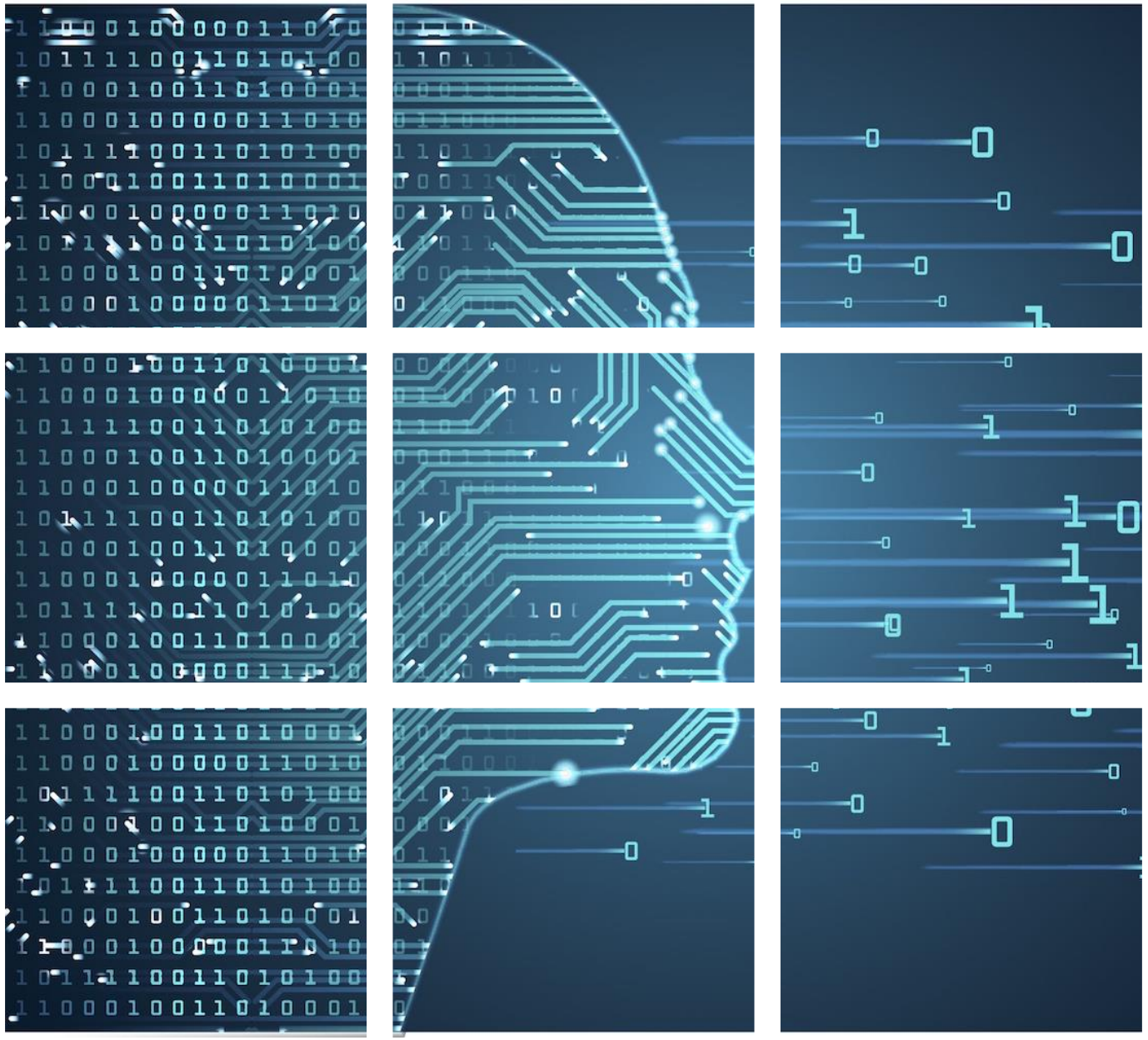
Predicting Traffic Accident Severity

Coursera IBM Data Science Capstone Project By: Mohamad Bouzi

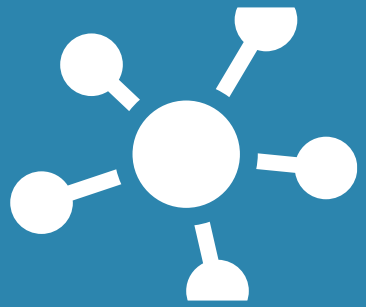
Traffic Accident

- 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.

