PREDICTING ACCIDENT SEVERITY

IBM DATA SCIENCE CAPSTONE PROJECT

BY : BLESSED MUTENGWA

GITHUB REPO LINK

INTRODUCTION

- Road traffic accident statistics are recorded everyday due to different reasons, which some will be reviewed in the following passages. According to a research conducted by the World Health Organization (WHO) there were 1.35 million road traffic deaths globally in 2016, with millions more sustaining serious injuries and living with long-term adverse health consequences. Globally, road traffic crashes are a leading cause of death among young people, and the main cause of death among those aged 15 to 29.
- Road traffic severity prediction will improve and assist in the following:

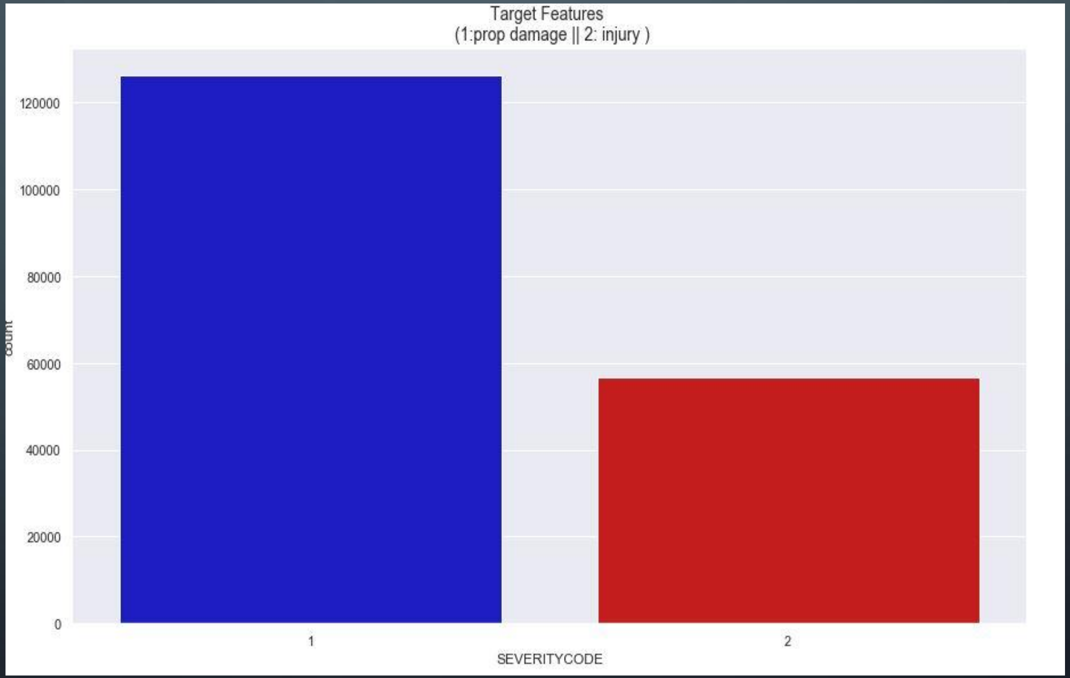
1. Allocation of medical resources to close target hospitals
2. Warning road users based on possible accidents zones based on features like weather
3. To improve road conditions and infrastructure in areas where severity % is very high

DATA ACQUISATION

- The original dataset for this project was downloaded from the Applied Data Science Capstone Project from the following link : https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv along with its metadata pdf. I selected the most relevant attributes for use as features for the prediction of accident severity.
- Data targets to observe and predict accident severity for the United States cities **Seattle.**

