IBM Data Science Professional Certificate

Capstone Project – Airport versus Stadium

Introduction

Introduction where you discuss the business problem and who would be interested in this project.

For my Capstone project, I used Foursquare to pull location data on areas in the United States with NFL Stadiums and International Airports. With the data, I will attempt to use machine learning to determine whether the type of venue (Stadium or Airport) can be determined using the venue categories within a half mile from the Stadium or Airport.

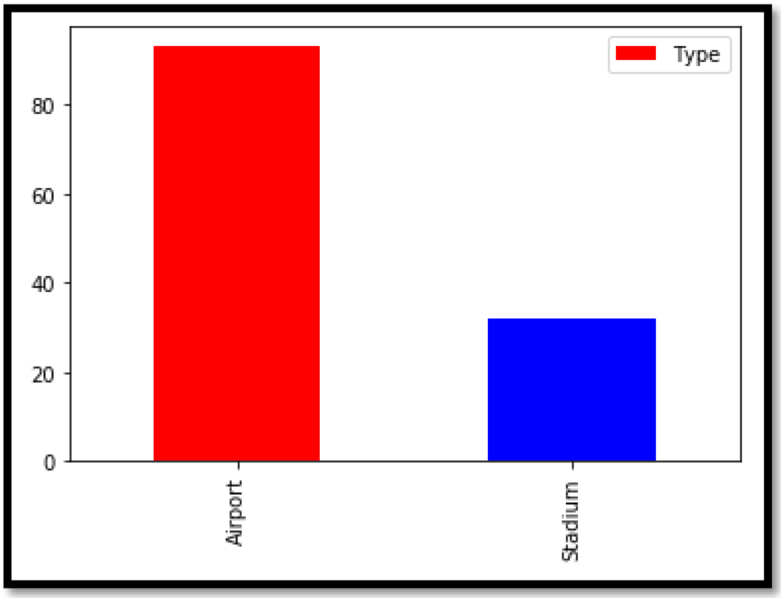
Data

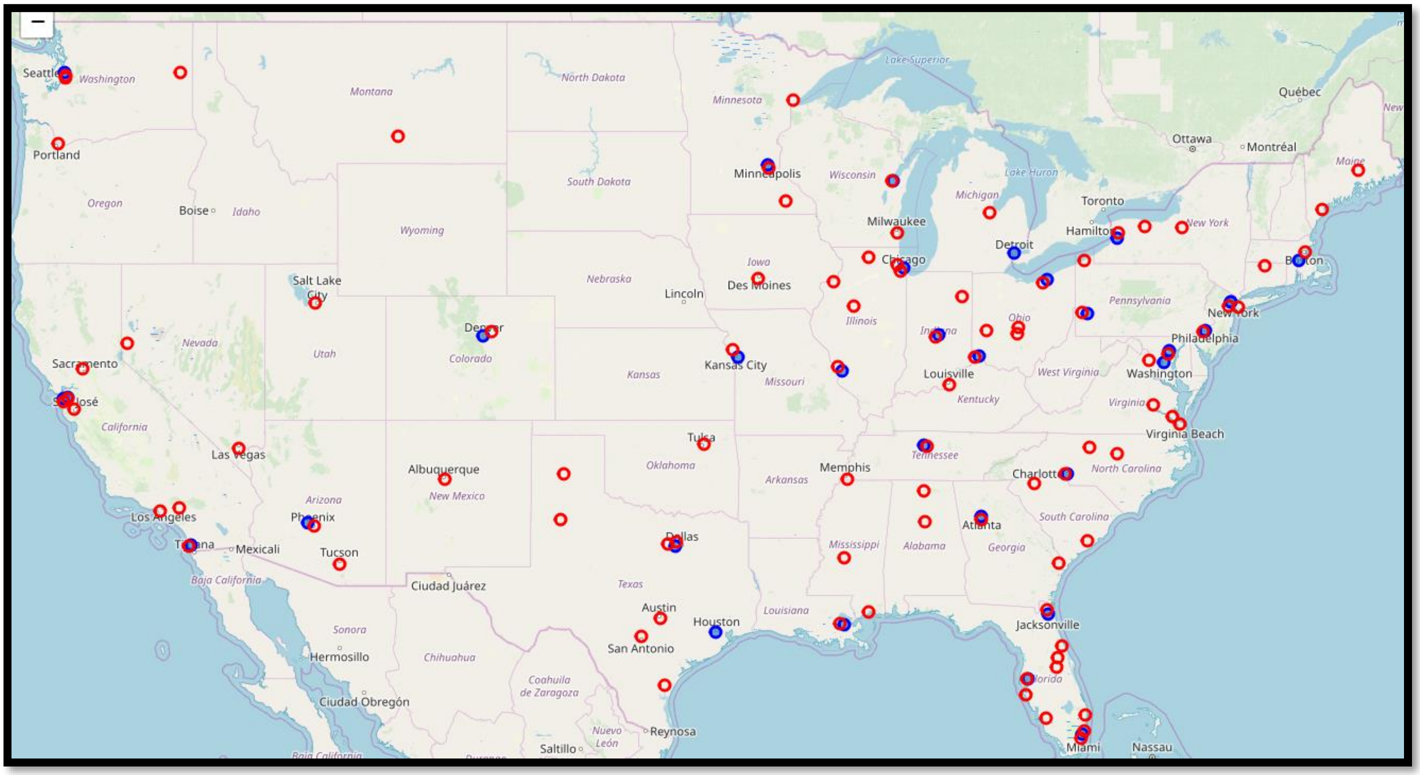
Data where you describe the data that will be used to solve the problem and the source of the data.

