Predicting the Severity of Traffic Accidents

Seshni Govender

September 2020

1 Introduction

1.1 Background

Traffic accidents are a significant cause of deaths, injuries, property damage and loss, and are major concerns for the public health and traffic safety. According to an article by the World Health Organization (WHO), approximately 1.35 million people die every year as a result of traffic accidents [1]. These accidents have negative effects on the direct participants of the incident physically and emotionally , and may incur considerable economic loss. Hence, given several factors, predicting the severity of such accidents can be useful since with this information a road user may be prompted to drive more carefully, use a different transportation route or possibly plan accordingly for such events .

1.2 Problem

Data that may determine the severity of a traffic accident may include road conditions, weather conditions, light conditions, alcohol or drug influences , whether the drivers were speeding or distracted, location and time, and number of participants and vehicles involved. This project aims to predict the severity (Injury or only property damage) of the traffic accident based on these data.

1.3 Interest

Road traffic safety institutions such as traffic police may be highly interested in the prediction of traffic accident severity to assist in improving overall road traffic safety. For example, if severe accidents are likely to occur at a certain time e.g. peak traffic hours, then more traffic police patrols can be dispatched to ensure that drivers adhere to the road rules. The general public may be highly interested in these predictions so that road users are likely to drive carefully, use different traffic routes or plan for a possible accident. Insurance companies may also be interested in these predictions to warn their clients of traffic accident possibilities.

2 Data

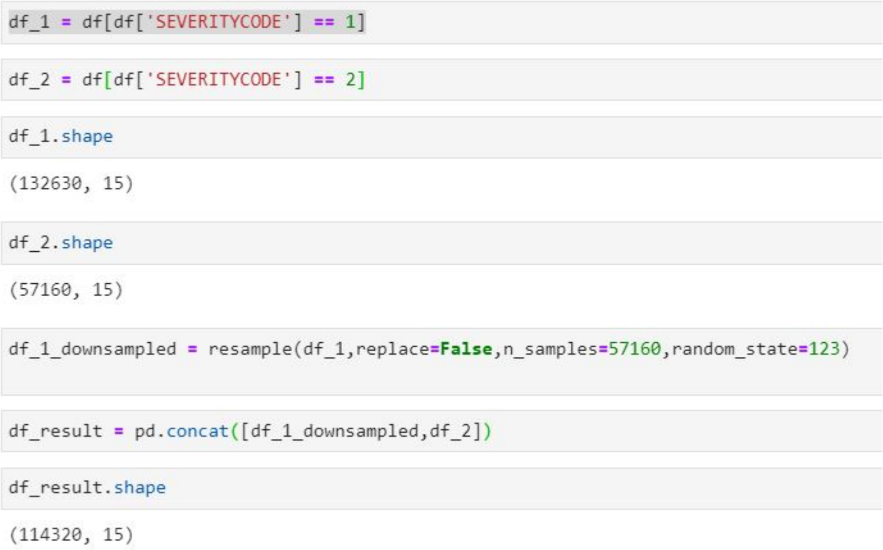
2.1 Data source

A shared data-source and metadata of the traffic accidents provided by the Seattle Police Department and recorded by Traffic Records for the Seattle city from 2004 to the present are used [2] [3]. The target or label is the accident ”severity” in terms of human fatality or injury. There are 37 attributes in the data-set, some of which will not be useful and some contain empty data. The data contains only two types of severity categories:” injury” or ”property damage”. The data-set also has unbalanced labeled data with 136 485 accidents with a ”property damage” severity and 58 188 accidents with a ”injury” severity. This will need to be balanced so that the model will be unbiased. Supervised machine learning will be used to predict accident severity.

2.2 Data Cleaning

The data-set contained a lot of missing data. Rows which had more than half of the values in the columns as empty (”NaN”) were dropped. For example, if the Feature set had 14 columns, then if a row had more than 7 of its column data as empty, it would be dropped.

