Predicting the correlations between driving conditions and traffic accidents

Sotheavatey Bong

October 7, 2020

Business Problem

Traffic accidents have taken away millions of people's lives each year. The number of death caused by traffic accidents have become equivalent to the number of deaths caused by diseases which bring into light the importance of drivers' traveling conditions. The objective of this project is to predict the correlation between the severity of accidents and the driver's traveling conditions. It aims to determine whether car accidents are caused by the weather, the condition of the road or speeding. The result from this analysis can raise more awareness for drivers to be extra careful as they hit the road.

Data

The data is provided by Coursera for IBM Data Science course. The data presents the number of accidents/collisions from 2004 to the present in Seattle, Washington. The total number of observations consists of about 194670 rows and 37 attributes; however, not all attributes are used for this analysis. The analyst will rightfully select variables based on his understanding of the business problem.

There are also missing values in the dataset which require data preparations and cleaning before the final model can be built.

Methodology

Two key variables will be selected to build a model to predict the correlations. They are road conditions [ROADCOND] and speeding [SPEEDING]. The target variable is severity code [SEVERITYCODE].

Libraries used are Pandas, Numpy, Scikit-learn.

The dataset will be split into training set (80%) and testing set (20%).

Model will be built and predicted using Decision tree, Linear Regression and KNN method.

Variable Selection

```
df_Data = ['ROADCOND','SPEEDING','SEVERITYCODE']
df_Data = Data[df_Data]
df_Data
```

Fill in missing values

```
df_Data['ROADCOND'] = df_Data['ROADCOND'].fillna('Unknown')
df_Data['SPEEDING'] = df_Data['SPEEDING'].fillna('N')
df_Data
```

Assigning new values

```
df_Data["ROADCOND"].replace(to_replace=['Wet','Dry','Unknown','Snow/Slush','Ice','Other','Sand/Mud/Dirt',
'Standing Water','Oil'],value =
['Dangerous','Safe','Safe','Dangerous','Dangerous','Safe','Dangerous','Dangerous','Dangerous'],inplace=True)
df_Data["SPEEDING"].replace(to_replace=['N','Y'], value=[0,1], inplace=True)
df_Data["ROADCOND"].replace(to_replace=['Dangerous','Safe'],value=[0,1],inplace=True)
testdf_Data = df_Data[['SPEEDING','ROADCOND']]
testdf_Data.head()
```

Training model

```
x = testdf_Data
y = df_Data['SEVERITYCODE'].values.astype(str)
x = preprocessing.StandardScaler().fit(x).transform(x)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=1234)

print("Training set: ", x_train.shape, y_train.shape)
print("Testing set: ", x_test.shape, y_test.shape)
```

Tree Model

```
Tree_model = DecisionTreeClassifier(criterion="entropy", max_depth = 4)
Tree_model.fit(x_train, y_train)
predicted = Tree_model.predict(x_test)
Tree_f1 = f1_score(y_test, predicted, average='weighted')
Tree_acc = accuracy_score(y_test, predicted)
```

Logistic Regression

```
LR_model = LogisticRegression(C=0.01, solver='liblinear').fit(x_train, y_train)
predicted = LR_model.predict(x_test)
LR_f1 = f1_score(y_test, predicted, average='weighted')
LR_acc = accuracy_score(y_test, predicted)
```

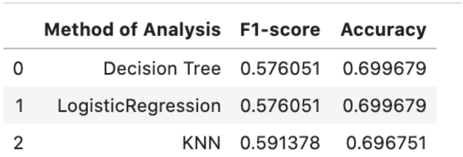
KNN model

[ ]:

```
KNN_model = KNeighborsClassifier(n_neighbors = 4).fit(x_train, y_train)
predicted = KNN_model.predict(x_test)
KNN_f1 = f1_score(y_test, predicted, average='weighted')
KNN_acc = accuracy_score(y_test, predicted)
```

Result

```
results = {
  "Method of Analysis": ["Decision Tree", "LogisticRegression","KNN"],
  "F1-score": [Tree_f1, LR_f1, KNN_f1],
  "Accuracy": [Tree_acc, LR_acc, KNN_acc,]
}
results = pd.DataFrame(results)
results
```



	Method of Analysis	F1-score	Accuracy
0	Decision Tree	0.576051	0.699679
1	LogisticRegression	0.576051	0.699679
2	KNN	0.591378	0.696751

Discussion

All models present consistent results which mean the model being created is likely to be statistically correct. After generating results, it can be seen that F1 score and accuracy score of KNN is better than the other two models.

Conclusion

Based on the result, the prediction result shows correlations between speeding, the weather conditions and the severity of accidents. In other words, dangerous road conditions and high-speed driving result in higher chance of traffic accidents.