Prediction of Severity of car accidents

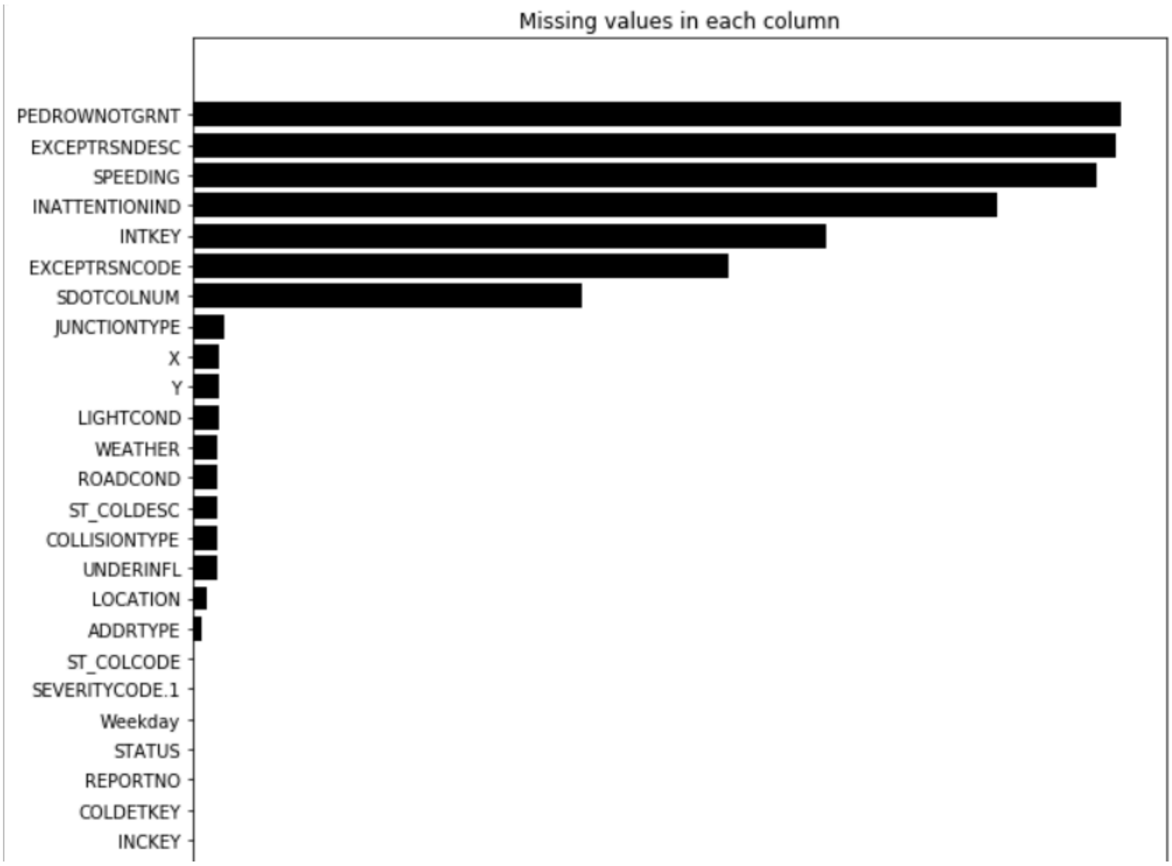
1. Introduction

Traffic accidents are a common occurrence that are due most of the time to a driver's actions but environment conditions and others factors related to road can have an impact of the occurrence. Generally, there are virtually an unlimited number of causes of car accidents. Weather and road conditions are a common cause of many car accidents, but many accidents are caused by failure of a driver to keep his attention to the road, and operation of his vehicle. Understanding the factors that contribute to car accidents can help drivers avoid them. Some of the most common causes of car accidents include: **Speeding, Using a Device, Driver Fatigue, Drunk Driving, Defective Auto Parts, Rubbernecking, Poor Weather Conditions** about the weather it states that Weather conditions that leave the roadway wet or icy, or reduce visibility, pose a danger to vehicles on the road, and require drivers to pay extra attention, and to slow down. High winds, blowing dust, fog, and torrential downpours are common causes of accidents.