Data Understanding



Recorded Accident in France from 2005 to 2016



Kaggle.com

is the Data Source for this Project

The Dataset was including **49 Feature** and **839,985 rows**

Redundant and not relevant features were dropped and **29 Features** pre-selected

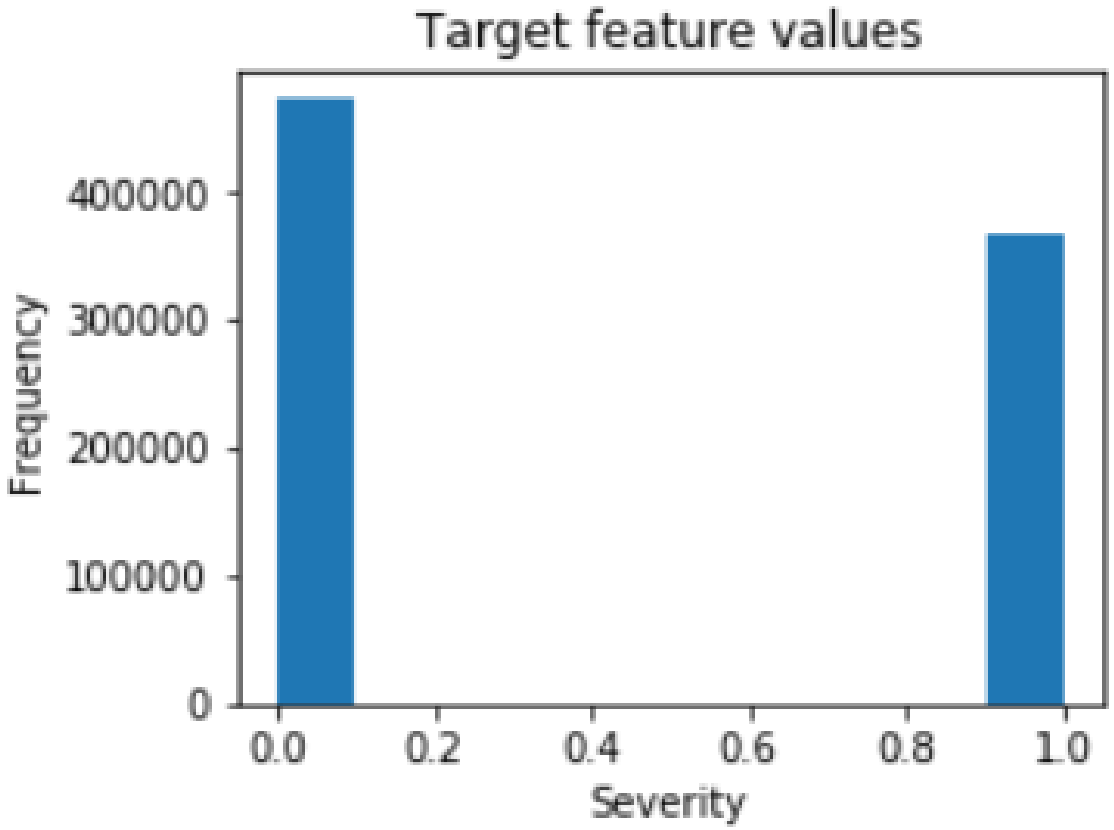
In **Data Cleaning** Phase missing values and outliers were replaced.

