Predicting Traffic Accident Severity

Applied Data Science Capstone

SWATHI R,

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.

In Section 2, Road accidents have been profiled by road category, type of impacting vehicle, type of collision, age of victim, gender and road user category which inter-alia bring out the following:

- National Highways which comprise of 1.94 percent of total road network, accounted for 30.2 per cent of total road accidents and 35.7 per cent of deaths in 2018. State Highways which account for 2.97% of the road length accounted for 25.2 percent and 26.8 percent of accidents and deaths respectively. Other Roads which constitute about 95.1% of the total roads were responsible for the balance 45 % of accidents and 38% deaths respectively.
- In impacting vehicle categories, two-wheelers accounted for the highest share (35.2%) in total accidents and (31.4%) in accident related killings in 2018. Light vehicles comprising cars, jeeps and taxis as a category, ranks second with a share of 24.3 per cent in total accidents and 20.3 per cent in total fatalities.

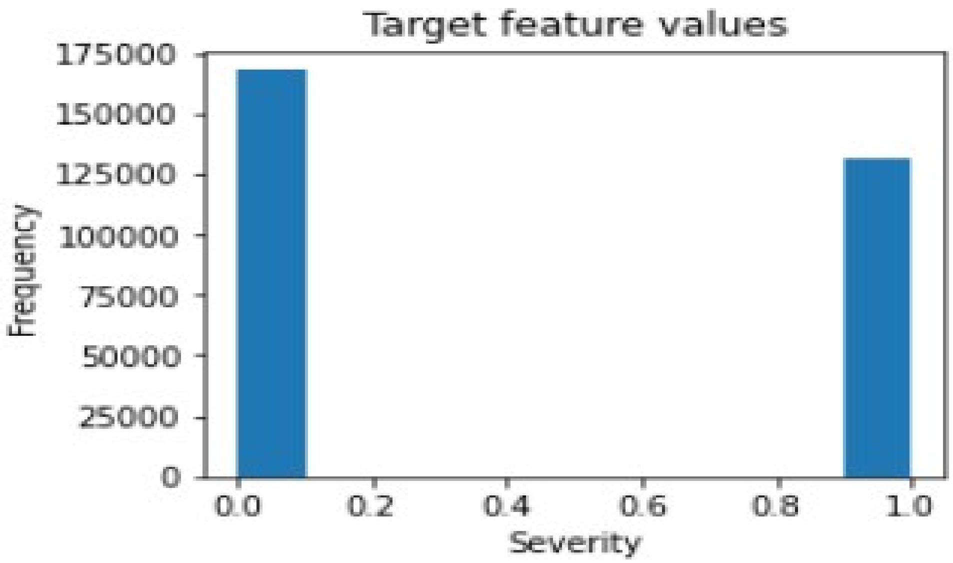
- In terms of accident related killings by type of road user, the number of Pedestrians killed accounted for 15%, the share of cyclists was 2.4% and that of Two wheelers was 36.5%. Together these categories explain 53.9% of the accident related killings and are the most vulnerable category quite in line with global trends.
- During 2018, like the previous two years, young adults in the age group of 18 - 45 years accounted for nearly 69.6 percent of road accident victims. The working age group of 18 – 60 accounted for a share of 84.7 percent in the total road accident deaths.
- The number of hit and run cases in 2018 accounted for 18.9% of the deaths compared to 17.5% in 2017. Head on collision , followed by Hit and run cases followed by Hit from the back accounted for almost 56% of persons killed in 2018. The category which registered the maximum increase in terms of persons killed in 2018 was collision with parked vehicles.
- The share of males in number of total accident deaths was 86% while the share of females hovered around 14% in 2018

Data

- All the recorded accidents in France from 2005 to 2016, both years included. Initial dataset from the Kaggle, here. Pre-selcted 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

