PREDICTING TRAFFIC ACCIDENT SEVERITY

ACCIDENTS DATASET FROM FRANCE (2005-2016) / IBM/COURSERA DATA SCIENCE CAPSTONE PROJECT

HTTPS://GITHUB.COM/MABRAHAMM/GRAND-FINALE-APPLIED-CAPSTONE

INTRODUCTION

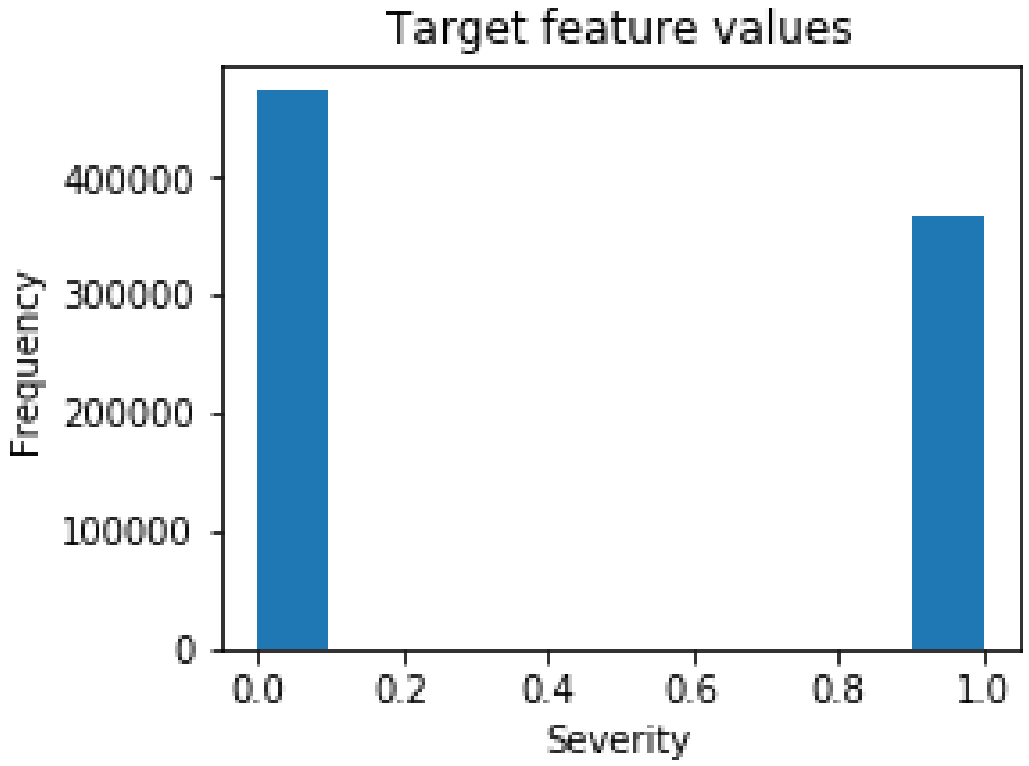
- Traffic accidents
  - Caused 1.35 million deaths worldwide in 2016
  - Main cause of death among youth (15-29)
  - Predicted to become 7th leading cause of death in 2030
- Predicting the severity of an accident in advance can allow emergency rescue teams to send the right number of staff to the accident scene hence saving many lives
- Reducing traffic accidents should be a priority for governments and private companies that invest in technologies to reduce road accidents

