Predicting Motor Vehicle and Pedal Cycle Accident Severity

Qazi Mussabbir

18th October 2020

Contents

1. Introduction

1.1 Background

Road accidents are a common problem in almost all large cities across the world. Capturing the data on road accidents will help government and local transport authorities to take appropriate steps to reduce the number of road accident injuries and fatalities. The data set used in this assignment provides Seattle city's accident severity for motor vehicles and pedal cyclists from January 2004 till May 2020.

1.2 Problem

There a number of data attributes that might contribute to determining the severity of motor vehicle and pedal cyclist's accident and they include - location, road condition, weather condition, car speeding, light, attention level, etc. The project aims to develop a predictor classifier for predicting accident severity in terms of either: injury (1), or, property damage (2).

1.3 Interested Parties

The output of the accident severity prediction model can be used by a number of interested stakeholders:

- Public Institutions - Seattle transport and road authorities for planning on putting in place preventive measures for reducing the number of road accidents, taking into consideration certain inputs like particular accident prone locations, weather, light and road conditions. The preventive measures could be multifaceted like placing warning or signs in particular streets or junctions where there are higher chances of accidents, or alerting driver (e.g. through SMS based services) when the road or weather conditions are not favorable for safe driving.
- Automotive and Cycling Manufacturers - Alert drivers/cyclists through in vehicles systems when the road, weather, light, and location leads to a high chance of injury based accidents

2. Data acquisition and cleaning

2.1 Data sources

The dataset used for this project is derived on the overall motor vehicle and pedal cycle accidents in the city of Seattle from the year 2004 till 2020. The data was captured by Seattle Police Department (SPD). The primary objective of the data is to provide a view on the severity of the road accidents based on a number of conditions (i.e. data attributes).

2.2 Data cleaning

The datasets in its original form needed a lot of cleaning and transformation to be rendered fit for use by a model. To start, there were many missing values (i.e. Null/NA) across the columns within the dataset. In total there were 38 columns and 194673 rows. The target variable SEVERITYCODE is unbalanced with a biasness towards the 'property damage' (value 1) versus 'Injury' (value 2) with a balance inequality of 42%.

2.2.1 Initial Feature Selection

As mentioned in section 1.2, the main purpose of the model is to predict the severity of a potential road accident by considering a number of factors like road conditions, weather conditions, driver attention level, location etc. Any attributes which are deemed to not influence the objective of the model has been dropped. In total 17 attributes/columns have been dropped and we have kept 21 attributes at an initial level for exploratory purposes and modelling purposes. It should be noted that the feature set chosen at this stage is set change following the exploratory analysis and modelling phases, from which further attributes could be deduced, or added and transformed as required for improving the model performance.

The columns which have been dropped are - 'OBJECTID', 'INCKEY', 'COLDETKEY', ‘REPORTNO', 'STATUS', 'INTKEY', 'LOCATION', ‘EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SDOT_COLCODE', ‘SDOT_COLDESC', 'PEDROWNOTGRNT', 'ST_COLCODE', ‘ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY'

2.2.1 Dealing with missing values

Any attribute which has missing values over 60% of its total column size has been dropped. From the initially chosen feature set, the INATTENTIONIND attribute has been dropped because of a very large percentage missing values.

2.2.2 Replacement of missing values

The following attributes missing values have been replaced the maximum frequency value of the column: ‘ADDRTYPE’, ‘SEVERITYDESC’, ‘COLLISIONTYPE’, ‘JUNCTIONTYPE’, ‘UNDERINFL’, ‘WEATHER’, ‘ROADCOND’, ‘LIGHTCOND’

2.2.3 Transformation of categorical variables

Some of the categorical variables had many levels of which quite a few are irrelevant. The levels within these attributes have been reduced by merging similar levels. The impacted attributes in these cases are: ‘JUNCTIONTYPE’, ‘WEATHER’, ‘ROADCOND’, ‘LIGHTCOND’

2.2.3 Transformation of date time objects

The two attributes – ‘INCDTTM’ and ‘INCDATE’ were initially object types. They needed to be changed to a datetime type object. The ‘INCDTTM’ and ‘INCDATE’ attributes were further transformed to int64 objects called ‘hourofday’ and ‘dayofweek’ respectively to indicate which hour (e.g. 14) or which day of the week (e.g. 4).

2.2.4 Drop rows

Geo location attributes ‘X’ and ‘Y’ had 5334 records missing records which were dropped.