Capstone Project
The Battle of the Cities and their restaurants¶

Applied Data Science Capstone by IBM/Coursera (Week 2)¶

Maurus Wüllner, 14.02.2021

1. Introduction: Decision Problem

The problem of this project is to help decide, where to drive with our new weekend car for the best food. The idea is to make a roadtrip to the best restaurant. But since parking is tough in each city, it should be just looked out for restaurant categories. It should be evaluated to which city to drive makes the most sense according to likes of restaurants in each city provided by Four Square. Then, it should be shown through data analysis for which restaurant type/class to look out for in this city. Is it the Kebab in Frankfurt, the Brauhaus in Cologne or the Biergarten in Munich? Let's find out!

2. Data

The dataframe will be named 'raw_dataset' as it is the most complete compiled form of the data before needing any processing for analysis via machine learning.

1. the geographical coordinates of the three cities (Munich, Frankfurt and Cologne) is obtained,
2. establish the Foursquare API in order to get the URLs that give the raw data in JSON form,
3. each respective URL is then scraped for the columns: 'name', 'categories', 'latitude', 'longitude', and'id' for each city,
    a. the city column will help us when separating where the restaurants are from.

The focus will be on those restaurants found within a 800km radius from the coordinates that were provided by the geolocator. The Foursquare API provides with

abundant venue categories only with that big radius. In real-life this makes no sense at all, since the cities are closer than 800 km from each other. But for the sake of the project and obtaining at least some data, we will continue like this. The results needs to be cleaned out by removing non-restaurant rows. Pulling the 'Likes' data is necessary to make final decisions. The 'id' column is used in order to pull the 'Likes' using the API and append the information into the dataframe. We concluded by naming the dataframe 'raw_dataset', which is used in the machine learning portion of the project.

3. Methodology

A dataframe called raw_dataset was set up. For the German citites of Munich, Frankfurt and Cologne the likes from FourSquare for restaurants were pulled. For later analysis purposes the restaurant were divided into categories: German, World, Others, Middleeast and European. This categorization shall help to better sort restaurants while driving through one of the cities. Furthermore the likes were categorized based on their quartiles into 4 categories (1 with the most likes 4 with the least). Right now, a decision tree should show, in which city it is most likely to obtain the highest likelihood for good food according to likes on FourSquare. The idea behind the Decision Tree is to give a help for possible decisions while driving through each city. Which restaurant category might provide the most likes and best food? Afterwards a regression analysis will be performed to get statistical insights regarding to which city to drive in the first place for the best restaurants. Secondly, the best cuisine category should be found through regression analysis.

4. Analysis

a. Let's create a decision tree

That's especially important while driving around with a car. For which factor in a restaurant place should we look at first?!

Neither the Decision Tree nor the Random Forest Classifier show high accuracies (both 29%).

Yet, the Decision Tree shows, for which type of cuisine first to look for, when driving through the three German cities.