In this study we will analyze whether environment conditions and others attributes such as road condition, light condition, junction type play a maggior role than believed, a machine learning approach was taken. Using a dataset from https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv ;
road accidents attributes were grouped and trained to classify the level of severity of an accident.

2. Preparing and visualize the data

There are 194,673 observations and 38 variables in this data set. Since we would like to identify the factors that cause the accident and the level of severity we use SEVERITYCODE as the target value. some data were missed and are represented in the below graph.

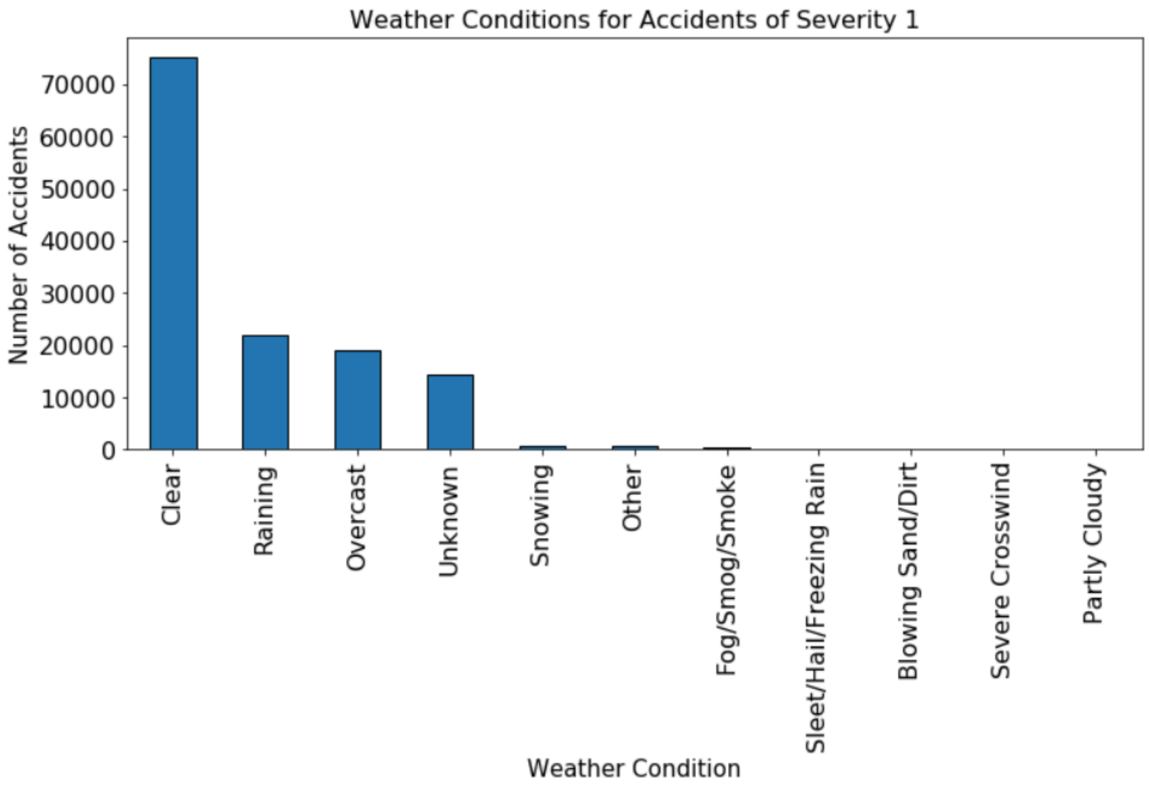


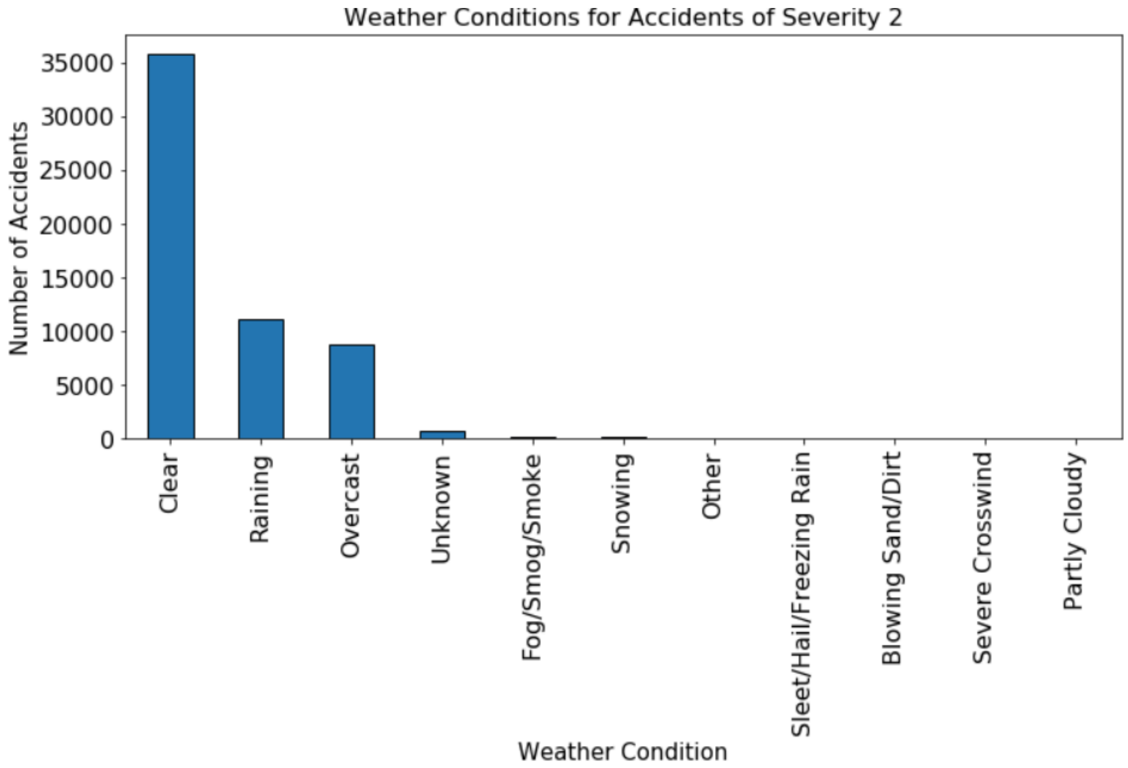
From this graph we can notice that the missing values for variable such as WEATHER, ROADCOND, LIGHTCOND, JUNCTIONTYPE seem insignificant (~2%) over the total value of the rows of the entire dataset so we will replace them with “Unknown” label instead of drop them. The others variable with high number of missing will not be considered for the analysis such as the **SPEEDING.**

