1. Introduction

1.1 Background

The economic and societal impact of traffic accidents cost U.K. citizens hundreds of billions of dollars every year. Thousands of car drivers are killed or seriously injured in car crashes every year in the United Kingdom. And for many people who are not seriously injured, the injuries they experience can cause symptoms that last for months and even year, sometimes for the rest of their lives.

1.2 Problem

There are several factors that can affect the possibility of people getting into a car accident and how severe it would be, such as the weather and the road conditions. Knowing those factors can provide drivers with a little insight into the severity of the accident, the drivers would drive more carefully or even change their travel if they are able to, and the government might be able to implement well-informed actions and better allocate financial and human resources to reduce the impact of car accidents.

1.3 Objective

The first objective of this project is to recognize key factors affecting the accident severity. The second one is to develop a model to predict accident severity based on the sophisticated traffic accident dataset. To be specific, for a given accident that just happened or a potential one, this model is supposed to be able to predict the likelihood of this accident being a severe one.

2. Data acquisition and cleaning

2.1 Data sources

The dataset “Road Safety Data – Accidents 2018” is downloaded from Open Data Platform UK. The dataset is published by the Department of Transportation and is being shared under Open Government Licence. The dataset captured road accidents in the UK in 2018 and has 32 features/columns and about 120K rows. All the data variables are coded rather than containing textual strings. The dataset provider also attached lookup tables that can guide people who is going to use the dataset to understand the data.

2.2 Data cleaning

Since all the data variables are coded rather than containing textual strings. The dataset is relatively clean and formatted. Data cleaning was first performed to detect and handle corrupt or missing records. Because the number of observations is over 100k, which is enough for generating a model, the observation with missing values will be deleted from the dataset. Since the objective of the project is to predict the impact of road condition and weather condition on accident severity. Therefore, not all the variables will be used in training the model, such as “Latitude”, “Local Authority (District)”, “Day of Week”, and “Number of Police Force”.

2.3 Data processing

EDA (Exploratory Data Analysis) will be done over most variables, including calculation of target variable, the relationship between accident severity and weather condition, the relationship between accident severity and road surface condition, the relationship between accident severity and light condition, and the relationship between accident severity and speed limit. Then, Logistic Regression, Decision Tree Classifier, Support Vector

Machine, K-Nearest Neighbors Algorithm, and Random Forest Classifier will be used to develop the predictive model. Finally, all models generated will be ranked based on their accuracy. The model with the highest accuracy will be selected as the final one to predict car accident severity.