Predicting Severity of collision on Seattle

Duman Zumadilov

11 October 2020

1. Introduction

1.1 Background

According to the annual United States road crash statistics by ASIRT, more than 38,000 people die every year in crashes on U.S. roadways. The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants. It is evident that with the increasing number of vehicles on urban and suburban roads, the cases of vehicle accidents are also increasing. This project aims to analyze vehicle collision data available in public domain provided by Seattle Geo Data (SDOT).

1.2 Problem

We or our friends/family members/mates have experienced the collision involving cars or bicycle. Most of cases are ends with “happy end” with non-significant injuries like scratches or bruise. Unfortunately, there are still accidents resulting fatalities despite the deployed state-of-art technologies that regulates lights or lights with deployed sound effect.
The modern technology does not consider simple people’s distraction or carelessness, so additional measures should be introduced. However, we are living in the world where the resources are limited and modern machine learning alorightms could assist us to wisely use resources by identifying crucial variables.

1.3 Interest

Seattle government will be interested in accurate prediction of collision fatalities which could help to determine where to put apply additional measure to decrease the fatalities.

2. Data acquisition and cleaning

2.1 Data Sources

Data provided by the Seattle Department of Transportation (SDOT) on vehicle collisions along with its severity might be useful to derive insights and may show some pattern with the environmental factors like weather, road conditions etc. The dataset consists of 40 columns having different kinds of data like, collision severity, road conditions, number of people involved, location of collision, weather etc. Meta-data of the dataset can be viewed https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

2.2 Data Cleaning

The data has some missing data but the main issue is different classification of the same feature like instead of “Yes” puts 1 or “No” puts 0.

One of the first step was to adjust feature that shows whether or not a driver involved was under the influence of drugs or alcohol (column “UNDERINFL”). It contains Y and 1, or N and 0 as a model could accept only text values we should convert “Y” to 1 and “N” to 0.

To fill missing values for features - ***LOCATION, JUNCTIONTYPE,COLLISIONTYPE, ST_COLCODE, ST_COLDESC***, would be difficult as has ***enormous kind of values***. However, LOCATION and ADDRTYPE could be indentified by Lonlitute and Latitude but it requires significant amount of time an, moreover, this feature will be considered any way by Lonlitute and Latitude.For other features, we could accept most ***logical value*** or assign to "Unknown", if applicable.

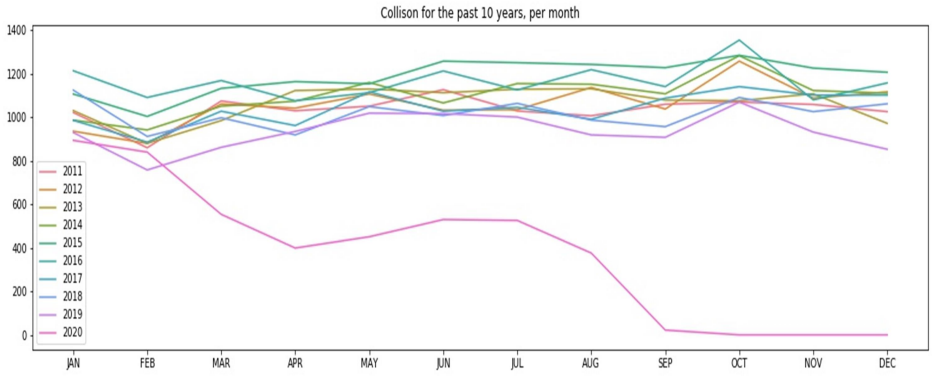
2.3 Feature Selection

This dataset too consist of a lot of missing values and useless datas .so before further processing the data it is important to clean the dataset¶

The following features(columns) should be drop as they have a lot of missing values and not usefull or irrelevant for predicting severity: STATUS, INTKEY,OBJECTID, INCKEY, COLDETKEY, REPORTNO, EXCEPTRSNCODE, EXCEPTRSNDESC, INCDATE, INATTENTIONIND, PEDROWNOTGRNT, SDOTCOLNUM, SEGLANEKEY, CROSSWALKKEY

