IBM Applied Data Science Capstone Project

Predicting Road Accident Severity

By: Ritvik Shyam

1. Introduction/Business Problem

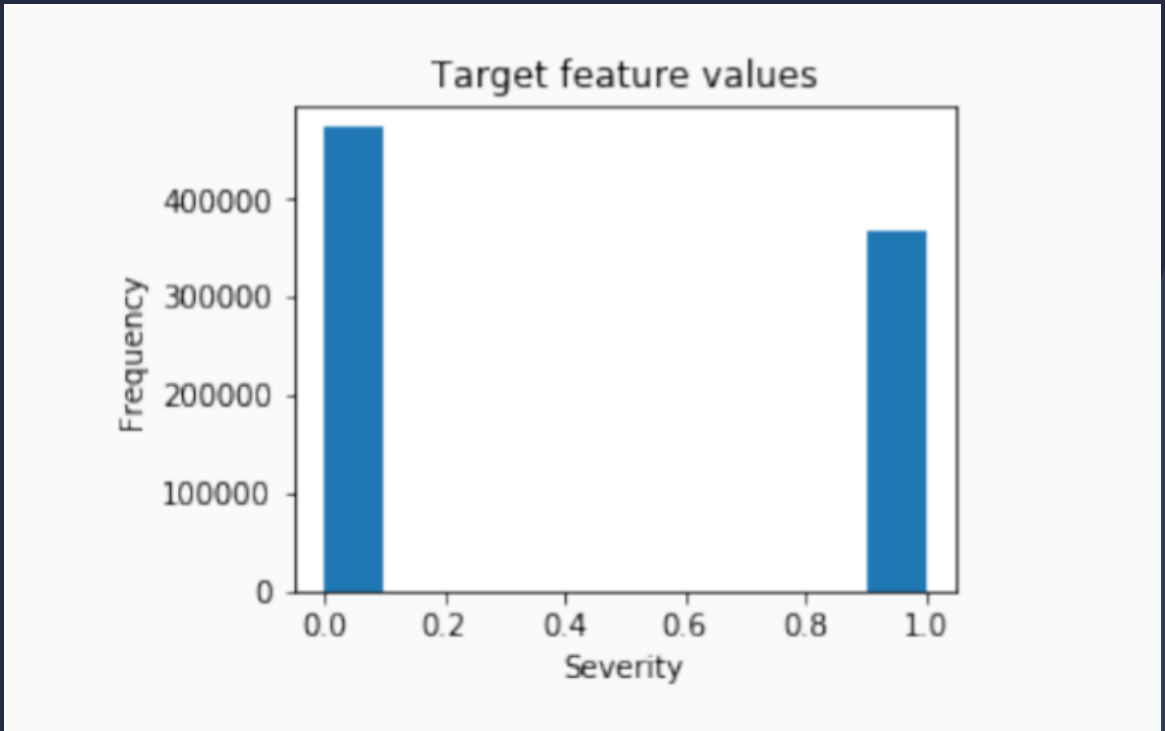
- Traffic accidents are the cause of 1.35 million deaths globally in 2016.
- Main cause of death among those aged 15–29 years.
- Predicted to become the 7th leading cause of death by 2030.

- The introduction and business problem is to create a predictive machine learning model intended for car drivers to enable the reduction in frequency of car collisions through the implementation of algorithms that predict the severity of an accident. The functional aim of the model is to alert the car driver to be cautious about safety when the conditions listed above are unfavourable by analysing a significant range of factors including weather conditions, public events, road-works and traffic congestions.
- The tangible outcome of this project is to utilize its insights and findings to enable relevant stakeholders such as law enforcement and private companies involved in road contracts to allocate their resources more effectively in advance to prevent potential accidents, thus significantly saving resources and alerting drivers accurately in order to save thousands of human lives.

2. Data

- All the recorded accidents in France from 2005 to 2016, both years included
- Initial dataset from the Kaggle
- Pre-selected features and data selection (check file on GitHub)
- 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

3. EDA-Target

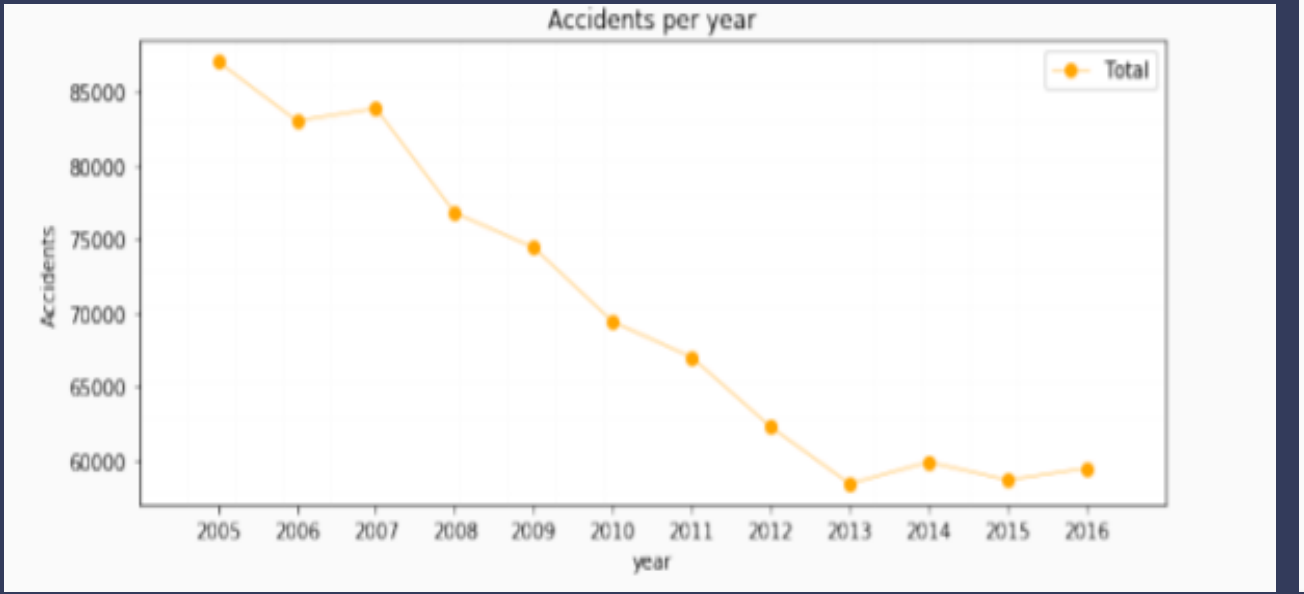


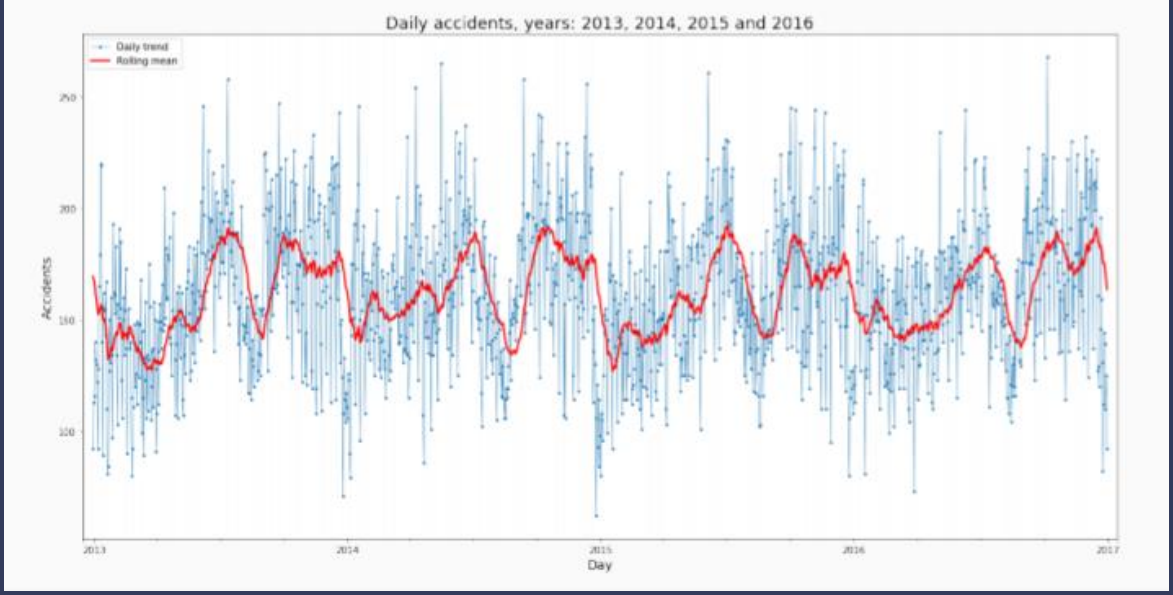
- The target feature a binary classifier, describing the accident severity.
- 0: low severity.
- 1: high severity, from hospitalized wounded injuries to death.

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

4. EDA-Seasonality

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

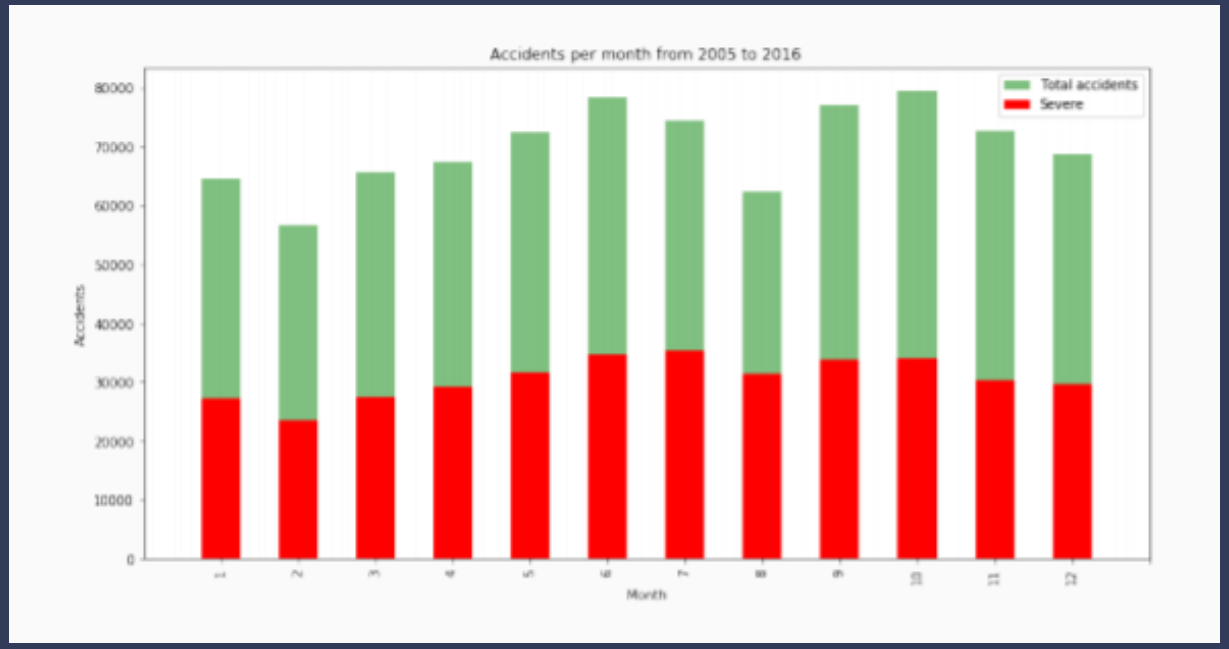


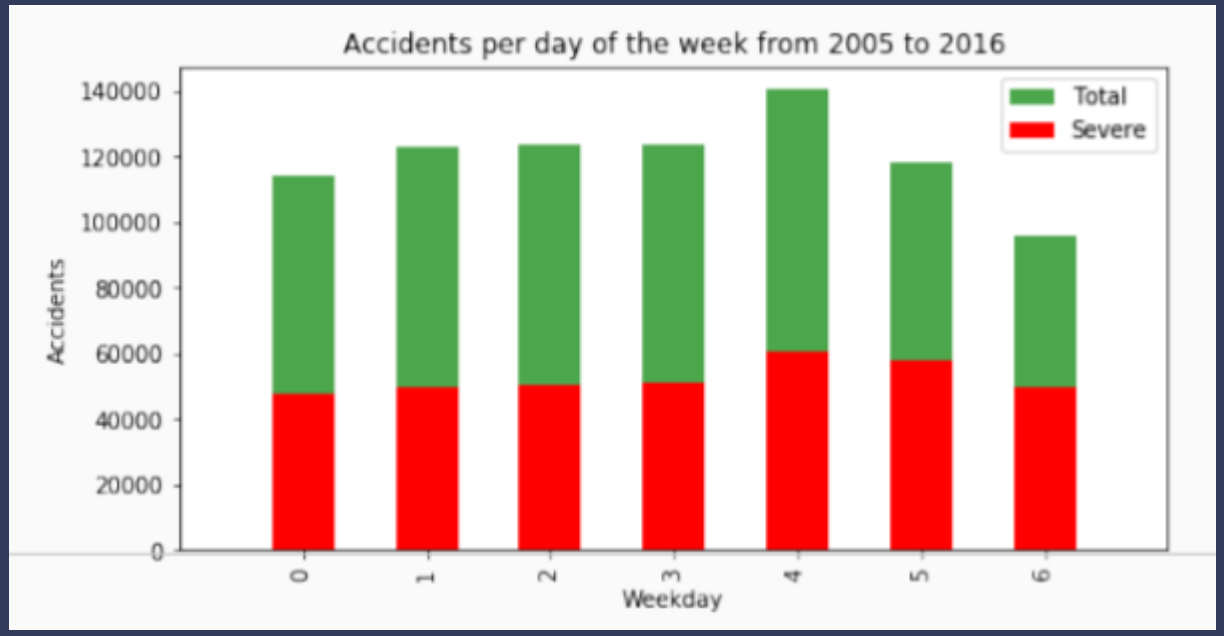


5. EDA-Seasonality (continued)

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

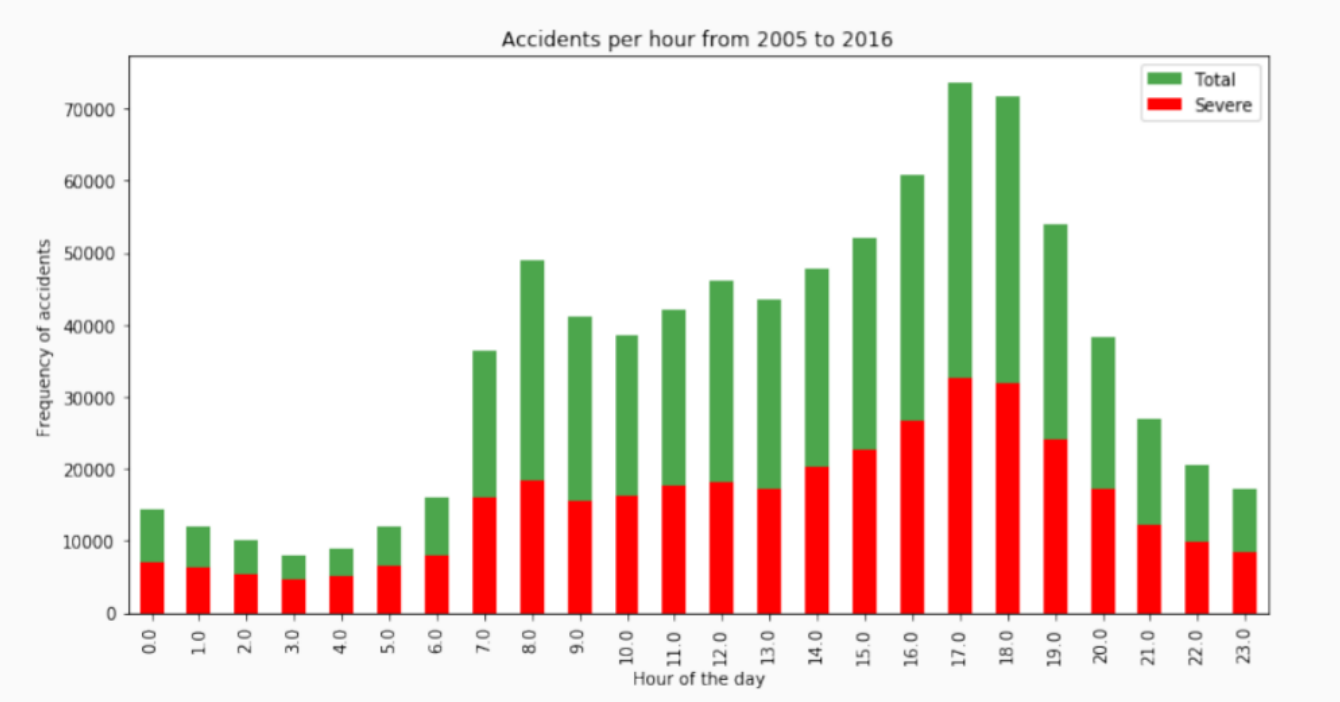
- Steady trend during the week. More accidents on Friday and less on Sunday





