Predicting Vehicle Collision Severity

**Carlos Fortin**

**September 2020**

# Introduction

## Background

Today, most people in the developed world drive some type of motor vehicle daily. They drive to and from work, take their loved ones on vacations trips, their children to school and often do daily chores such as going to the supermarket utilizing their vehicles. Unfortunately, the more vehicles the more accidents take place. Some of these accidents are severe and can cost lives.

For an accident to occur, there are many factors involved such as traffic violations (speeding, running a red light, etc.), mechanical failures such as a flat tire causing the driver to lose control and sometimes weather or road conditions might be poor causing the driver to struggle to keep control.

## Problem

The purpose of this report is to identify key features that increase the likelihood of an accident to occur and use them to create a model that can predict the severity of that accident so that future drivers might be able to understand the risk associated with their driving under specific conditions.

# Data Acquisition and Cleaning

## Data Source

For this analysis and the model, the data from <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv> was used. This data originally contained +190,000 rows and 38 columns (features).

## Data Cleaning

Before any analysis and modeling, the data was first studied and cleansed. At first glance, there were several features that seemed to be redundant and missing so those were removed. Additionally, extreme outliers or data that seemed to be entered in mistake were also removed.

The features that were dropped from this data frame were: 'ADDRTYPE','INTKEY','X','Y','INCDATE','PEDROWNOTGRNT','INATTENTIONIND','ST\_COLCODE','ST\_COLDESC','HITPARKEDCAR','JUNCTIONTYPE','COLDETKEY','EXCEPTRSNCODE','EXCEPTRSNDESC','SEVERITYCODE.1','SDOT\_COLCODE','SDOT\_COLDESC','SDOTCOLNUM','SEGLANEKEY', 'CROSSWALKKEY','STATUS','REPORTNO'.

The features above were removed from the analysis since they describe more the conditions after the accident occurred and describe consequences of the accident after it occurred. For this analysis, the objective is to predict the severity of an accident given certain conditions prior to an accident so the features above were not needed.

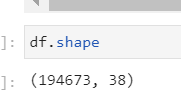
The following features were kept for further analysis:

'SEVERITYCODE', 'OBJECTID','INCKEY', 'LOCATION', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT',’'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDTTM','UNDERINFL', 'WEATHER', 'ROADCOND',

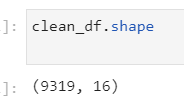
'LIGHTCOND', 'SPEEDING'.

Additionally, the data was further cleaned by dropping blanks from Weather, Road conditions, lighting, and speeding and driving under the influence. Some features contents were also modified to simplify analysis. For example, certain entried contained 0,1,N,Y. The N were all converted to 0 and the Y were all converted to 1.

Starting data frame:



Cleaned data frame:



# Exploratory Data Analysis

In this section of the report, I take a deeper dive into the data set. Key features are plotted to better visualize and understand how the data is behaving and to spot any interesting characteristics that might not be obvious at a first glance.

After the data was cleaned, a total of 9,319 accidents were kept. As the graphs shows below, 3524 are considered severe, that is, a person(s) was injured and 5795 were not severe, that is, only damage to property occurred.

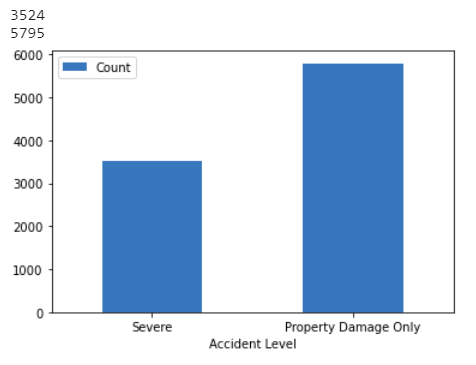


Figure 1.1 Count of Severe vs. non-Severe Accidents

Of the severe accidents recorded, the chart below summarizes the conditions reported during the accidents.

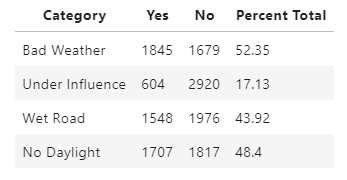


Chart 1.1 Severe Accident Summary

As it is clearly shown on the graph above, 52.35% of the severe accidents happened under bad weather conditions followed by the second highest at 48.4 which was when there was no daylight.

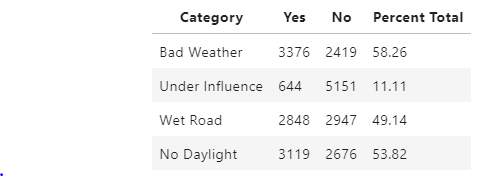


Chart 1.2 Severe Accident Summary

Again, as it is clear in the summary chart above, the two highest conditions that are true during an accident are bad weather (58.26%) and non-daylight conditions (53.82%). These features will be key in the development of the prediction algorithms in the next sections.

# Methodology

For this report and analysis, we will attempt to predict whether an accident will be severe or not under the following conditions: Weather, driving under the influence, wet road condition and lack of daylight.

The first step was collecting the data which can be found on the data section of this report. The data initially contained +190,000 rows and 38 features. The second step, was cleaning the data and visualizing it to have a better understanding of what can be used and what should be discarded for this analysis.

Third, after having a good understanding of the data, it will be modeled using machine learning algorithms used to predict the likelihood of an accident being severe given certain conditions. These will be compared and validated using standard accuracy tests. Given that the variable we are trying to predict, whether an accident will be severe or not, is categorical, Logistic Regression is the ideal machine learning algorithm to be used.

Finally, the results and conclusions will be summarized and reported.

