Car Accident Severity Analysis

Lei Wang
https://github.com/lwanglitle/Coursera_Capstone

Contents

1. Introduction

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. This project is in order to help reducing the loss in car accidents by finding the key factors in an accident and try to avoid that.

2. 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.

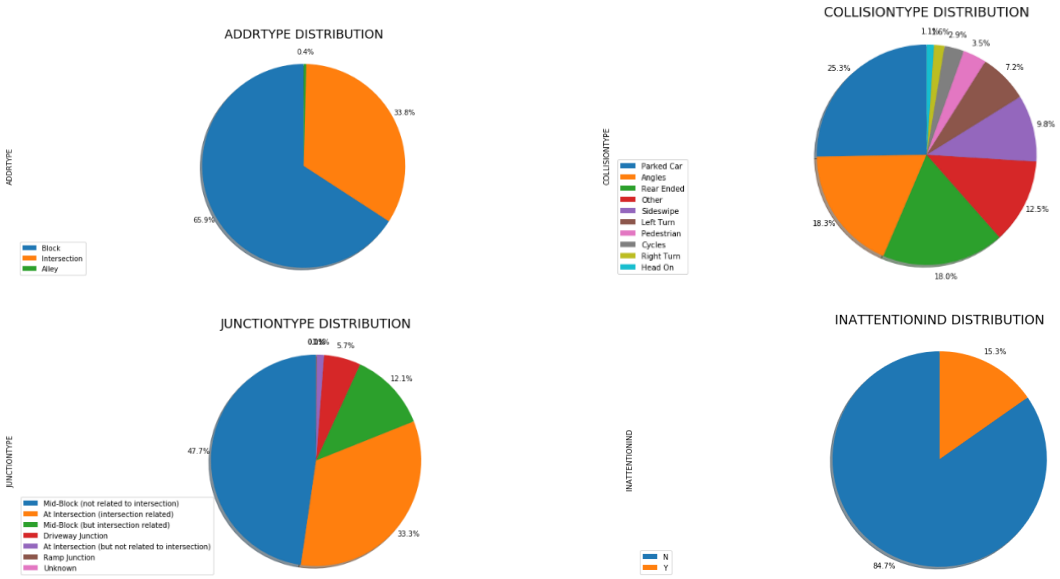
2.1 Data Cleaning

First, we must go through the data and do the preprocessing step. In this step, data will be cleaned so that no missing or unusual values are eliminated. And some fix will be applied to the data. For example, the INATTENTIONIND column has only 'Y' in some rows but leave the other rows blank, which we assume the blank should be 'N', and data would be fixed in this way. Some columns are duplicate information (eg. SDOT_COLCODE and SDOT_COLDESC), so one of them will be dropped too. Columns that have identity values for the incident will not help on the data analysis, so that they will be dropped as well. The date and time columns will be transferred to month/dayofweek information, since particular datetime does not indicate car

accident information but statistical information of them would be useful. About 16% of the data do not have the time of the incident, so the hour info will not be retrieved from the rest data.

2.2 Feature Selection

Once we have the clean data, we can pull the feature distribution of the data, which will help us getting deeper understand of the data, some examples are:



After we get the distribution, the next step we will encode the features for retrieving the correlations between the features and severity code, this will help us figure out which features are more important for the machine learning model to learn, and which features are not that relevant. With pandas corr() calculation, we get the correlations as below:

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
```

Based on the result, we select features 'ADDRTYPE','COLLISIONTYPE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','JUNCTIONTYPE','WEATHER','ROADCOND','LIGHTCOND','PEDROWNOTGRNT','HITPARKEDCAR'] for our ML model to learn.

3. 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. The Decision Tree breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable. KNN is an algorithm that stores all available cases and classifies new cases based on a similarity measure based on distance. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall. 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.

4. Results

After iterating the hyper-parameters for each model, the best ones have been found out and the models are created with them as:

```
KNeighborsClassifier(n_neighbors = 10).fit(x_train,y_train)
```

```
DecisionTreeClassifier(criterion = 'entropy', max_depth = 7)

SVC(C = 0.01, gamma = 'auto', kernel = 'rbf')

LogisticRegression(C = 0.001, solver = 'newton-cg')
```

The f1-scores and accuracy scores are:

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

5. Discussion

F1-score is a measure of accuracy of the model, which is the harmonic mean of the model's precision and recall. The highest value for the f1-score is 1, which means perfect precision and recall. The other metric we use is accuracy score. In binary and multiclass classification, accuracy_score function is equal to the jaccard_similarity_score function, which is the Jaccard similarity coefficient score. Jaccard similarity coefficient 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.
By comparing all the models by their f1-scores and accuracy_scores, we can get a clearer picture in terms of the accuracy of the models individually as a whole and how well they perform Car Accident prediction. 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.

6. Conclusion

Based on the data analysis and exploration on the models, we found out the [placeholder] 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,
- The number of pedestrians involved in the collision.
- The number of bicycles involved in the collision.
- Category of junction at which collision took place,
- Description of the weather conditions during the time of the collision,
- The condition of the road during the collision,
- The light conditions during 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.