Target Classes

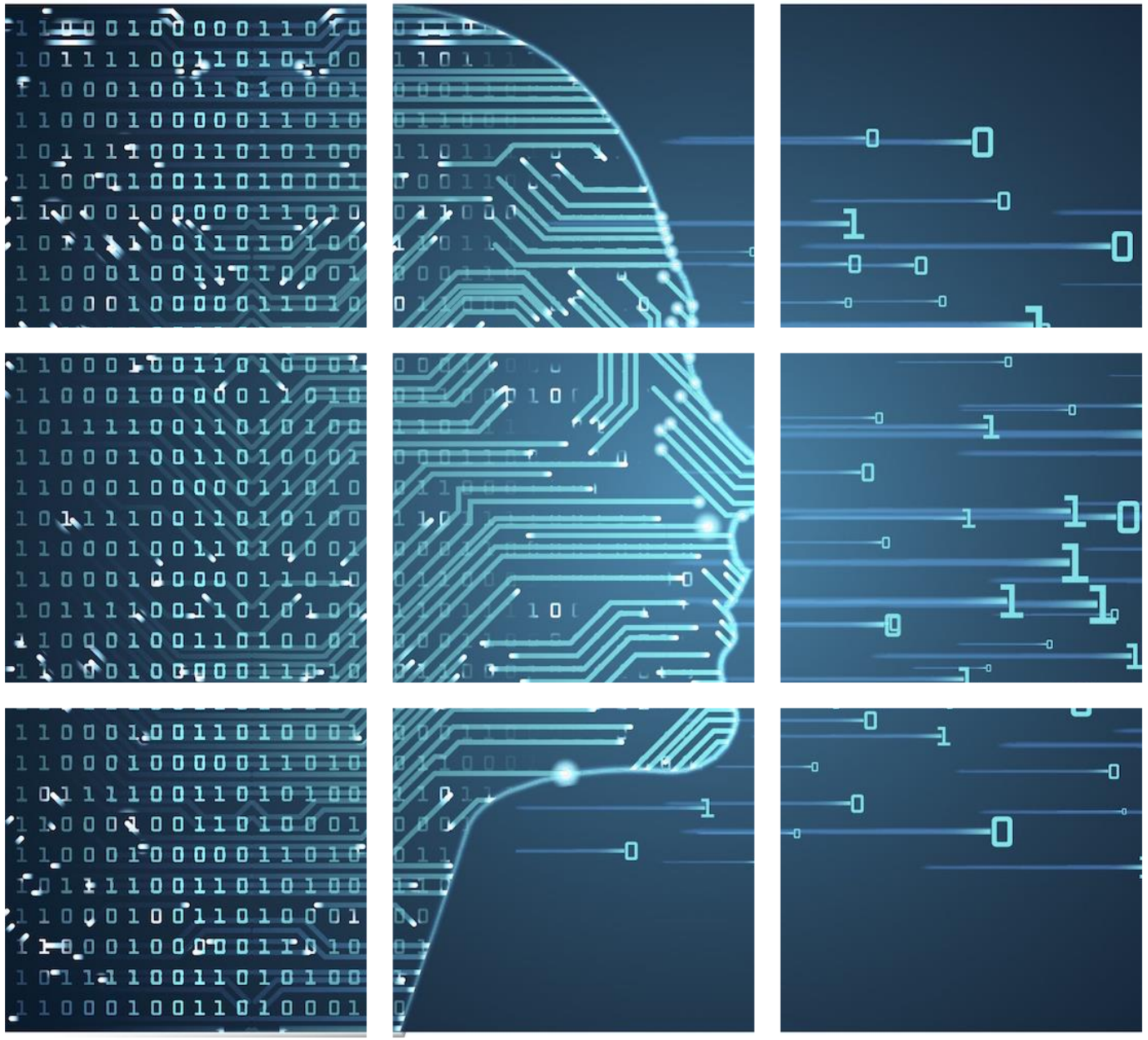
The **Target** feature is a binary classifier, describing the **accident severity** as following;

