Predicting Traffic Accident Severity

Applied Data Science Capstone

Ravi Singh

https://github.com/rasingh6/IBM-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.

Accident severity Prediction in advance is important for effective and efficiency of resolving accident related delays.

Road safety should be a prior interest for governments, and non-government organizations by exploring technologies which can help reduce accidents and improve overall driver safety.

DA
TA

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

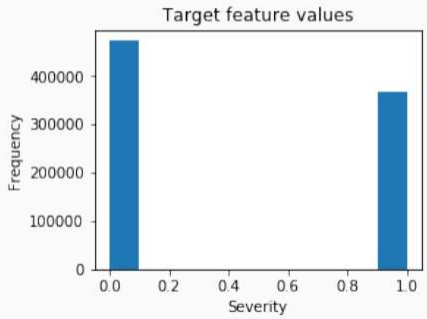
- Initial dataset from the Kaggle, here.
- Pre-selcted features on my GitHub, here
  - In total 49 features, 839,985 rows in the Kaggle dataset
  - Redundant and not relevant features were dropped
- 29 features pre-selected
- Cleaning data missing values and outliers were replaced.

EDA-TARGET

The targeted binary classifier, describe accident severity.

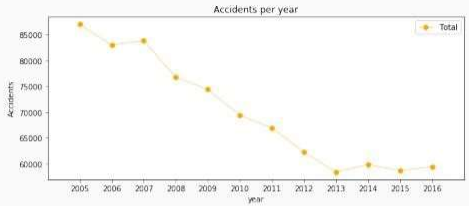
0: low severity.

1: high severity, from hospitalized wounded injuries to death.



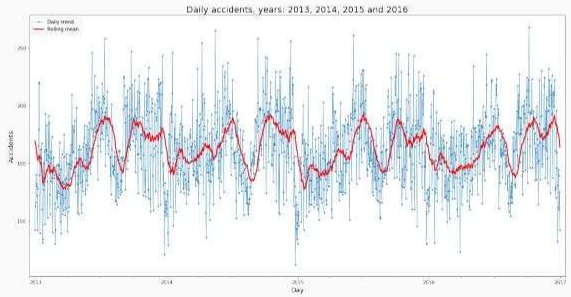
A balanced labeled dataset with more cases of lower severity.

EDA-SEASONALITY



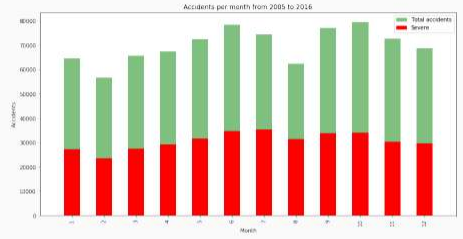
Number of traffic accidents decreased between years 2005 to 2013,

After which the trend became stable.

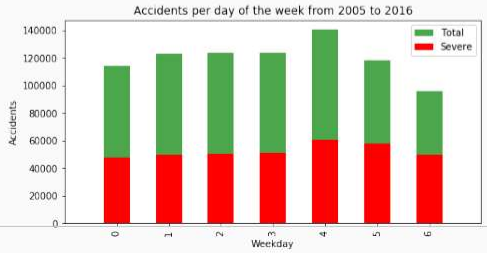


EDA-SEASONALITY

Accidents increase March to June and later on in September,
Decreases at the end of the year.

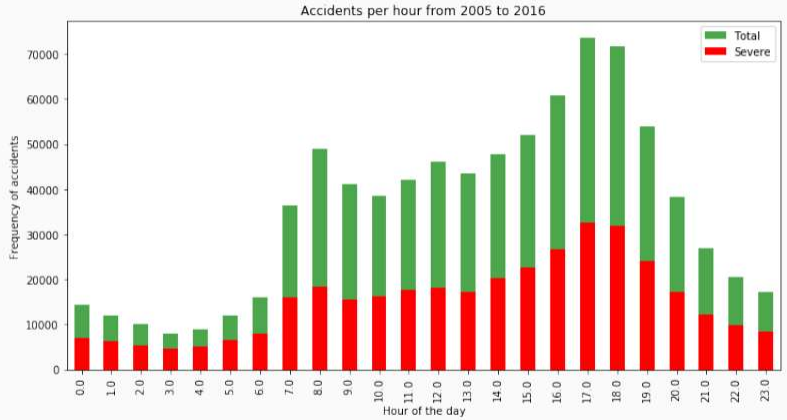


Steady Accident during the **week**. More accidents on Friday and lest on Sunday



EDA-SEASONALITY

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



Spikes:

8:00 am: People on way to work

5:00-6:00 pm: People return from home.

CLASSIFICATION MODELS

Random Forest:

10 decision trees

maximum depth for 12 features

Logistic Regression

c=0.001

K-Nearest Neighbor

K=16

Supervised Vector Machine

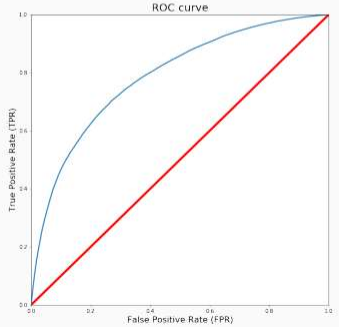
For computation training size reduced to 75,000 samples.

Results

Table reports the results of the evaluation of each model.

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

Random Forest is the best model, Same way 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 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.