# IBM Applied Data Science Capstone

## Predicting Car Accident Severity

Daniel Torrecampo

September 18, 2020

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Background

According to the National Highway Traffic Safety Administration (NHTSA), Over 36,000 people were killed in traffic accidents in 2018 in the United States. Of those deaths, 6,283 were pedestrians and 857 were bicyclists. As these numbers trend upwards, serious concerns arise from the statistics.

## 1.2 The Problem

Many car accidents can be avoided if sufficient warning is provided in advance. With historical collision data from Seattle Department of Transportation, we can predict the probability of getting into a car accident and provide actionable information to drivers and improve overall safety for the public.

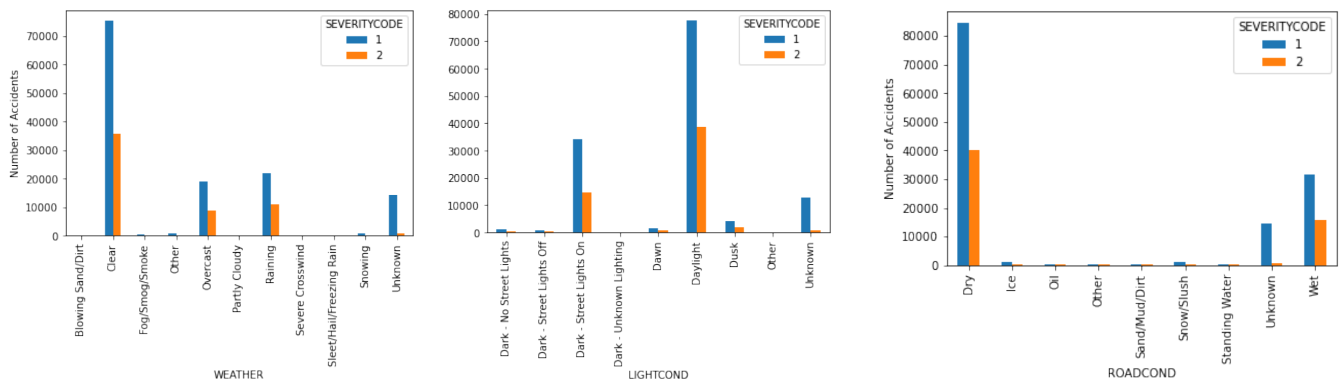
# 2 Data Acquisition

## 2.1 Seattle Department of Transportation

For this case study, we will use accident data from Seattle's Department of Transportation from 2004 to the present. This includes all types of collisions recorded by traffic records which is updated on an a weekly basis. This dataset contains 38 attributes such as location, collision type, injuries, weather, road conditions, light conditions, speeding, vehicle count, pedestrian count, and severity code. We will condition the data and prepare it for a number of analytical models. This data will ultimately help predict the severity of an accident based on the environmental conditions and driver behavior.

## 2.2 Data Exploration

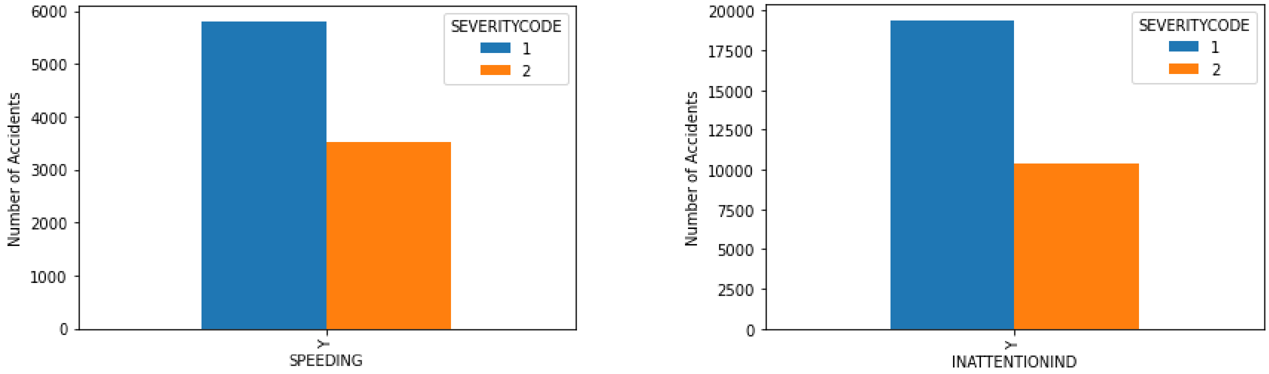
In this report, car accident severity is categorized as either 1 or 2 where 1 is property damage and 2 is injury. Car accident severity will be the target variable for this classification problem. If we can classify the severity, we can provide a model than can predict higher risk situations for drivers and pedestrians.