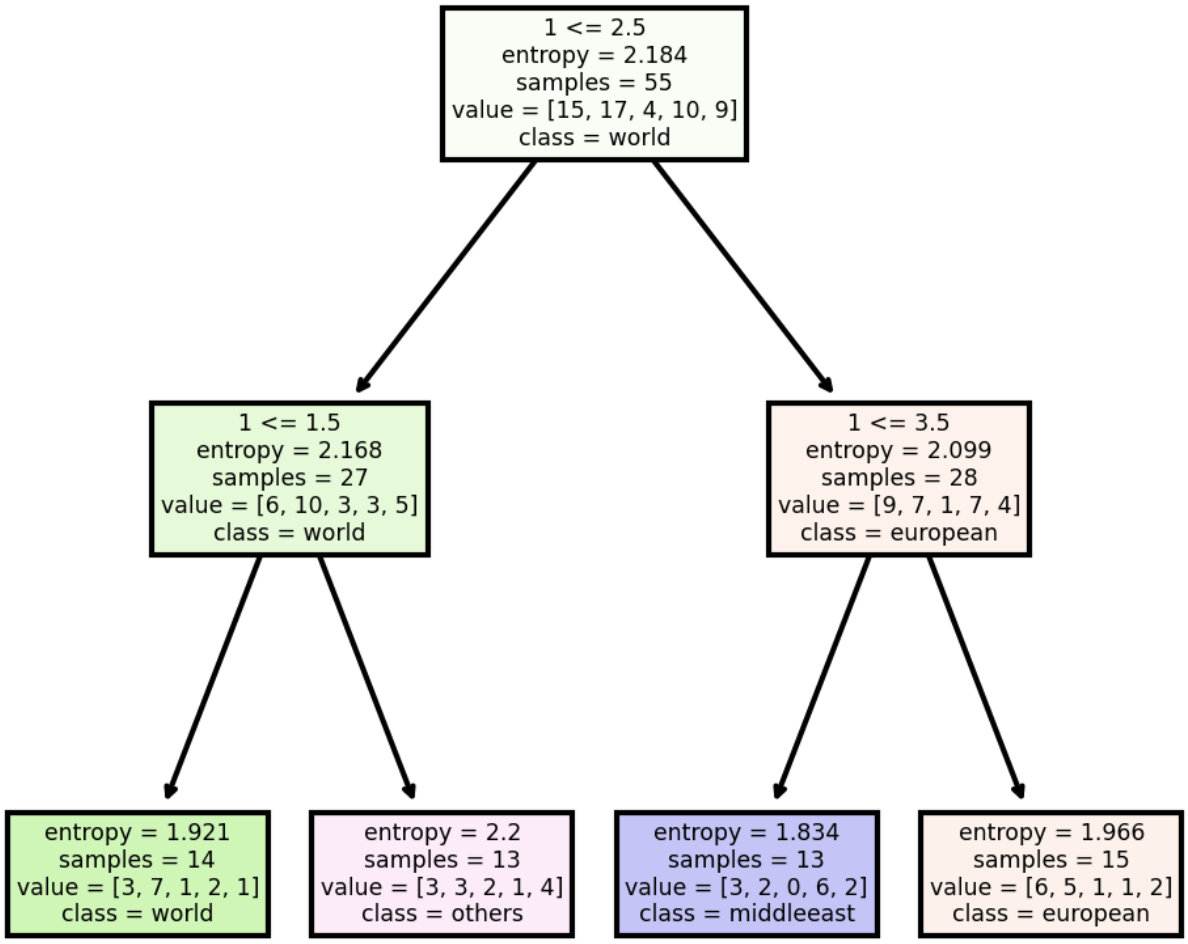

Figure 1 Decision Tree Cuisine Ranking in all three German cities

Figure 1 shows, that to look for the cuisine “world” shows the lowest entropy while scoring the highest, here the lowest since category 1 has the most likes. The worst kitchen to look for while driving through the cities is “European”. The order would be hence “world”, “others”, “middleeast” to “European”. The Decision Tree for each city only showed improved accuracy for Cologne with 44%.

b. Multiple linear regression

The multiple linear regression has the likes as the dependent variable and the cuisine and the city categories as the independend variables ('european', 'world', 'german', 'middleeast', 'others', 'Munich', 'Frankfurt', 'Cologne'). A linear regression model was

trained on a random subsample of 70% and then the other 30% was used for testing purposes. The residual sum of squares and variance score were calculated. Residual sum of squares: 38273.71 Variance score: -0.80

The variance score is negative. The linear regression is not a good way of modeling our data. Therefore, a logistic regression was performed for our analysis.

c. Multinomial ordinal logistic regression

The multinomial ordinal logistic regression model was trained on a random subsample of 70% and then tested on the remaining 30%. The jaccard score and log-loss were both calculated. Jaccard-Score: 0.15942028985507245 Log-Loss: 1.4008828664119835

Although the prediction is not too well, a jaccard score of 15% is somewhat reasonable. The classification report is included in the analysis.

Given the modestly accurate ability of this mode, model was run on the complete dataset. The coefficients show, that most factors show negative relations regarding the likes of a restaurant. Yet, the best combination to find a good restaurant is related to the factors "German" and "Munich".

5. Discussion

The Decision Tree showed, that, generally, in the three cities Munich, Frankfurt and Cologne the World kitchen is the go-to option, which consists mainly of Asian restaurants. This is true also just looking only on the cities of Munich and Cologne. In Frankfurt European cuisine, so Italian food shows the best ranking due to likes.
So, we already have some sort of sense to what to look out for, when approaching each city. But given, that the best result should be obtained through planning, it was set out to do a regression analysis. The linear regression analsis showed truly poor accuracy. Yet, the logistic regression was somewhat better, still with low accuracy (15%). This is probably due to a high variability in the data set. A way to obtain a higher Jaccard scores is by having more data. This is due to the low usage of FourSquare in Germany difficult. To find another data provider is out of scope for this project.

The logistic regression showed, that a restaurant in the model would most likely fall into the 1st category of more than 154 likes (40%). It has a 0% to fall into the second likes class and 38% and 25% to fall into the 3rd and 4th ranking class respectively. So which city to drive to? Munich shows most promise since it has a positive correlation to likes (r=0.61). Also German cuisine has the highest correlation to the highest ranking class in the model (r=0.69). The other cities have negative correlations to the highest ranking class. So do the other cuisines, but "European" (r=0.06).

6. Conclusion

The data analysis of restaurants in Munich, Frankfurt and Cologne shows, that Munich and German restaurants are to look out for, when searching for the best food in these towns. So Weißwurst and Weizen beer wins the comparison. Yet, the analysis shows low accuracies throughout each (!) step. This is due to few data provided by FourSquare. For "real-life" decision where to eat which cuisine, FourSquare seems unsuitable to provide sophisticated venue data - at least in major German cities. In a closer to real-life analysis better data should be used by more prominent providers in Germany such as Google maps. This was not possible for this project, though, due to fees.

Check out the GitHub repository:
https://github.com/maruk101/Capstone-Project-Coursera-ML