Car Accident Severity Prediction

Applied Data Science Capstone

Hey, I am Ayush Kothari; a keen learner on the path of gaining many varied skills and integrating them in the domain of Mechanical Engineering!

Business Understanding

With the increase in number of cars and the performance of cars, there has been a big rise in the frequency of car collisions. With this Machine Learning Model, the severity of a car accident if it happens can be predicted by taking the various factors into account. Besides the aforementioned reasons, weather, visibility, or road conditions are the major uncontrollable factors that can be prevented by revealing hidden patterns in the data and announcing warning to the local government, police and drivers on the targeted roads.

This algorithm must be made so as to predict the severity of an accident given the weather condition, road and visibility conditions, the amount of light or darkness. When conditions are such that an accident is prone to happem, this model can alert the drivers and the road safety department about this likelyhood so that they can prepare and be cautious.

Data Understanding

The data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present. The data was updated weekly

The data consists of 37 independent variables and 13846 rows.

Features used to calculate the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

Furthermore, because of the existence of null values in some records, the data needs to be preprocessed before any further processing.

Severity codes corresponds to the severity of the collision:

- 3 : fatality
- 2b : serious injury
- 2 : injury
- 1 : prop damage
- 0 : unknown

Link for the metadata

Data Preprocessing

The dataset has many columns which we will not be using in this project. Also, the dataset in its current form s not fit for building a model. So, we'll be performing some operations and making the dataset cleaner.

In target 'SEVERITY CODE', the number of rows in class 1 is almost three times bigger than the number of rows in class 2. It is possible to solve the issue by downsampling the class 1.

```
raw_df_major = raw_df[raw_df['SEVERITYCODE']==1]
raw_df_minor = raw_df[raw_df['SEVERITYCODE']==2]
```

```
[10] raw_df_major = raw_df_major.sample(n=4258,random_state=7)
     raw_df_major['SEVERITYCODE'].value_counts()
```

```
1    4258
Name: SEVERITYCODE, dtype: int64
```

```
equal_df["ROADCOND"] = equal_df["ROADCOND"].astype('category')
equal_df["LIGHTCOND"] = equal_df["LIGHTCOND"].astype('category')
equal_df["WEATHER"] = equal_df["WEATHER"].astype('category')

equal_df.dtypes
```

```
[113] equal_df["ROADCOND_cat"] = equal_df["ROADCOND"].cat.codes
      equal_df["LIGHTCOND_cat"] = equal_df["LIGHTCOND"].cat.codes
      equal_df["WEATHER_cat"] = equal_df["WEATHER"].cat.codes
      equal_df
```

```
[119] from sklearn import preprocessing
      X = preprocessing.StandardScaler().fit(X).transform(X)
      X[0:5]
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.28)
print("Training Data : ", X_train.shape,y_train.shape )
print("Test Data : ", X_test.shape, y_test.shape)
```

Methodology

Our data is now ready to be fed into machine learning models. We will use the following models:

K-Nearest Neighbor (KNN) : It'll will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

Random Forest : It is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Decision Tree : It'll give us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

Logistic Regression : Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

Results & Evaluation

The final results of the model evaluations are summarized in the table on the right:

Based on the above table, KNN is the best model to predict car accident severity.

algorithm	Jaccard Similarity Score	F1-Score	Log Loss
KNN	0.521174	0.604571	N/A
Decision Tree	0.553459	0.445023	N/A
Logistic Regression	0.523690	0.431431	0.684034
Random Forest Classifier	0.560587	0.433514	N/A

Conclusion

Based on the dataset provided for this capstone from weather, road, and light conditions pointing to certain classes, we can conclude that particular conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2)

Thanks!

Ayush Kothari

GitHub : Ayushft

ayush.kotharift@gmail.com