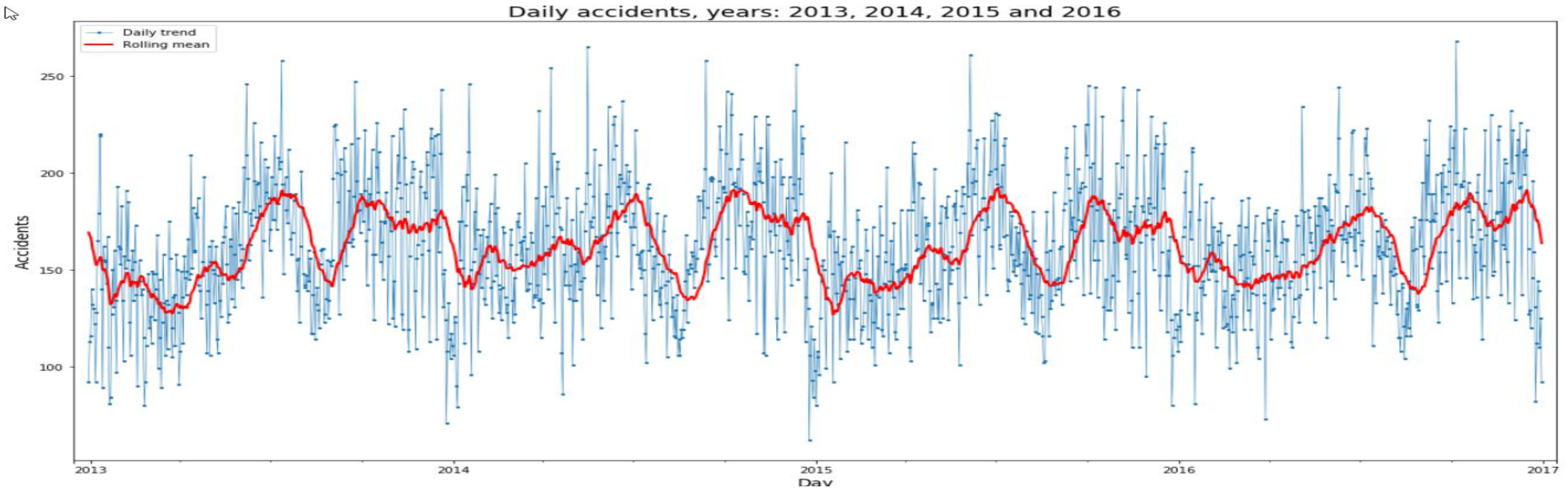
- Value **0** mean **Low severity**
- Value **1** mean **High severity**, and it varied from hospitalized wounded injuries to death cases.

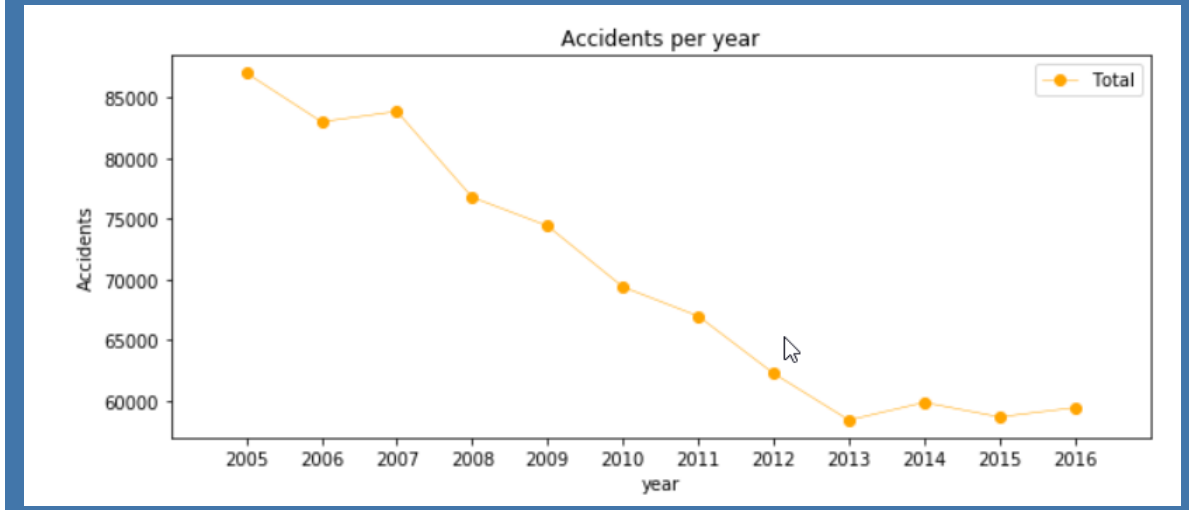
And To avoid Bias I balanced labeled dataset with more cases of **lower severity**.



