Predicting Severity of Car Accidents

IBM Data Science Capstone

David Wang

August 31, 2020

Introduction

Every year approximately 1.35 million people die in road crashes. On average 3,700 people lose their lives every day. This statistic is quite large and can be dramatically reduced. This project will involve working on a case study where we will predict car accident severity. The stakeholders will be the drivers of ridesharing companies. This is important for employee safety and their customers. Drivers usually don't pay attention to how carefully they should be driving given certain weather and road conditions. Developing a model to predict the severity of an accident can help drivers make better decisions on how cautious to drive and assess their travel plan.

Data

I chose to use the accident severity data from Coursera for this project. They provide a metadata file that has a description of each column. The data labels each accident with a severity code of 1 for property damage and 2 for injury collision. It is classified based on different features such as weather, road conditions, lighting conditions, and more. Categorical variables were hot encoded to fit the machine learning algorithm. For this study, I looked at which combination of the different features are more likely to lead to property damage or injury collision.

Some features that are included:

Severity code (1 property damage, 2 injury collision)

Collision type

Address type (Intersection, Block, Alley)

Time of incident

Number of vehicles

Number of injuries

Location

Weather

Number of pedestrians

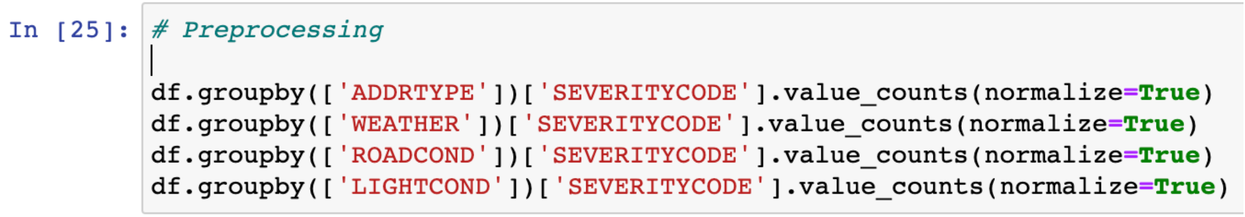
Number of cyclists

Road conditions

Light conditions

Methodology

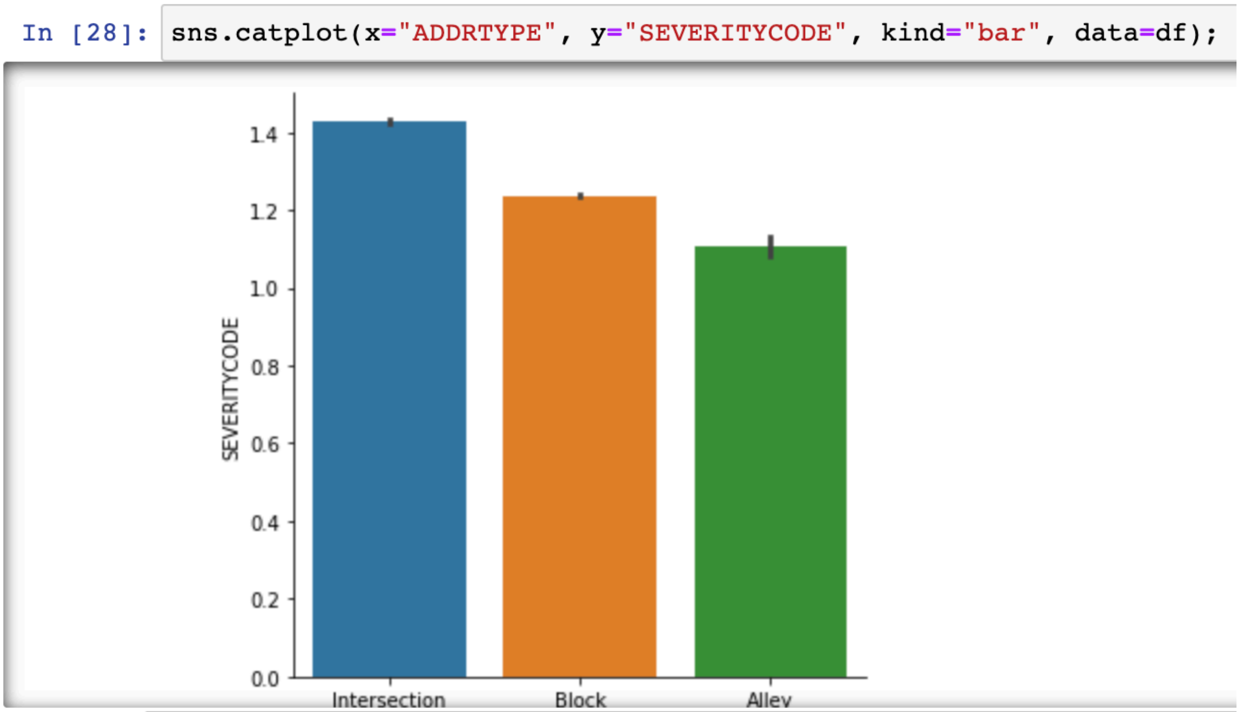
First, step after importing the data was to hot encode it so machine learning algorithms are able to read it.