**(a)** Weather **(b)** Lighting **(c)** Road Conditions

**Figure 1:** Environmental conditions during accidents

Preliminary research appears be inconclusive as most accidents occur during clear weather, broad daylight, on dry roads as shown in **Figure 1**. Based on environmental conditions alone, accidents resulting in property damage occur twice as often as accidents resulting in bodily injury. Both types of accidents seem to spike in number with either of the following conditions: overcast, rain, dark lighting (with street lights on), and wet road conditions. Environmental conditions alone will not provide sufficient context for accidents; therefore, we shall explore driver behavior.



**(a)** Speeding

**(b)** Inattention

**Figure 2:** Driver Behavior that Resulted in Accidents

From the behavioral data presented above, the accidents involving property damage (1) are nearly twice as common than that of injuries (2) from speeding and inattention.

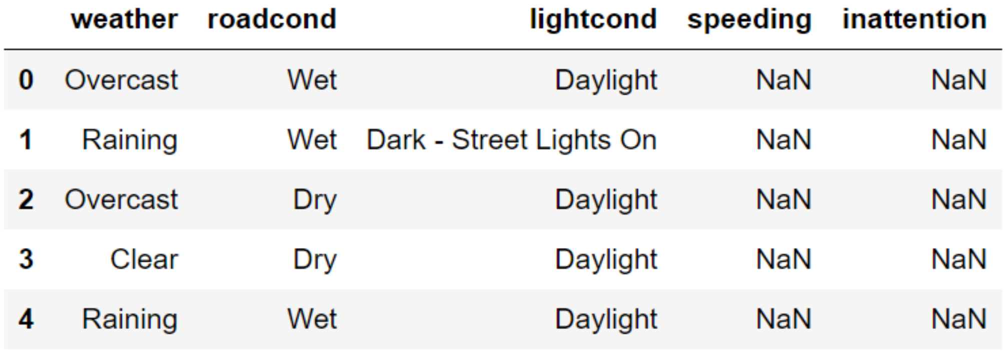
## 2.3 Pre-processing and Data Management

### 2.3.1 Target Variable

We want to predict the severity of accident based on environmental conditions and driver behavior; therefore, our target variable is ”severity” and our indepdendent variables are: weather,road conditions, light conditions, speeding, and inattention. These variables are presented below in **Table**