6. EDA-Seasonality (continued)

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



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

7. Classification Models

- Random Forest:
  - 10 decision trees
  - maximum depth of 12 features

- Logistic Regression
  - c=0.001

- K-Nearest Neighbor
  - K=16

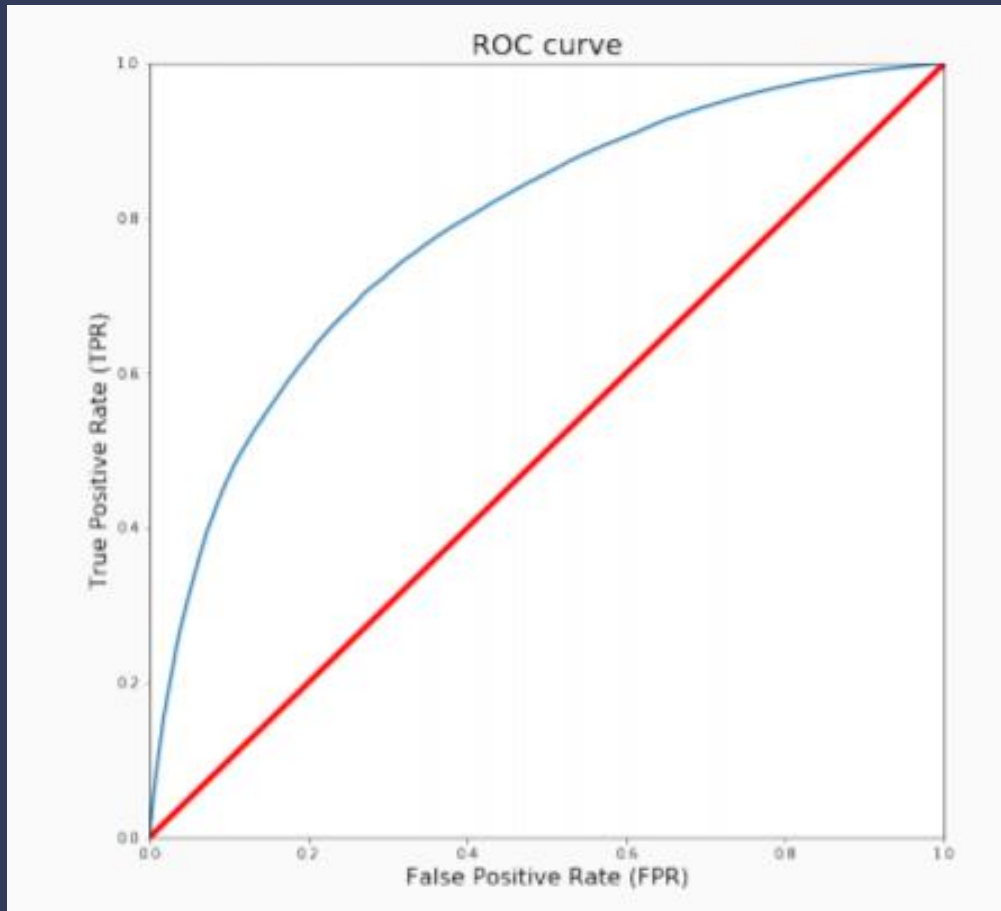
- Supervised Vector Machine
  - Due to computation inefficiency, training size was reduced to 75,000 samples.

8. Results

- This 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

- 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.



9. Conclusion and possible extensions

- 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.

10. Acknowledgements

- Kaggle (for the data source)
- IBM Watson Studio for providing a platform to generate Jupyter Notebooks

And last but not the least,

IBM Coursera for the providing an opportunity to develop my skills and equip myself for the most on-demanding knowledge and application of this century!

Thank you,

Ritvik Shyam