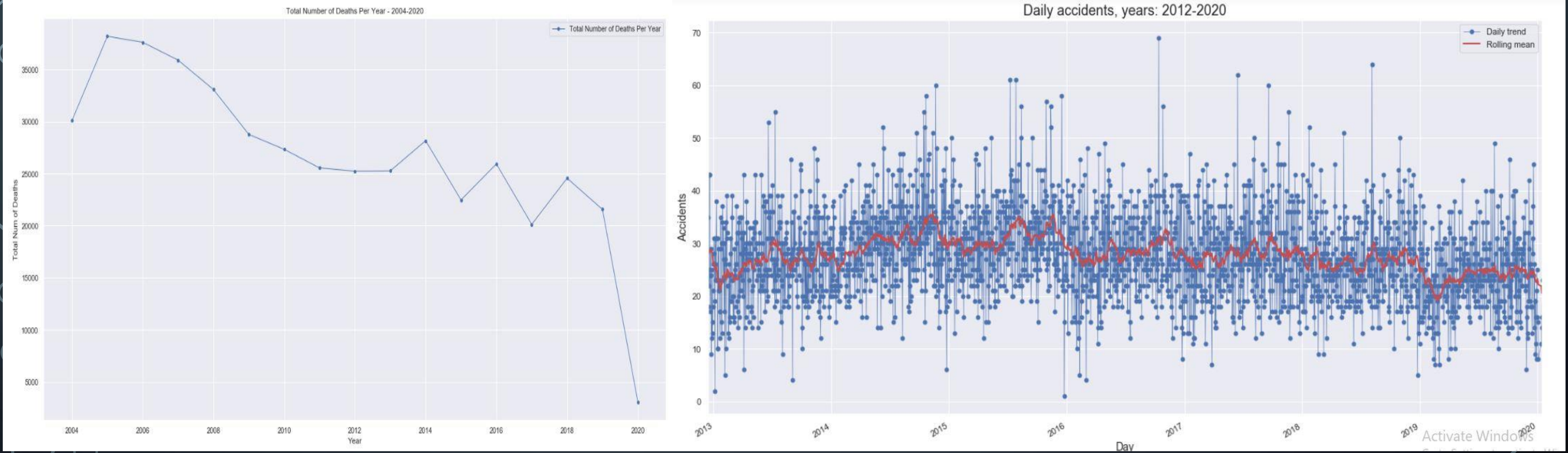
EXPLORATORY DATA ANALYSIS -TARGET

- There are two target features:
- -prop damage and injury



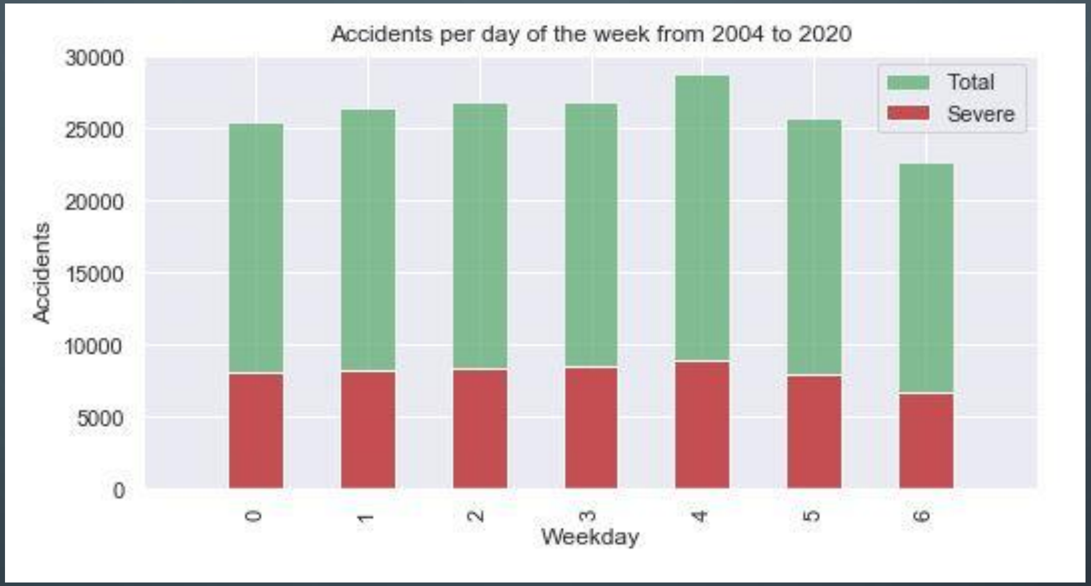
It is a well balanced labelled dataset

EXPLORATORY DATA ANALYSIS - SEASONALITY



Visualisation of the total number of accidents from 2004-2020 and a trend which was Stable in 2016 to date

