Seattle Car Accident Severity Analysis Report

Michael Sheppler 11/1/2020

1. Introduction/Business Problem

The World Health Organization estimates that 1.35 million people die each year because of road traffic crashes. It is the first cause of death among children aged 5-14 and among young adults aged 15-29. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities. By using machine learning (ML) predictive modelling, I will measure the influence of independent variables on the severity of accidents.

Several factors contribute to road crashes and resulting deaths and severity of injuries. By understanding these factors, accidents could be prevented, and insights gained which can be shared with the respective drivers of the vehicles so that they can make informed road safety decisions. They are the target audience of this report using Machine Learning (ML) predictive modelling.

2. Data

The data were downloaded from the Example Dataset from the Coursera Applied Data Science Capstone Project. The data are a subset derived from the Seattle Accident Traffic Records Department. The URL is https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.The data consists of 38 independent variables and 194,673 rows.

The dependent variable, “SEVERITYCODE”, contains data that correspond to different levels of severity. This data is a broader categorization of the data found in the Metadata file that accompanies the dataset.

Attribute	**Description**
SEVERITYCODE	A code that corresponds to the severity of the collision: • 1—property damage • 2—injury

The data collection also includes attributes which include **unsafe road user behaviors**, such as whether the driver was speeding or under the influence of drugs or alcohol and **environmental attributes** such as the weather, road conditions and the light conditions reported at the time of the accident. I have selected the following environmental attributes to accurately build a model to prevent future accidents or reduce their severity.

Attribute	Description
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision

This dataset will be used to analyze WEATHER, ROADCOND and LIGHCOND to determine whether they have any relationship on the Severity of the accident.

3. Methodology

3.1 Data / Preparation

Data columns not needed for data analysis were dropped from the dataset.

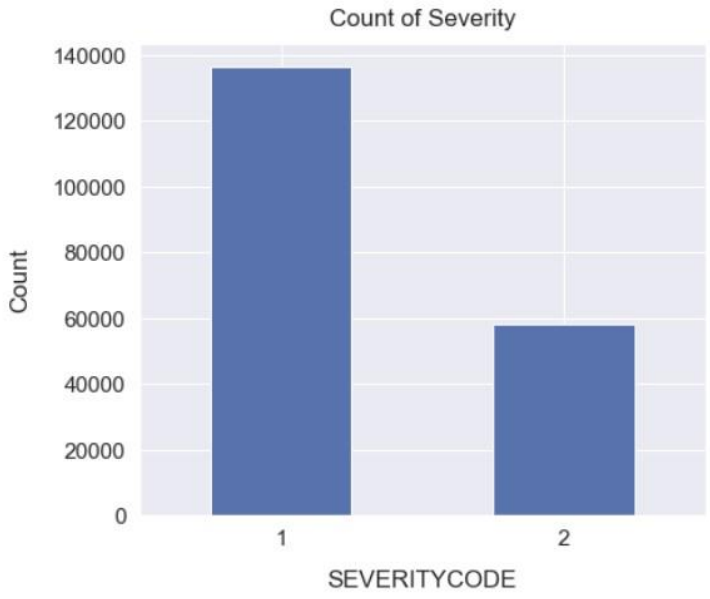
Dropped columns	Kept columns
'OBJECTID', 'SEVERITYCODE.1', 'REPORTNO', 'INCKEY', 'COLDETKEY', 'X', 'Y', 'STATUS','ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYDESC', 'INCDATE', 'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC', 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR', 'PEDCOUNT', 'PEDCYLCOUNT', 'PERSONCOUNT', 'VEHCOUNT', 'COLLISIONTYPE', 'SPEEDING', 'UNDERINFL', 'INATTENTIONIND'	'SEVERITYCODE', 'WEATHER', 'ROADCOND', 'LIGHTCOND',

The remaining columns were then converted to categories and variables assigned for model analysis.

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	WEATHER_CAT	ROADCOND_CAT	LIGHTCOND_CAT
0	2	Overcast	Wet	Daylight	4	8	5
1	1	Raining	Wet	Dark - Street Lights On	6	8	2
2	1	Overcast	Dry	Daylight	4	0	5
3	1	Clear	Dry	Daylight	1	0	5
4	2	Raining	Wet	Daylight	6	8	5

3.2 Data / Exploration

I created a histogram of the SEVERITYCODE values.



Bar 1 property damage Bar 2 Injury

I also examined the value counts for the “WEATHER”, ” ROADCOND” and “LIGHCOND” columns. They are as follows:

```
WEATHER
Clear                     111135
Raining                    33145
Overcast                   27714
Unknown                    15091
Snowing                      907
Other                        832
Fog/Smog/Smoke               569
Sleet/Hail/Freezing Rain     113
Blowing Sand/Dirt             56
Severe Crosswind              25
Partly Cloudy                  5
```

```
ROADCOND
Dry               124510
Wet                47474
Unknown            15078
Ice                 1209
Snow/Slush          1004
Other                132
Standing Water       115
Sand/Mud/Dirt         75
Oil                   64
```

```
LIGHTCOND
Daylight                  116137
Dark - Street Lights On    48507
Unknown                    13473
Dusk                        5902
Dawn                        2502
Dark - No Street Lights     1537
Dark - Street Lights Off    1199
Other                        235
Dark - Unknown Lighting       11
```

3.2 Data / Standardization

Because the dependent variable, Severity, was unbalanced and would cause problems with the modeling, I resampled the data to produce a balance data set.

```
2    58188
1    58188
Name: SEVERITYCODE, dtype: int64
```

I defined the two variables X (WEATHER, ROADCOND and LIGHCOND) and Y (SEVERITYCODE). I then normalized the dataset and initiated a Train/Test Split.

3.3 Data / Predictive Modeling

I used 3 machine learning models for my project.

- KNN will be used to predict the severity of an outcome by finding the most similar data point within k distance.
- Decision Tree will give us an outcome of environmental conditions on the severity of the accident.
- Logistic Regression will be used to predict severity outcomes.

4. Results

The following chart represents the accuracy of the models.

Algorithm	Jaccard	F1 - Score	Log Loss
KNN	0.55	0.53	NA
Decision Tree	0.55	0.47	NA
Logistic Regression	0.52	0.51	0.68

Based on the accuracy percentages, we can see that by having median values, the environmental attributes (weather, road conditions and light conditions) have some, but not an overwhelming influence, on the severity of the accidents.

5. Discussion

Given my understanding of the data, I would further analyze other combinations of the data, such as spatial data, time data and unsafe road user behaviors to create a better understanding of the factors that contribute to accidents in the Seattle area. I would recommend providing that information to the public to promote a safer driving experience.

6. Conclusion

I followed the CRISP-DM process model for my Seattle Car Accident Severity Analysis Report. The initial phase began with stating the object of the project which was to predict the severity of accidents in the Seattle by analyzing the environment data found in the database example provided for the Capstone Project. I extracted the dataset and took the following steps.

a) determined what attributes were needed to train my machine learning models
b) dropped superfluous data that were not relevant to the project's objective

c) converted the remaining environmental attributes to categories
d) assigned variables to the categories
e) balanced the severity data set
f) transformed the data for ingestion into the ML predictive models
g) ran the models
h) created a chart to illustrate the accuracy of the models using the data

This is my first project in which I have utilized the principles and procedures I learned from my Coursera classes for the IBM Data Science professional certificate. I realize it is an introduction to the field, and I hope with more experience, I will gain a deeper understanding of the principles of data science.