Predicting car accident severity

1. Introduction

1.1 Background

Car accidents are one of the most frequent cause of serious injuries and casualties in modern societies. There are many potential approaches which can be considered in order to reduce their occurrence. However, the main objective should be to reduce the ones which are more serious.

There are several variables which might have some influence when trying to predict the probability of an accident. But if the objective is to predict its severity, it might not be intuitive to determine which ones are the most relevant.

1.2 Problem

Data with a list of accidents which have occurred in the city of Seattle is available, labelled in terms of their severity. Severity is expressed as a binary variable: 2 when the accident caused some injure and 1 when there is only property damage. This project aims to predict severity of the accident in terms of the different variables which can be found in the dataset.

1.3 Interest

As it was said before, reducing the number of severe accidents is very important. Building a model which can predict the severity of an accident would be valuable for that purpose.

In addition, this model could be used in navigation applications. This would improve the journey which these applications recommend, based on the probability that a severe accident might occur.

2. Data acquisition and cleaning

2.1 Data Sources

The data which will be used to elaborate this report is the .csv file provided by IBM. Some databases on the website Kaggle were acquired as well as a complement.

2.2 Data cleaning

The dataset has a total of 194674 rows. Some of them have empty values.

As a first approach *dropna()* function was applied to the whole dataframe. Only 29943 rows were deleted. This is considered reasonable, so this was the chosen method to delete empty values.

Another problem is that the dataset is not balanced: severity 2 accidents (which caused injuries) are undersampled in the dataset. In order to solve this, the method *sample()* was used. Two dataframes were created from the original (severity-1 and severity-2), a sample with fraction 0.5 was taken from severity-1 dataframe and finally the two dataframes were concatenated.

2.3 Feature selection

There are some features which are not valuable for our problem. For instance, there are several features which are codes, just useful to identify the accident. OBJECTID, INCKEY, SHAPE, COLDKEY are some of these.

Other variables might be valuable to describe the characteristics of the accident, but not to predict its severity. For example, COLLISIONTYPE, SDOT_COLCODE. In fact, if these features were introduced in the model, they might improve the accuracy. However, this would be misleading. The only reason this might happen is that these features *describe* the severity of the accident, they do not predict it.

The features which were used for the analysis can be classified as the following: time variable (date), coordinates (spatial), weather and road conditions and finally features which describe the location of the accident (ADDRTYPE for instance).