I used 3 data sets for this exercise.

- NFL Stadiums and their locations: https://sites.google.com/site/32nflteamsabsoluteandrelative
- International Airports and their locations: https://www.latlong.net/category/airports-236-19.html
- Foursquare location data.

The stadium and airport data will be combined, categorized, and then merged with the Foursquare data so we can apply a categorical analysis. There are 32 NFL Stadiums and 62 Airports.




−
Seattle
Washington
Portland
Oregon
Montana
North Dakota
Minnesota
Lake Superior
Québec
Ottawa
Montréal
Maine
Minneapolis
Wisconsin
Lake Huron
Michigan
Toronto
South Dakota
Boise
Idaho
Wyoming
Milwaukee
Hamilton
New York
Detroit
Chicago
Boston
Iowa
Des Moines
Nebraska
Lincoln
Salt Lake City
Illinois
Pennsylvania
New York
Nevada
Utah
Denver
Colorado
Indiana
Ohio
Philadelphia
Sacramento
Kansas City
Missouri
Kansas
Louisville
West Virginia
Washington
San José
Kentucky
Virginia
California
Virginia Beach
Tulsa
Las Vegas
Oklahoma
Tennessee
Albuquerque
Memphis
Arkansas
Charlotte
North Carolina
New Mexico
Arizona
Los Angeles
Phoenix
Atlanta
South Carolina
Tijuana
Mexicali
Dallas
Mississippi
Alabama
Georgia
Tucson
Baja California
Ciudad Juárez
Texas
Austin
Louisiana
Houston
Jacksonville
Sonora
San Antonio
Hermosillo
Chihuahua
Coahuila de Zaragoza
Florida
Ciudad Obregón
Nuevo León
Reynosa
Saltillo
Miami
Nassau

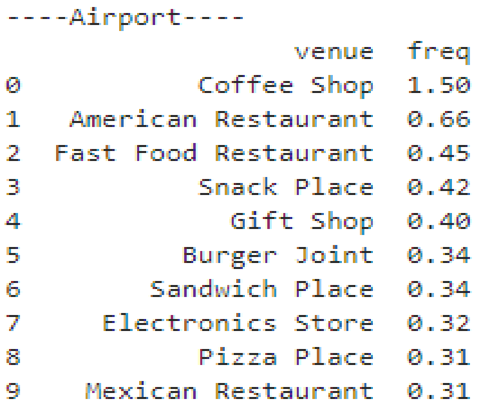
Methodology

Discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why

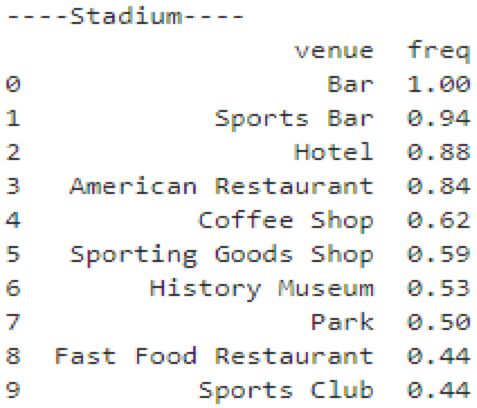
First, I excluded venues that would easily identify the location as an NFL Stadium or Airport. The excluded categories are as follows:

- Football Stadium
- Stadium
- Athletics & Sports
- Airport
- Airport Service
- Airport Gate
- Airport Terminal
- Airport Lounge
- Duty-free Shop
- Airport Food Court
- Rental Car Location

After excluding the venues, I pulled the frequency of the categories by location. The table on the right displays the top 10 venues by their mean. Coffee Shops lead the venues at Airports with 1.5 per location where Bars/Sports Bars lead the categories for Stadiums. There are overlapping categories however like Coffee Shops, American Restaurants, and Fast Food Restaurants.



```
----Airport----
                  venue  freq
0           Coffee Shop  1.50
1   American Restaurant  0.66
2  Fast Food Restaurant  0.45
3           Snack Place  0.42
4             Gift Shop  0.40
5          Burger Joint  0.34
6        Sandwich Place  0.34
7     Electronics Store  0.32
8           Pizza Place  0.31
9    Mexican Restaurant  0.31
```

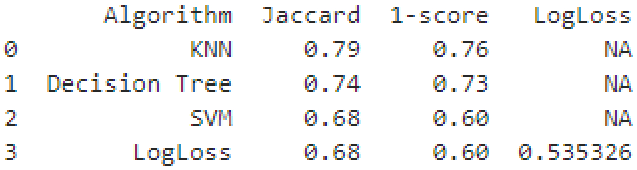


```
----Stadium----
                  venue  freq
0                   Bar  1.00
1            Sports Bar  0.94
2                 Hotel  0.88
3   American Restaurant  0.84
4           Coffee Shop  0.62
5   Sporting Goods Shop  0.59
6        History Museum  0.53
7                  Park  0.50
8  Fast Food Restaurant  0.44
9          Sports Club  0.44
```

Machine Learnings

I selected 4 algorithms to identify Airports and Stadiums from the surrounding venues – KNN, Decision Tree, Support Vector Machin, and Logistic Regression. I used a test size of .20 of the data set.

The most successful algorithms using the test set were KNN and Decision Tree.

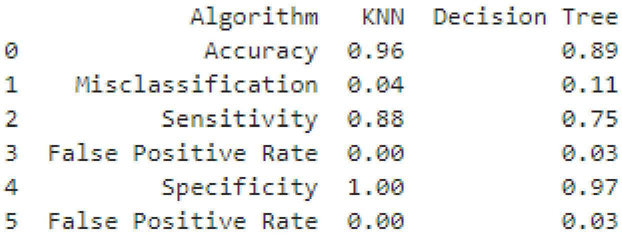


```
          Algorithm  Jaccard  1-score   LogLoss
0               KNN     0.79     0.76        NA
1     Decision Tree     0.74     0.73        NA
2               SVM     0.68     0.60        NA
3           LogLoss     0.68     0.60  0.535326
```

Results

Discuss the results

After applying KNN and Decision Tree algorithms to the full data set, the results were as follows:



```
             Algorithm   KNN  Decision Tree
0             Accuracy  0.96           0.89
1    Misclassification  0.04           0.11
2          Sensitivity  0.88           0.75
3  False Positive Rate  0.00           0.03
4          Specificity  1.00           0.97
5  False Positive Rate  0.00           0.03
```

KNN performed slightly better. It was able to identify all stadiums correctly and identified 28 of the 32 stadiums correctly. 4 stadiums were misidentified as airports.

KNN Actual

		Stadium	Airport
KNN Predicted	Stadium	28	0
	Airport	4	62

Decision Tree Actual

		Stadium	Airport
Decision Tree Predicted	Stadium	24	2
	Airport	8	60

Discussion

Discuss any observations you noted and any recommendations you can make based on the results

The algorithm worked well however it could be improved by better categorizing the surrounding venues. The categories provided by foursquare are specific (i.e. type or restaurant, bar). It might help to generalize the categories. The radius could also be revised to include/exclude venues.

This exercise was limited to specific locations. It would be interesting to expand the analysis to use general locations to see if it can be determined whether a stadium or airport exists there.

Conclusion

It seems that, for these examples, foursquare data can be used to identify if a type of venue is in the vicinity.