EXPLORATORY DATA ANALYSIS - SEASONALITY



A trend showing the frequency of how accidents occurred during the week, with the most Recorded on the 4th Day(Friday) and the least on the 6th Day (Sunday)

CLASSIFICATION MODELS

- Logistic Regression

- C=0.01

- K-Nearest Neighbour

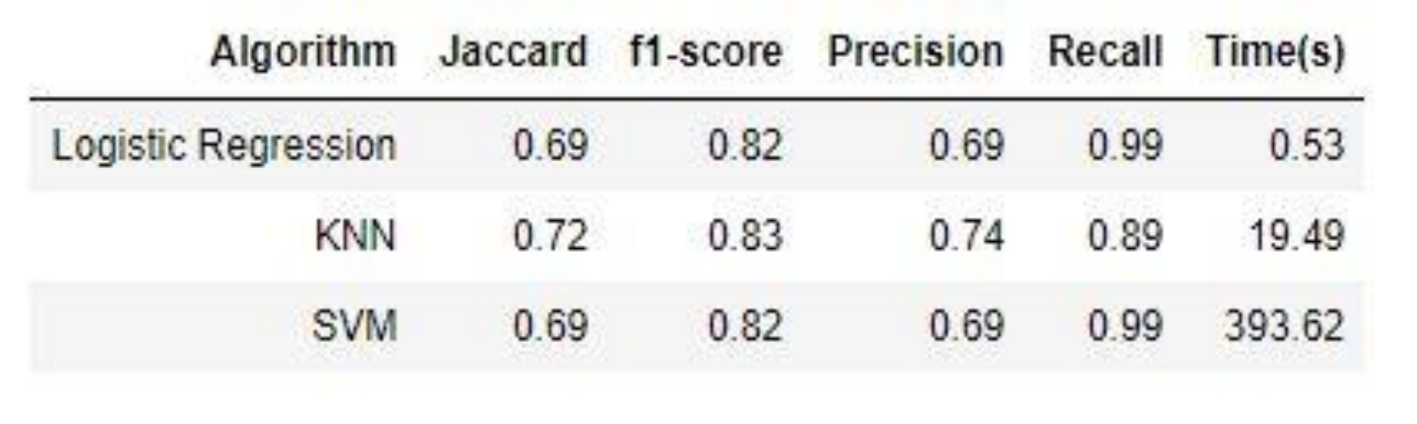
- K = 14

- Supervised Vector Machine

- Due to computation inefficiency, training size was reduced to 75000 samples.

RESULTS

- The results with Jaccard Score, f1-score, Precision, Recall and Time of execution are as displayed in the table below:



Algorithm	Jaccard	f1-score	Precision	Recall	Time(s)
Logistic Regression	0.69	0.82	0.69	0.99	0.53
KNN	0.72	0.83	0.74	0.89	19.49
SVM	0.69	0.82	0.69	0.99	393.62

With no doubt the logistic regression is the best model.
It has the accuracy of 0.69 and a total execution time of 0.53 seconds.

CONCLUSION AND FURTHER IMPROVEMENTS

- The designed models were quite competent and performed with a better accuracy at classifying accidents based on the input features.
- There is still further improvement to increase on the prediction accuracy of the models and enhance on the computational time as well.

- The model could be further improved to predict on a particular geographical space outputting the co-ordinates where an accident is bound to happen.
- The model can be made to integrate with mobile applications for hospital and rescue companies for alerts on possible locations of accidents severity on a particular day.