Exploratory Data Analysis - EDA

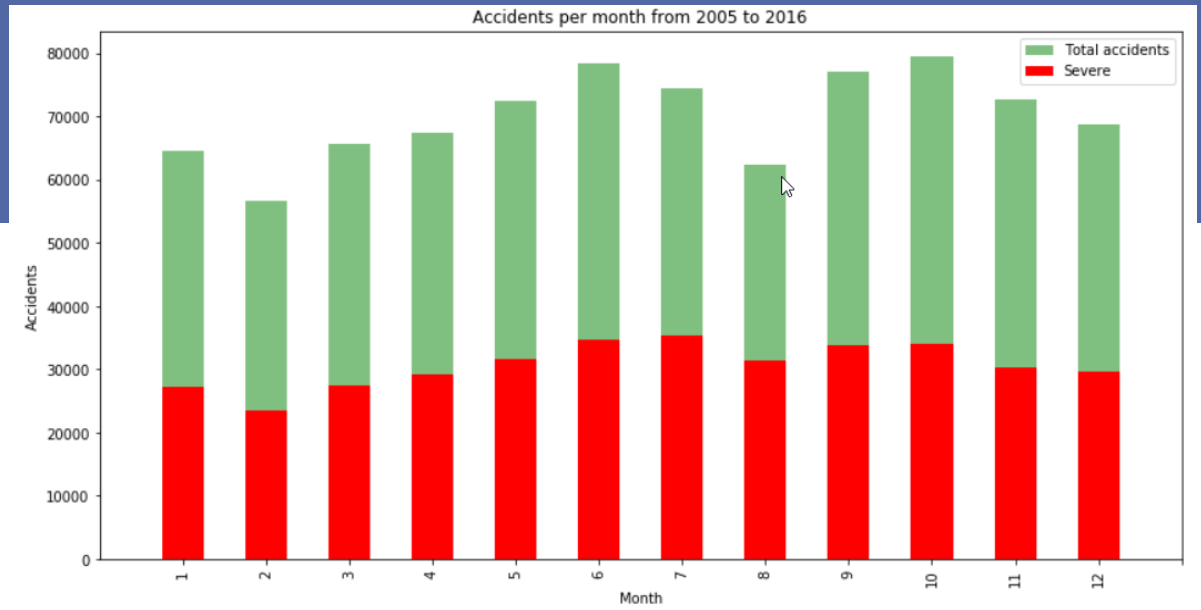






The number of traffic accidents decreased over the years 2005 to 2013 after which the trend became stable

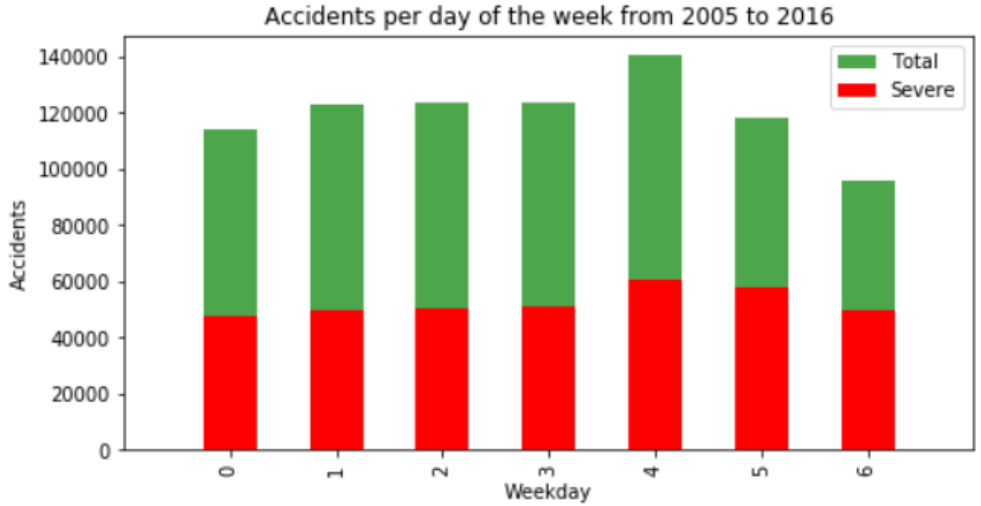
EDA



Accidents Per Month Shows that:

Accidents increase from **March to June** and in **September** it Suddenly decreasing at the end of the year.

