Car Accident Severity

Introduction

Motivation

- Traffic accidents are severe concern for most of the countries
- Approx. 1.25 million people deaths caused because of road accident injuries in a year [1]
- Complexity of dataset

Objective

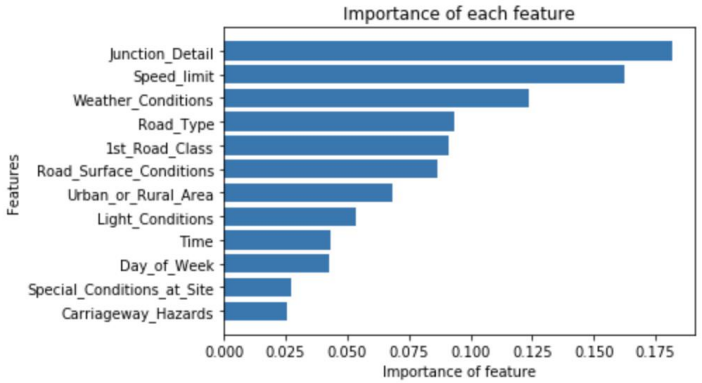
- To help traffic control authorities predict the accident severity
- Effectively able to predict “Serious” accidents

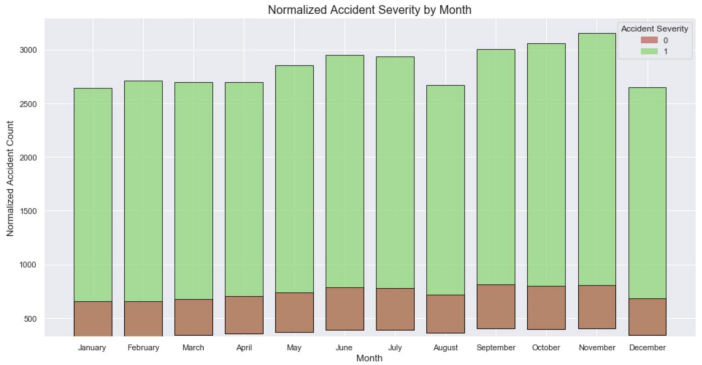
Dataset

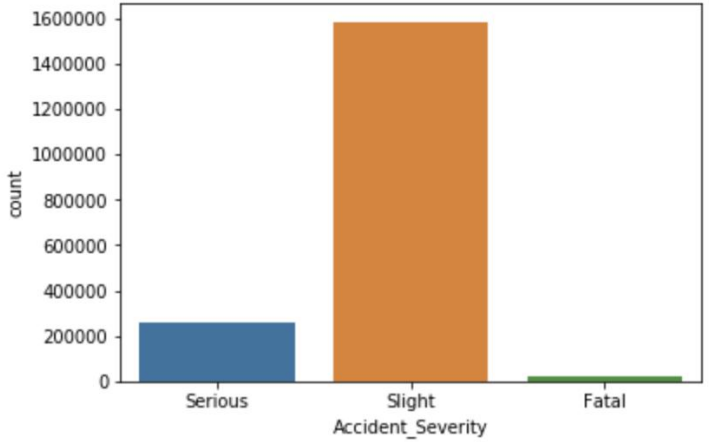
- Size of Dataset: ~630 MB
- Number of records: ~2 Million rows
- Number of columns: 34 Columns
- Source : https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

Data Pre-processing

- Data missing values are imputed by the most frequent value of the column
- Categorical data labelled with numerical values
- Merged similar categorical values
- SelectKBest: provides the k best features by performing various statistical tests i.e., chi squared computation between two non-negative features
- RFE(Recursive Feature Elimination): Recursively eliminates the features which does not in target variable values
- Merged Serious and Fatal classes as Serious class



