1. Introduction

More than 1.2 million people worldwide die each year in road traffic accidents (RTAs) and another 20-50 million are injured. The rise in the number of fatalities in road traffic accidents is an increasingly serious problem. Indeed, the number of road traffic participants and the number of vehicles in the world are growing every year. Of course, the main factors in road accidents are driving under impact of alcohol and drugs and speeding. These causes of accidents are prevented by means of education for road users and fines. In turn, the negative impact of bad weather conditions, lighting and road quality can be analyzed and preventive measures (information board) for participants can be taken to prevent accidents.

2. Business Background

The main purpose of road accidents analyzing is to save human lives. Hence, the task of collecting, processing and analyzing data on road traffic accidents is important for the following reasons:

- reduction of fatal accidents to save human lives
- reduction in the total number of accidents reduces the number of payments for a number of insurance policies
- reduces the cost of hours of work for police and other rescue services
- reduces the number of traffic difficulties caused by an accident.

In this project only an impact of wether, road and light conditions on collision are investigated.

3. Data

Accident registration data is provided by the City of Seattle and has been recorded since 2004. Data is updated weekly. The data is presented as a table in hard currency format and contains 38 columns and 194673 lines. The target parameter is an accident "severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact. To predict the class of accident, the following parameters of the road traffic situation were selected:

- *SEVERITYCODE* (a code that corresponds to the severity of the collision: 2—injury, 1—prop damage);
- *WEATHER* (A description of the weather conditions during the time of the collision)
- *ROADCOND* (The condition of the road during the collision)

- *LIGHTCOND* (The light conditions during the collision)

4. Data Preprocessing

All data are checked for the presence of missing information. Any rows with missing information will be removed from the data frame to create a Machine learning model.

1. Remove all severity data if driver involved was under the influence of drugs or alcohol or speeding or inattention were factors in the collision. There are 4 values for driving under the influence of alcohol and drugs. In this case it was decided to consider 0 and H as 0 (there is no influence of alcohol or drug intoxication) and 1 and Yes as 1.

2. Remove all rows with «Unknown» and «Other» categories from data frame.

3. Remove all NaN rows.

4. Replace text in string format which describes wether, road and light conditions with a series of binary elements.

5. Balance data frame.

```
In [35]: dftd['SEVERITYCODE'].value_counts()

Out[35]: 2    40857
         1    40857
         Name: SEVERITYCODE, dtype: int64
```

After Balancing

6. Normalize data frame.