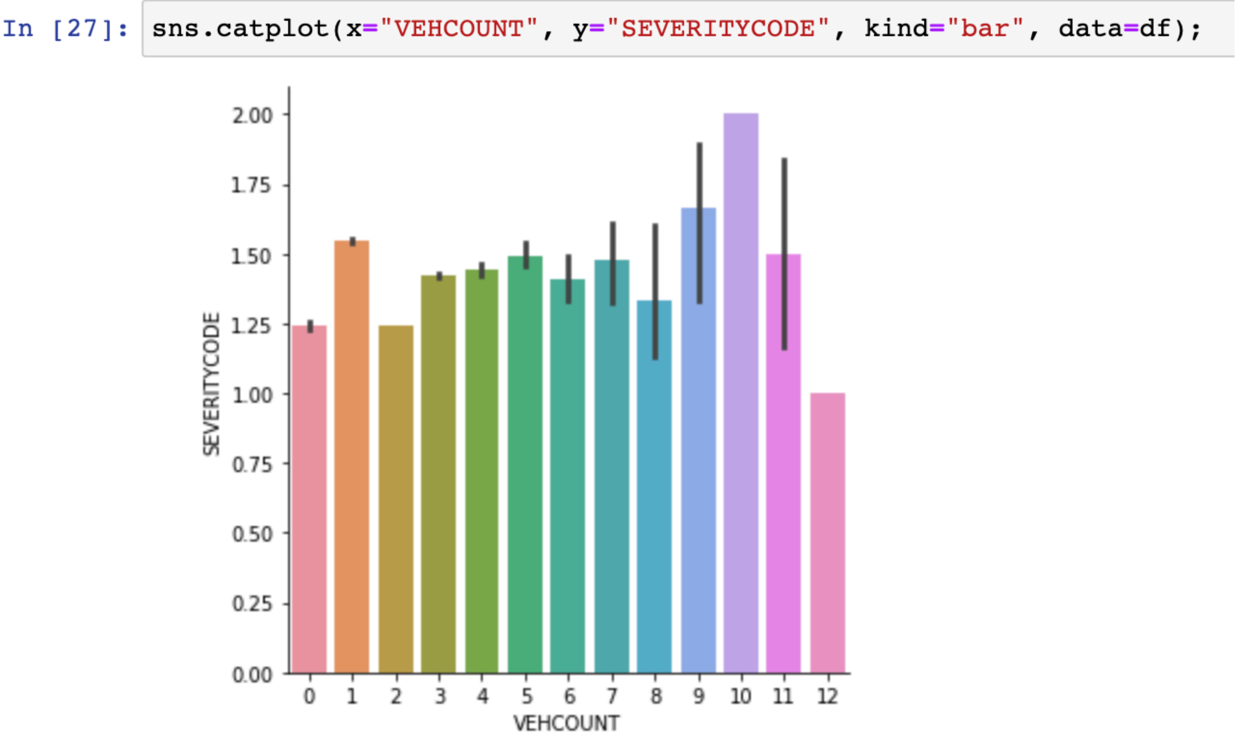
```
In [25]: # Preprocessing

         df.groupby(['ADDRTYPE'])['SEVERITYCODE'].value_counts(normalize=True)
         df.groupby(['WEATHER'])['SEVERITYCODE'].value_counts(normalize=True)
         df.groupby(['ROADCOND'])['SEVERITYCODE'].value_counts(normalize=True)
         df.groupby(['LIGHTCOND'])['SEVERITYCODE'].value_counts(normalize=True)
```

This normalized the data so the categorical data would be read as floats.

