Predicting Seattle Accident Severity

Babu Konnayil

Predicting accident severity is valuable for our society

- Traffic accidents impacts individuals, families, corporates and government emotionally and economically.
- Eliminating fatal, injury accidents and reducing other types of accidents will help our society emotionally and economically.
- Drivers, commuters, insurance companies, hostpitals, government, etc. have an interest in avoiding accidents

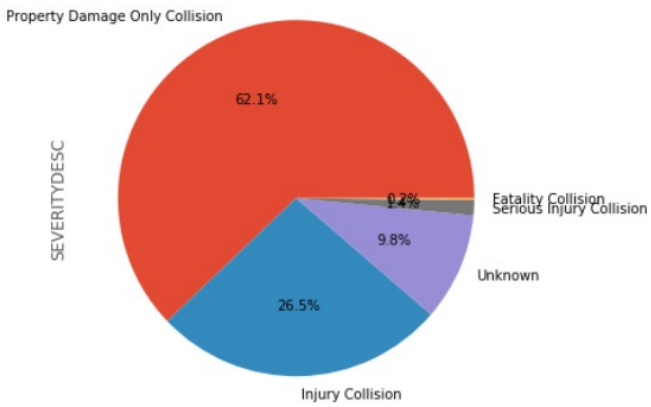
Data acquisition and cleaning

- Accident data (Collisions.csv) directly downloaded from Seattle Department of Traffic (SDOT)

https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d

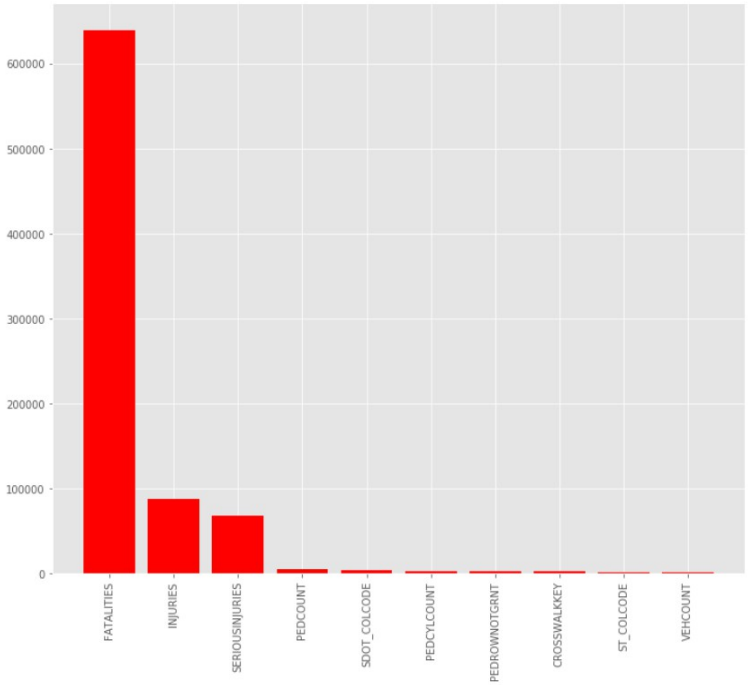
- Data contained 221266 samples and 40 features from 2004 till 2020 download time.
- Dataset was unbalanced and required treatment to balance the dataset.
- Features required treatment for missing values, data type corrections.
- Features contained both numeric and categorical data and needed separate treatment for feature selection.

