Coursera – Applied Data Science Capstone

Roman Semin

“Car Accident Alert System”

Week 1

Business Problem

Cars and motorcycles are the most dangerous way of transportation resulting in around 219.85 fatalities per billion passenger miles according to Statista [1]. In the US over 6 million people get into car accidents every year, and on average 90 people die every day as a direct or indirect result of them [2]. Most often the accidents occur because of conditions that prevent drivers from communicating together well on the road, including bad weather, road and light conditions. Considering that these are the factors not controllable by drivers who take a certain path, it makes it extremely dangerous to drive when such conditions are bad. Conditions like slippery roads, heavy fog, or even rain take away control from the driver, putting lives of people on the road of a roulette, which many people lose and end up either in a hospital or dead.
The purpose of the project is to find out the correlation between conditions such as road, weather or light conditions, and severity of potential accidents that might happen. The report is answering two questions. The first question is how likely you are to get into an accident.
The second question is what type of accident you are most likely to get into depending on the conditions mentioned above.

Data Analysis

Through inspection it was found that the dataset contains 38 columns. Most of the columns are qualitative (ex. text, object). Each of 194673 records describes details about the accident recorded by police. The data that needed for analysis of the problem included four columns: 'ROADCOND', 'WEATHER', 'LIGHTCOND', 'SEVERITYCODE'. For the ‘ROADCOND’ column, the data is significantly skewed (60% of the dataset) towards *Dry* value. The dataset also contained ~15’000 unknown values, which had to be removed. For ‘WEATHER’ ‘LIGHTCOND’ columns, similar observation as 60% of data is skewed towards one value such as *Clear* and *Daylight* respectively. For all records, the 'SEVERITYCODE' has ⅔ ratio of a minor accident and only ⅓ values for accidents that caused some traumas. It is quite an amount. The approach to solving the problem is finding the relationship in severity code increase and one of the conditions. The data from the three road, weather and light condition columns would be inspected to find the best combination with the highest probability of the severity code and develop a machine learning model that based on the conditions predicts the severity code of the accident.

[1] M. Armstrong and F. Richter, “Infographic: The Most Dangerous Ways to Travel in the U.S.,” *Statista Infographics*, 04-Jun-2019. [Online]. Available: https://www.statista.com/chart/18264/the-most-dangerous-ways-to-travel-in-the-us/. [Accessed: 29-Sep-2020].

[2] “Car Accident Statistics in the U.S.: Driver Knowledge,” *DriverKnowledge*, 25-May-2019. [Online]. Available: https://www.driverknowledge.com/car-accident-statistics/. [Accessed: 29-Sep-2020].