Capstone Project - Car accident severity (Week 1 & 2)

1. Introduction : Business Understanding

1.1.Background a discussion of the background

Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. [1]

France too suffers of road accidents with 3244 fatalities on French mainland roads in 2019, ten more than 2018. The numbers of accident for 2019 were 56016 and this is a very important number.

1.2.Problem A description of the problem

We have to predict the severity of accidents using a dataset that should contain sever information about the weather and road condition, human fatalities, traffic delay, property damage and so an.

The aim of this project is to determine the possibility we get into a car accident and how severe it would be using these data

1.3. Stakeholders

French road safety observatory and French government would be very interested to predict the severity of an accident, in order to improve road safety to be able to reduce the number of accidents and fatalities.

2. Data

2.1.Data Sources

The data can be founded on kaggle : https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016

2.2.Feature selection

The data consist of the recorded accidents in France from 2005 to 2016. On Kaggle we have 5 data sets; I decided to use 3 of them only: Characteristics, places and users.

I determine the features that I will use to train the model. I kept the feature that I judges will help me to train the model to have a better results.

“Characteristics” section describes the general circumstances of the accident.

“Places” section which describes the main location of the accident even if it took place at an intersection.

“Users” section which describes the users involved.

Set	Kept features	Dropped features
Characteristics	Num_Acc, mois, jour, hrmn, lum, agg, int, atm, col, com, dep	an, adr, gps, lat, long.
Places	Num_Acc, catr, circ, nbv, prof, surf, infra	voie, v1, v2, vosp, pr, pr1, plan, lartpc, larrout, situ, env1.
Users	Num_Acc, grav	num_veh, place, catu, sexe, an_nais, trajet, secu, locp, actp, etatp.

After dropping all the features that I don’t need I merged the 3 data set onto one. The data set resulted from the feature selection has 839985 samples and 17 features.

2.3.Data cleaning

For dealing with missing data I have dorped rows for “com” column and I replace by frequency for the rest:

- Drop the whole row:
    - Com : 2 missing data
- Replace by frequency:
    - Atm: 55
    - Col: 11
    - Catr: 1
    - Circ: 798
    - Nbv: 1790
    - Prof: 1061
    - Surf: 1018
    - Infra: 1278

After data cleaning we have a data set with 839983 rows and 17 features.

3. Methodology

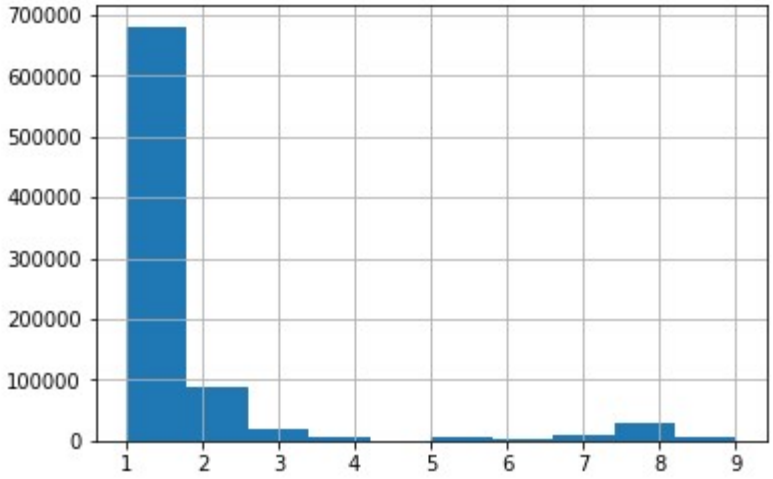
3.1.Exploratory data analysis

In this suction I used to do some exploratory data analysis to better understand data I will use to predict the severity of accident in the Model section.