To balance the data, a down sampling was performed on the majority class severity of type ”property damage” or 1. This left a total sample of 114 320 accident records.



```
df_1 = df[df['SEVERITYCODE'] == 1]

df_2 = df[df['SEVERITYCODE'] == 2]

df_1.shape

(132630, 15)

df_2.shape

(57160, 15)

df_1_downsampled = resample(df_1,replace=False,n_samples=57160,random_state=123)

df_result = pd.concat([df_1_downsampled,df_2])

df_result.shape

(114320, 15)
```

Figure 1: Down sampling of collision data-set in Python

In the ”ADDRTYPE” column which specifies the location of the accident, any unknown or missing row values were imputed using ”Block” since the mass of the accidents occur at a ”Block” address type.

The ”SPEEDING”column had 9000 accidents where the driver(s) were speeding. The rest of the accidents did not have a value for this attribute. Either 1 of 2 options could be selected: (1) The empty values could be treated as a non-speeding accident (2) The ”SPEEDING” column could be dropped altogether. Even though speeding may be a significant factor in predicting a car accident severity, it was decided that option (2) will be selected. It cannot be assumed to take the empty rows in the ”SPEEDING” column as a non-speeding accident since there is insufficient data to make the assumption.

For the ”UNDERINFL” attribute which is whether the driver was under the influence or drugs or alcohol, the data contained a mixture of ”Y” or ”1” for ”Yes” and ”N” or ”0” for ”No”. To be consistent, all string values were changed to integers correspondingly. Any empty values were imputed using the most common class which was ”1” representing the most accidents happening with the driver under the influence of alcohol or drugs.

The date and date-time columns were converted to Pandas date-times so that these values can be used to determine which day of the week the accident occur when selecting features.

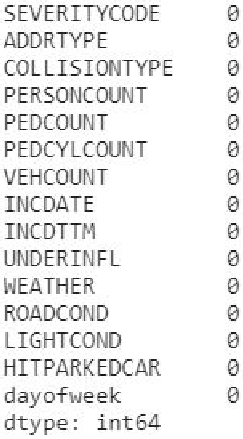
For the weather conditions, road conditions and light conditions columns, any row with a column value of ”Unknown” was also imputed using the most common class in that column.

A similar method was applied to all other categorical columns.

2.3 Feature Selection

After cleaning, the data-set contained 114 320 rows with 14 features. The first step was to determine which are the relevant attributes which can be used to predict the accident severity. There were several columns that have no meaning in predicting the accident severity. For example, the codes used to classify the intersections at which the accidents took place have no meaning in predicting accident severity. The ”LOCATION” and ”JUNCTIONTYPE” columns were removed because these columns contained similar information to the

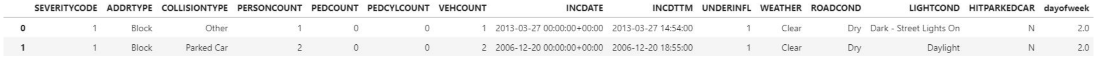
”ADDRTYPE” column. The ”STCOLDDESC” was also dropped because it contained similar information to the ”COLLISIONTYPE” column. These columns were dropped which left the data-set with:



```
SEVERITYCODE      0
ADDRTYPE          0
COLLISIONTYPE     0
PERSONCOUNT       0
PEDCOUNT          0
PEDCYLCOUNT       0
VEHCOUNT          0
INCDATE           0
INCDTTM           0
UNDERINFL         0
WEATHER           0
ROADCOND          0
LIGHTCOND         0
HITPARKEDCAR      0
dayofweek         0
dtype: int64
```

