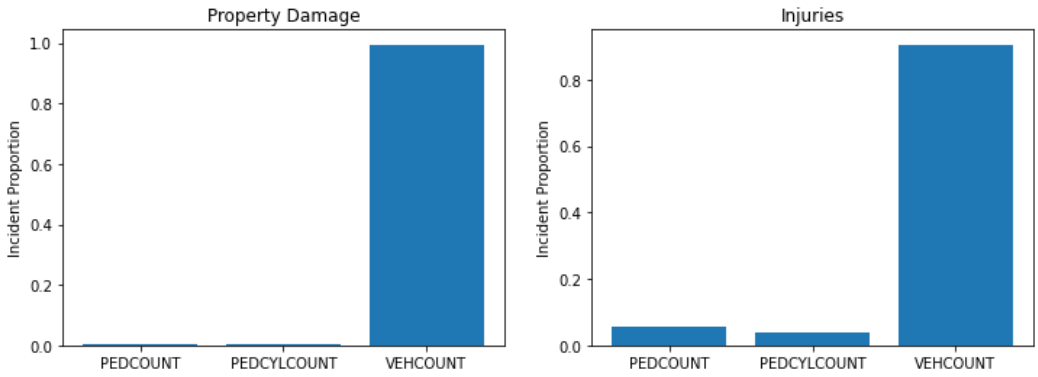
There are 194,673 records in the dataset. It has 37 attributes which classify the severity into 2 cases, Injury Collision and Property Damage Collision.

Exploratory Data Analysis

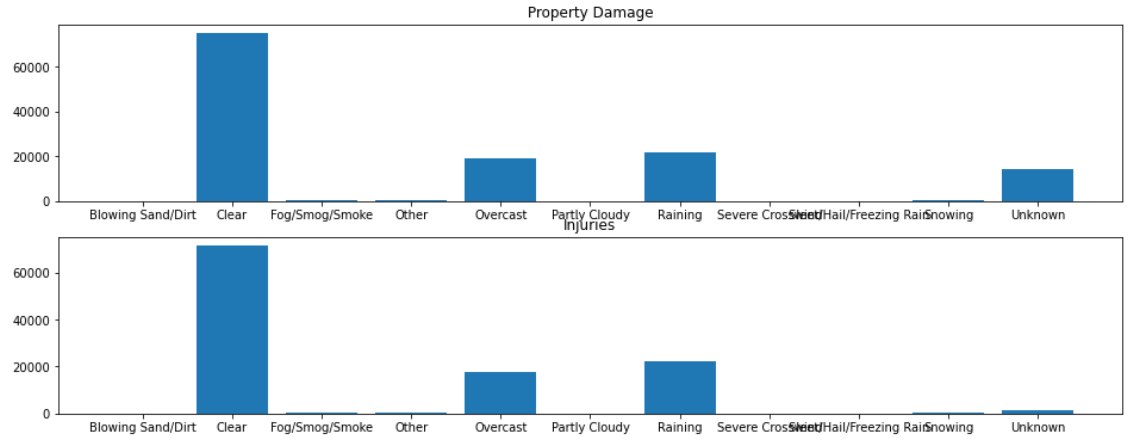
Based on the severity of the accidents, the data is divided into Injury Collision (injury) and Property Damage Collision (property damage).



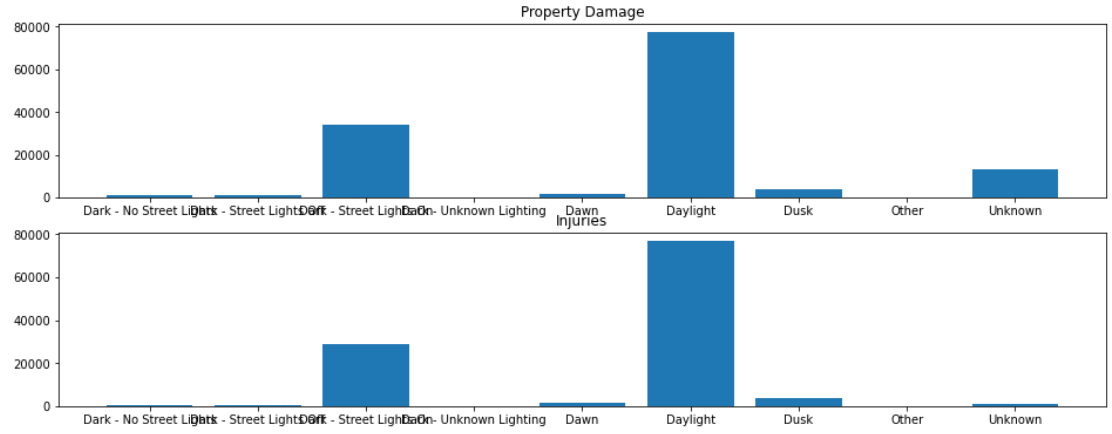
Accidents locations are evenly spread out across the map, with the exception of the center of the map where Injury Collisions take place the most frequent.



For both types of accidents, it involves major vehicle damage. However, pedestrians and cyclists are more likely to be involved in injury accidents.



For both types of accidents, the top 3 common weather conditions are Clear, Raining and Overcast.



For both types of accidents, the most common lighting conditions are Daylight, Dark – Street Lights and Dusk.

Methodology

1. Preprocessing

In this phase, missing values are dropped along with foreign key columns. Columns that serve as explanation to other columns, such as INCAT, SDOT_COLDES, ST_COLDES are dropped as well.

2. Parsing INCDTTM Column

Since drivers may have different physical and/or mental conditions at different times of a day, the hour information during which the accidents take place is extracted from INCDTTM column.

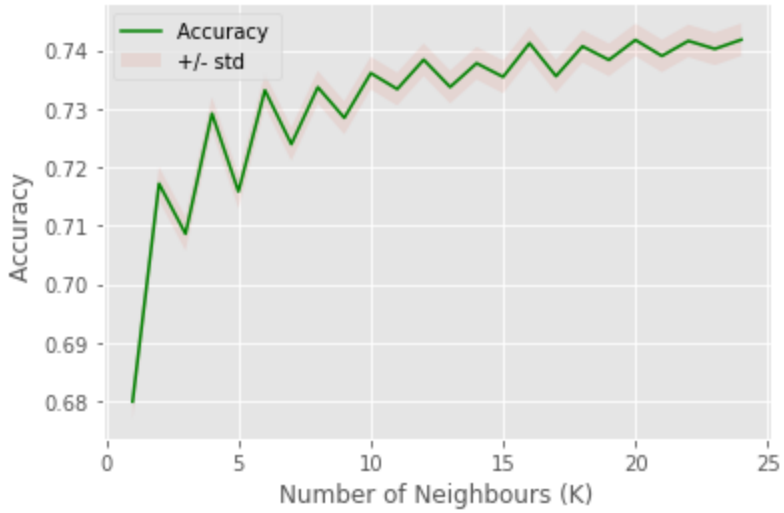
3. Parsing Categorical Columns

There are 10 categorical columns after data preprocessing: UNDERINFL, HITPARKEDCAR, ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, SDOT_COLCODE, WEATHER, ROADCOND, LIGHTCOND, ST_COLCODE. Among them, UNDERINFL and HITPARKEDCAR are converted to Boolean type while others are converted to dummy variables.

Model Selection and Training

1. KNN

Since there are location features (X, Y), and weather condition, road condition and even culture (carpooling or not, popular to DUA or not, etc.) are somewhat related to geographic distribution. Therefore, by using KNN, it may give us some interesting result. Because there is a hyper parameter K which governs the outcome of the model, the training set is split into another training and test set to get the best K value before fit the model with the whole training set and get the test result.



(K = 10 is at the sweet spot between accuracy and complexity)

2. SVM

Because of the high dimensionality of the data, SVM is a good choice to train the model.

Result

KNN's evaluation F1-score: 0.71

SVM evaluation F1-score: 0.71

KNN and SVM in this project produced a similar level of accuracy.

Discussion

About CRISP-DM after finishing this capstone project, I found out that it is better to do EDA prior to data preprocessing. dropping N/A values or converting categorical columns into may sometimes make it more difficult to have a first intuitive perception against the raw data

About data preprocessing it may produce a better learning if there are more information associated to each column. Correlation analysis is not nearly enough. And, for example, I only extracted hour information from INCDTTM column since the date may not carry enough information. However, there may be some implicit day or month and severity relationship which would have increase the model accuracy

About KNN it is slow and not effective against high dimensional data. therefore, I took out the ST_COLCODE and SDOT_COLCODE features. While I was tempted to convert other remaining categorical columns into numerical values instead of dummy columns, I didn't do that implementation because the distance between 1 and 2 and 1 and 9 should be treated the same, in the aspect of categorical value

Conclusion

Based on the limited data, the study indicates that based on the hours, junction type, road condition, weather and other driving condition, people can have a basic prediction on what kind of accident it is more likely to occur.