The “atm” attribute represent the atmospheric condition and after analysis we can find that we have more accident when atmospheric condition is normal.

```
accidents["atm"].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x20db99c6908>
```

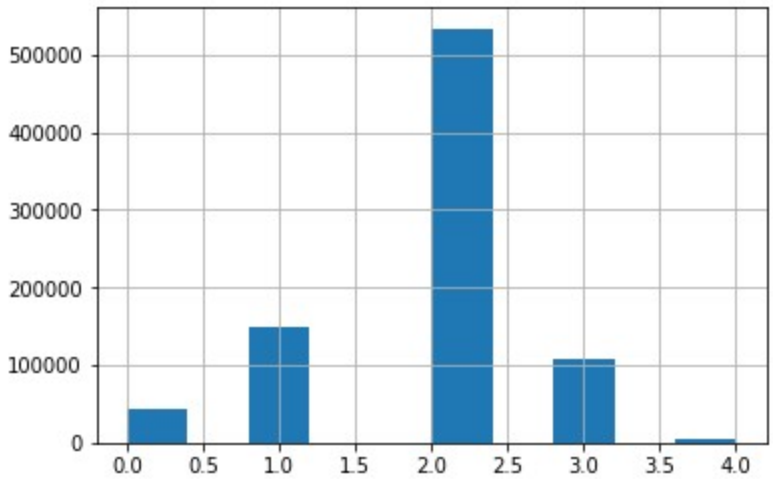


Atmospheric condition “atm”

The “circ” attribute represent the traffic regime, after analysis we can find that we have more accident when the traffic regime is bidirectional.

```
accidents["circ"].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x20d8023cdd8>
```

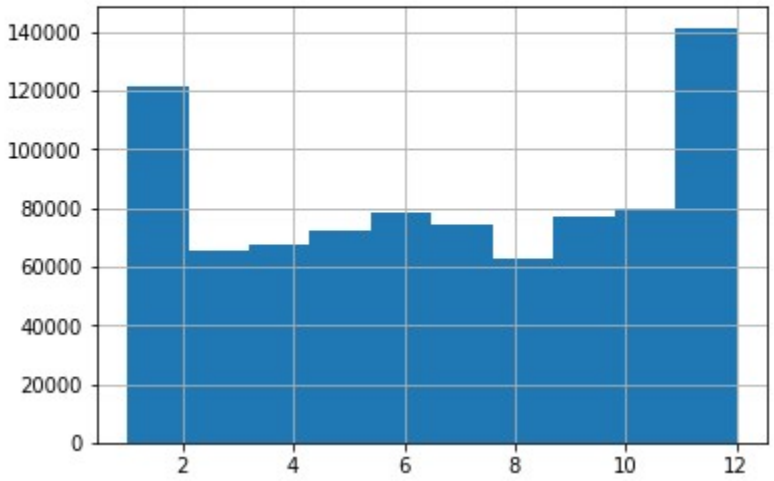


Traffic regime “circ”

The “mois” attribute represent the month of the accident, after analysis we can see that we have more accident on January and December.

```
accidents["mois"].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x20d800ab438>
```



Months “mois”

3.2.Model development

In this section I tried to predict the “grav” attribute using different algorithm:

- Random Forest
- K-nearest neighbors
- Logistic Regression

I tried to choose the best parameters to get the best accuracy:

- Random Forest : 100 decision trees
- K-nearest neighbors : K=13
- Logistic Regression : C = 0.01, solver = liblinear

In the end I calculate the accuracy of each model:

- Accuracy of Random Forest model : 0.7324
- Accuracy of K-nearest neighbors model : 0.6796
- Accuracy of Logistic Regression model : 06626

4. Results

Algorithm	Accuracy
KNN	0.6796
Random Forest	0.7324
Logistic Regression	0.6626

Like we can see the accuracy of the random Forest model equal to 0.732 is the better one comparing it to the accuracy of KNN equal to 0.6796 and Logistic Regression equal to 0.6626.

5. Discussion

We have a good result for the first study with accuracy of 0.7324. It is a very good result that we can improve with modifying the preprocessing section and the parameters of the models.

6. Conclusion

In this study, I tried to predict the severity of accident using different attributes. In first time I tried to analyze and understand the data set I choose to make my study. In second time I tried to makes models and to find the best model presenting a best accuracy.

This is an important study to predict the probability of an accident.

7. References

[1] World Health Organization https://www.who.int/news-room/fact-sheets