### 2.3.2 Cleaning Variables and Using Dummy Variables

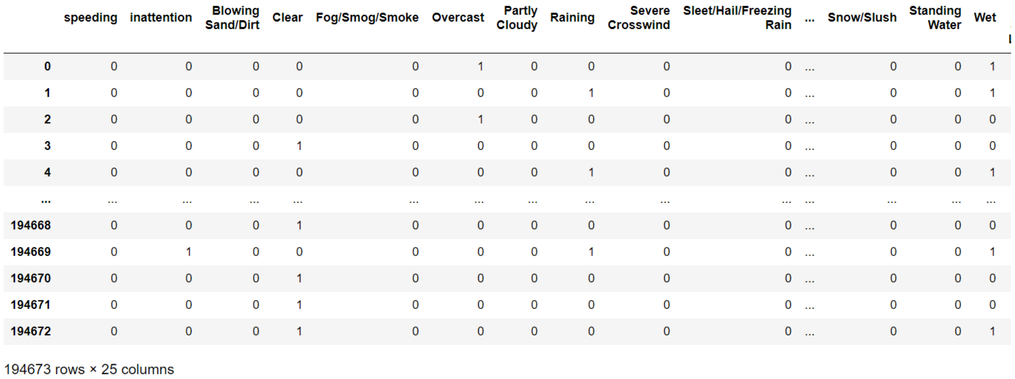
The data must be prepared in such a way that our machine learning model can digest. More specifically, the raw csv data from Seattle’s Department of Transportation, shown in **Figure 3**, contains categorical object datatypes such as ”overcast” or ”dry” that need to quantified with dummy variables. These dummy variables will encode the data and apply tags to each condition (weather/behavioral) for computational methods.



| | weather | roadcond | lightcond | speeding | inattention |
|---|---|---|---|---|---|
| **0** | Overcast | Wet | Daylight | NaN | NaN |
| **1** | Raining | Wet | Dark - Street Lights On | NaN | NaN |
| **2** | Overcast | Dry | Daylight | NaN | NaN |
| **3** | Clear | Dry | Daylight | NaN | NaN |
| **4** | Raining | Wet | Daylight | NaN | NaN |

**Figure 3:** Sample Data from SDOT

The method used to convert categorical object datatypes into dummy variables was *The One Hot Encoding* method. The resulting dataset is shown below in binary like format where the object attributes are added to columns; one indicates that the condition was true and 0 was false. Note, that the number of columns increased from 5 to 25 in this dataset.



| | speeding | inattention | Blowing Sand/Dirt | Clear | Fog/Smog/Smoke | Overcast | Partly Cloudy | Raining | Severe Crosswind | Sleet/Hail/Freezing Rain | ... | Snow/Slush | Standing Water | Wet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 1 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | ... | 0 | 0 | 1 |
| **2** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | ... | 0 | 0 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **194668** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| **194669** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | ... | 0 | 0 | 1 |
| **194670** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| **194671** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| **194672** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 1 |

194673 rows × 25 columns

**Figure 4:** Sample Data from SDOT Encoded with Dummy Variables

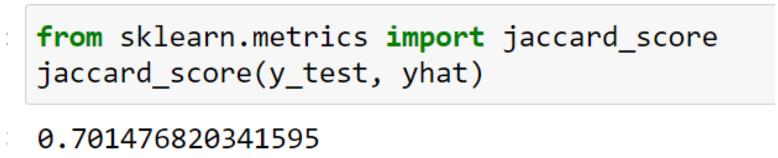
# 3 Methodology

## 3.1 Classification Method: Logistic Regression

To predict the severity level of car accidents, we will need to use a machine learning classifier. The classification method for this case study was logistic regression. This method predicts the outcome of binary events. For example, in our case, whether or not the accident be mild or serious. Logistic regression also calculates the probability of outcome which is critical for our intent - to inform the driver whether he/she will be involved in a serious accident based on his/her driving conditions.

During the training phase, 20 percent of the data was used for testing while 80 percent was used for training the model such that 155,738 elements were processed in the training model and 38,935 elements were used for testing the model. The liblinear model was used with a C value of 1E-9.