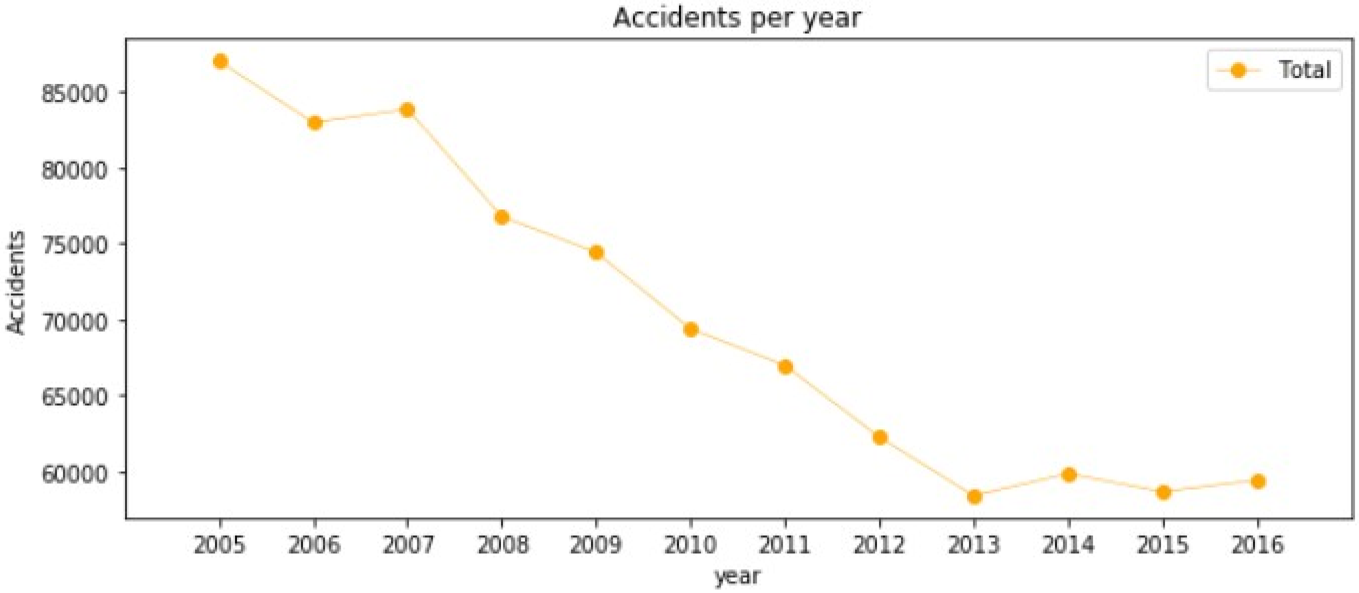
EDA-Target

- 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



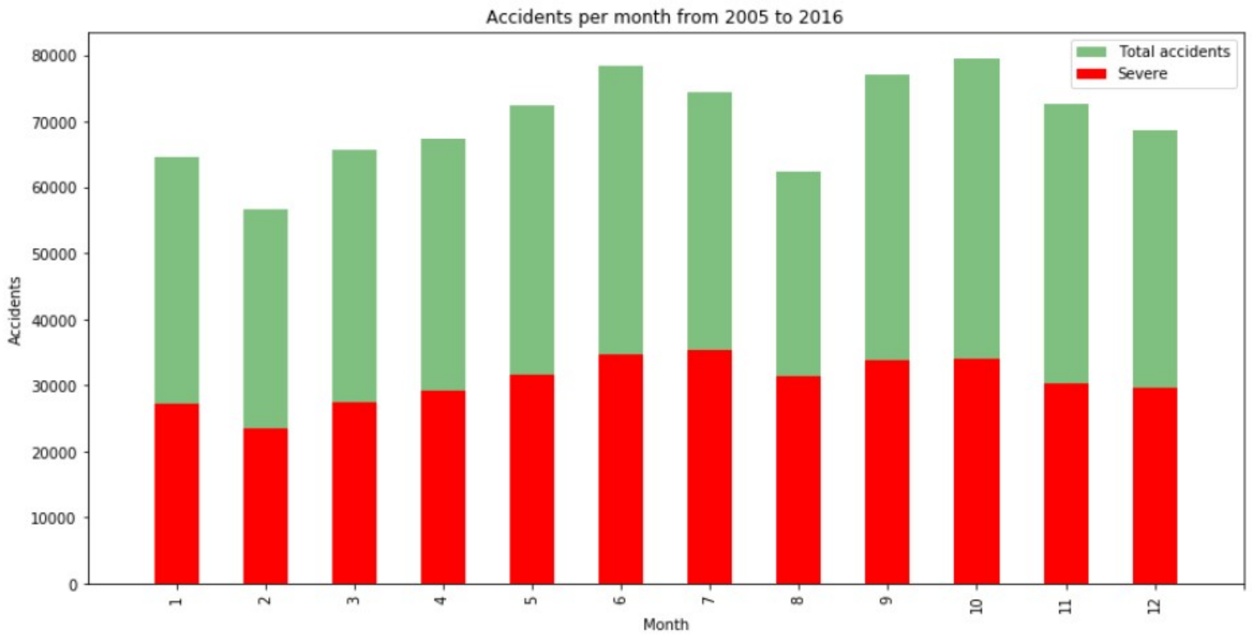
EDA-Seasonality

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

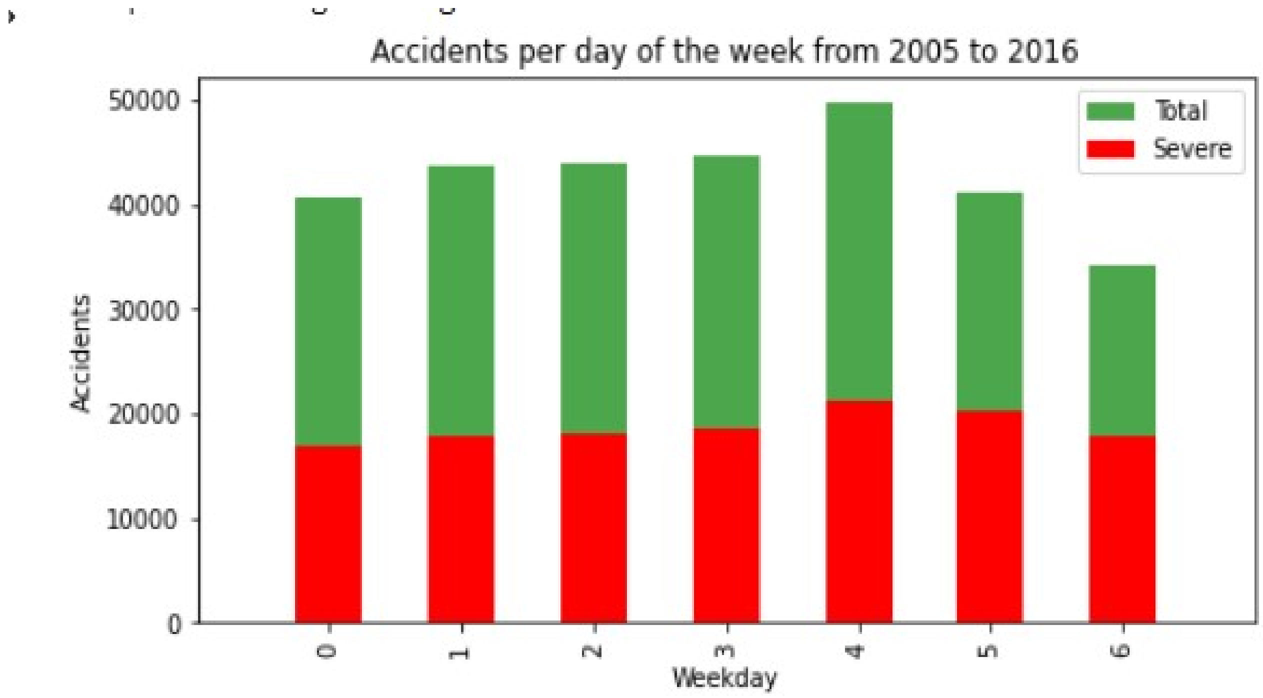


