Analysis of Road Accident Severity

Anil Sasidharan

1. Introduction

1.1 Background

Road safety has always been a concern for any city's governing bodies, law enforcement officers, analysts and the inhabitants. Ideally, it's a constantly improving metric and the level of improvement would always be a success rating for the administration. Each year, new methodologies and technologies come into the fore predicting and ensuring better road safety. Some promise and also deliver results, whereas some innovations don't generate the numbers expected. The measure of the success of these methods and technologies are always done with data, and its analysis. Data science is proven to be a huge asset in this field, among many others, and will be a strong tool in improving road safety in the future.

The Seattle Department of Transportation (SDOT) is a municipal government agency in Seattle, Washington that maintains the city's transportation systems, including roads, bridges and public transportation. The SDOT has an estimated $20 million of transportation assets, with a citywide street car network and a good bicycle infrastructure. The SDOT provides data on the collisions that took place inside the department's limits along with the details such as location, nature of collision, people and aspects involved, environment conditions and the collision severity. This is the data I will be using to perform an analysis so as to unearth patterns which will provide important inferences and also shed light on how effectively it is possible to predict the severity of future collisions.

1.2 Problem

The aim of this study is to show that road accident data can be used in tackling road safety in the future. The history of road accidents are classed by Severity index. In the future, the successful tackling of safety should result in a measurable decrease of severity index across road incidents. The principle followed in this report will be to analyze the available data of road incidents with their severity codes and study the relation with factors like road and environment conditions, the person and vehicle conditions and similar other factors. Using this proper study, I can build a model which will be able to predict the severity of a road incident using the prevailing conditions.

1.3 Stakeholders

Safety should be of utmost concern for any governing body. The guarantee of safe movement of population increases the confidence of the general public, their capability in work and decreases individual and collective stress, not to mention the pride in a safe locality. These leads to a rise in general productivity of a region. Due to these factors, a study of road safety will be of interest not only to road transport and maintenance authorities, but also to state administrations and economic players.

With the results of this study, we can expect to gain a valuable insight into what a severe road accident depends on, and work towards mitigating those concerns. This will result in bringing a new, safer environment for general public travel.

2. Data for the project

2.1 Source

The data for this study is provided by The Seattle Department of Transportation data portal. It includes all the collisions over the years from 2004 to present, recorded by the Seattle Police Department (SPD) and by Traffic Records. This database is updated weekly.

The database consists of several descriptive information about a collision, including a Severity Code or Severity Index. This will be my target variable, whose value I will be able to predict for future collisions, at the end of the analysis. The Severity Code for each collision is classified 1 to 3, with 3 being the most severe case.

2.2 Data preparation and Feature selection

The attributes of the database included codes given by the SDOT for identification, location information, description of the collision and severity, number of actors involved, number of injuries, date and time, and factors causing the collision, including road and weather conditions and driver error situations.

To start, I discarded the featured used for identification which do not make much sense in analysis. The following features were identification information:

- ObjectID
- INCKey
- ColdetKey
- IntKey
- ExceptRSNCode
- SDOT_ColCode

- SDOTColNum
- SegLaneKey
- CrosswalkKey

The features with information on number of injuries and their details can be considered to be a duplicate of the Severity code as they do not provide any extra relevant information applicable to the analysis. The date of collisions too is not required for the analysis, but the time of day is an important indicator, so it is kept. For this, the feature Date & Time is used to extract Time information and saved as a new attribute. The remaining are the features involving factors causing the collisions. These are the most important for this analysis and are retained.

Some features required a bit of data cleaning, detailed as follows:

- Time of collision information was part of a Date & Time attribute. So I extracted it to form a different feature to enable easier analysis.
- Each factor contributing to a collision had a number of values. So average values of the severity code have to be used to compare for the different values of the factors. So for each factor, I split the data for each value, calculating average values of Severity index.
- The feature 'Driving under the Influence' had 4 different values which actually denoted only 2 conditions. I performed data cleaning for this feature to denote 2 values.
- The attributes showing Driver error conditions had many values as NA. This will cause a difficulty in calculation of severity index in the analysis. So I replaced all the NA values to show 'N' or negative, as these are Boolean attributes.

2.3 Principle of use

To give a brief idea of the methodology I will follow in the data analysis, I will first identify and classify the more significant factors which affect the severity index of a collision. These factors will be considered individually at first, and the relation with the severity index will be analyzed. The second step will be to identify which of these factors would have a combined effect. Some factors could be accumulative and others could be cancelling. It is also possible that the same factor can show a different relation when combined with another factor. After the analysis of the combined factors, I will have a better idea of the most significant factors and their relations. The third step will be to build a model using this information which can predict the value of severity code.

3. Methodology
3.1 Study of available data
3.2 Initial inferences
3.3 Relationship between factors
3.4 Model building
3.5 Predictions
4. Results
4.1 Observed trends
4.2 Predicted trends
5. Discussion
6. Conclusion