# 4 Results



```
from sklearn.metrics import jaccard_score
jaccard_score(y_test, yhat)
```

```
0.701476820341595
```

**Figure 5:** Jaccard Index Score

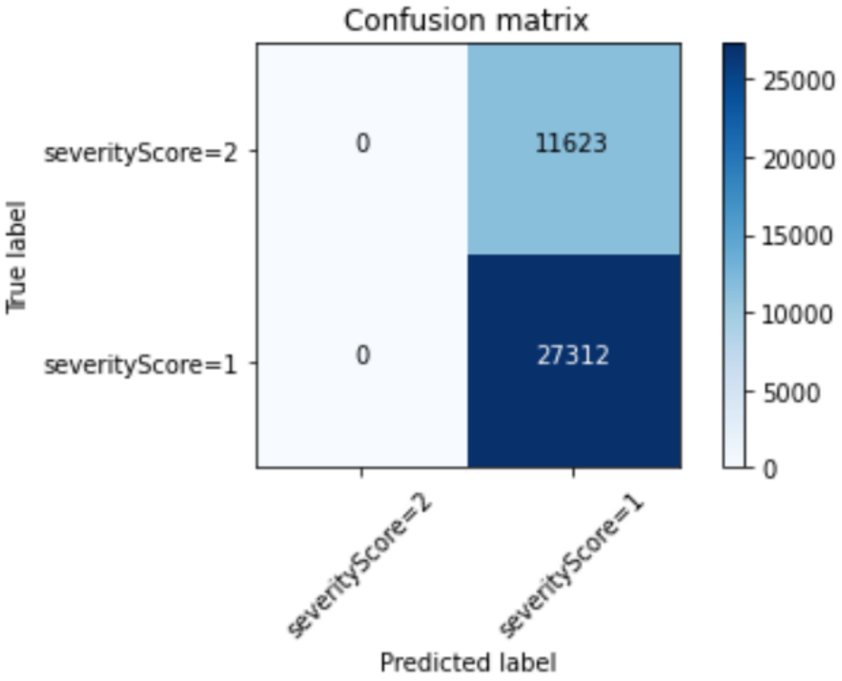
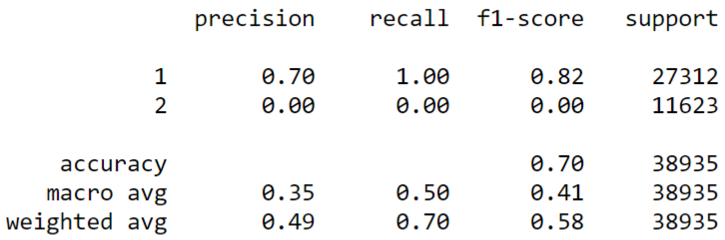


**Figure 6:** Confusion Matrix

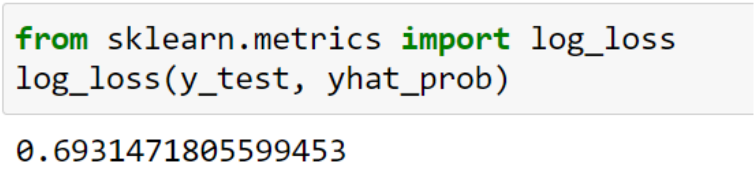


```
              precision    recall  f1-score   support

           1       0.70      1.00      0.82     27312
           2       0.00      0.00      0.00     11623

    accuracy                           0.70     38935
   macro avg       0.35      0.50      0.41     38935
weighted avg       0.49      0.70      0.58     38935
```

**Figure 7:** Confusion Matrix Table of Results



```
from sklearn.metrics import log_loss
log_loss(y_test, yhat_prob)
```

```
0.6931471805599453
```

**Figure 8:** Log Loss Results

# 5 Discussion

## 5.1 Model Evaluation

The Jaccard index score for the machine learning model we produced was approximately 0.70. When plotted on the confusion matrix, the model correctly predicted 27,312 of 38935 cases where the car accidents had a severity of 1, meaning only property damage. However, the model was unable to predict any accidents with a severity of 2. This is likely because no severity 2 incidents were present in the testing training set. When the Model was evaluated based on log loss, the greatest score was approximately 0.69.

# 6 Conclusion

The machine learning model needs more development to accurately and reliably predict the severity and probability of accidents based on environmental and behavioral factors. This model would best be served for informational purposes only and should not be used for life threatening situations.