CAPSTONE PROJECT

CAR ACCIDENT SEVERITY

INTRODUCTION / BUSINESS PROBLEM

* Traffic accidents are **one of the major causes of disability and mortality in many different countries**.
* Road traffic crashes cost most countries 3% of their GDP.
* It does not only causes deaths, but also millions of non-fatal injuries (between 20 and 50 million more people suffer it, according to the WHO), with many incurring a disability as a result of their injury.
* **Approximately 21% of vehicle crashes in a year are weather-related.**

DATA SOURCES

In this Project we will be using a shared dataset for **Seattle** city. It contains different characteristics about the collisions in this city **between the years 2004 and 2020**. The label for the dataset is **severity**, which describes the fatality of an accident. The severity can be an “injury collision” or a “property damage only collision”.

The dataset can be found at: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

DATA CLEANING AND FEATURE SELECTION

The shared data has unbalanced labels. So we have balanced the data, otherwise, we would have created a biased ML model.

First we have selected the relevant variables for our analysis, and cleaned the data as follows:

* Change "Unknown" and "Other" values in 'WEATHER', 'ROADCOND' and 'LIGHTCOND' to "NaN".
* Replace "NaN" in 'SPEEDING' and 'INATTENTIONIND' by "N".
* Replace "NaN" in 'UNDERINFL' by "0".
* Replace "N" by "0" and "Y" by "1" in 'UNDERINFL' and 'SPEEDING'.
* Drop the "NaN" rows.

DATA CLEANING AND FEATURE SELECTION

Then we have converted to date time object the 'INCDATE' column, so we can see easily if a collision happens in a weekday or weekends.

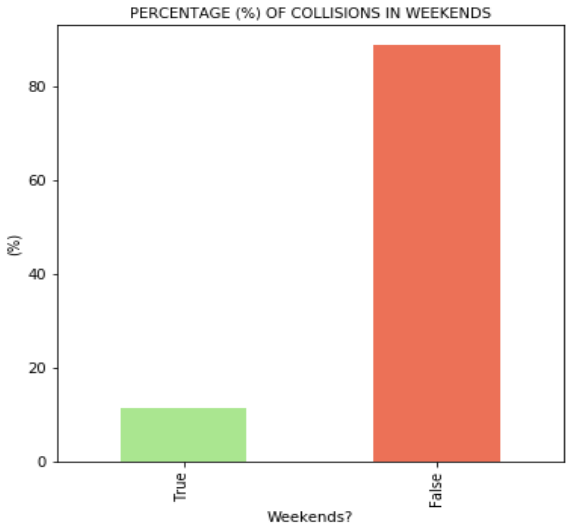
Finally, we have converted categorical features to numerical values. We noticed that the columns of our dataset have a mix of variables. So, first, we converted categorical variables into binary variables and append them to the feature Data Frame.

EXPLORATORY ANALYSIS

Most of the collisions are Property Damage Only. According to our dataset, around 33% are Injury Collisions.

COLLISIONS ON WEEKENDS:

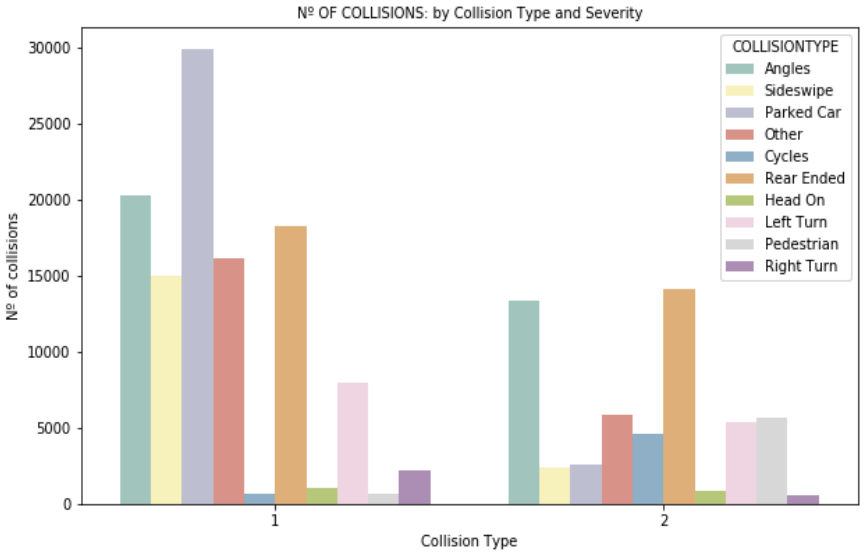
As wee can see in the plot below, it is clear that most of the collisions occur on weekdays instead of weekends.



