CAR ACCIDENT SEVERITY ANALYSIS

Lei Wang

https://github.com/lwanglitle/Coursera_Capstone

Background

According to WHO data, every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash.

Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole.

These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured.

Road traffic crashes cost most countries 3% of their gross domestic product.

Data

The dataset being used here is a collection of all collisions in Seattle area provided by SPD and recorded by Traffic Records, during timeframe from 2004 to 2020. This includes all types of collisions. This data is regarding the severity (1-Property Damage Only Collision, 2-Injury Collision) of each car accidents along with the time and conditions under which the accident happened.

Data Cleaning

Data with missing or unusual values will be cleane.

Some fix will be applied to the data.

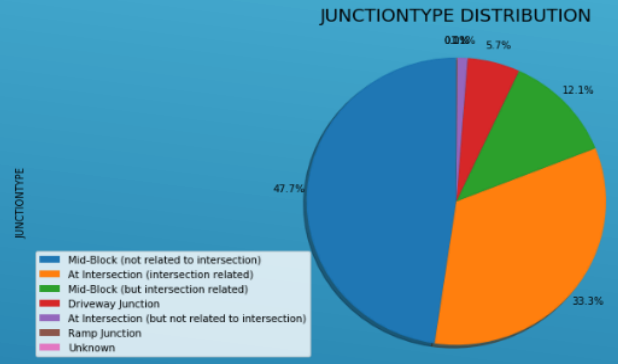
Some columns are duplicate information, one of them will be dropped too.

Columns that have identity values for the incident will be dropped.

The date and time columns will be transferred to month/dayofweek information.

Feature Selection

Pull the feature distribution of the data, e.g.



Retrieve the correlations between the features and severity code:

SEVERITYCODE	1.000000
PEDCOUNT	0.246338
PEDCYLCOUNT	0.214218
COLLISIONTYPE	0.208529
PEDROWNOTGRNT	0.206283
PERSONCOUNT	0.130949
INATTENTIONIND	0.046378
UNDERINFL	0.044377
SPEEDING	0.038938
MONTH	0.004730
DAYOFWEEK	-0.015246
VEHCOUNT	-0.054686
HITPARKEDCAR	-0.101498
LIGHTCOND	-0.139913
WEATHER	-0.146811
ROADCOND	-0.151747
JUNCTIONTYPE	-0.172874
ADDRTYPE	-0.196399

Methodology

Decision Tree, Logistic Regression, k-Nearest Neighbor (KNN) and Support Vector Machine (SVM) are the machine learning models used in the data analysis and prediction.

For each kind of model, we provide a range of hyper-parameters and iterate thought them to figure out which hyper-parameter would return best result. Then we compare the models with their best performance.

Result

Algorithm	f1-score	accuracy_score
KNN	0.7178	0.7419
Decision Tree	0.7226	0.753
SVM	0.6883	0.749
Logistic Regression	0.7012	0.7466

Discussion

F1-score is a measure of accuracy of the model, which is the harmonic mean of the model's precision and recall.

Accuracy_score function is equal to the jaccard_similarity_score function in binary and multiclass classification. jaccard_similarity_score represents Jaccard similarity coefficient, which is the size of the intersection divided by the size of the union of two label sets and is used to compare set of predicted labels for a sample to the corresponding set of labels in y_true.

Among the four models, Decision Tree has best performance regarding both f1-scores and accuracy_scores, so it is chosen to be our model for predicting car accident severity.

Conclusion

Based on the data analysis and exploration on the models, we found out the Decision Tree model is the best one to predict the car accident severity.

The conditions that would be relevant to the accident severity are:

- Collision address type,
- Collision type,
- The total number of people involved in the collision,

  ……
- Whether or not the pedestrian right of way was not granted,
- Whether or not the collision involved hitting a parked car.

After the study of the severity data, the traffic management department can do things to improve the conditions that are critical, and drivers can be notified and pay more attention on certain conditions to avoid severe accident happen.