Figure 2: Summary of column names and count of null values

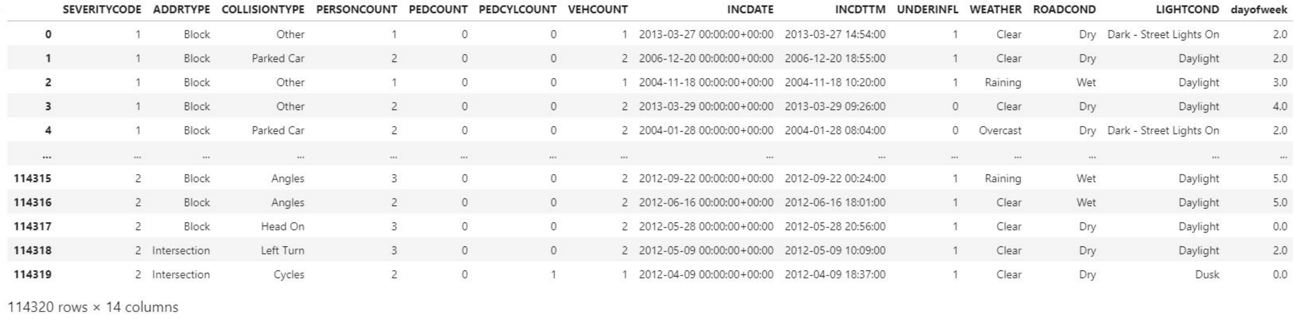
It can be seen that no null values are included in the final dataframe. For the Feature data-set, a new column ”dayofweek” was added which will be used to determine which days of the week an accident is most likely to be severe. The date and date-time columns will therefore not be used in the Feature set. The column ”HITPARKEDCAR” was also removed because the data did not line up with the collision type. For example, many of the collision types were ”Parked Car” but the ”HITPARKEDCAR” column did not classify it as a ”Y” for ”Yes”.



	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INCDATE	INCDTTM	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	HITPARKEDCAR	dayofweek
0	1	Block	Other	1	0	0	1	2013-03-27 00:00:00+00:00	2013-03-27 14:54:00	1	Clear	Dry	Dark - Street Lights On	N	2.0
1	1	Block	Parked Car	2	0	0	2	2006-12-20 00:00:00+00:00	2006-12-20 18:55:00	1	Clear	Dry	Daylight	N	2.0

Figure 3: Example of collision type correspondence with ”Hit parked car” class

Finally, this will leave 11 features or attributes used to predict the accident severity. A sample of the final data-set is shown below.



	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INCDATE	INCDTTM	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	dayofweek
0	1	Block	Other	1	0	0	1	2013-03-27 00:00:00+00:00	2013-03-27 14:54:00	1	Clear	Dry	Dark - Street Lights On	2.0
1	1	Block	Parked Car	2	0	0	2	2006-12-20 00:00:00+00:00	2006-12-20 18:55:00	1	Clear	Dry	Daylight	2.0
2	1	Block	Other	1	0	0	1	2004-11-18 00:00:00+00:00	2004-11-18 10:20:00	1	Raining	Wet	Daylight	3.0
3	1	Block	Other	2	0	0	2	2013-03-29 00:00:00+00:00	2013-03-29 09:26:00	0	Clear	Dry	Daylight	4.0
4	1	Block	Parked Car	2	0	0	2	2004-01-28 00:00:00+00:00	2004-01-28 08:04:00	0	Overcast	Dry	Dark - Street Lights On	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
114315	2	Block	Angles	3	0	0	2	2012-09-22 00:00:00+00:00	2012-09-22 00:24:00	1	Raining	Wet	Daylight	5.0
114316	2	Block	Angles	2	0	0	2	2012-06-16 00:00:00+00:00	2012-06-16 18:01:00	1	Clear	Wet	Daylight	5.0
114317	2	Block	Head On	3	0	0	2	2012-05-28 00:00:00+00:00	2012-05-28 20:56:00	1	Clear	Dry	Daylight	0.0
114318	2	Intersection	Left Turn	3	0	0	2	2012-05-09 00:00:00+00:00	2012-05-09 10:09:00	1	Clear	Dry	Daylight	2.0
114319	2	Intersection	Cycles	2	0	1	1	2012-04-09 00:00:00+00:00	2012-04-09 18:37:00	1	Clear	Dry	Dusk	0.0

114320 rows × 14 columns

Figure 4: Final dataframe

3 References

1 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

2 https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

[3] https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf