Predicting “Severity” of Car Accidents in Seattle City

Bhanu Bommidi

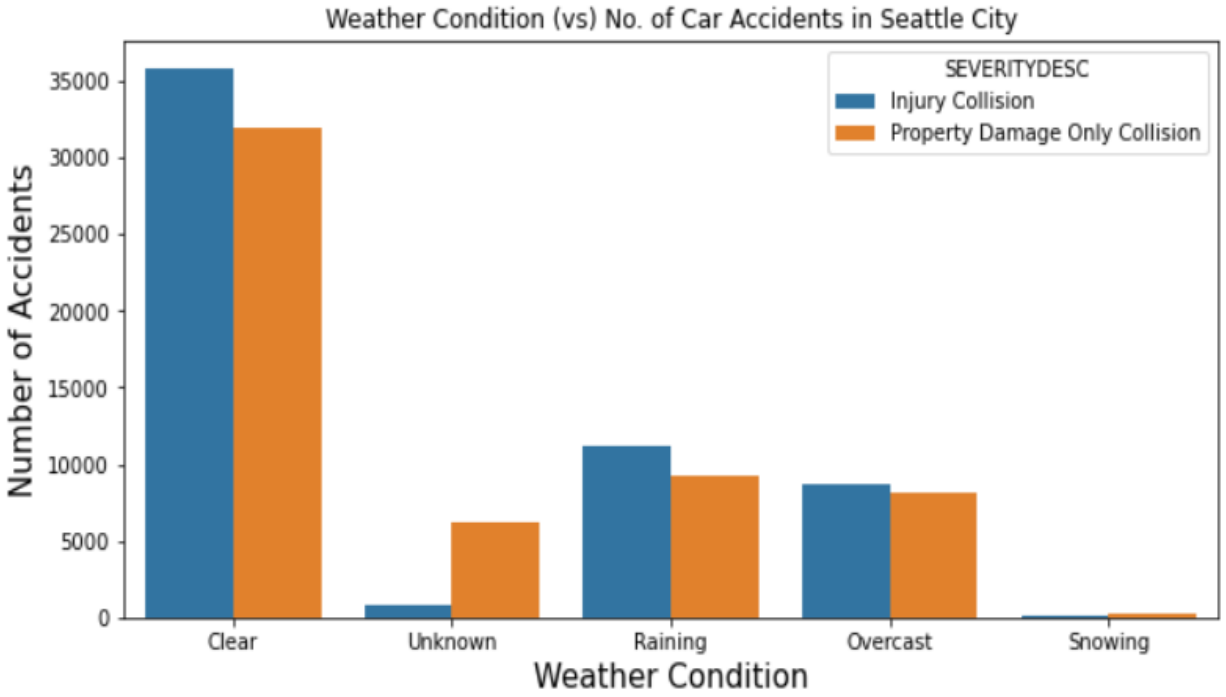
Brief Overview

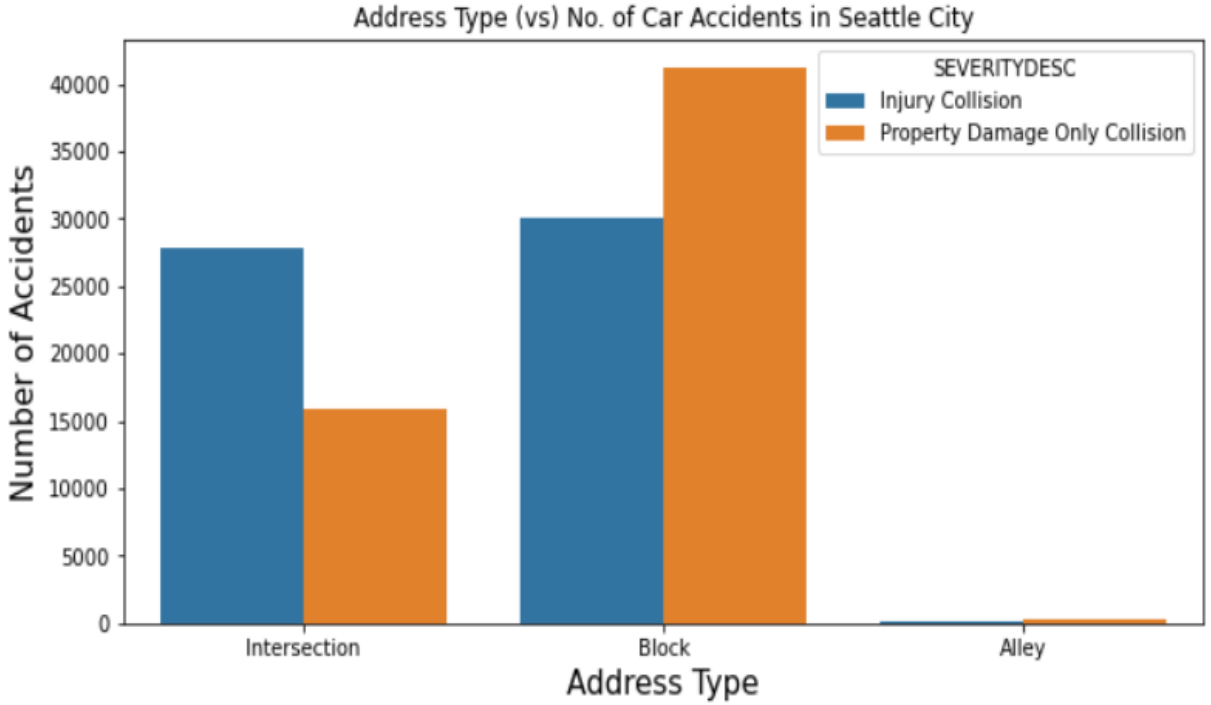
- Car Accidents are one of the most common hazards we all face in daily life. Such accidents often result in injury, disability, death and property damage as well as financial costs and therefore change the lives of the person involved and the those closer to them forever.
- Predicting the possibility of somebody getting into an accident and even gauge the severity, based on the weather and road conditions can directly and indirectly results in saving valuable lives.
- This will enable car commuters to maybe drive carefully or change their travel route or even avoid it, thereby minimizing the possibility of them getting involved in serious car accidents.