Explorative Data Analysis



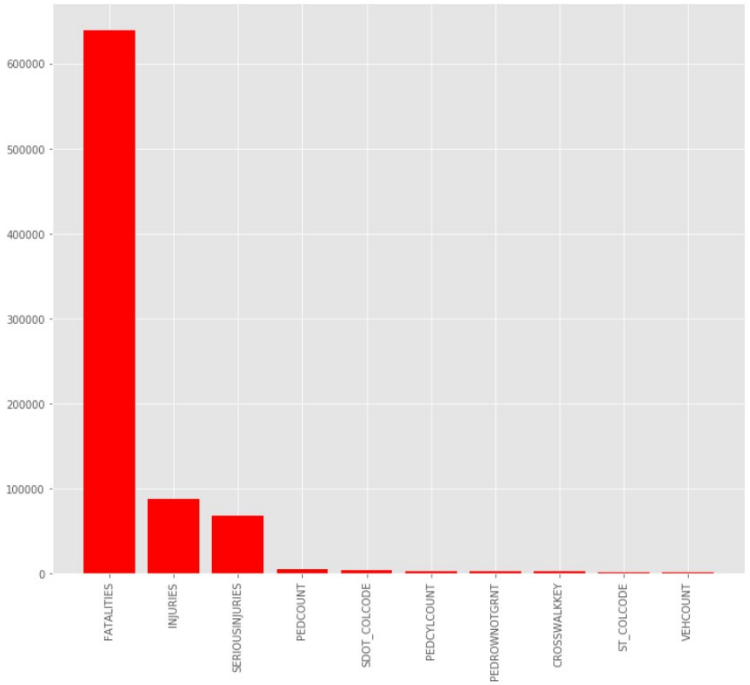
- 62% of accidents resulted in property damage
- 26.5% of accidents resulted in injury collotions
- Discovered 'unknown' type of non-useful samples

Feature Selection and Reduction



- Numeric and Categorical input features put against separate statistical tests to select best features
- SelectKBest and ANOVA Test for top 10 features
- Chi2 test for categorical features after encoding them

Feature Selection and Reduction

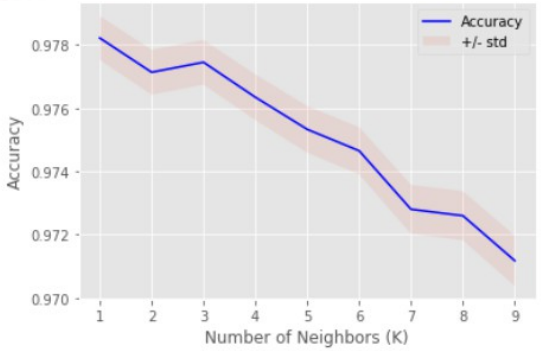


- Numeric and Categorical input features put against separate statistical tests to select best features
- SelectKBest and ANOVA Test for top 10 features

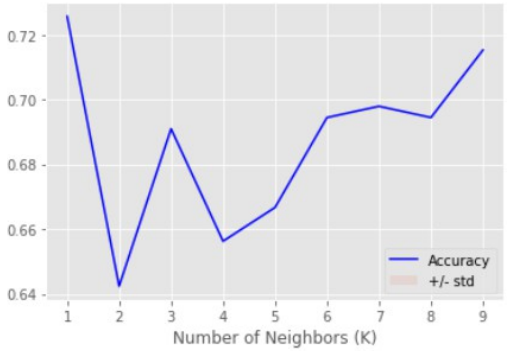
Kept all categorical features after testing it agains Chi2

Modeling - KNN

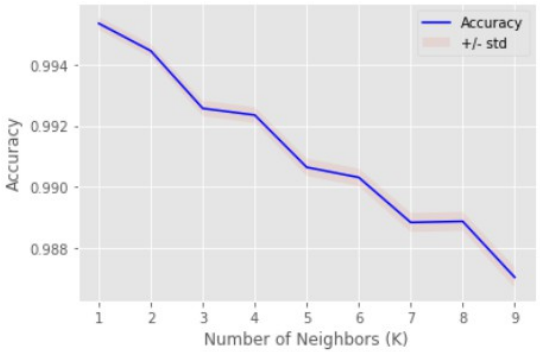
Unbalanced



UnderSample



Oversample



- KNN classification model performed best when K=1
- Model on oversampled data provided best accuracy

Model Performances

Sample Type \ Model Used	KNN	SVM	Logistic Regression
Unbalanced	97.80%	66.71%	98.01%
Balanced (under-sampled)	72.56%	29.86%	79.16%
Balanced (Over-sampled)	99.53%	30.39%	76.42%

- KNN provided overall best accuracy
- SVM performance was low with all three types of dataseet

Conclusion & Future Direction

- K-Nearest Neighbors model with K fold at 1 provided highest accuracy among other models tested
- Further work need to be done to see whether any additional features can be reduced so that when a model is deployed you only need minimum inputs for prediction
- Deploy a model with minimum input and maximum accuracy
- Build model to predict probability of an accident in a given route given certain features