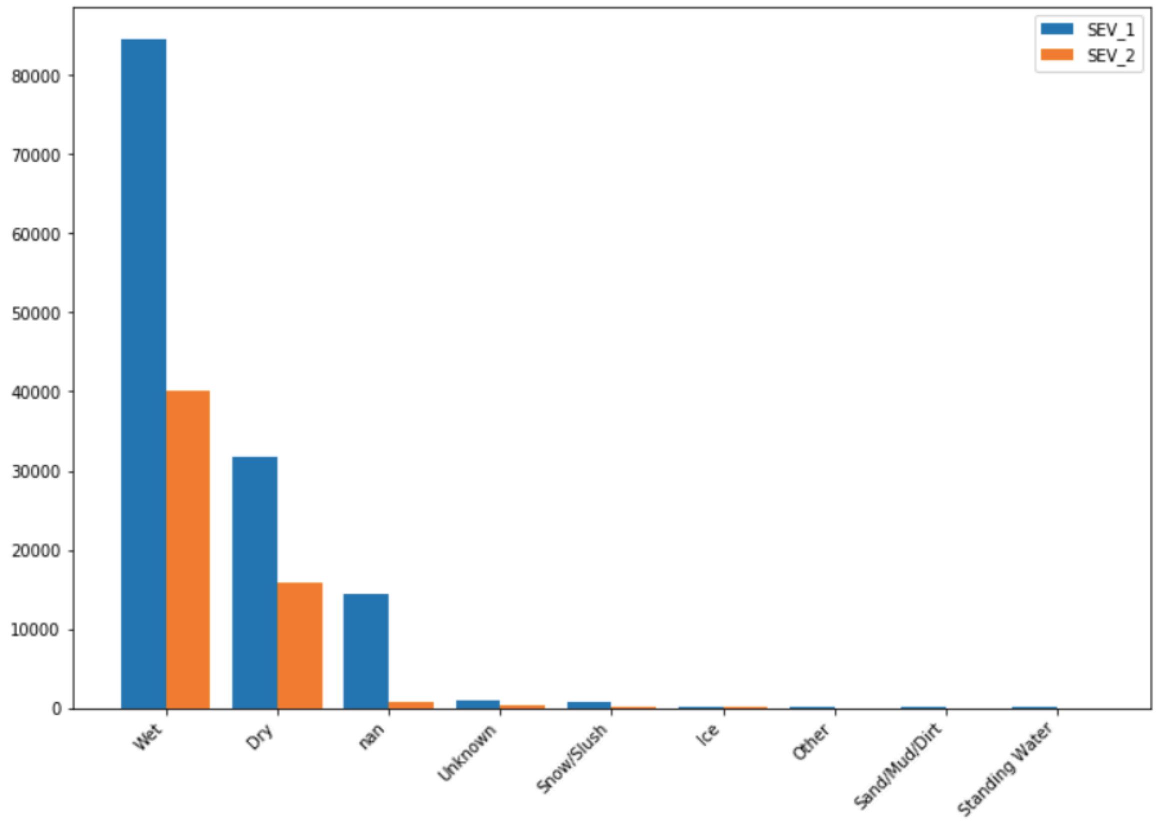
Data visualization

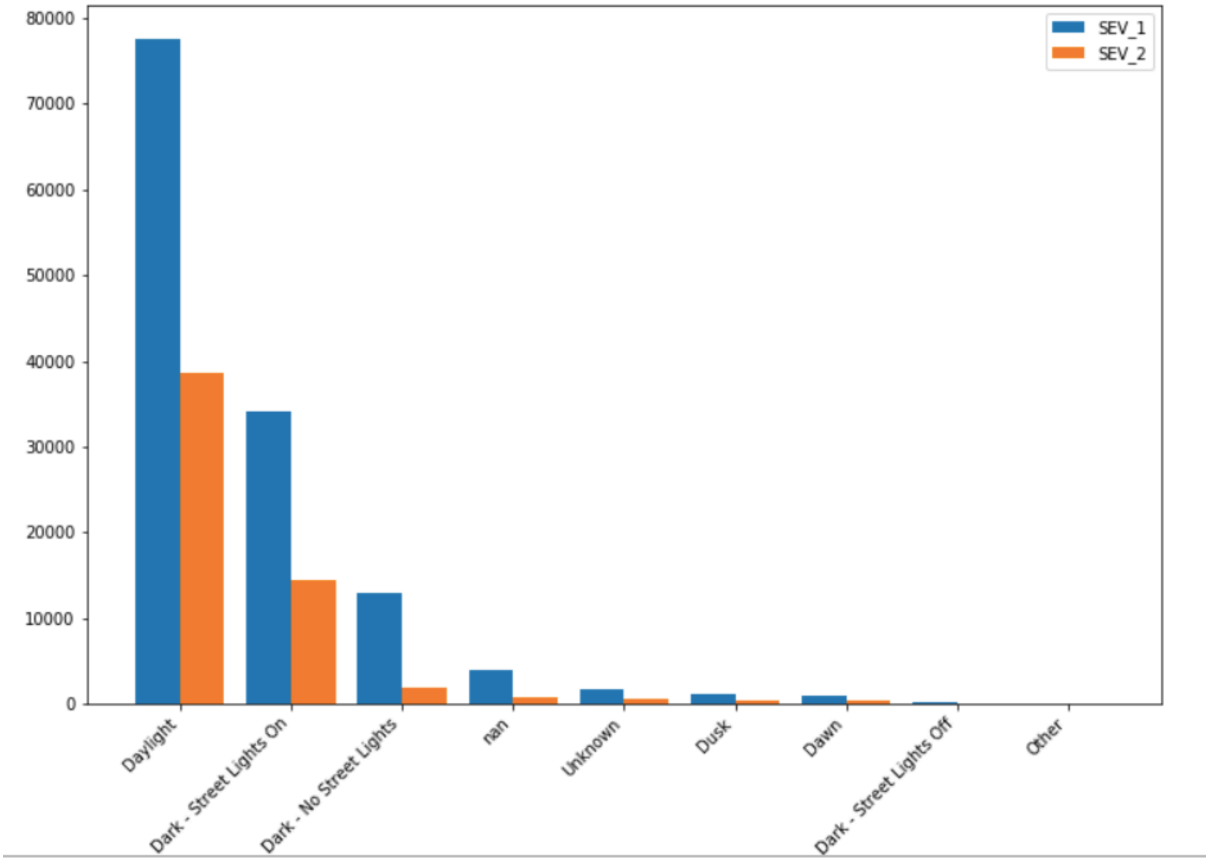
We will visualize different kind of plot in function of severity