# Analysis

This section of the report contains the code and raw results from the analysis and modeling of the data.

The target variable is Severity of Collision. After further cleaning the data, we arrive to the summary chart as show below:

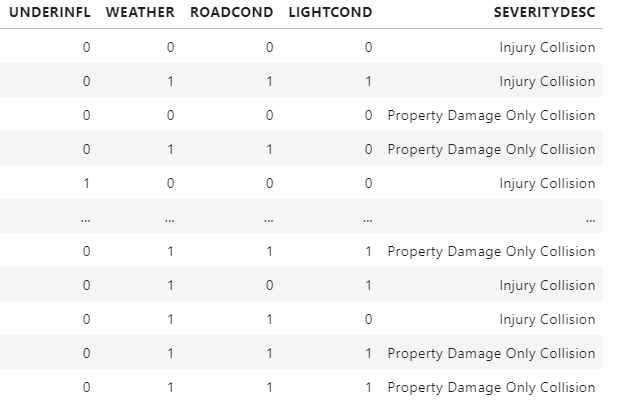


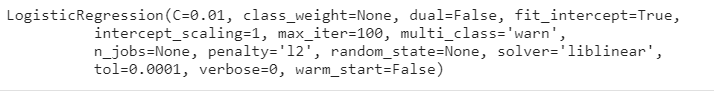
Figure 2.1 (1 indicates poor condition, 0 good condition)

The features above were used in a logistic regression model to predict whether an accident would be severe or not.

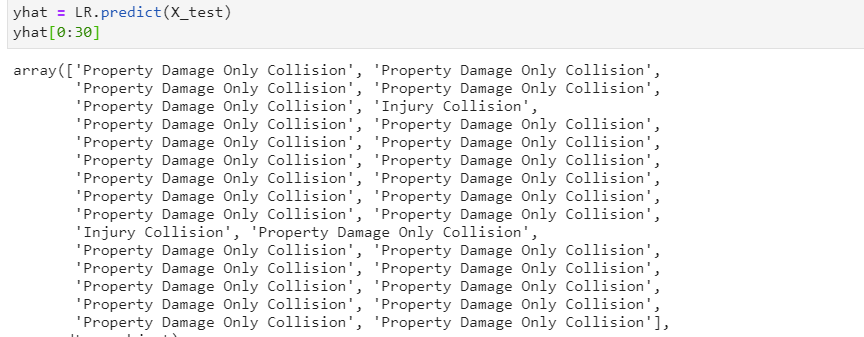
The data was Trained/Tested/Split at a 20/80. The results of the data sets below:



Following the Train/Test/Split, the Logistic Regression Model was built.

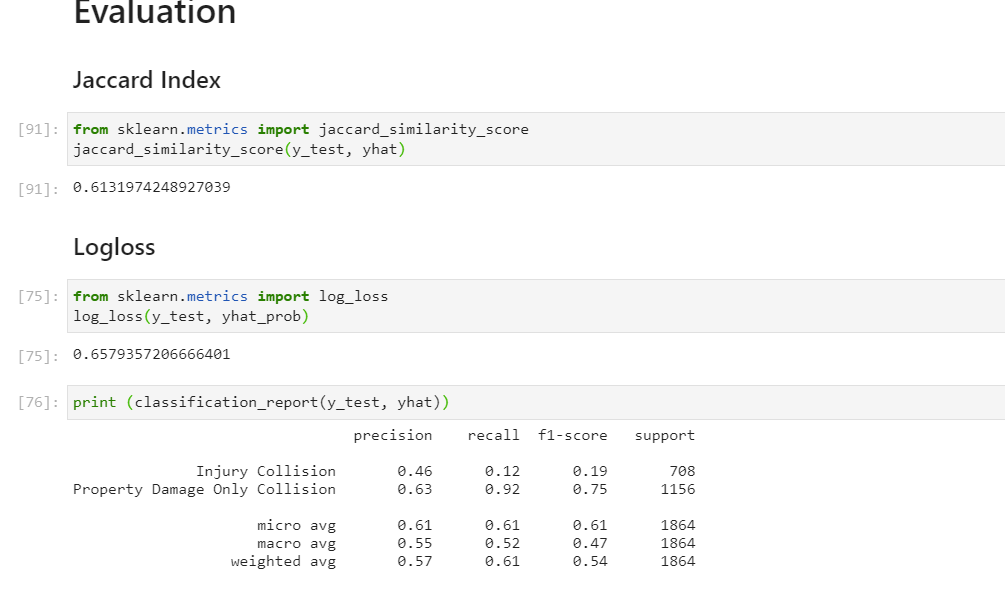


After the model was built, the following 30 predictions were made:



# Evaluation

In order to validate the model, several evaluation techniques and metrics were performed. Below, a summary of those metrics



# Discussion and Results

As the data, analysis and evaluation show, Logistic regression is good at predicting the severity of an accident when given the correct features. Since the variable of interest is categorical, logistic regression is also a good model for this prediction.

From general data observation and further analysis, I can note that most accidents occur during poor weather conditions and poor light conditions. In fact, these two factors are the highest determinants of causing an accident.

# Conclusion

In conclusion, there are several factors involved in an accident. Some of these factors are impossible to take into considerations so it is important to remember that the even the best prediction model cannot guarantee 100% accuracy. That being said, however, based on my results the two factors that contribute most to an accident happening are weather condition and light condition. In fact, these two factors also contribute to the severity of the accidents as well.

As a recommendation to all people who drive, I would suggest to reconsider travelling when the weather and light conditions are not optimal. If that is not possible, I would encourage extra caution to avoid any collision.