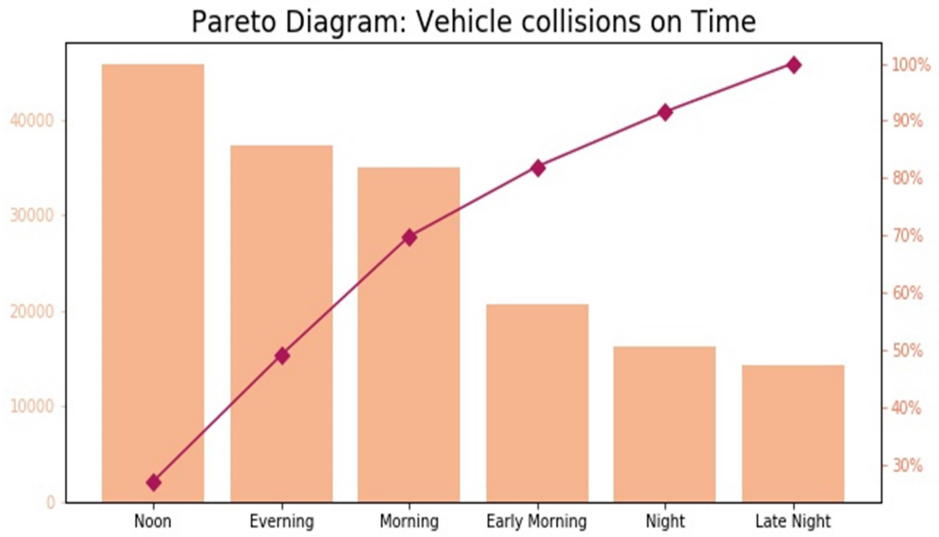
3. Exploratory Data Analysis

3.1 Trend of accidents by month and year



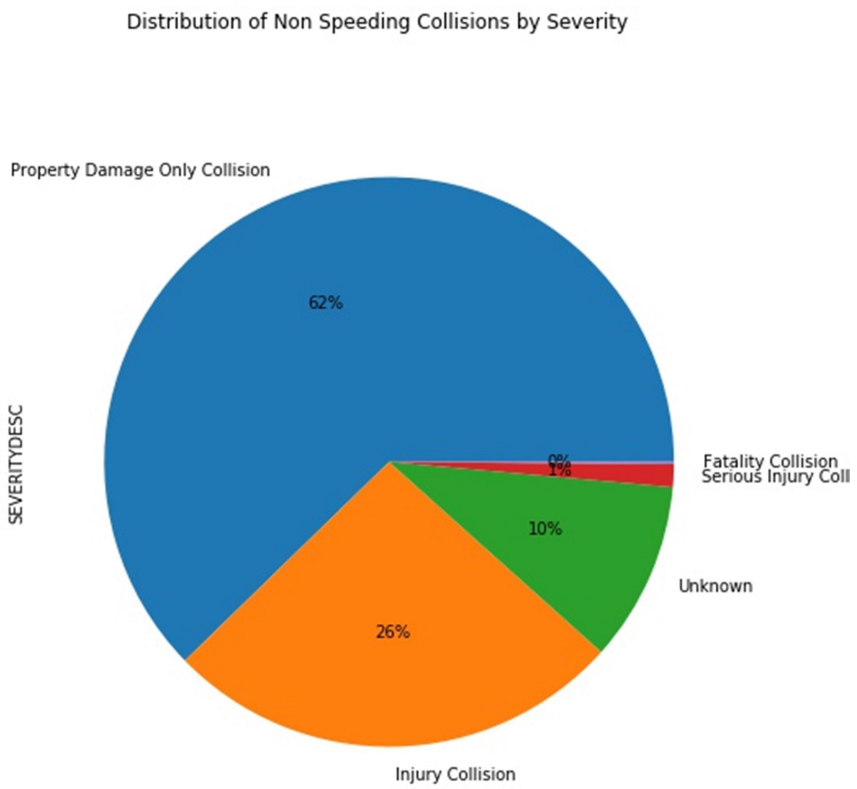
You can observe that number of accidents are steadily decreasing over the years without taking into account 2020 as COVID 19 quarantine locked population in their homes. For example 2019, number of collisions per month is below for previous years. Year 2020 is not considered as it was unique situation as COVID 19 almost suspended all means of transportation for few months during quarantine.

