

# Capstone Project

## Car accident severity

Applied Data Science Capstone

**Satvik**

[github.com/SatvikSC/IBM-Data-Science-Capstone](https://github.com/SatvikSC/IBM-Data-Science-Capstone)

### **Traffic accidents are...**

- **Cause of 1.35 million deaths globally in 2016.**
- **Main cause of death among those aged 15-29 years.**
- **Predicted to become the 7th leading cause of death by 2030.**

**Predicting the accident severity in advance could be used to send the exact required staff and equipment to the place of the accident, thus saving a significant amount of lives each year.**

**Road safety should be a prior interest for governments, local authorities and private companies investing in technologies that can help reduce accidents and improve overall driver safety.**

# Data

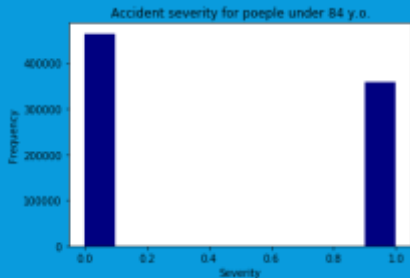
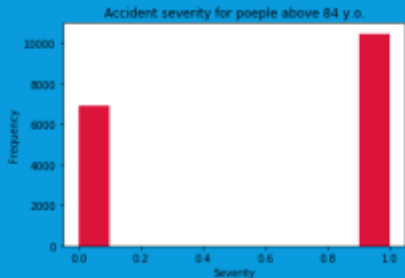
**All the recorded accidents in France from 2005 to 2016, both years included.**

- **Initial dataset from the Kaggle, [here](#).**
- **Pre-selected features on my GitHub, [here](#)**
- **In total 49 features, 839,985 rows in the Kaggle dataset**
- **Redundant and not relevant features were dropped**
- **29 features pre-selected**
- **On the data cleaning missing values and outliers were replaced.**

# EDA-TARGET

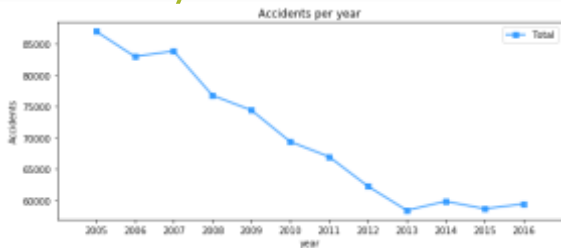
The target feature a binary classifier, describing the accident severity.

- **0:** low severity.
- **1:** high severity, from hospitalized wounded injuries to death.

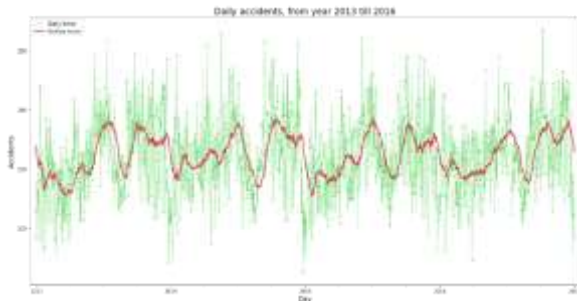


**It is a balanced labeled dataset with more cases of lower severity.**

# EDA-Seasonality



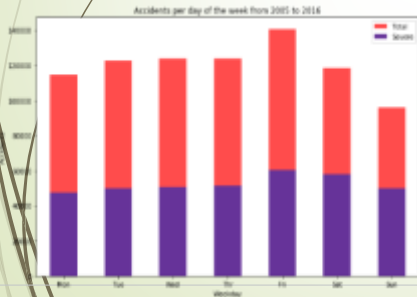
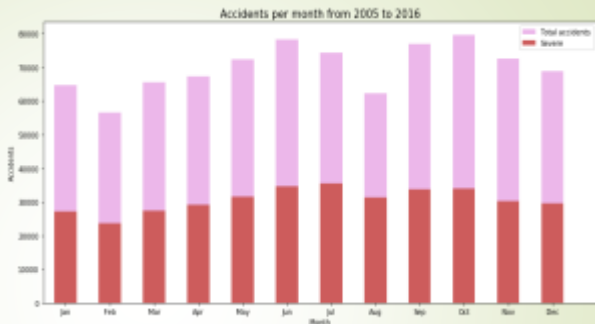
**The number of traffic accidents decreased over the years from 2005 to 2013, after which the trend became stable.**



# EDA-Seasonality

6

**Accidents increase from March to June and then again in September, decreasing at the end of the year.**



**Steady trend during the week.  
More accidents on Friday and  
less on Sunday**

# EDA-Seasonality

The trend of highly severe accidents is proportional to the global trend.



Spikes:

- 8am: people go to work
- 5-6pm: people return home.

# Classification Models

## Random Forest:

- 10 decision trees
  - maximum depth of 12 features
- 

## Logistic Regression

- $c=0.001$

## K-Nearest Neighbor

- $K=16$

## Supervised Vector Machine

- Due to computation inefficiency, training size was reduced to 75,000 samples.



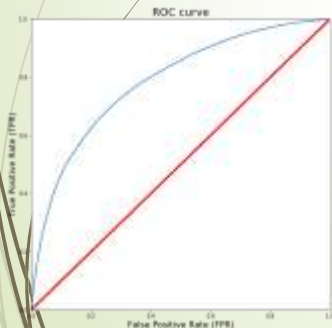
# Results

9

This table reports the results of the evaluation of each model.

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

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.



- Built useful models to predict the severity of a traffic accident.
- Accuracy of the models has room for improvement.
- Future projects:
  - Add features such as vehicle speed and time of uninterrupted traveling.
  - Prediction of potential accident, critical spots and time.