Data Acquisition and Cleaning

- Data has been sourced from one of the example datasets here. Along with the data, the detailed description of the metadata can be found here.
- The example dataset consists of various Incidents in Seattle City since 2004 and consists of information pertaining to various types of collisions, Types of Address location like Alley, Block, Intersection etc.;, the Types of Collision like Angles, Parked Car etc.;, Number of persons involved etc.;.
- Input dataset consists of 194,673 rows and 38 features with unbalanced target labels i.e.; 136,485 Severity 1 & only 58,188 Severity 2 rows. Dataset has been balanced to avoid creating a biased prediction model.
- Data has been cleaned to consider only 11 features, by removing various Key & Description / Label features, along with features which have very low correlation coefficient (> -0.05 and < +0.05).

Data Acquisition and Cleaning (contd.)

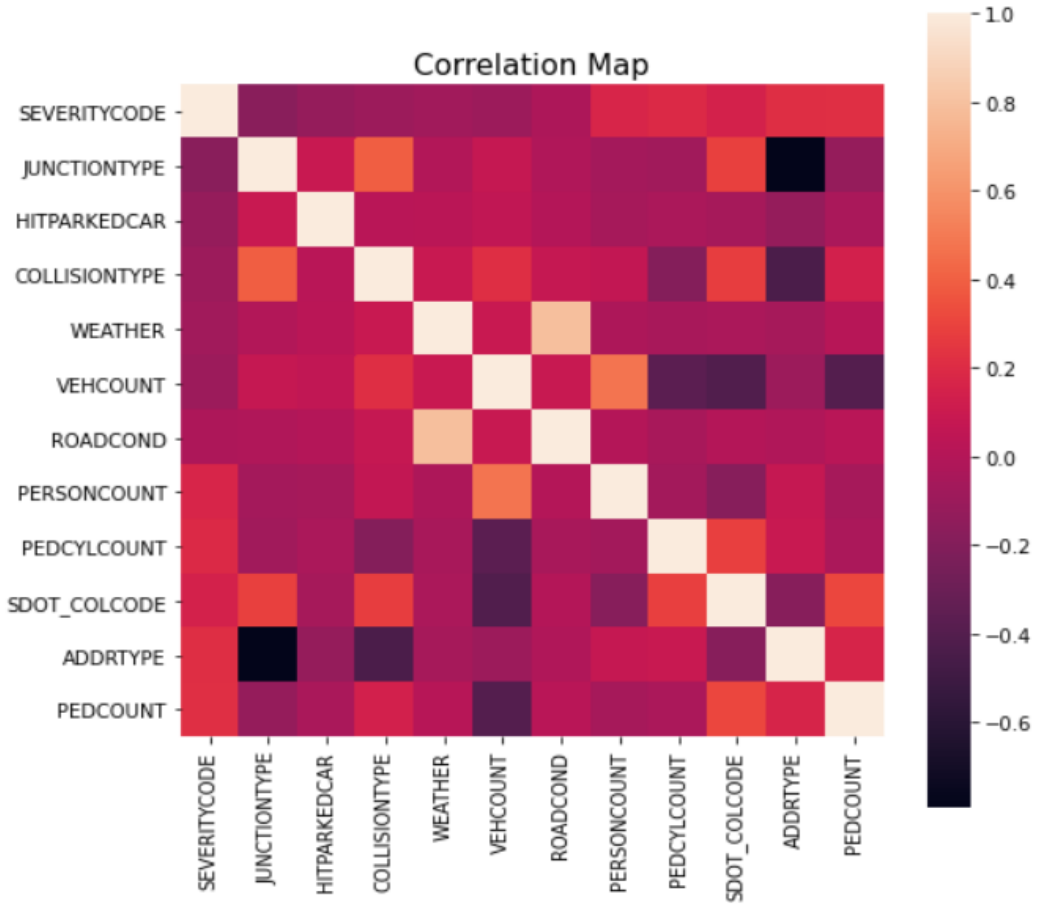




Correlation between the “Weather Condition” & “Address Type” respectively with the “Severity” of the Car Accident

Data Acquisition and Cleaning (contd.)