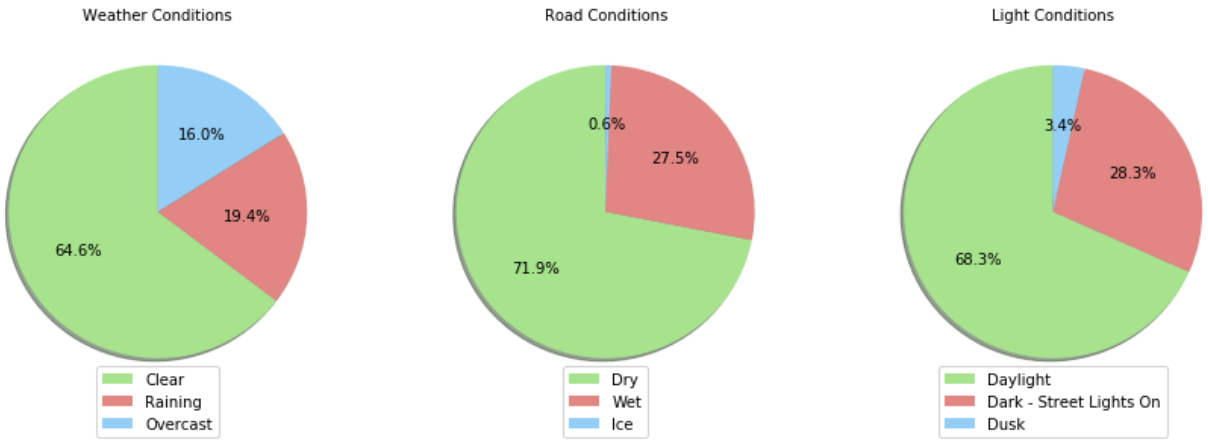
NUMBER OF COLLISIONS: SEVERITY VS. COLLISION TYPE

Property Damage collision mostly occurs with parked cars being hit, followed by far by angles, and rear-ended impacts.

In Injury collision, most of the accidents are rear-end and angle impacts.



WEATHER, ROAD AND LIGHT CONDITIONS



DRIVERS SPEEDING / UNDER THE INFLUENCE

The severity of a collision is higher when the driver is speeding or under the influence.

SPEEDING	SEVERITYCODE	
0	1	0.672375
	2	0.327625
1	1	0.617752
	2	0.382248

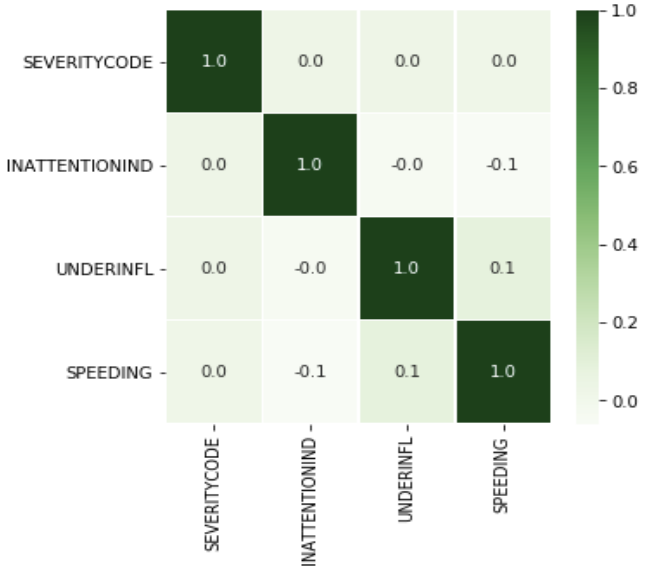
Name: SEVERITYCODE, dtype: float64

UNDERINFL	SEVERITYCODE	
0	1	0.672913
	2	0.327087
1	1	0.607633
	2	0.392367

Name: SEVERITYCODE, dtype: float64

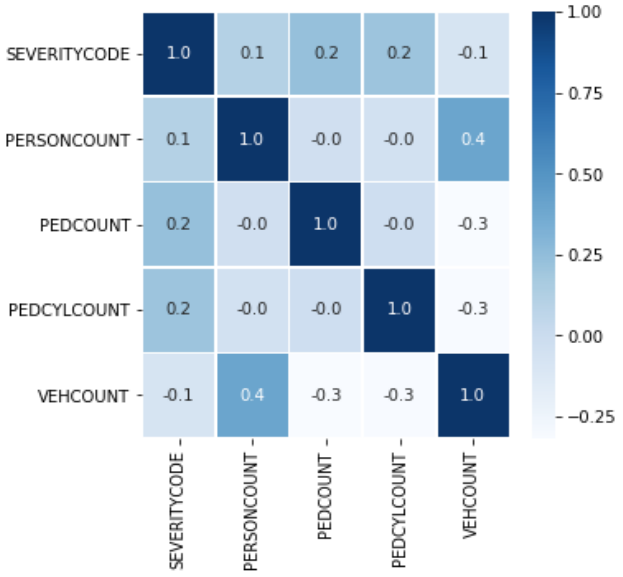
SEVERITY / INATTENTION / UNDER INFLUENCE / SPEEDING

Whether the driver is under the influence and their inattentiveness are not correlated to the severity of a collision. There is just a very slightly correlation of a driver speeding and being under the influence.



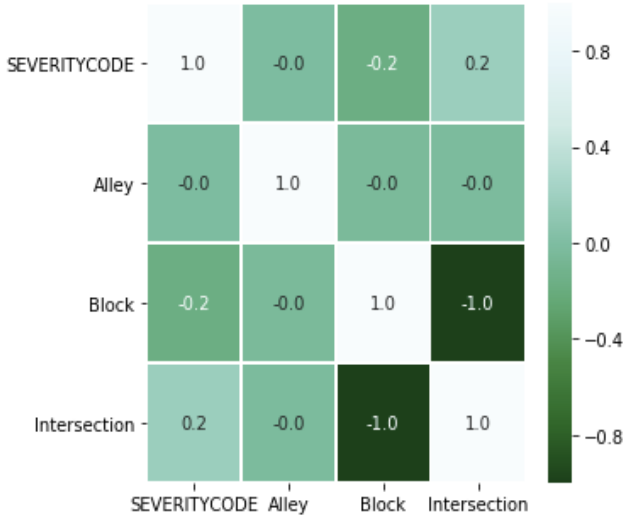
PERSON / PEDESTRIAN / BICYCLE / VEHICLE

The severity of a collision is higher when pedestrians and/or bicycles are involved. We also see the correlation between the number of vehicles and persons involved in an accident.



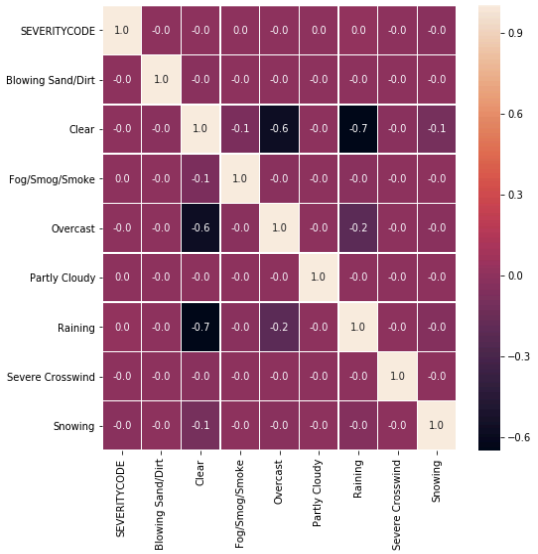
TYPE OF LOCATION

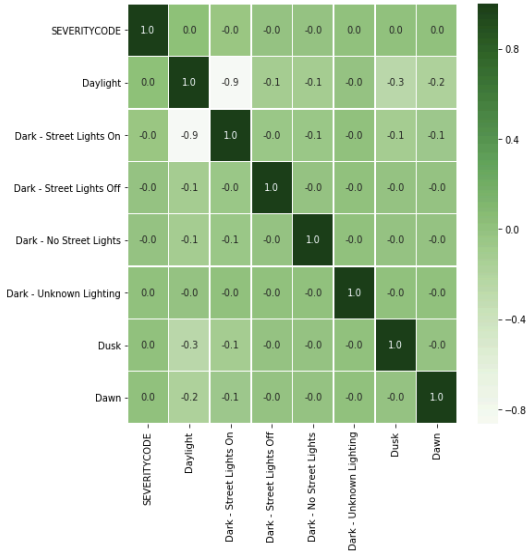
It shows us the slight correlation that exists between severity and intersection. A collision is more likely to be more severe when it takes place in an intersection, instead of an alley or a block.

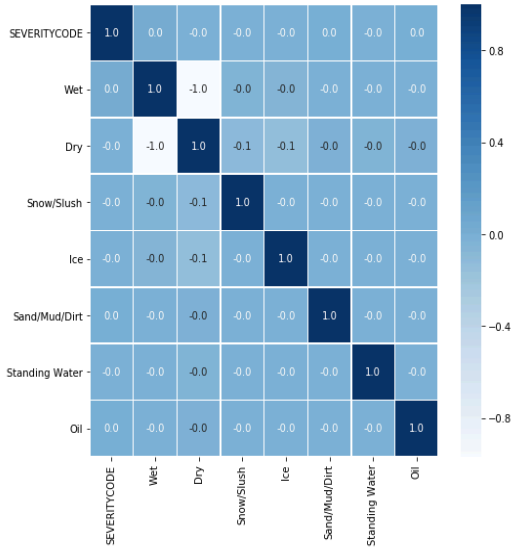


WEATHER, LIGHT AND ROAD CONDITIONS

There is no correlation between Weather / Light / Road Conditions and the Severity of a collision.







PREDICTIVE MODELING

First we have normalized the data, as data standardization give data zero mean and unit variance. We also have defined the Target Value variable ‘SEVERITYCODE’.

Later in this part, we had to create the Training and Test Set.

Machine Learning Models

Now we have used the Training Set to build an accurate model. Then we are going to use the Test Set to report the accuracy of the model.

For this project we used the following:

* **Decision Tree**
* **Logistic Regression**
* **XGBoost**
* **Random Forest Classifier**

RESULTS

Here we see the accuracy of the built model with the different evaluation metrics previously used:

Algorithm	Jaccard	F1-score	Precision	LogLoss
Decision Tree	0.7199	0.6522	0.7067	NA
Logistic Regression	0.7285	0.6882	0.7268	0.5264
XGBoost	0.7304	0.6910	0.7283	NA
Random Forest Classifier	0.7104	0.6902	0.7379	NA

DISCUSSION

As mentioned before, we have used 4 different algorithms:

* **Decision Tree** (accuracy of 71.99%)
* **Logistic Regression** (accuracy of 72.85%)
* **XGBoost** (accuracy of 73.04%)
* **Random Forest Classifier** (accuracy of 71.04%)

The Jaccard Index is one of the simplest accuracy measurements. According to this model, the XGBoost has a 73.04% match between the training set and the test set, being the algorithm with the best accuracy. This means that, using this algorithm; we could predict correctly the severity of almost 3 out of 4 collisions in Seattle.

Following the F1-score, XGBoost also has the highest harmonic average of the precision and recall. But it is the Random Forest Classifier the one with a better precision, only 1% far from the XGBoost and Logistic Regression.

CONCLUSION

The main goal of this Capstone Project was to focus on the analysis of the shared dataset for Seattle city, in order to create a machine-learning model able to predict accident "severity". So drivers can be warned and have more information about the possibility of getting into a car accident and how severe it would be.

According to the results, we are able to predict the severity of almost 3 out of 4 collisions. Taking into account road, weather and light conditions; location; speeding, inattentiveness and if the driver is under the influence; number of pedestrians, bicycles and vehicles involved.

This means that the number of sever accidents could be reduced if drivers have and use this information when traveling. So they can be more careful when driving, and changing their route if possible, in order to arrive safe and on time.