Then, I explored some of the data using plots with the seaborn library.



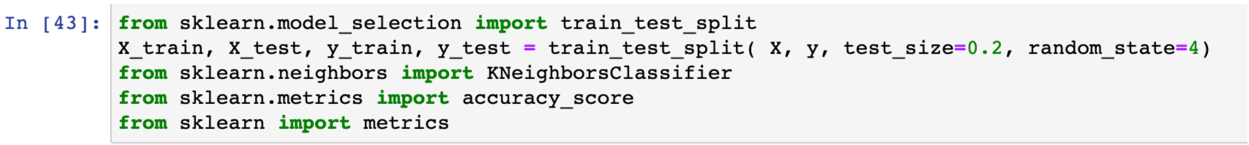
```
In [28]: sns.catplot(x="ADDRTYPE", y="SEVERITYCODE", kind="bar", data=df);
```



```
In [27]: sns.catplot(x="VEHCOUNT", y="SEVERITYCODE", kind="bar", data=df);
```

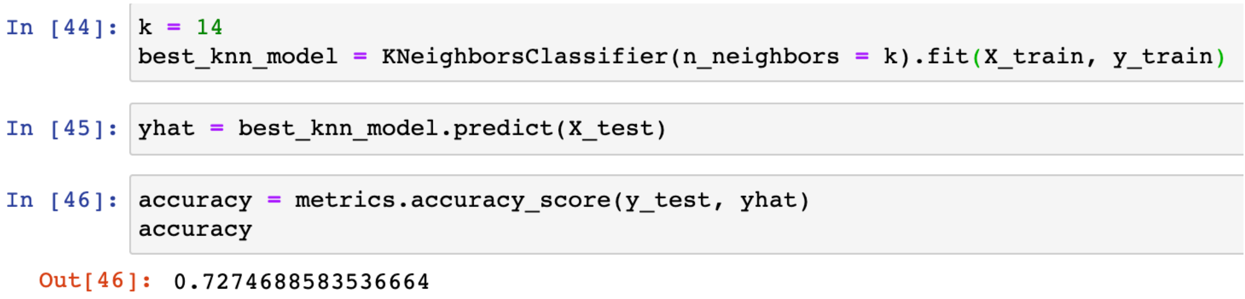
After examining the different plots, I selected the features used to build a model. I chose K-Nearest Neighbors (KNN) as the machine learning model to predict accident severity.

To train the model, I imported the libraries required and split the data into training and test data.



```
In [43]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=4)
         from sklearn.neighbors import KNeighborsClassifier
         from sklearn.metrics import accuracy_score
         from sklearn import metrics
```

Finally, I trained the model and used it to predict accident severity of the test data set.



```
In [44]: k = 14
         best_knn_model = KNeighborsClassifier(n_neighbors = k).fit(X_train, y_train)

In [45]: yhat = best_knn_model.predict(X_test)

In [46]: accuracy = metrics.accuracy_score(y_test, yhat)
         accuracy

Out[46]: 0.7274688583536664
```

Results

Using the KNN model gave me an accuracy of 73%. This demonstrates that the features I selected are able to give a good estimate of how severe an accident can become given the address type, weather conditions, road conditions, light conditions, and whether or not a car is speeding.

Discussion

From the exploratory analysis, I was able to recognize that most car accidents happened on clear days with dry roads during the day. Although these conditions may be the most prominent, it is important to note that this may be due to people driving more careful when conditions are bad and there are many more drivers during the day when the weather and road conditions are nice.

Conclusion

This project helps people understand the factors contributing to the severity of road accidents so drivers of ridesharing companies can make smarter decisions while driving such as taking different routes or driving more cautious than usual to limit the risk of accidents. Although this is targeted towards ridesharing company drivers, as this is their everyday job, any driver can find this information useful to assess their risk of driving during certain conditions and in specific locations.