Predicting the Severity of a Car Accident Using Seattle Collision Data

Pablo Sanhueza Toro

September 27, 2020

2. Data acquisition and cleaning:

1.1 Data sources

The collisions data is recorded by Seattle Department of Transportation, Traffic Management Division, Traffic Records Group and maintained by ArcGIS and can be found in the Kaggle dataset here and the metadata can be found from here. The collisions dataset includes all types of collisions from year 2004 to present. The collisions dataset is what I‟m going to be working with in this project.

1.2 Data Cleaning

The data downloaded from the above mentioned source was saved in a table. There were several columns (features) in the dataset that were added by various authorities to mainly identify the incident instance and were irrelevant to our final goal of developing a predictive model. Those columns were dropped. Our aim is to build a predictive model which can foretell or predict accurately the severity of any future accident given the conditions (features) that can be pre specified. There were several columns in our dataset which only made sense only when a collision had already taken place; so those columns were also dropped. Also we want our model to be independent of the exact locations of the collisions, therefore columns indicating the exact locations of the collisions were dropped and rather the columns describing the properties of such locations were kept. Below is a table of the dropped columns and the reason for their dropping:

Table 1: Dropping of irrelevant columns (features) during data cleaning

Dropped Column	**Reason for dropping**
OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SDOTCOLNUM	Irrelevant. Added by authorities only for identification purpose.
ST_COLCODE, ST_COLDESC, PEDROWNOTGRNT, COLLISIONTYPE, INJURIES, SERIOUSINJURIES, FATALITIES, SDOT_COLCODE, SDOT_COLDESC, HITPARKEDCAR	Can be known only after a collision.
SEVERITYDESC	Irrelevant. SEVERITYCODE kept instead.
INCDATE	Irrelevant. Prediction should work for any given day.

INCDTTM	Model to be independent of time. LIGHTCOND kept instead.
X, Y, LOCATION, SEGLANEKEY, CROSSWALKKEY	Model to be independent of exact location. ADDRTYPE, JUNCTIONTYPE kept instead.

There were a lot of missing values as well as data inconsistencies in many of the kept features fields. For example, the „ADDRTYPE" feature which is a categorical variable, describes the types of the addresses where the collisions took place had 3721 missing values. I did not drop this column rather replaced the missing values with the label "others", as it seems a potential candidate in determining the severity of an accident and hence could be a valuable input to our model. In the exact similar way I also handled the missing values for the categorical features „JUNCTIONTYPE", „WEATHER", „ROADCOND", „LIGHTCOND" by replacing the missing values with the label "others" for the same reason.

The „INATTENTIONID" feature which is a categorical variable and supposed to have a „Y" for „yes" and „N" for „no" values, but instead it had data inconsistency where the „yes" values were labelled by „Y" but the „no" values were kept as missing values. I replaced all the missing values with 0"s and replaced the „Y" labels with 1"s.

The „UNDERINFL" feature had data inconsistency too. Some of the „yes" values were labelled as „Y" whereas the others as 1"s and in a similar fashion some „no" values were labelled as „N" and the others as 0"s. I handled the situation by converting all the 1"s to „Y"s and all the 0"s to „N"s. Moreover this feature also had 26479 missing values which I labelled as „UN" which stands for „unknown".

The „SPEEDING" feature had data consistency where the „yes" values were correctly labelled as „Y"s but the „no" values were indicated using missing values. I replaced the missing values with 0"s to indicate a „no" and the „Y"s to 1" to indicate a „yes".

At last I checked whether the data-types of the features were consistent with the values they contained or not. I found out the „INATTENTIONID" feature had a data-type inconsistency where the data-type of this feature was „object" but the values it contained were 0"s for a „no" and 1"s for a „yes". Therefore I converted the data-type of this feature from „object" to „int64" to make it consistent with the values it contained.

1.3 Feature Selection

After data cleaning there were 221265 samples and 13 features in the cleaned data. To better select the features the Pearson correlation coefficient was calculated for the dataset to get an idea about the extent to which the features were correlated to the target variable „SEVERITYCODE". As Pearson correlation function only works on numeric features I had to encode the categorical features into numeric values and then calculate the Pearson correlation.

Examining the Pearson correlation of different features with the target variable I found out that none of the features displayed a high positive or negative correlation, with
„WEATHER“ being the most negatively correlated with a Pearson correlation coefficient of negative 0.46. This is understandable because none of the features contained continuous values and most of them were categorical and they were encoded into numerical values which were discrete in nature.

Depending on the above observation none of the features were dropped and all of them were kept to be better understood during the Exploratory Data Analysis phase.