3.2 Relation between accidents and time of the day



From the Pareto diagram, we can see that about 50% of accidents tend to happen during the day time i.e. between 12:00 P.M. to 8:00 P.M.

3.3 Relationship between severity and speeding



We can see percentage of Injury Collision **increases** from 26.52% to 36.3%. Serious Injury Collision **increases** from 1.39% to 3.92% and Fatality Collision **increases** from 0.15% to 0.89% because of speeding. While the percentage of Property Damage Only Collision **decreases** from 62.14% to 58.87%

4. Modeling

4.1 Classification models

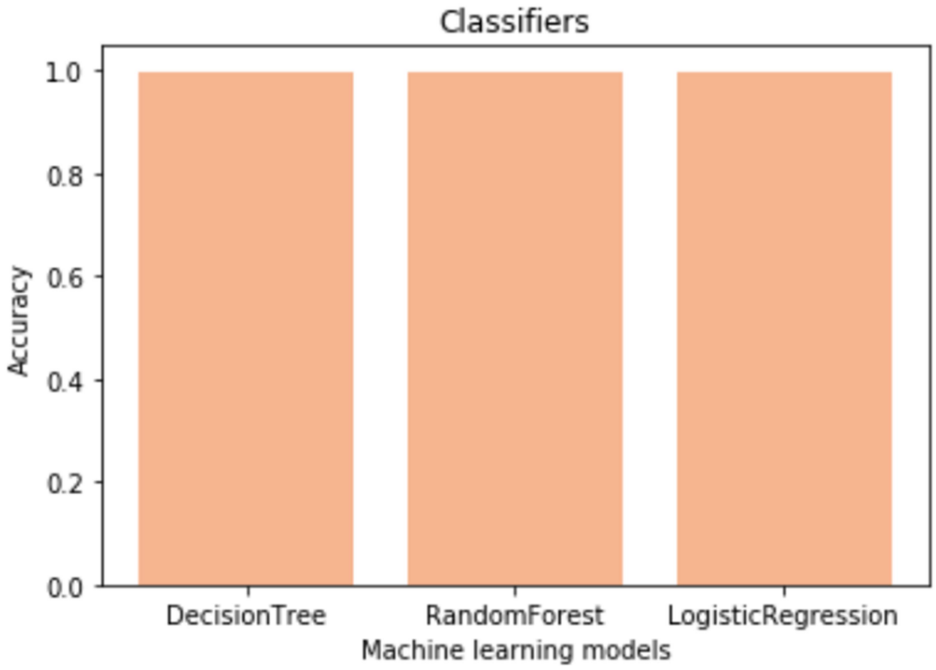
The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data.The output from modeling is a trained model that can be used for inference, making predictions on new data points.¶

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. To avoid overfitting, both methods use a test set (not seen by the model) to evaluate model performance

Due to type of output is expect (Fatality - Yes/No), classification model should be applied. there is a lot of different classifiers - Logistic regression, Decision tree, and so on. I will test all of them to find a model showing the best accuracy on both test and train

4.2 Results

The accuracies of all the models is around 99.8% which means we can accurately predict the fatality of an accident.



5. Conclusions

The accuracy of the classifiers is great and all models almost show 100%. However, the Decision tree classifier shows less FalseTrue than other models. In our case, False True means that it predicted incorrectly that there is no fatality however there was 56 fatal incidents. Despite this fact, overall, the model has trained well and fits the training data and performs well on the testing set as well as the training set. We can conclude that this model can accurately predict the Fatality of car accidents in Seattle.