EDA-Seasonality

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

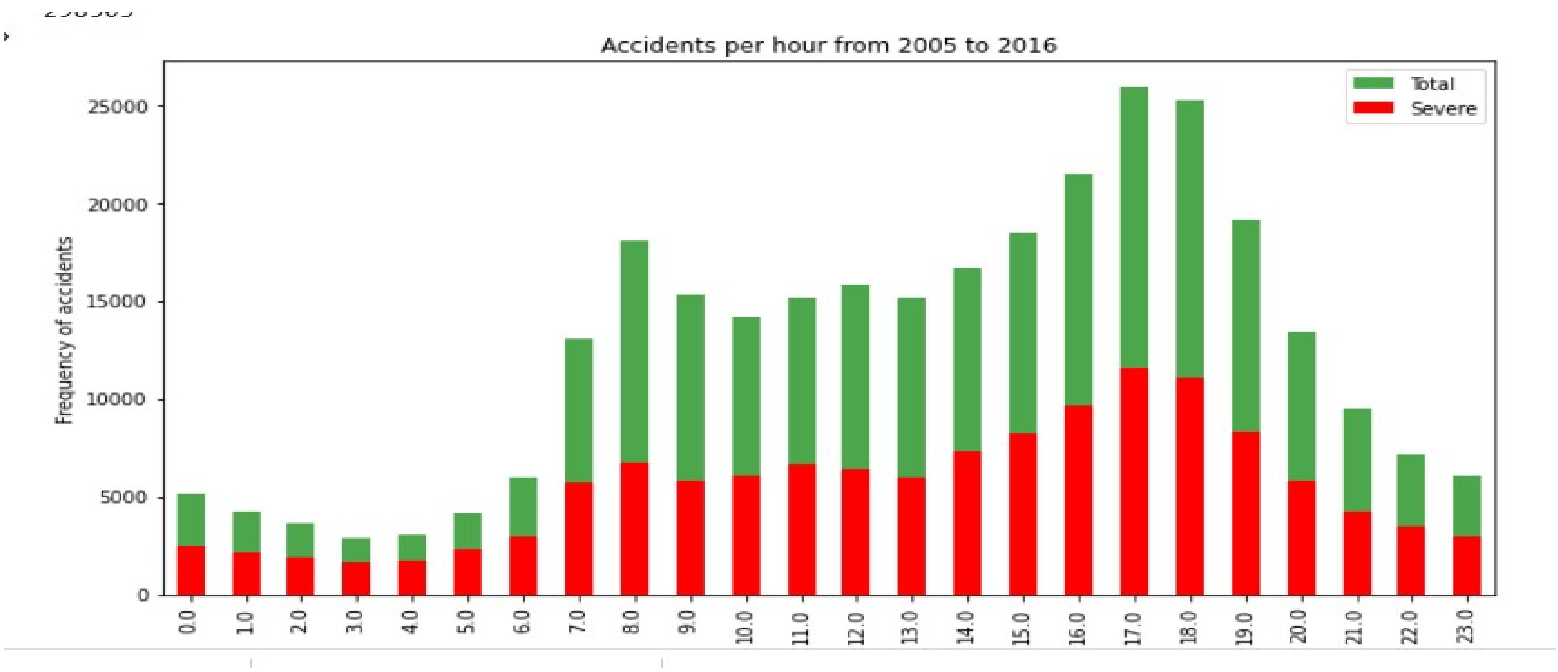


- Steady trend during the week. More accidents on Friday and less on Sunday



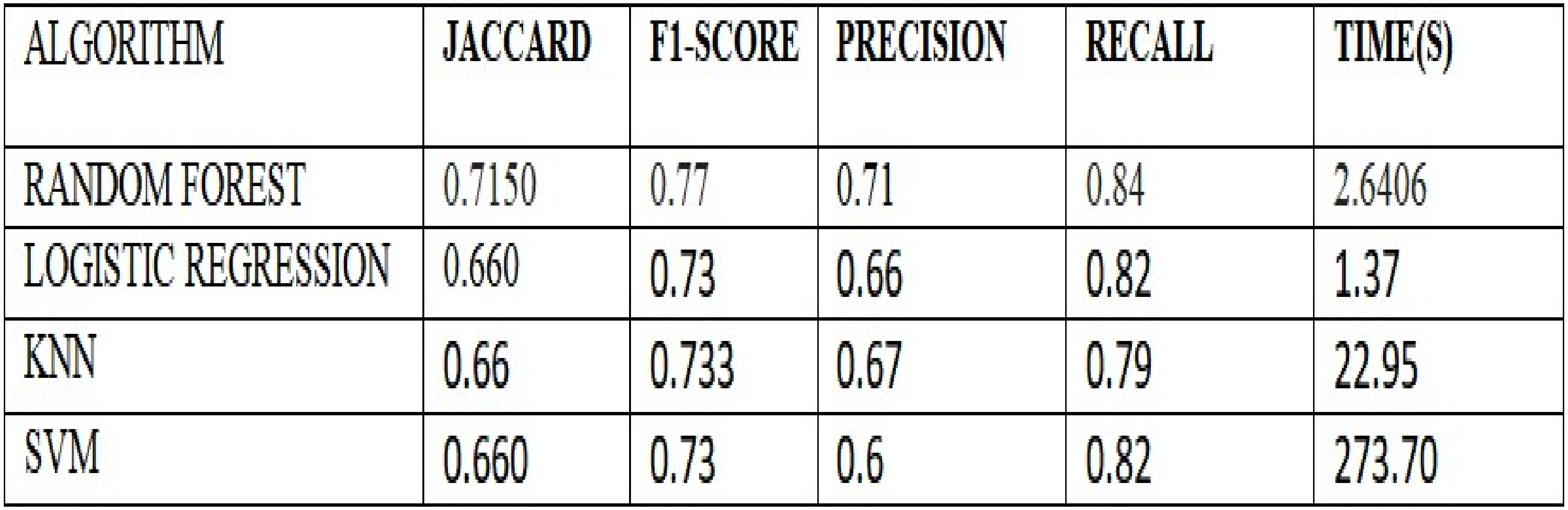
EDA-Seasonality

