Predicting Traffic Accident Severity

Coursera Final Project

Arjun Pandey

Background

Traffic accident key facts:

- Approximately 1.35 million people die each year as a result of road traffic crashes.
- The2030 Agenda for Sustainable Development has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2020.
- Road traffic crashes cost most countries 3% of their gross domestic product.
- More than half of all road traffic deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists.
- 93% of the world's fatalities on the roads occur in low- and middle-income countries, even though these countries have approximately 60% of the world's vehicles.
- Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.

Who is at risk?

- Socioeconomic status: 90% road traffic deaths occur in low- and middle countries
- Age: Childern and young adults ages 5-29 years
- Sex: Young male are more likely involved in road traffic crashes than females

Risk factors:

- The safe system approach: accommodating human error
- Speeding
- Driving under the influence of alcohol and other psychoactive substances
- Nonuse of motorcycle helmets, seat-belts, and child restraints
- Distracted driving
- Unsafe road infrastructure
- Inadequate post-crash care
- Inadequate law enforcement of traffic laws

Data Preparation

The data set were downloaded from the source Kaggle
Data sets: France from year 2005-2016

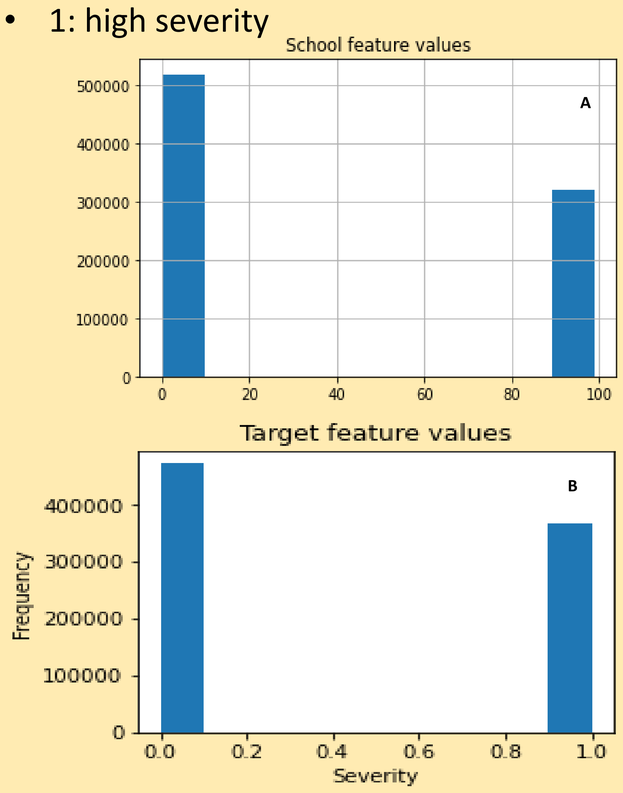
Initial methods includes:

- Combine of all the features
- Dropping irrelevant features
- 29 features were selected
- Missing values and outliers were replaced

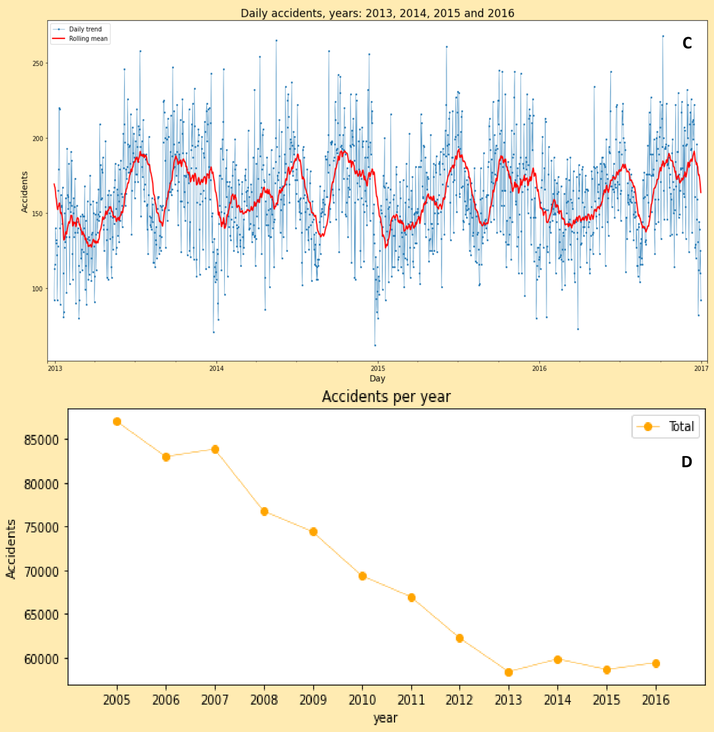
Exploratory Data Analysis:

