Part 1: Introduction/Business Problem

Car Accident Severity Analysis

(Data Science Coursera Capstone Project)

Prepared By:

Rajkumar Bathani

1. Introduction

1.1 Background

National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to *$871 billion* in a single year. Seattle, a city on Puget Sound in the Pacific Northwest, is surrounded by water, mountains and contain thousands of acres of parkland. Washington State's largest city, it's home to a large tech industry, with Microsoft and Amazon headquartered in its metropolitan area. There was a significant increase in the total number of vehicles in 2016 than in 2010. An increase in number of vehicles and car population rates can lead to higher numbers of accidents on the road.

1.2 Problem

The given dataset consists of 37 attributes that is associated with the car accident happened in Seattle, Washington, in which some or all can be used to train the Machine Learning model. Data that might contribute in predicting accident severity, Now, it would be great if you were warned of weather conditions and the road conditions, and about the possibility of you getting into a car accident and how severe it would be, So that you would drive more carefully, or even change your travel if you are able to.

2. Business Problem

Due to the increasing number of accidents and number of fatalities and damage, National Highway Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. Apart from that huge number of fatalities has been recorded due to the accidents. The project aims to predict how severity of accidents can be reduced based on a few factors. The reduction in severity of accidents can be beneficial to the Public Development Authority of Seattle which works towards improving those road factors and the car drivers themselves who may take precaution to reduce the severity of accidents.

2.1 Data Science Methodology to solve a business problem

The CRISP-DM methodology is a process aimed at increasing the use of data mining over a wide variety of business applications and industries. The intent is to take case specific scenarios and general behaviors to make them domain neutral. CRISP-DM is comprised of six steps with an entity that must be implemented in order to have a reasonable chance of success. The six steps are shown in the Figure 1. This section will associate each stages of CRISP-DM methodology with the problem of predicting car accident severity.

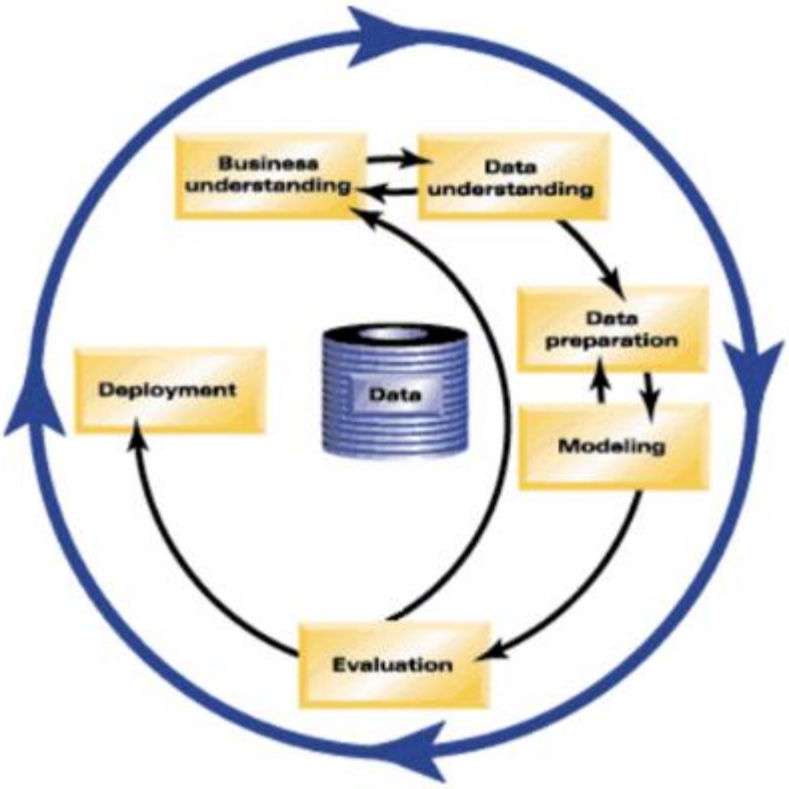

Figure 1: Data Science Methodology

The Business Understanding stage is the most important because this is where the intention of the project is outlined, the team of data scientists would communicate with the stakeholders and clarify the requirements. Stakeholders might have different objectives, biases, and modalities for relating information. Data Scientist will understand that how predicting the accident severity help the local city Police and the corresponding departments.

The Data Understanding relies on business understanding. Data is collected at this stage of the process. Once the goals are estimated after communication with stakeholders, the team of data scientists figure out the understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods. This combines the stages of Data Requirements, Data Collection, and Data Understanding from the Foundational Methodology outline.

Data Preparation once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. Many descriptive statistical tests are performed at this stage to obtain the statistical information. Also, exploratory data analysis is the integral part of data preparation process, where visualization plot and feature engineering might play a vital role in exploring the data and give business insights.

Modeling Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest. Models are selected on a portion of the data and adjustments are made if necessary. Data Scientist will determine what approach to use for modelling (supervised or unsupervised based on the dataset).

Evaluation The selected model must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow you to see the effectiveness of the model on a set it sees as new. Results from this are used to determine efficacy of the model and foreshadows its role in the next and final stage.

Deployment In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders. The new interactions at this phase might reveal the new variables and needs for the dataset and model. These new challenges could initiate revision of either business needs and actions, or the model and data, or both.

This is how the CRISP-DM methodology can be applied on real life problems and each stages of this methodology helps in solving the problems step by step.