Predicting the Severity of an Accident

Lyndall Fourie

September 2020

1. Introduction

1.1 Background

Every year millions of people across the world are involved in road accidents. The severity of these accidents differ and an understanding of which factors impact the severity of an accident may assist emergency response units, hospitals and road side assistance companies to plan accordingly. The insight may also assist motorists to plan their journeys.

1.2 Problem

Many factors may influence the severity of a road accident. The selection of factors and the way each factor contributes to the severity of an accident are often difficult to assess. Machine Learning (ML) will assist in understanding the patterns and make predictions that will assist the target audience to predict the severity of an accident in order to plan accordingly.

1.3 Scope

The focus of this study is to develop a ML model to assist Seattle City in predicting the severity of accidents within its border. Various factors may influence the severity of the accidents and will be assessed for use in the prediction model, but only factors that may influence the severity of the accident before it happens will be included in the scope. For instance, both weather and the angle of impact may influence the severity of the accident, but the target audience can only act according to the weather.

2. Data acquisition and cleaning

2.1 Data source

The data source is the IBM Capstone Project dataset. The following is important to note:

- The data was provided by the Seattle Traffic Department and contains information on accidents during the period of 1 January 2004 to 20 May 2020;
- The data set contains data of accident classes based on the severity of accidents. The label for the data set is therefore severity, which describes the fatality of an accident;
- The columns contains various factors (features) that may influence the severity of an accident.
- The data set has a total of 194674 cases; and

- The original data set contained 5 classes namely: *3 – fatality, 2b – serious injury, 2 – injury, 1 – property damage and 0 – unknown.* The data set provided for the Capstone project contains only two classes namely: *2 – injury and 1 – property damage*.

2.2 Data limitations

The following are potential data limitations:

- The data has unbalanced labels which will lead to a biased ML model. It should therefore be balanced during the course of the project;
- The data set contains incomplete records and fields with incorrect formats. Data cleaning will be required to ensure the data set contains information that will be useable by the ML algorithm;
- Features engineering will be required to increase the predictability of the model; and
- The exclusion of 3 of the 5 severity classes from the data set may have an impact on certain statistical features of the data i.e. statistical significance between the features of severity classes 3 and 1 would have been high whereas the statistical significance between the features of severity classes 2 and 1 may be statistically insignificant. This may have an impact on the usability of the ML model.

2.3 Feature engineering

The sourced data set contained 38 features (columns) that were both relevant and irrelevant to the scope of the project namely, *to predict the severity of accidents within the Seattle City border*. The features that are relevant to the scope of the project are listed in column 1 of Table 1 and were included in a new data set. The features that are not relevant to the scope of the project are listed in column 2 of Table 1, with column 3 describing the reason why it is not relevant to the scope of the project.

It may be useful to understand the impact of the 'Day of the Week', 'Month of the Year' and 'Time of Day' features on the severity of an accident. These features were derived from the INCDATE and INCDTTM columns and listed in column 1.

Table 1: Features included and excluded from the project data set

Columns kept in the data set	Columns dropped (excluded) from the data set	Reason for exclusion from the data set
SEVERITY CODE	X	More than 10 000 unique entries
OBJECTID	Y	More than 10 000 unique entries
SEVERITYDESC	INCKEY	More than 10 000 unique entries
PERSONCOUNT	ADDRYPE	Similar to JUNTIONTYPE
PEDCOUNT	SEVERITYCODE	Duplicate column
PEDCYLCOUNT	STATUS	No value for the analysis
VEHCOUNT	COLDETKEY	More than 10 000 unique entries
INCDATE	REPORTNO	More than 10 000 unique entries
INCDTTM	INTKEY	Too many empty fields (>50%)
JUNTIONTYPE	LOCATION	More than 10 000 unique entries
WEATHER	EXCEPTRSNCODE	Too many empty fields (>50%)
ROADCOND	EXCEPTRSNDESC	Too many empty fields (>50%)

Columns kept in the data set	Columns dropped (excluded) from the data set	Reason for exclusion from the data set
LIGHTCOND	SDOT_COLCODE	Will not predict the severity of an accident before it happens.
Day of the Week	SDOT_COLDESC	Will not predict the severity of an accident before it happens.
Month of the Year	INATTENTIONID	Too many empty fields (>50%)
Time of Day	PEDROWNOTGRNT	Too many empty fields (>50%)
	SDOTCOLNUM	More than 10 000 unique entries
	SPEEDING	Too many empty fields (>50%)
	COLLISSIONTYPE	Will not predict the severity of an accident before it happens.
	UNDERINFL	No indication of what 0 and 1 refer to.
	SEGLANEKEY	Too many fields with '0' (>50%)
	CROSSWALKKEY	Too many fields with '0' (>50%)
	HITPARKEDCAR	Too many fields with 'N' (>50%)
	INCDATE	Replaced by 'Day of the Week' and 'Month of the Year'.
	INCDTTM	Replaced by 'Time of Day'.

2.4 Data Cleaning

The following data cleaning activities were performed:

- Created a new data frame with the selected features (16 columns);
- Identified the columns with missing values:
    - CCOLLISIONYPE (4904) - 2.5% of dataset
    - JUNCTIONTYPE (6329) - 3.25% of dataset
    - WEATHER (5081) - 2.6% of dataset
    - ROADCOND (5012) - 2.6% of dataset
    - LIGHTCOND (5170) - 2.7% of dataset
- Further analysis on the csv dataset indicated that in 99.57% of the cases, where the COLLISSIONTYPE field is empty, the WEATHER, ROADCOND and LIGHTCOND fields are also empty. The empty COLLISIONTYPE fields will therefore not contribute to the scope of the analysis and were deleted. Further analysis indicated similar scenarios for WEATHER, ROADCOND and LIGHTCOND. All rows with empty fields in these columns were therefore deleted.
- The empty values in the JUNCTIONTYPE column were replaced with 'Unknown'.
- In order to derive the 'Day of the Week', 'Month of the Year' and 'Time of Day' the INCDATE and INCDTTM column formats were changed to 'datetime' format.

Number of cases were reduced from 194 673 to 189 316.

The formats of the columns, which will be used for further analysis, are in the format of 'int64', 'object' and 'datetime64'. This will be sufficient for exploratory analysis. To enable machine learning, all columns will be transformed to the appropriate formats at the specific stage in the project.

As discussed in section 2.2, the data has unbalanced labels which will lead to a biased ML model. It will therefore be balanced during the further course of the project.

3. Exploratory Data Analysis

4. Predictive Modelling

5. Conclusion and future work