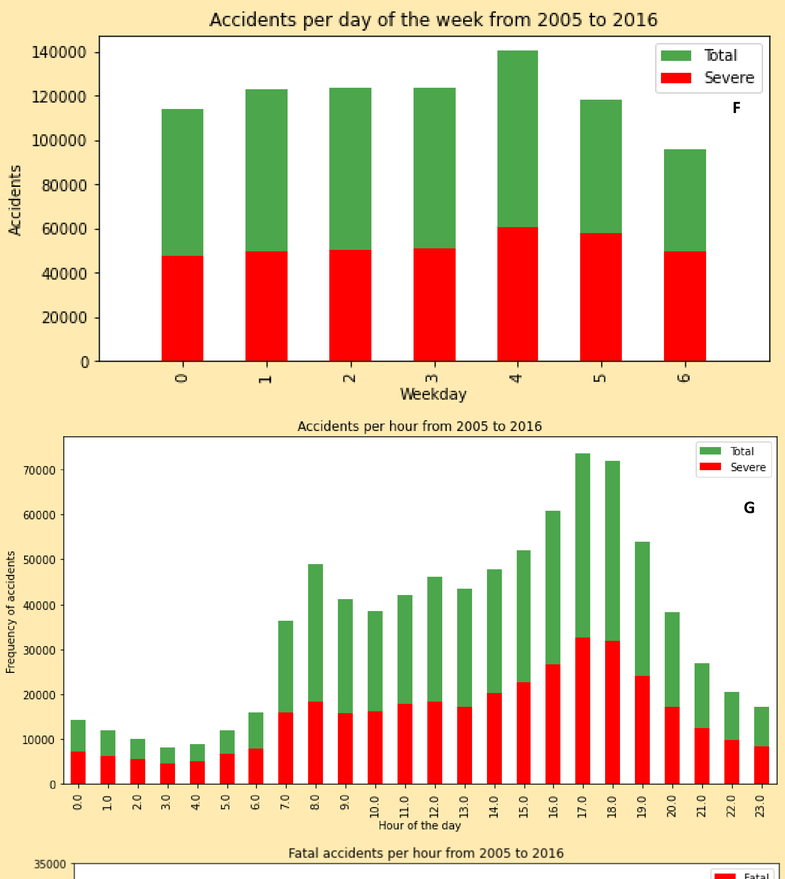
Classified into binary

- 0: low severity
- 1: high severity

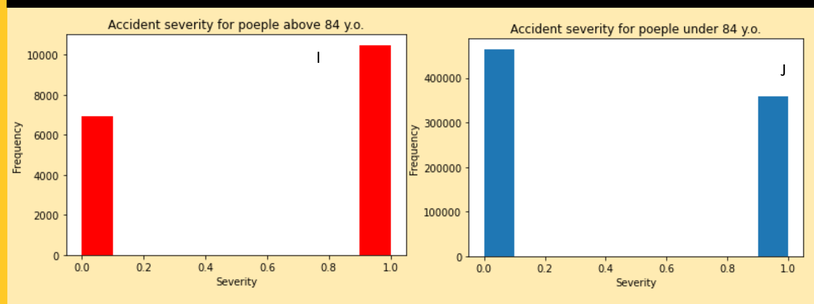


Data analysis- Seasonality





Severity among age factor



Model development

Random Forest:
Decision tress: 10
Maximum depth of features : 12

Logistic Regression
C= 0.001

K_Nearest Neighbour
K=15

Supervised Vector Machine
Training size=75,000 samples

Results of each model

Random Forest:
Decision tress: 10
Maximum depth of features : 12

Logistic Regression
C= 0.001

K_Nearest Neighbour
K=15

Supervised Vector Machine
Training size=75,000 samples

Result and Discussions

Figures:

A&B) More cases of lower severity within one feature with balanced dataset can be observed
C&D) Number of accidents over the year
E) More number of accidents were observed on the March and September of the year
F) More accidents was observed on Friday (4) and less on Sunday (6)
G) More accidents was observed on morning 8 am and 5-6 pm, which is office hours, that make sense and is proportional to the global trend

Algorithm	Jaccard	F1-score	Precision	Recall	Time(S)
Random Forest	0.469875	0.63934	0.734314	0.566119	4.736231
Logistic Regression	0.400583	0.572023	0.665856	0.50137	5.833542
KNN	0.375631	0.546122	0.665315	0.463148	181.549
SVM	0.356917	0.52607	0.691429	0.424539	295.6989

From the above table, we can see that model Random Forest shows better performance in compared to the other models. With reference to the time, recall and precision, random forest consider as best model. However, logistic regression shows comparable performance.

References

World Health Organization (WHO). Global Status Report on Road Safety 2018. December 2018. [cited 2019 April 8]. Available from URL: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/external icon

Centers for Disease Control and Prevention (CDC), National Center for Injury Prevention and Control (NCIPC). Web-based Injury Statistics Query and Reporting System (WISQARS). [cited 2019 November 4]. Available from URL: http://www.cdc.gov/injury/wisqars

https://www.cdc.gov/injury/features/global-road-safety/index.html#:~:text=Each%20year%2C%201.35%20million%20people,on%20roadways%20around%20the%20world.&text=Every%20day%2C%20almost%203%2C700%20people,pedestrians%2C%20motorcyclists%2C%20and%20cyclists.

Acknowledgements

Thank you so much for the Coursera providing the course which help me a lot to figure out real data, their analysis.