Accident Per Week shows Steady trend during the week where we find More accidents accrues on **Friday** and less accident on **Sunday**



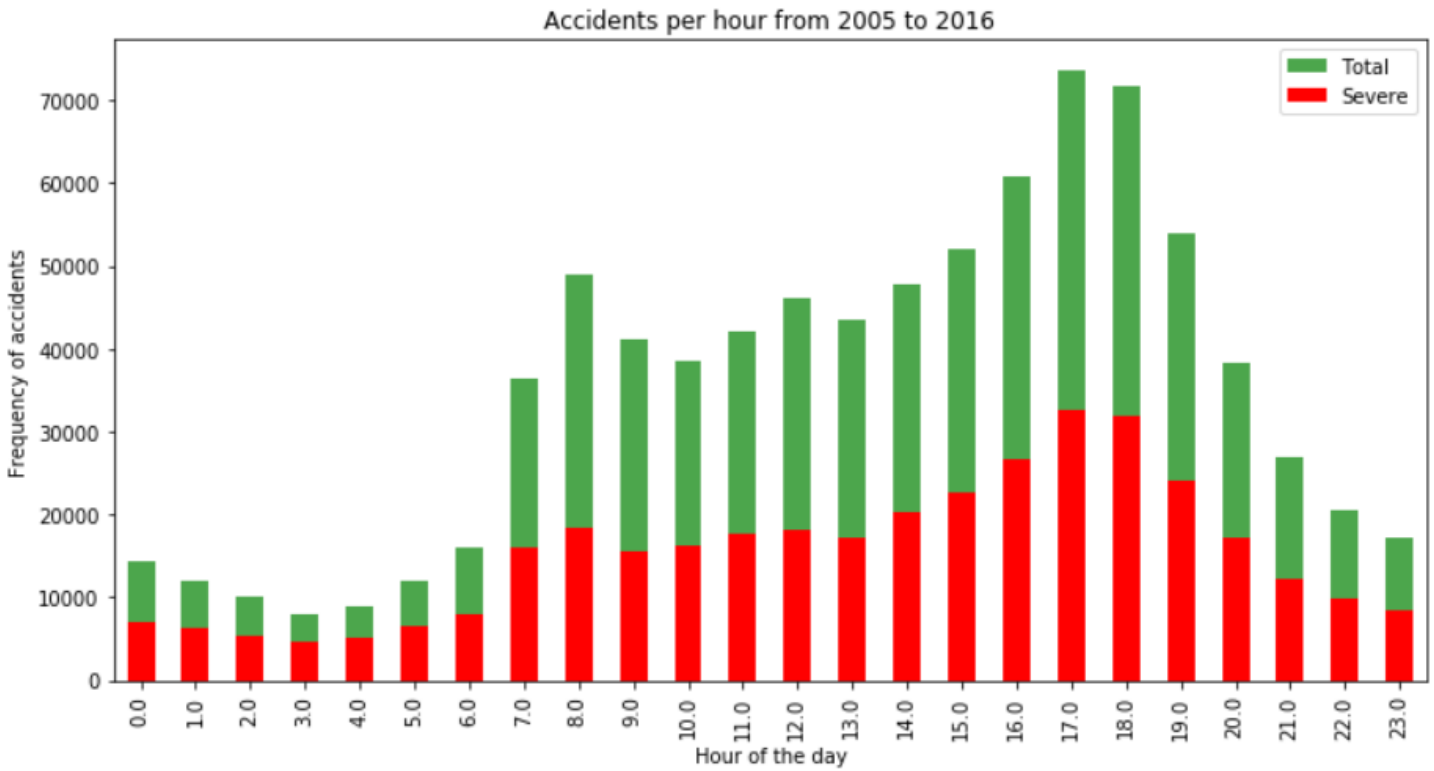
Accident per Day Hours

For Accidents per Hours we found a pattern with Two **Spikes**:

