Predicting the severity of accidents in Seattle

Leyla Yunis

September 4, 2020

1. Introduction

1.1 Background

Currently, the vast majority of countries population have one or more vehicles as means of transportation. This leads to more vehicles on the streets, and more accidents. In countries with large population, the amount of vehicles, and collisions, are a common thing. For the population, it is important to avoid collisions at all costs, since they could end up badly hurt. But the probability of being involved in a collision is unknown for people. It is necessary to know the probability of being involved in a collision and the severity of it, so people can take action and drive more carefully or try to avoid it in high probability cases.

1.2 Problem

Data that might contribute to determining the probability and severity of a collision might include weather, specific streets, date, time, alcohol or drugs consumption, condition of the road, amount of light on the street, speed, and more factors that describe the general condition in which the collision took place. This project aims to predict whether a collision will happen or not and the severity of it based on these data.

1.3 Interest

People that live in Seattle would be very interested in accurate prediction of collision probability and severity of it, for safety reasons. Others interests could be the mayor of the city and the government, so they could create safety legislations based on this, improve roads, or make awareness campaigns that aims to reduce accidents.

2. Data

2.1 Data Source

The data on collisions and severities can be found in Seattle GeoData dataset here. This dataset is complete and it has information from 2004 to the present, recorded by Seattle Department of Transport (SDOT). This dataset differs from the sample dataset provided in the IBM Data Science Capstone. Specifically, the dataset used has more columns and rows, and the severity target variable consists on five discrete values instead of the binary values in the sample dataset. The description of the severity values are shown in the table below.

Severity code	Meaning
0	Unknown
1	Property Damage
2	Minor Injury
2b	Serious Injury
3	Fatality

2.2 Data Cleaning

The dataset has the following peculiarities that makes it not suitable for performing quantitative analysis:

- Dataset contains several missing values ($NANs$), where key variables are absent. This is the case for the variable that indicates whether speeding was a factor in the collision, that has 95% of missing entries. It is necessary to discards rows with that are missing crucial data.
- Dataset contains some categorical variables, which are not suitable for performing quantitative analysis. Severity, Weather and Road Condition are examples of this variables. It is necessary to cast this variables to numerical fields.
- Dataset contains unnecessary or redundant variables that replicate information from other variables. ObjectId, SeverityDesc and IncKey are examples of this variables. It is necessary to discard this columns.
- Dataset is unbalanced. There are much more collisions with severity code = 1 than with severity code = 3. It is necessary to balance the dataset.