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



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

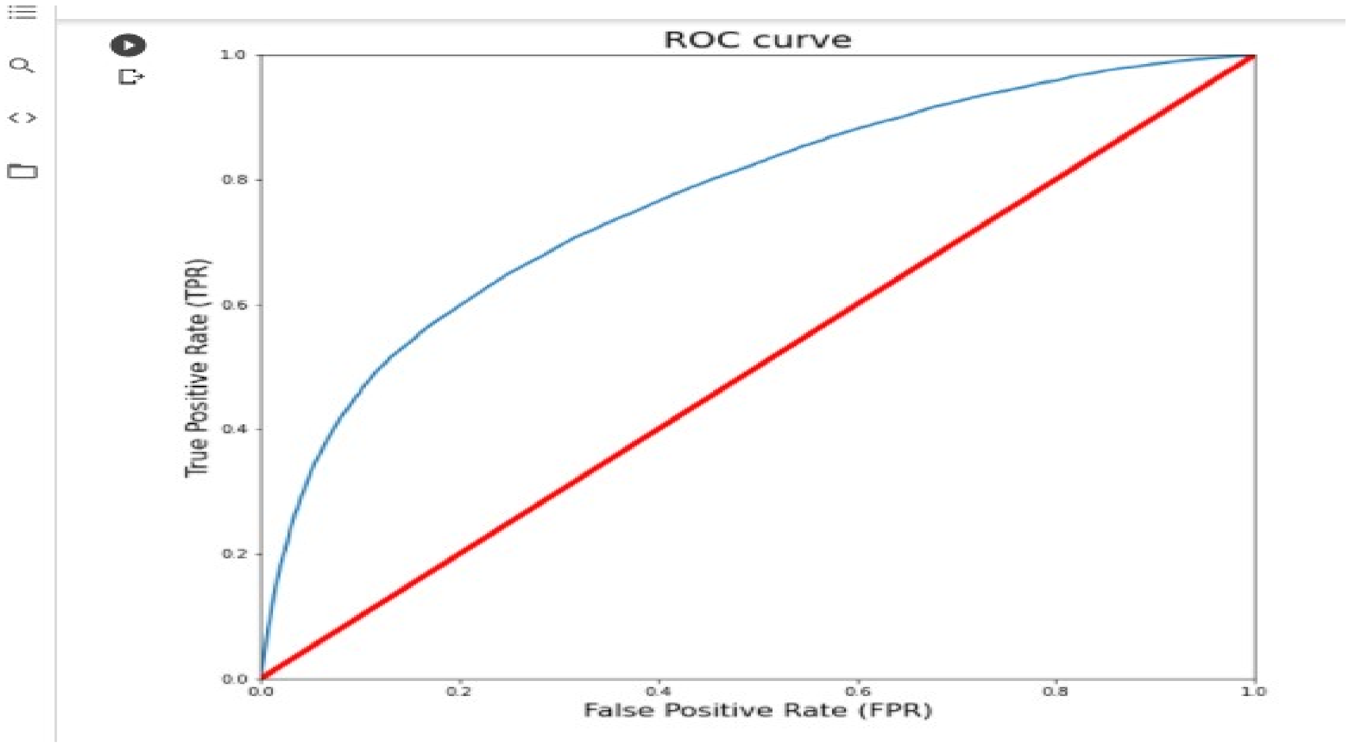
Classification Models



ALGORITHM	JACCARD	F1-SCORE	PRECISION	RECALL	TIME(S)
RANDOM FOREST	0.7150	0.77	0.71	0.84	2.6406
LOGISTIC REGRESSION	0.660	0.73	0.66	0.82	1.37
KNN	0.66	0.733	0.67	0.79	22.95
SVM	0.660	0.73	0.6	0.82	273.70

Results

- 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 0.71 and the recall from 0.79 to 0.84.



Conclusion and future projects

- 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.