DATA

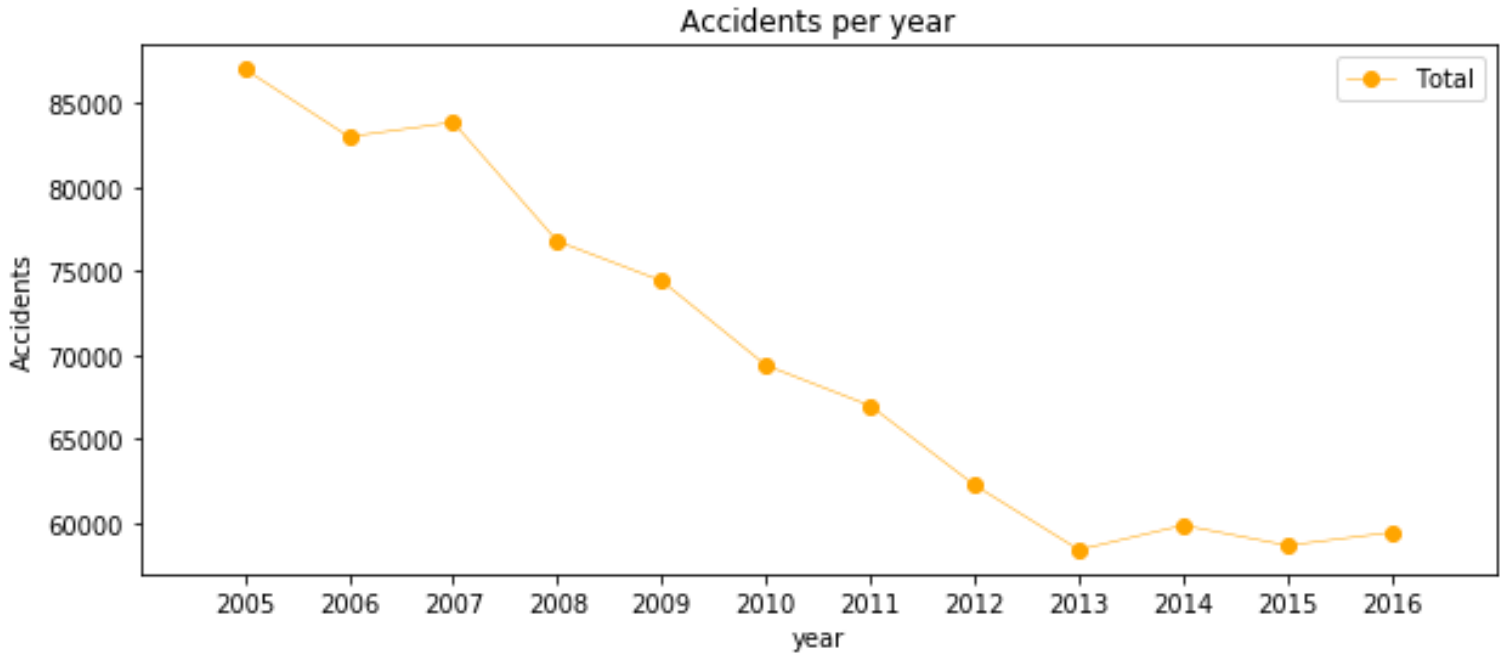
- I have obtained a dataset of all accidents that occured in France from 2005 to 2016
- Dataset was obtained from Kaggle
- Preselected features can be found in Github
  - Dataset from Kaggle has 839,985 rows and 49 features
  - Irrelevant and redundant features were dropped
- 29 features were preselected
- During data cleaning missing values and outliers were replaced

EXPLORATORY DATA ANALYSIS

- Target feature is a binary classifier, describing the accident severity
  - 0: low severity
  - 1: high severity (wounded/death)
- Clearly a balanced data with more cases for low than high severity

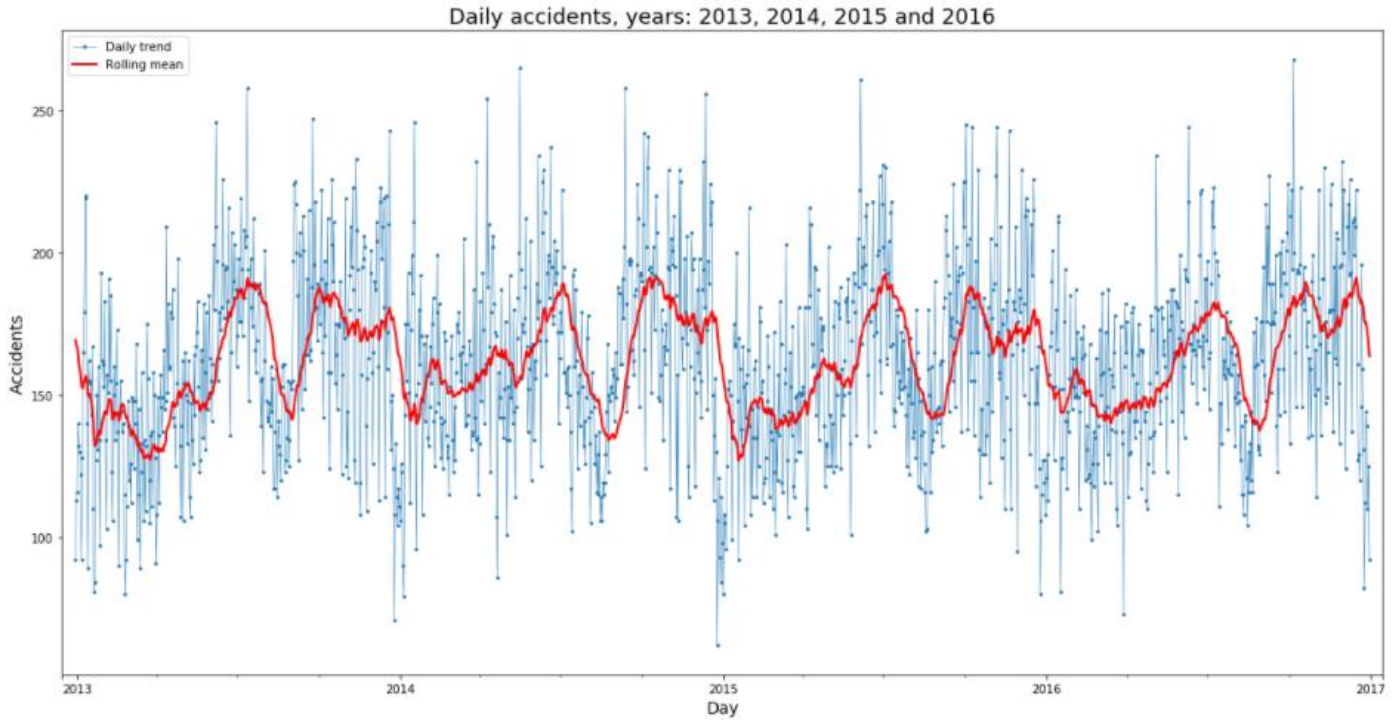


EXPLORATORY DATA ANALYSIS - SEASONALITY



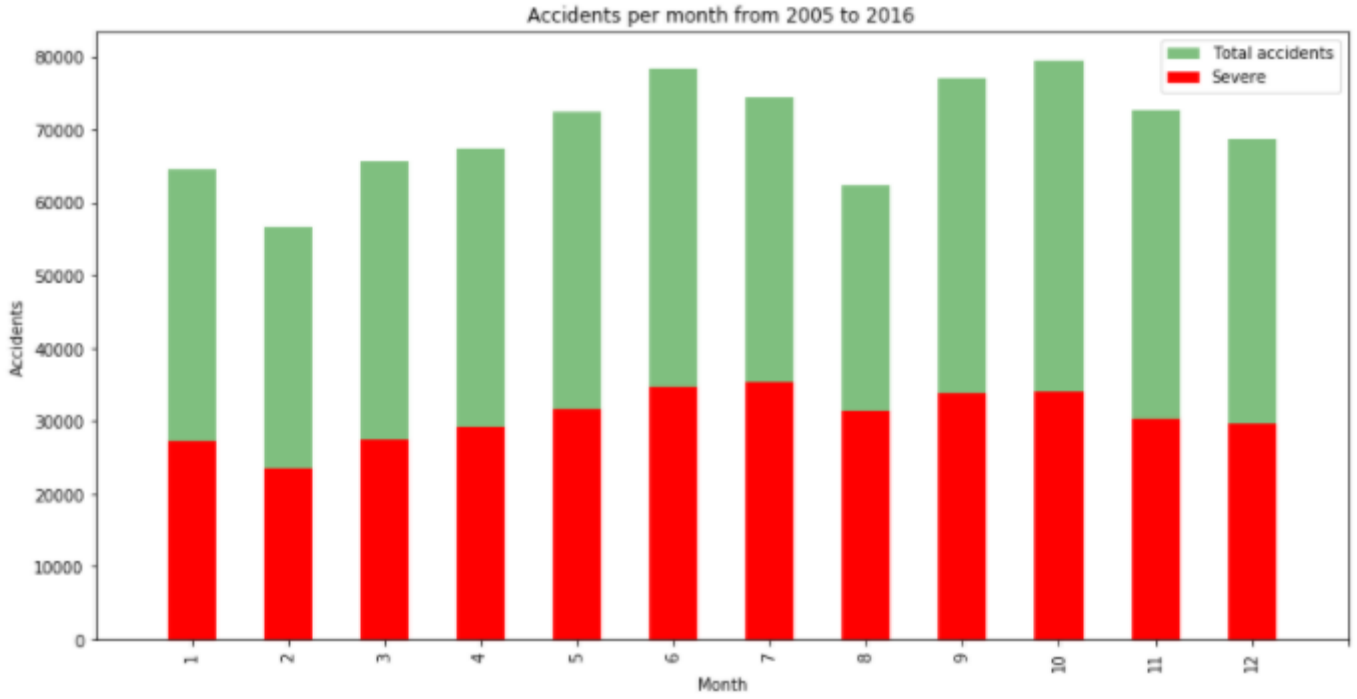
- The number of accidents has decreased over the years from 2005 to 2013 after which there was a stabilization

EXPLORATORY DATA ANALYSIS - SEASONALITY



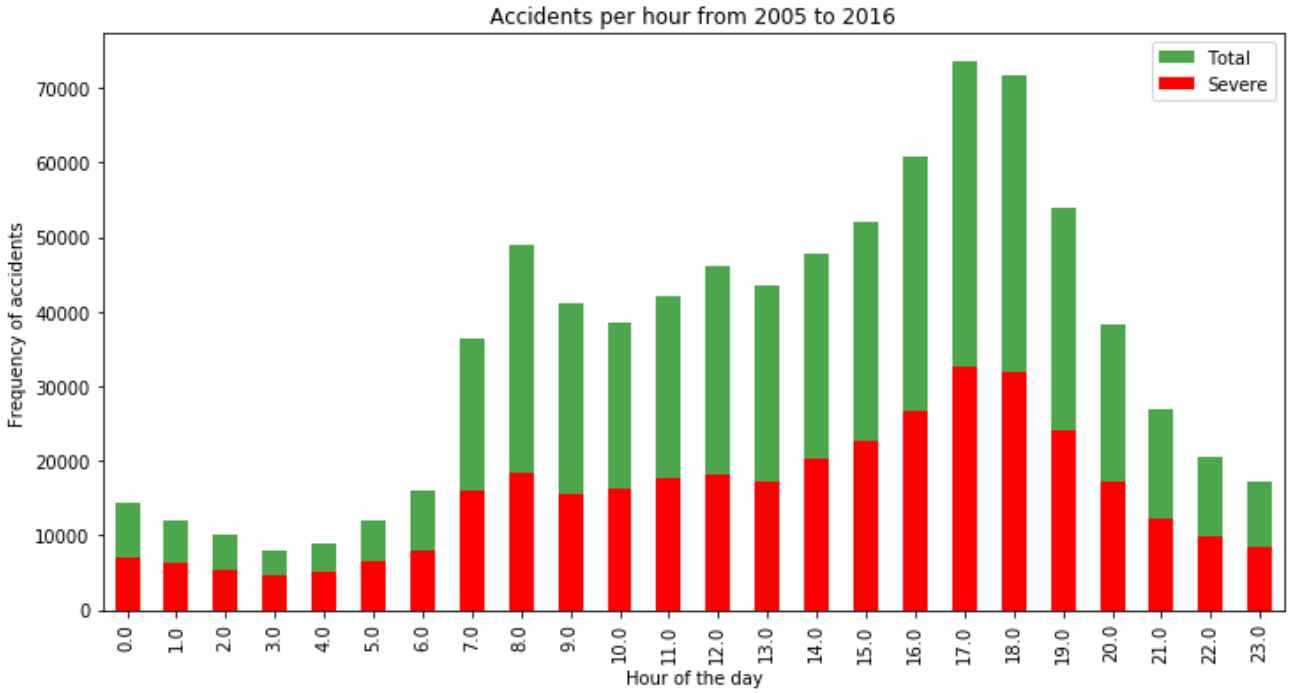
- The number of accidents per annum has stabilized from 2013 to 2016

EXPLORATORY DATA ANALYSIS - SEASONALITY



- Accidents increase from March to June then again in September before decreasing towards the last two months of the year

EXPLORATORY DATA ANALYSIS - SEASONALITY



- The trend of accidents during certain hours maps the global reality
  - Spike in accidents at 8 am when people go to work
  - Spike in accidents between 5 pm and 6 pm when people are returning home from work

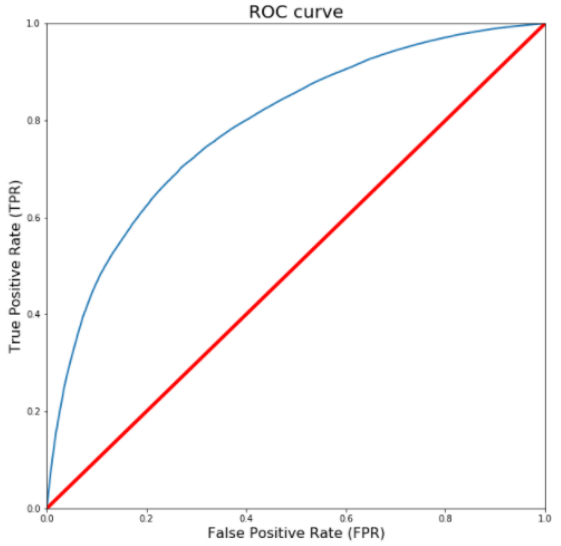
CLASSIFICATION METHODS

- Random forest: 10 decision trees, maximum depth of 13 features and maximum of 8 features compared for the split
- Logistic Regression: c=0.001
- K-Nearest Neighbour: k=16
- Supervised Vector Machine: size of training set = 75,000 samples (reduced due to computational inefficiency)

RESULTS

- The table represents the evaluation of each model
- Random Forest is the best model with an accuracy of 0.72 and computational time efficiency of 6.588 seconds

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



CONCLUSION AND DISCUSSION

- Build useful models for prediction of accident severity
- Accuracy of model could be improved with the inclusion of speed as a factor
- Model could also be useful if drivers could be forewarned about hot spots where accidents are likely to occur