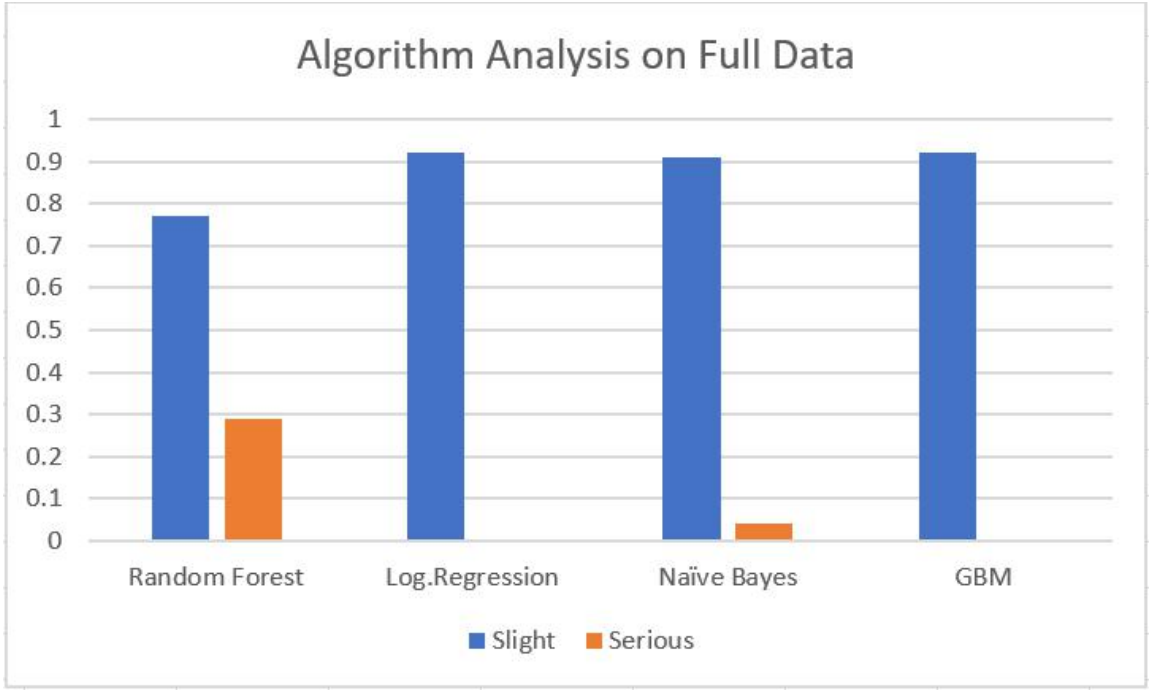


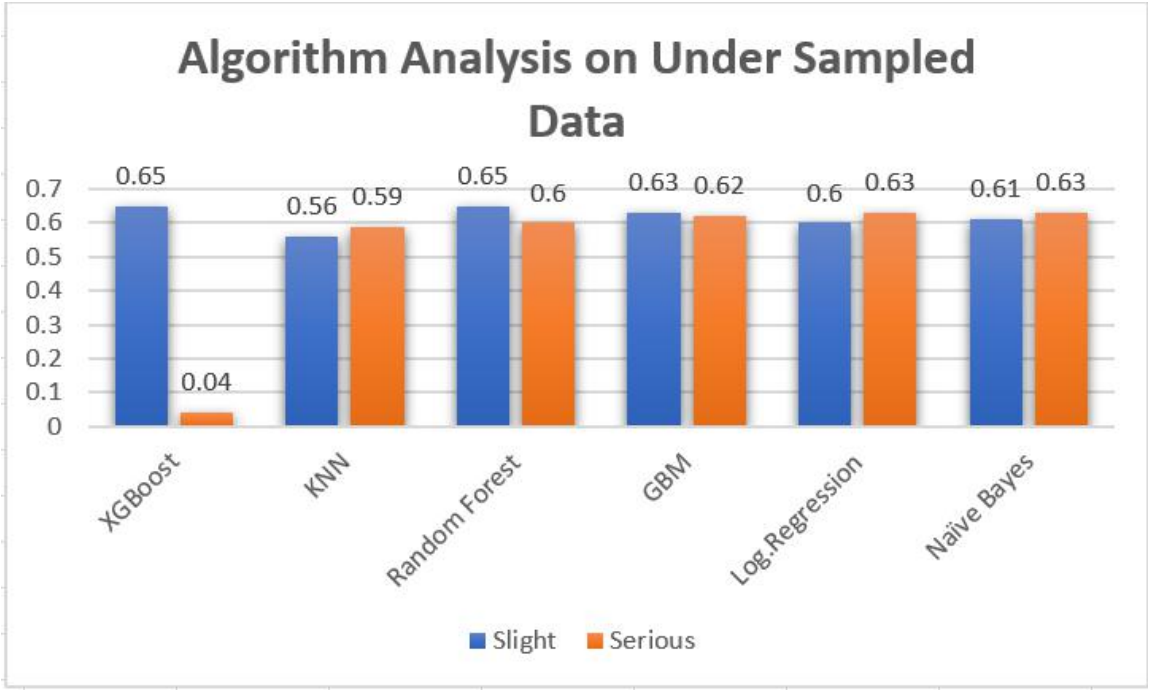


Data Visualization

Algorithms Used

- K- Nearest Neighbor
- Naïve Bayes
- XGBoost
- Random Forest
- GBM
- SVM
- Logistic Regression





Comparative Analysis

Handling Imbalanced Data

- Over Sampling
- Under Sampling
- Mis-classification penalty
- Ensemble methods

Challenges

- Cannot run most of the algorithms on local machines
- Not able to test over sampling
- Highly imbalanced classes

What worked

- Under Sampling
- Fine tuning the parameters
- Data Preprocessing

What not worked

- Over Sampling
- Certain popular ensemble methods did not work well

Conclusion

In conclusion, most of the algorithms are biased towards most frequent class. However, efficient pre-processing and corresponding imbalanced data techniques should give optimal results.

References

[1] Global Status Report on Road Safety 2015

http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/