- Total number of incidents based on Weather condition
- Total number of accidents based on Road condition
- Total number of accidents based on Light condition


Weather Conditions for Accidents of Severity 1
Number of Accidents
0
10000
20000
30000
40000
50000
60000
70000
Clear
Raining
Overcast
Unknown
Snowing
Other
Fog/Smog/Smoke
Sleet/Hail/Freezing Rain
Blowing Sand/Dirt
Severe Crosswind
Partly Cloudy
Weather Condition

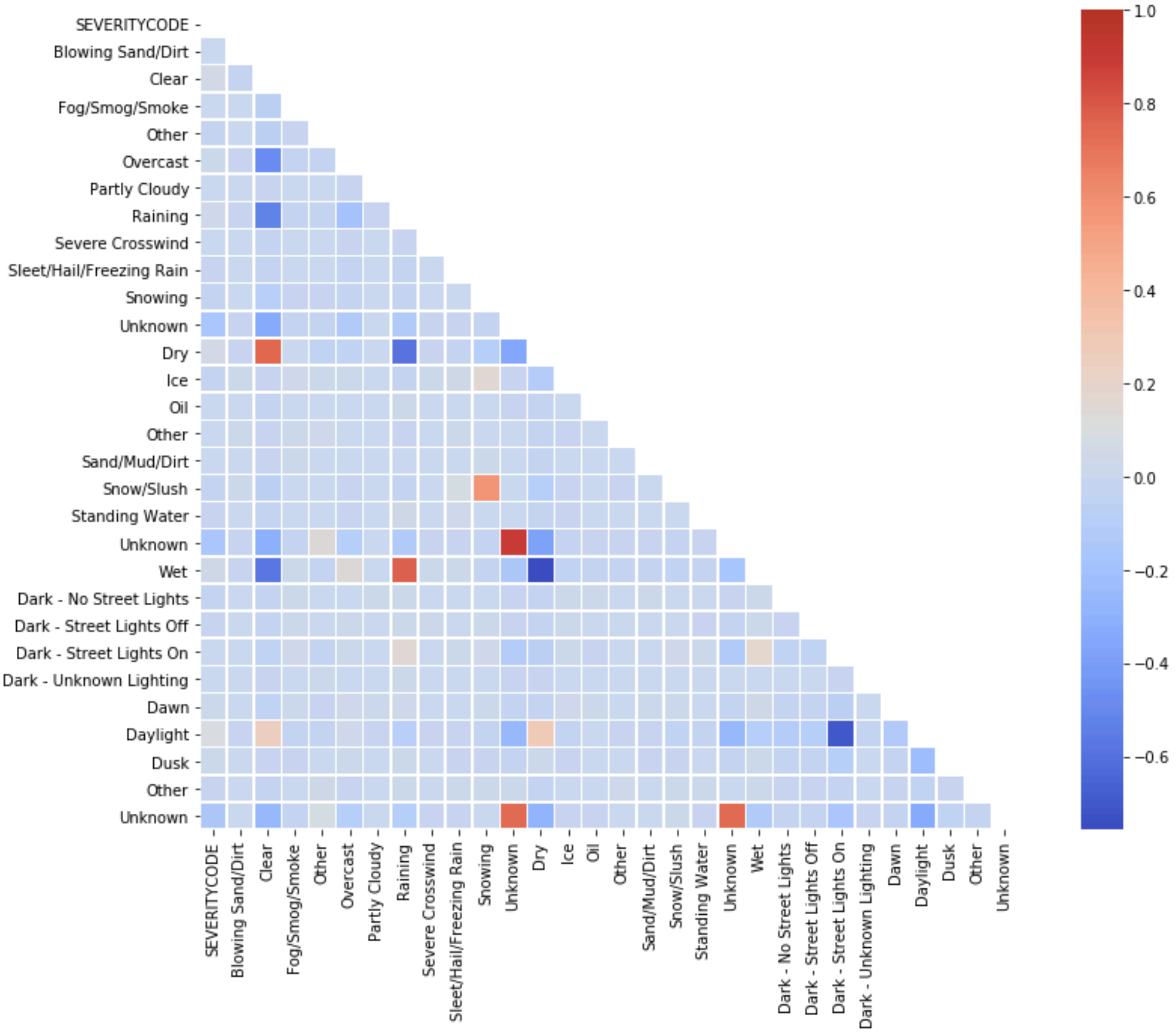

Weather Conditions for Accidents of Severity 2
Number of Accidents
0
5000
10000
15000
20000
25000
30000
35000
Clear
Raining
Overcast
Unknown
Fog/Smog/Smoke
Snowing
Other
Sleet/Hail/Freezing Rain
Blowing Sand/Dirt
Severe Crosswind
Partly Cloudy
Weather Condition


SEV_1
SEV_2
80000
70000
60000
50000
40000
30000
20000
10000
0
Wet
Dry
nan
Unknown
Snow/Slush
Ice
Other
Sand/Mud/Dirt
Standing Water


SEV_1
SEV_2
80000
70000
60000
50000
40000
30000
20000
10000
0
Daylight
Dark - Street Lights On
Dark - No Street Lights
nan
Unknown
Dusk
Dawn
Dark - Street Lights Off
Other

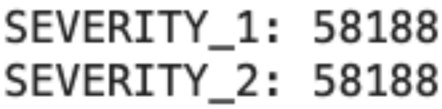
From these 3 graphs we can affirm that most of accidents happen in Daylight when road is wet and clear weather.

Since all the features we wanted to used are all nominal categorical, we decide to use the one hot encoding to convert data into numerical and to ensure to have different level of the labels. and perform a correlation matrix to visualize the relation between feature and target variable.



We can notice small correlation between some light condition and severity as well as road condition. we can see a strong correlation between road and light condition.

finally we balance the data for the SEVERITYCODE in order to avoid bias on the result.



```
SEVERITY_1: 58188
SEVERITY_2: 58188
```

3. Building Predictive Model

for this part we basically have to answer to the question What are the main factors causing an accidents, and can we predict the severity based on these factors? as seen before we decide to perform the analysis on road condition, light condition and weather.

We will use the following algorithms:

K-Nearest Neighbor (KNN)

KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

Decision Tree

A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

Logistic Regression

Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

we split the data by 30 % for test and 70% for training and we look after the accuracy of each algorithm. below a recap of the result.

Algorithm	KNN	Decision Tree	Logistic Regression
Accuracy	0,558	0,556	0,559
Jaccard simility score	0,56	0,56	0,56
F1-score	0,55	0,48	0,49
Log loss	n.a	n.a	0,67

From this table we can conclude that the Logistic regression has a better accuracy among others.

4. Conclusion

Based on the above results it seems that the prediction of the Severity using Road condition, light condition and weather is not perfectly what we expected. We decide to use those feature only by logic. This conclusion is based on the accuracy percentage being around mid-60%. Further analysis for improvement can be made on others feature such as the junction type, the location and so on.