Predicting Car Accident Severity: A Machine Learning-Based Approach

Sumanta Sannigrahi

28 september 2020

1. Introduction

This section outlines the problem context. Although every licensed driver will have been exposed traffic accidents somewhere down the line, we describe the problem background and interest here. Furthermore, we formalize the problem within a problem statement

1.1. Background

With the increasing use of motorized vehicles, came a vast pressure on the network of highways. In large cities such as Seattle, this often leads to traffic congestion and as a result, accidents. Those incidents may leave other road users in peril, requiring action from emergency services to restore safety. Each accident is different in terms of severity, it is therefore often unknown how many emergency vehicles should be sent out. In high-severity cases, emergency aircrafts and fire brigade involvement may be required too. It is advantageous for those institutions to accurately predict how many vehicles and manpower are required to mitigate the impact of car accidents. A pre-emptive assessment of car accident severity leads to more complete information on the means required to escort each casualty to safety and restore the infrastructure.

1.2. Problem

This study aims to predict the severity of a car accident according to historical data. Emergency services often receive incomplete information when informed about collisions. Therefore, they struggle to assess how many resources they should allocate to mitigate the accident. Based on historical data, machine learning tools may aid them in resource allocation by predicting collision severity according to data provided by civilians. This study aims to bridge the gap between past accidents and future emergency assistance.

1.3. Interest

Emergency services would reap large benefits from early information on the severity of incoming accidents. These benefits mainly apply to more efficient resource planning, optimization of personnel use and more effective mitigation. Furthermore, road users are stakeholders as well, since they benefit from faster accident resolution and experience enhanced safety when involved in accidents themselves.

2. Data

This section describes the data requirements for this study. The data set of choice is introduced, as well as the steps for data cleaning and feature selection. The latter consists of a list of features that serve as input to predictive modelling, to be used by machine learning algorithms later.

2.1. Data source

Data on collisions within the city of Seattle is available through an Open Data initiative hosted by Seattle GeoData (Collisions, n.d.). It is noted this dataset is the example dataset provided for this

capstone project. Since it provides an extensive set of attributes, it was decided to use it as main data source for this project.

2.2. Data cleaning

The data retrieved from Seattle's Open Data bank contained an abundance of null values, along with various other issues. It was chosen to drop every row containing null values for the attributes chosen for feature selection. Due to insufficient record keeping, it was undesired to drop all rows with null values for each attribute, as this led to dropping rows that may include values we can use for model building.

Several issues arose during data cleaning. Besides null values, various columns include 'Unknown' entries. For some attributes, the occurrence of missing values, labelled as unknown, is as high as 12000. However, it was chosen to not drop these values. This mainly stems from the fact that emergency services may not always receive perfect information ahead of coming to the rescue. In order to increase predictability of the model for cases with limited information, unknown values were chosen to be left in place.

Secondly, the inclusion of categorical data required a variety of data transformation operations. Attributes such as junction type, weather, road condition and light condition hold categorical values, rather than numerical ones. In order to prepare those attributes for machine learning algorithms, they were transformed through one-hot encoding.

Thirdly, speeding data contained a significant amount of null values, as only 9339 out of 188617 rows were filled out. A first exploration of the attribute resulted in finding no entries were recorded in case speeding did not occur. That is, speeding data only concerns cases in which speeding occurred, rather than stating 'no' if speeding did not apply. In order to surmount this issue, each NA value was replaced by 'No', as to represent speeding did not occur.

Finally, collision location was transformed into a feature representing whether the accident occurred at a high risk location. Originally, the location attribute yields a description of the collision location. The top 20 most occurring descriptions were labelled as a high risk location, as where others were not.

After fixing these problems, attributes with numerical values were checked for outliers. For vehicle count, collisions between seven cars or more amounted for <0.05% of the total data and were therefore dropped. Likewise, collisions involving more than 10 people were also relatively rare and were therefore dropped as well (<0.2% of the total dataset).

2.3. Feature selection

Data cleaning resulted in a data set with 188,617 rows and 40 attributes. Only a smaller subset of those features are viable for machine learning algorithms. First, the data set includes metadata for each recorded collision such as an incident key and a unique identifier. Those were all dropped. Features regarding date and time were dropped, as they incorporated several problems. Date and time were often not fully filled out, causing time to often be incomplete. If one were to be interested in investigating the relationship between time of day and the likelihood of collisions, this would not be possible with this data. As where it would be interesting to investigate which day of the week, the transformation from dates to weekdays was deemed too expensive for this particular research.

Since emergency services are the problem owner for this study, we assume the machine learning model will have to forecast collision severity based on incoming civilian reports. Therefore, attributes related to codes and descriptions provided by officials are not viable. Furthermore, civilians are

unable to distinguish between injuries, serious injuries and fatalities. Therefore, they are not included within the selected features.

Some attributes were redundant taking into consideration other features. For example, including the amount of people involved, renders the count of pedestrians and bicycles rather obsolete. Those features are significantly narrow focused. Here, a more general approach to the amount of people involved is assumed.

The data set yields a multitude of features with binary data. This includes whether the collision occurred due to speeding or inattention, whether the person was under the influence of drugs or alcohol or whether a parked car was hit. It was chosen to include speeding and intoxication/drug use as features, since data exploration unveiled those cases are prone to higher accident severity. Severe cases require more extensive effort and resources by emergency services. It is therefore paramount our model encapsulates such information. Intoxication or drug use may not always readily be assessed by civilians, yet it is assumed high levels of substance abuse reflect in driver behaviour.

The final set of features includes nine attributes, which yield the following characteristics: 1) no expert view is required to assess their value and 2) they are not highly correlated to other features within the selection. Table 1 provides an overview of the feature selection process.

Kept features	Dropped features	Reason for dropping features
High risk location, JUNCTIONTYPE	X, Y, LOCATION, CROSSWALKKEY	Aggregated into binary high risk location (Y/N), included JUNCTIONTYPE as additional predictor.
PERSONCOUNT, VEHCOUNT	PEDCOUNT, PEDCYLCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES	Dropped features that cannot be assessed by civilians. Aggregated people into PERSONCOUNT.
WEATHER, ROADCOND, LIGHTCOND	-	Included to assess environmental conditions.
-	INCDATE, INCDTTM	Incomplete, too expensive to transform.
SPEEDING, UNDERINFL	INATTENTIONIND, PEDROWNOTGRNT, HITPARKEDCAR	Opted for two binary variables that imply higher severity.
-	COLLISIONTYPE, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY	(Meta)data provided by Seattle's emergency services, excluded as civilians cannot assess them.

References

Collisions. (n.d.). Retrieved September 28, 2020, from https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions