Capstone Project - Car accident severity

1. INTRODUCTION / BUSINESS PROBLEM

1.1. BACKGROUND, PROBLEM AND INTEREST

Traffic accidents are **one of the major causes of disability and mortality in many different countries**. According to the World Health Organization (WHO) “approximately 1.35 million people die each year as a result of road traffic crashes”, and “road traffic injuries are the leading cause of death for children and young adults.”

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.

It is also important to note that “more than half of all road traffic deaths are among vulnerable road users”, so it is not only about ourselves as drivers, but also about the passengers, pedestrians, cyclists and motorcyclists that could be involved in an accident.

Those injuries cause economic losses to individuals, their families, and to nations. These losses are materialized as the cost of treatment and lost of productivity for those killed or disabled, and for family members who have to take time off work to care for them. According to WHO, “road traffic crashes cost most countries 3% of their GDP”.

Speeding and an unsafe road infrastructure are two key risk factors of a road traffic accident, influencing both the risk of a crash and the severity of the injuries suffered.

On the other hand, the US Department of Transportation explains that **approximately 21% of vehicle crashes in a year are weather-related**.

Let’s set an example of a group of friends or a family that is planning to travel by car to visit their relatives / friends. Wouldn’t it be great if there was something in place that could warn them, given the weather and the road conditions about the possibility of getting into a car accident and how severe it would be, so that they would drive more carefully or even change their travel if they are able to? Or imagine that you have a very important business meeting in another city and you are a bit in a hurry, it would be great if there were something to help you to know the places where you have to be more careful when driving, and changing your route if possible, so you arrive both safe and on time.

In this Car Accident Severity Capestone Project we will focus on the analysis of a 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.

2. DATA ACQUISITION AND CLEANING

2.1. 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”.

Some of the attributes or features from this dataset are: the exact moment and location where accidents take place and the junction type (intersection, driveway junction…); weather (raining, snowing, fog…), road (wet, dry, sand…) and light (daylight, dark with street lights on / of…) condition at that moment; whether the car involved in an accident was speeding or not; if drivers were under the influence of drugs; as well as the number of people and vehicles involved.

We are going to use the dataset described above in order to create a machine-learning model able to predict accident "severity", and see which of the variables mentioned above have a higher/lower effect in the severity of a traffic accident. For example, we will see which weather conditions are more likely to "cause" a traffic accident, if it is during daylight or in dark conditions, in which junction type and how many cars could be involved, among others. This will provide the driver more information about in which conditions he/she should drive more carefully, take another route, or even drive any other day (or time of the day) in better conditions.

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

2.2. DATA CLEANING AND FEATURE SELECTION

This dataset is about collisions. The **Data-Collisions.csv** data set includes details of collisions in Seattle between the years 2004 and 2020. It includes following fields:

Field	Description
SEVERITYCODE	A code that corresponds to the severity of the collision (1 is Prop. Damage; 2 is Injury)
X	-
Y	-
OBJECTID	ESRI unique identifier
INCKEY	A unique key for the incident
COLDETKEY	Secondary key for the incident
REPORTNO	Collision report number
STATUS	Collision status (Matched or Unmatched)
ADDRTYPE	Collision address type (intersection, block, alley)
INTKEY	Key that corresponds to the intersection associated with a collision
LOCATION	Description of the general location of the collision
EXCEPTRSNCODE	-
EXCEPTRSNDESC	-
SEVERITYCODE	A code that corresponds to the severity of the collision (1 is Prop. Damage; 2 is Injury)
SEVERITYDESC	A detailed description of the severity of the collision
COLLISIONTYPE	Collision type (i.e. Sideswipe, parked car, left turn...)
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision
PEDCYLCOUNT	The number of bicycles involved in the collision
VEHCOUNT	The number of vehicles involved in the collision
INCDATE	The date of the incident
INCDTTM	The date and time of the incident
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to the collision by SDOT
SDOT_COLDESC	A description of the collision corresponding to the collision code
INATTENTIONIND	Whether or not collision was due to inattention (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol
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
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted (Y/N)
SDOTCOLNUM	A number given to the collision by SDOT
SPEEDING	Whether or not speeding was a factor in the collision (Y/N)
ST_COLCODE	A code provided by the state that describes the collision
ST_COLDESC	A description that corresponds to the state's coding designation
SEGLANEKEY	A key for the lane segment in which the collision occurred
CROSSWALKKEY	A key for the crosswalk at which the collision occurred
HITPARKEDCAR	Whether or not the collision involved hitting a parked car (Y/N)

The label for our data set is Severity, which describes the fatality of an accident. We have noticed that 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.

With this we have reduced the shape of our data frame from (194673, 38) to (167323, 15). Then we check that all “NaN” values have been removed:

```
SEVERITYCODE       0
ADDRTYPE           0
COLLISIONTYPE      0
PERSONCOUNT        0
PEDCOUNT           0
PEDCYLCOUNT        0
VEHCOUNT           0
INCDATE            0
JUNCTIONTYPE       0
INATTENTIONIND     0
UNDERINFL          0
WEATHER            0
ROADCOND           0
LIGHTCOND          0
SPEEDING           0
dtype: int64
```

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.

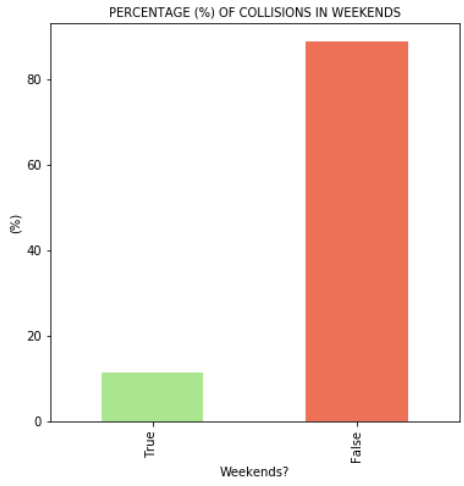
3. EXPLORATORY DATA ANALYSIS

```
1    112014
2     55309
Name: SEVERITYCODE, dtype: int64
```

Above we can see that most of the collisions are Property Damage Only. According to our dataset, around 33% are Injury Collisions.

3.1. COLLISIONS IN WEEKENDS

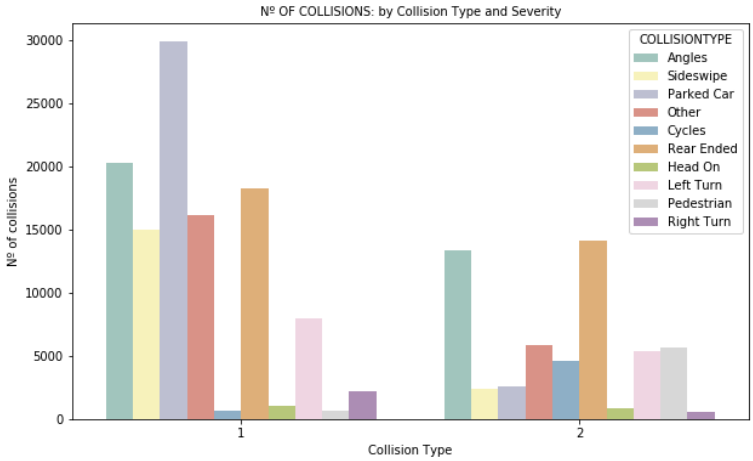
As wee can see in the plot below, it is clear that most of the collisions occur on weekdays instead of weekends.



3.2. NUMBER OF COLLISIONS: SEVERITY VS. COLLISION TYPE

As mentioned before, SEVERITYCODE takes value in 1 or 2, being **1: Property Damage Only Collision** and **2: Injury Collision**. Looking at this plot, we can see the following:

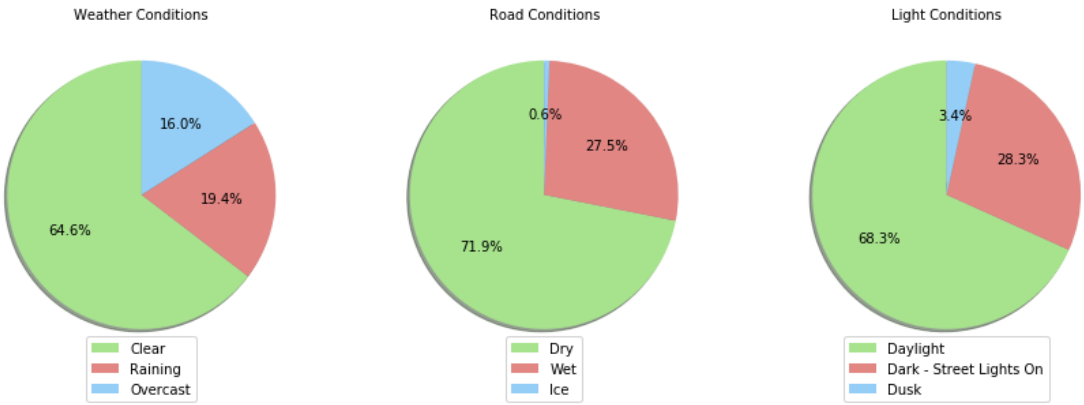
- Property Damage collision mostly occurs with parked cars being hit, followed by far by angles, and rear-ended impacts.
- In Injury collision, most of the accidents are rear-end and angle impacts.



3.3. WEATHER, ROAD AND LIGHT CONDITIONS

Here we have the top 3 Weather, Road and Light Conditions in which most of the collisions occur:

- More than 64% of the collisions occur with a clear weather, and less than 20% under the rain.
- More than 70% of the collisions take place in a dry road, followed by wet conditions with 27% of collisions.
- Most of the collisions occur in daylight (68%), followed by dark conditions with streetlights on.



3.4. DRIVERS SPEEDING / UNDER THE INFLUENCE

```
SPEEDING  SEVERITYCODE
0         1              0.672375
          2              0.327625
1         1              0.617752
          2              0.382248
Name: SEVERITYCODE, dtype: float64
```

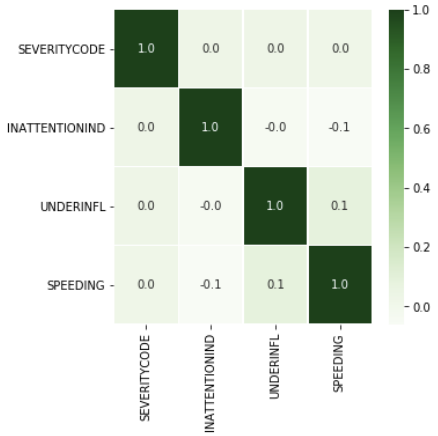
```
UNDERINFL  SEVERITYCODE
0          1              0.672913
           2              0.327087
1          1              0.607633
           2              0.392367
Name: SEVERITYCODE, dtype: float64
```

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

3.5. HEATMAPS

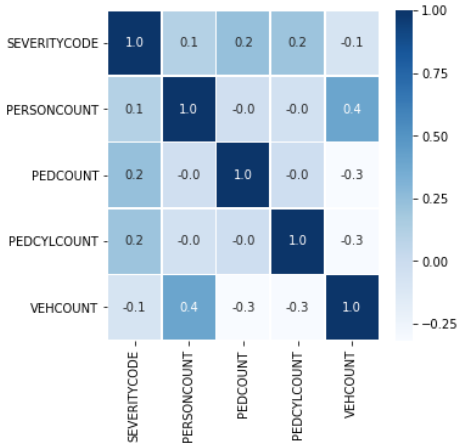
We have created 6 different heatmaps to indicate the correlation between each of the variables with one another. This correlation heatmap gives us a good overview of how the different variables are related to one another and, most importantly, how these variables are related to the Severity of a collision.

3.5.1. SEVERITY / INATTENTION / UNDER INFLUENCE / SPEEDING



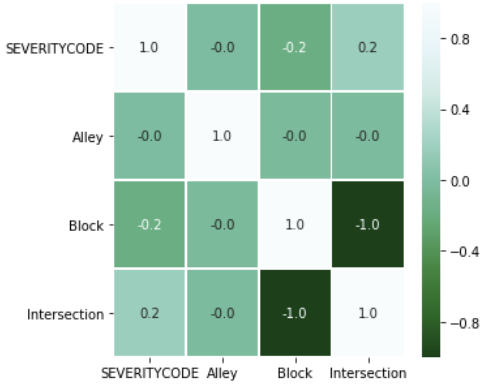
According to this heatmap, we see that whether the driver is under the influence and their inattentiveness are not correlated to the severity of a collision. There is just a very slightly correlation of a driver speeding and being under the influence.

3.5.2. PERSON / PEDESTRIAN / BICYCLE / VEHICLE



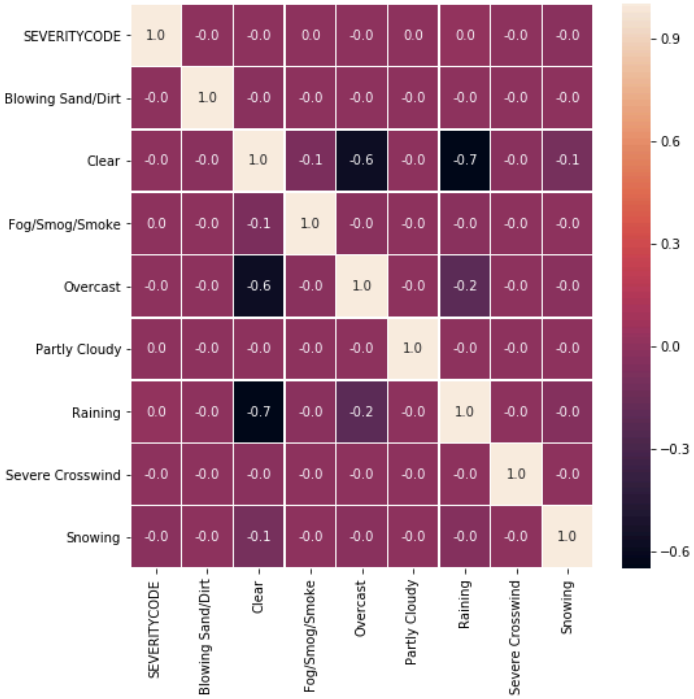
This heatmap tells us that the severity of a collision is higher when pedestrians and/or bicycles are involved. We also see the correlation between the number of vehicles and persons involved in an accident.

3.5.3. TYPE OF LOCATION



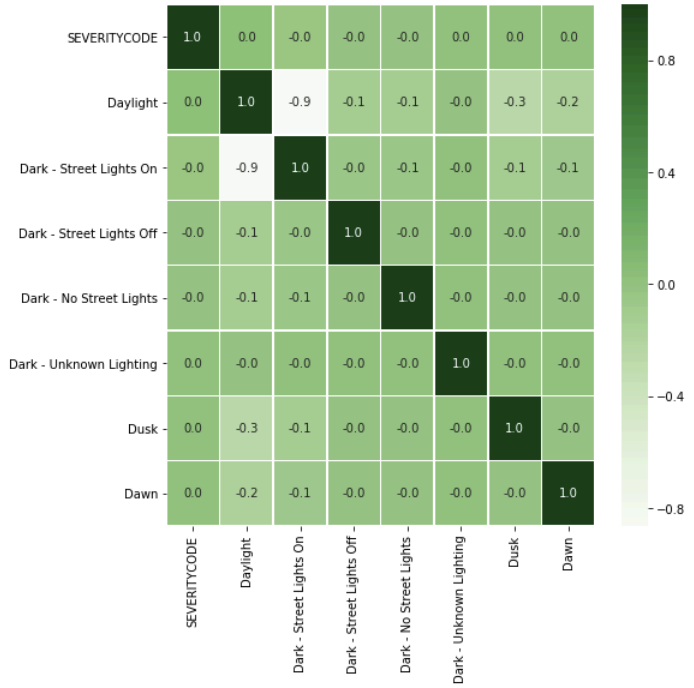
The heatmap by type of location shows us the slight correlation that exists between severity and intersection. A collision is more likely to be more severe when it takes place in an intersection, instead of an alley or a block.

3.5.4. WEATHER CONDITIONS



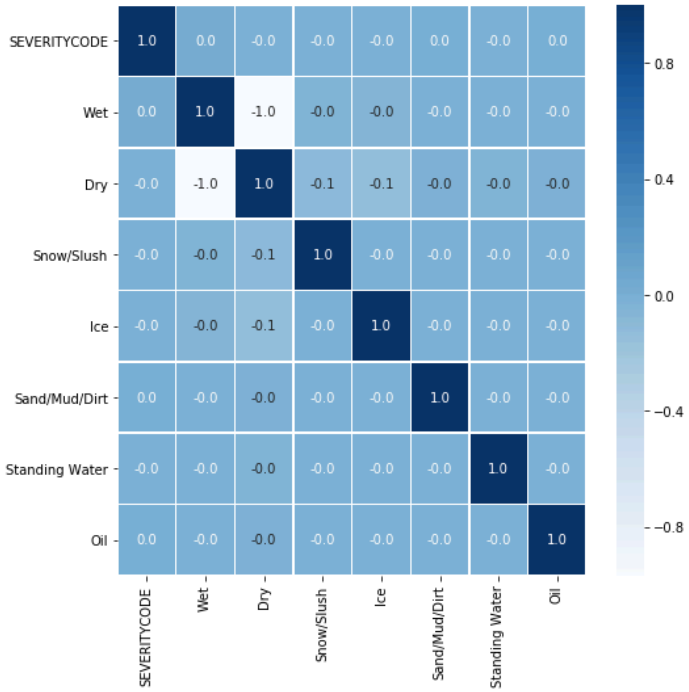
According to this heatmap, there is no correlation between Weather Conditions and the Severity of a collision.

3.5.5. LIGHT CONDITIONS



As it happens with Weather Conditions, there is no correlation between Light Conditions (daylight, dark...) and the Severity of a collision.

3.5.6. ROAD CONDITIONS



Finally, this last heatmap shows us that there is no correlation neither between Road Conditions and the Severity of a collision.

4. 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.

4.1. 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**: are built by splitting the training set into distinct nodes, where one node contains all of or most of one category of the data.
- **Logistic Regression**: it fits a special s-shaped curve by taking the linear regression and transforming the numeric estimate into a probability with the sigmoid function.
- **XGBoost**: is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. XGBoost provides a parallel tree boosting that solve many data science problems in a fast and accurate way.
- **Random Forest Classifier**: it fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

5. RESULTS AND DISCUSSION

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

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%)

And also 4 different model evaluation metrics:

- **Jaccard Index**
- **F1-score**
- **Precision**
- **LogLoss**

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.

On the other hand, the Logistic Regression has a Logarithmic Loss (or a performance) of 52.6%.

With the analysis made in this project, we can say that drivers should be a bit more careful when driving from Monday to Friday, during daylight, with clear and dry weather and road conditions, when there are pedestrians or bicycles around, and when driving in an intersection (a collision is more likely to be more severe if it takes place there). And, of course, not to speed or drive under the influence.

6. 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.