Seattle Car Crash Severity Prediction

IBM Coursera Applied Data Science Capstone

The problem

- Car crash is the most common accidents
- May cause serious injuries and even deaths
- Local government needs to know the cause of crashes to reduce number of casualties

Solution

Using machine learning algorithms to predict severity of collisions based on address type, road condition, light condition, and weather.

Implementation

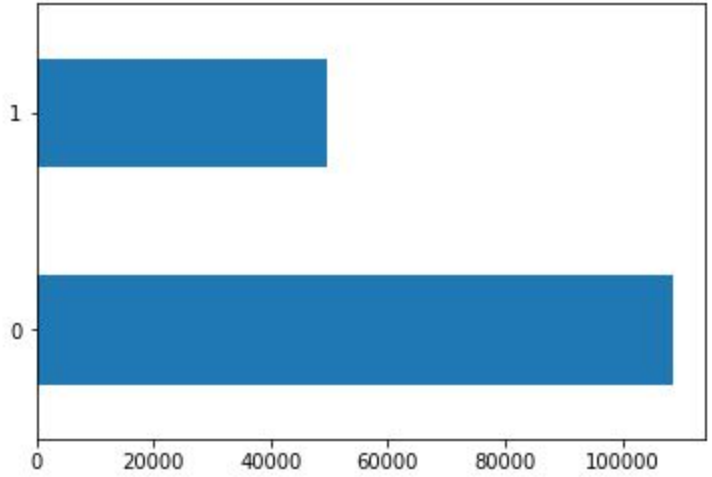
Data Acquisition and Preprocessing

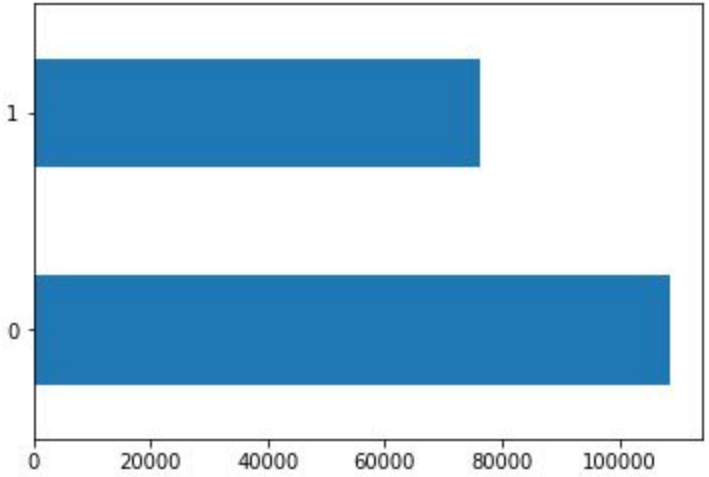
- Collisions data that happened in Seattle (from 2004 to September 2020) is acquired from Seattle GeoData website.
- There are 221,738 collision cases recorded in the dataset.
- 4 features selected as predictors and used the severity code feature as the target variable.
- Missing values are either dropped or put into the ‘Unknown’ category.
- Perform one-hot encoding to transform categorical features into binary-valued features.
- Re-group the data into two classes (severe and not severe collisions).

Handling Imbalanced Dataset

- Most of the data are collisions with low severity (property damage, no injuries)
- Imbalanced dataset make model biased towards the majority class, bad prediction on the minority class.
- Perform SMOTE to balance the dataset.

Data Distribution Before and After Using SMOTE





Models Performance

- Decision Tree: 0.6292 accuracy
- Random Forest: 0.6293 accuracy
- AdaBoost: 0.6287 accuracy
- KNN: 0.5478 accuracy
- SVM: 0.6289 accuracy

Random Forest from scikit-learn with default parameters achieved highest accuracy on validation set.

Model Tuning

- GridSearchCV with 5-fold cross validation to tune Random Forest model.
- The search range for number of estimators parameter of Random Forest is from 40 to 60 estimators.
- 55 estimators is the optimal value
- Best model able to achieve ~65% accuracy on test set.

Conclusion

- Prediction based on address type, weather, road condition, and light condition is not enough to accurately predict the severity of collision.
- Should gather more data (the current dataset contains too many missing values in some columns).