Correlation Map depicting the correlation between various features and “Severity” of the Car Accident.

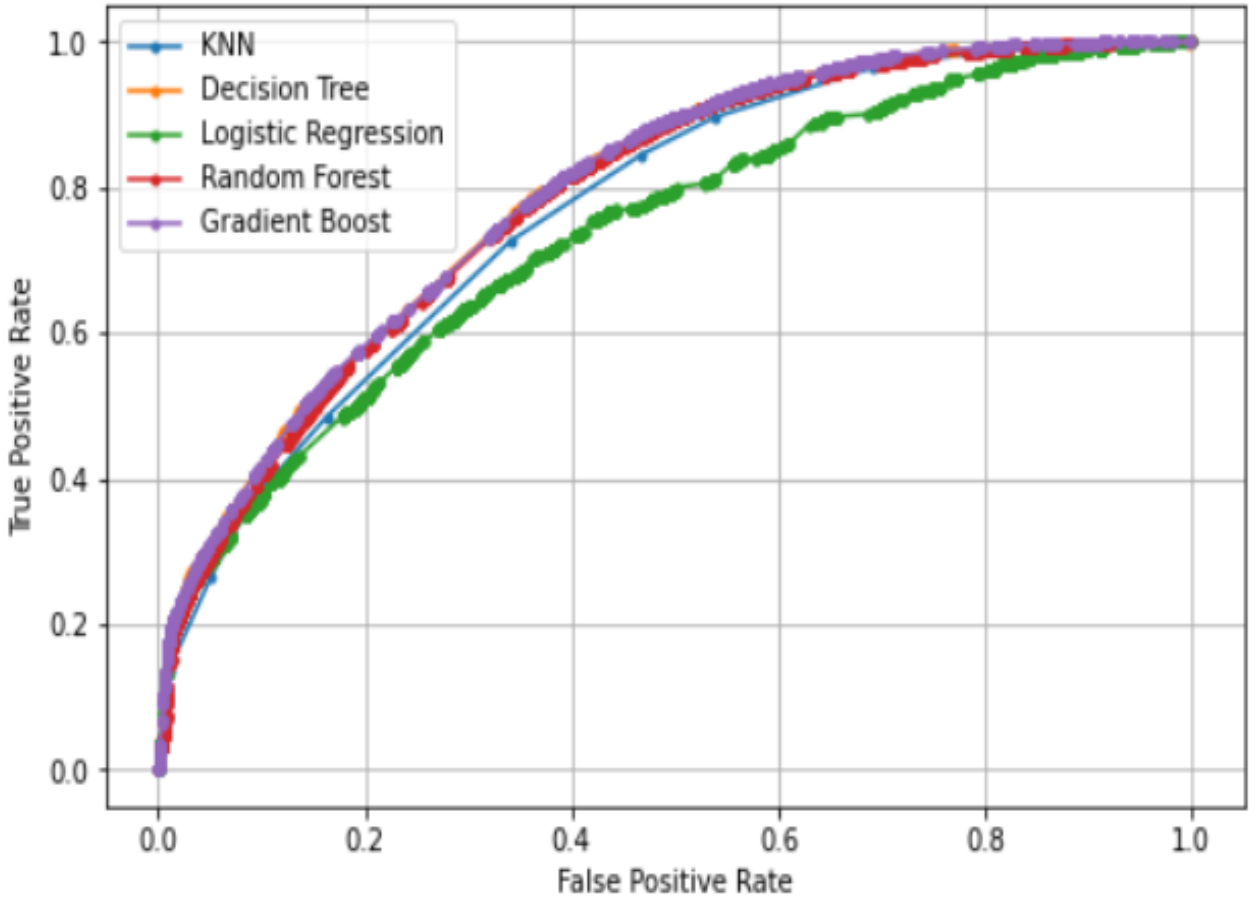


Predictive Modelling

- Various “Classification” models like K Nearest Neighbor (KNN), Decision Tree, Logistic Regression, Random Forest & Gradient Boost were explored and performance computed based on various metrics like Logarithmic Loss, Jaccard Similarity Score, F1-Score etc.;.
- “Gradient Boost” & “Decision Tree” performed the best followed by “Random Forest” & “KNN” and lastly “Logistic Regression” performed slightly less. Overall difference between the various models are comparatively small.
- Logarithmic Loss
  - ✓ 0.535 – 0.626 across 5 Classification models
- Accuracy (Jaccard Similarity Score)
  - ✓ 0.516 – 0.524 across 5 Classification models

Predictive Modelling (contd.)

ROC Curves of various Classification models



Conclusion and Future directions

- Built useful models to predict the predict the Severity of an Accident based on various features like Weather, Road Condition, Address Type, Collision Type etc.;.
- These models can be extremely useful in keeping the various car commuters informed on the severity of accidents and warn them.
- Accuracy of models have scope for improvement and having further classification / combinations of “Severity Codes” may possibly enhance prediction accuracy.