Prediction of Accident Severity

Capstone Project

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]

Objective

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

Dataset

Size of Dataset: ~70 MB

Number of records: 194673

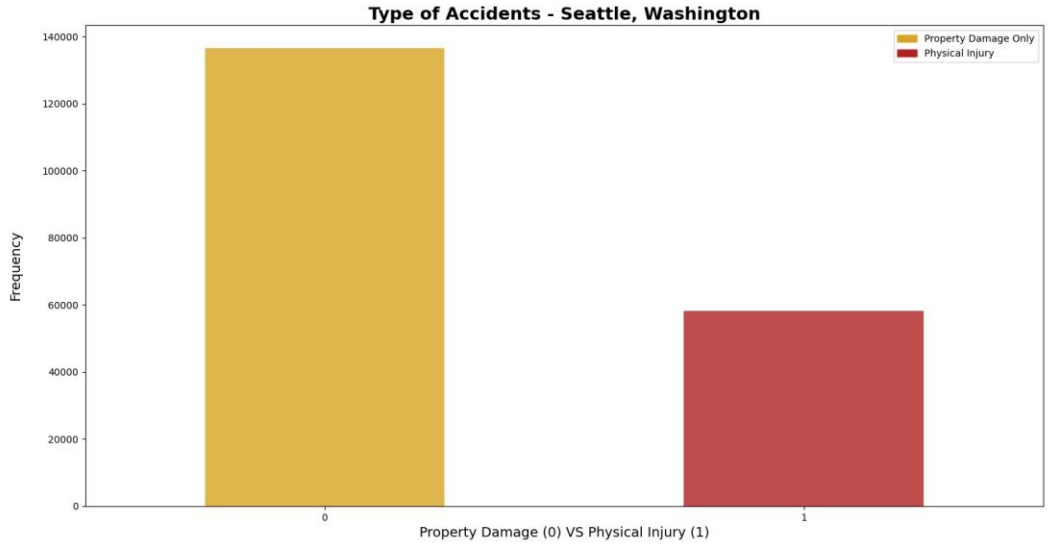
Number of columns: 38 Columns

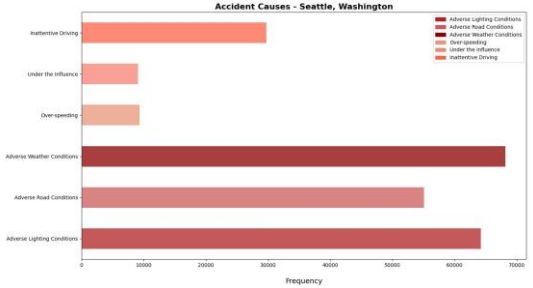
Source : Seattle city car accident data from 2004-2020

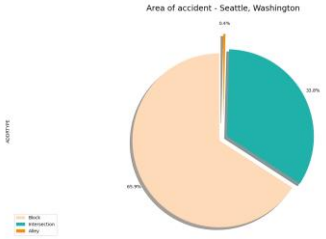
Data Pre-processing

- 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







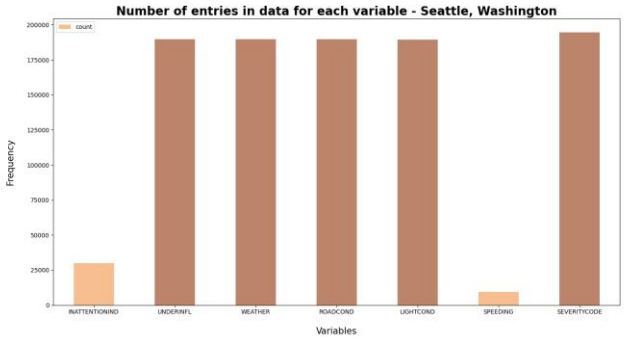
Algorithms Used

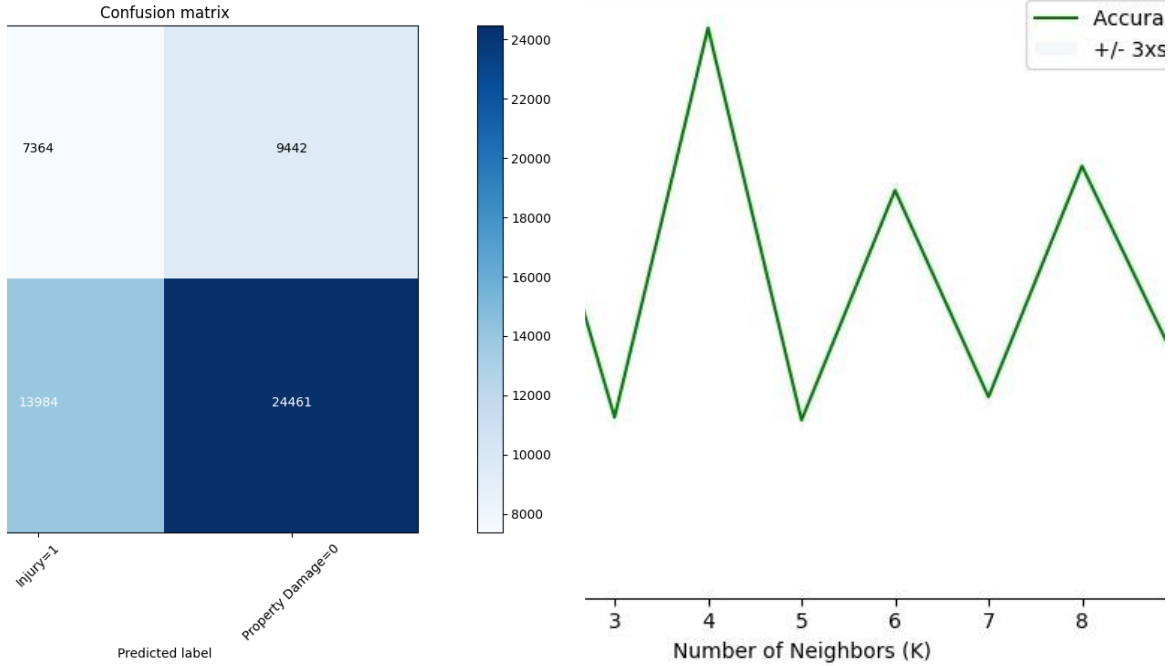
K- Nearest Neighbor

Decision Tree Analysis

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

What not worked

- Under Sampling
- Fine tuning the parameters
- Data Preprocessing

- 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.