**Introduction or Business Problem:**

Road accident has become one of the most cause of deaths in world. It is 8th leading cause of death in the world and estimated to become 7th leading cause by year 2030. It's time when we should analyze the factor that are causing road accident and we can ensure to take precautionaries measures to minimize the road accident in coming days. Governments should be highly intrested in getting the factors causing the road accidents and try to implementing the rules which can prevent those accidents. Private companies too should take interest in this analysis and aim in improving techonologies ensuring the road safety. In one line, here we'll analyze the several factors that are the reasons of causing road accidents.

**Data**

Data Source: The Data for this project comes from kaggle. You can use this("[https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016"](https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016%22)) link to download the data from kaggle. Categories in data: The data has five different data sets which consists of all the record of accidents from 2005 to 2016.

1. The Characteristics Dataset contains the information about time,places, type of collision, weather, lightening condition and type of intersection.

2. The places dataset contains the information about the road specifics like gradient , shape , category of road,the tra c regime, surface condition and infrastrcture.

3. The user dataset contains the information about place occupied by the user of the vehicle, user involved in accident, reason of travelling, severity of accident, the use of safety equipment and information on the pedestrains.

4. The Vehicle dataset contains the information about type of vehicle.

5. The holiday dataset contains the information about one label for the accidents occuring in which holidays.

The features of the datasets:

In the characteristics dataset, We'll keep the following features , "lighting", "localisation"(agg), "type of intersection", "atmospheric conditions", "type of collisions", "department", "adress", "time" and the coordinates. We added two new features from this original dataset, "date" and "weekend" indicating if the accident occurred during the weekend or not.

In the places dataset, we'll keep the following features: "road categorie", "traffic regime", "number of traffic lanes", "road profile", "road shape", "surface condition", "situation", "school nearby" and "infrastructure".

We have created the following features from the user datasets:

num\_us : total number of users involved in the accident. ped : Wether there are pedestrians involved or not. critic\_age : If there is any user in between 17 and 31 y.o. sev : maximum gravity suffered by any user involved in the accident: 0 = Unscathered or Light injury 1 = Hospitalized wonded or Death I used the holiday dataset to craft a new feature indicating the accident occurred during a holiday.

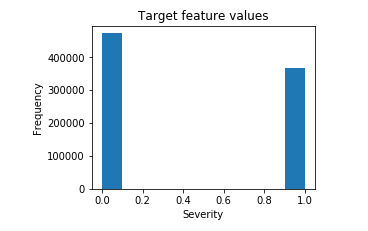
**EDA:**

Accidents classified in each level of severity:

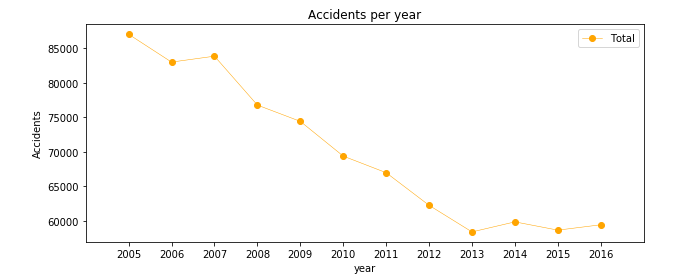
0 471695

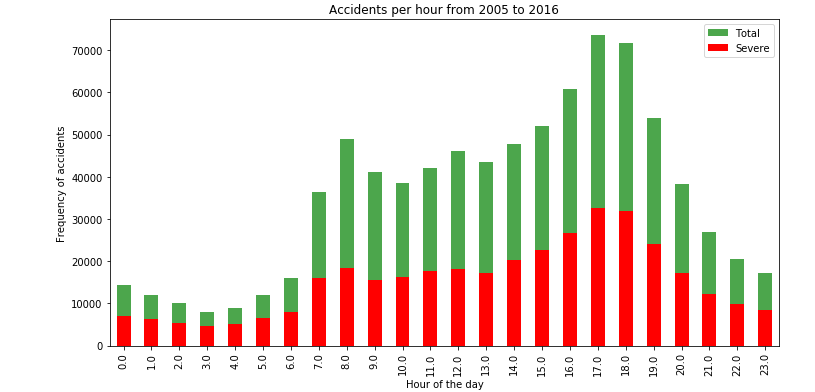
1 368290

Name: sev, dtype: int64



Seasonality The number of traffic accidents decreased over the years from 2005 to 2013, after which the trend became stable.\ Analyszing the yearly trend there is a seasonal pattern where the number of accidents increase around March and then again in September.\ Regarding the day of the week there is not a significant difference between them. There is a steady trend during the week with more accidents on friday, and sunday is the day with less accident of all.\ Lastly analyzing the accidents per hour, there are clearly two spikes, one at 8am, the time people go to work and another one between 5 and 6pm, time when people return home. The number of accidents decreases between these two spikes, nothing unusual but it proves there is a pattern here.





**Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Jaccard | f1-score | Precision | Recall | Time(s) |
| Random Forest | 0.722 | 0.72 | 0.724 | 0.591 | 6.588 |
| Logistic Regression | 0.661 | 0.65 | 0.667 | 0.456 | 6.530 |
| KNN | 0.664 | 0.66 | 0.652 | 0.506 | 200.58 |
| SVM | 0.659 | 0.65 | 0.630 | 0.528 | 403.92 |

For this specific problem precision means the % of predicted severe accidents that were truly severe. The recall instead, is the % of truly severe accidents that were properly predicted. For this specific problem, the recall is more important than the precision as a high recall will favor that all required resources will be equipped up to the severity of the accident.\ In this case, the recall is more important than the precision as a high recall will favor that all required resources will be equipped up to the severity of the accident. The logistic regression, KNN, and SVM models have similar accuracy, however the computational time from the regression is far better than the other two models. With no doubt the Random Forest is the best model, in the same time as the log. res. it improves the accuracy from 0.66 to 0.72 and the recall from 0.45 to 0.59.

**Conclusion**

In this study, we analyzed the relationship between severity of an accident and some characteristics which describe the situation that involved the accident. Initially We thought that features such as atmospheric conditions, the lighting or being a holiday would be the most relevant ones, yet we identified the department, the date and the time of the accident, the road category and type of collision among the most important features that act to severity of the accident. We built four classification models to predict whether an accident would have a high or low severity.

**Observation**

We were able to achieve 68% accuracy, However there was still significant variance that could not be predicted by the models in this study. We think other features like speed or uninterrupted time of travelling could be used to predict a more accurate classification. The next step on this problem could be to add a accident prediction model able to not just predict the accuracy but also the critical time and spots where potential accidents can occur.