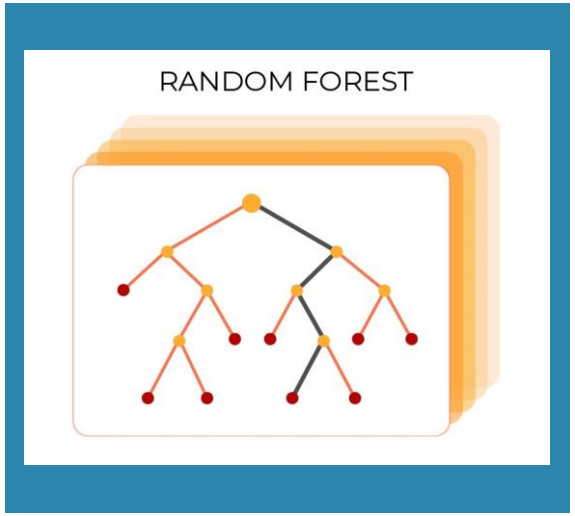
- At the Morning 08:00 when people usually go to work.
- Evening between

17:00 – 18:00

When the comeback

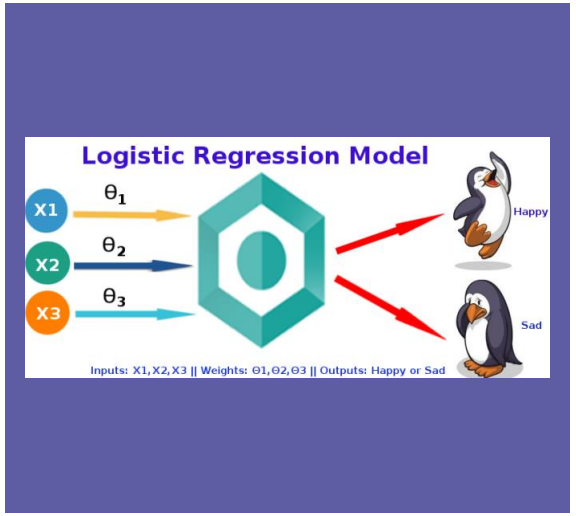


Classification Models



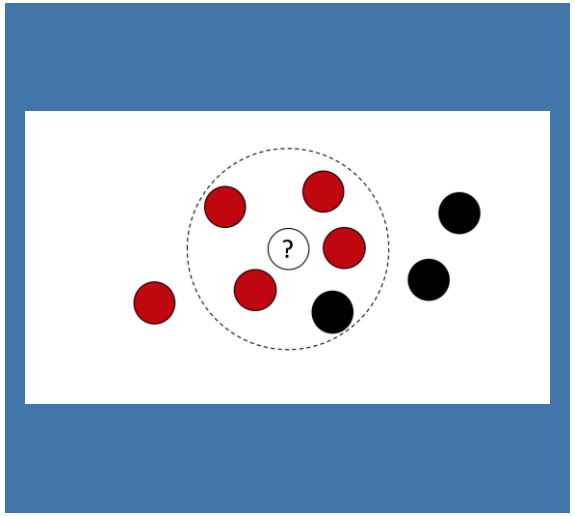
10 Decision Trees

Maximum Depth of 12 Features



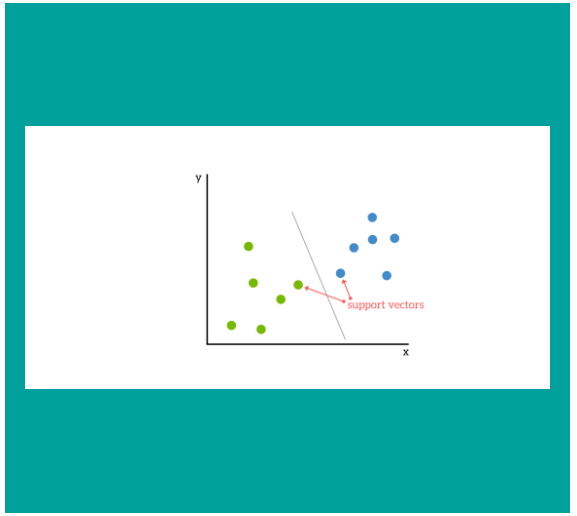
Logistic Regression Model

C = 0.001



K-Nearst Veighbor

K = 16



Supervised Vector Machine

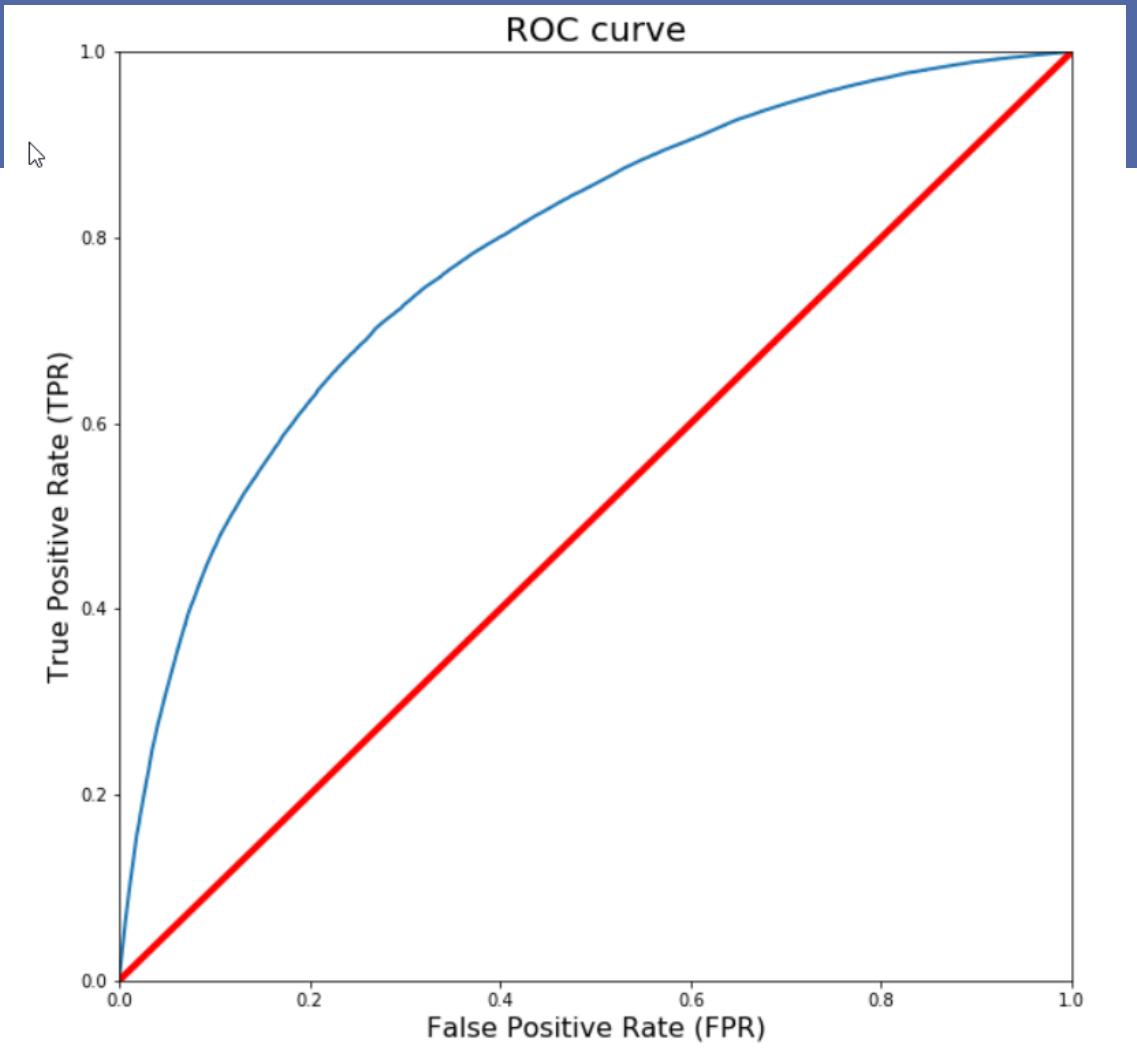
Due to computation inefficiency, training size was reduced to 75,000 samples.

Modeling Results

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 choice model according to the table

The Second Choice could be Logistic Regression



Conclusion and Possible improvements

Now we have useful models to predict the severity of a traffic accident. But, Still the accuracy of the models has room for improvement.

The improvement could be:

- Additional features such as vehicle speed and time of uninterrupted traveling.
- Prediction of potential accident, critical spots and time (Hard data to collect).

Predicting Traffic Accident Severity

- Coursera IBM Data Science Capstone Project By: Mohamad Bouzi