# Capstone Project

## Applied Data Science Capstone

## Week 5

## Problem Description

When deciding to open a restaurant, location is one of the biggest key factors influencing the success of the business. Given this small detail has such a high impact on the restaurant survivability, every businessman would be interested to know a priori where they should build their business. Which places have a higher tendency to lead to success and which ones are doomed to fail.

To have a general idea of what the problem entails, let us think about what type of location would be ideal. The new installations need to be in a good neighbourhood, one which is relatively quiet, but simultaneously crowded and with a sizeable crowd in order to increase the chances to find costumers. Moreover, it is best not to have too many restaurants around in order to decrease the competition between the different food establishments, but while also avoiding empty areas, or areas that are not associated with the business at hand.

Overall, there is no ideal neighbourhood and there are no hard constraints. One needs to find the equilibrium between the different requirements. So as to solve the problem present, in this project we will develop and approach that relies on machine learning and data extraction tools in order to figure out whether a certain neighbourhood has potential for us to build our restaurant.

## Data description

The study will focus on the area of Lisbon.

We will extract data from the listo of top 30 restaurants in Lisbon along with their addresses, from the zomato website (https://www.zomato.com/), as well as the list of bottom 30 restaurants (in ordert o have a means of comparison).

We will be using the tool foursquare, in order to get the location data from each restaurant.

## Methodology

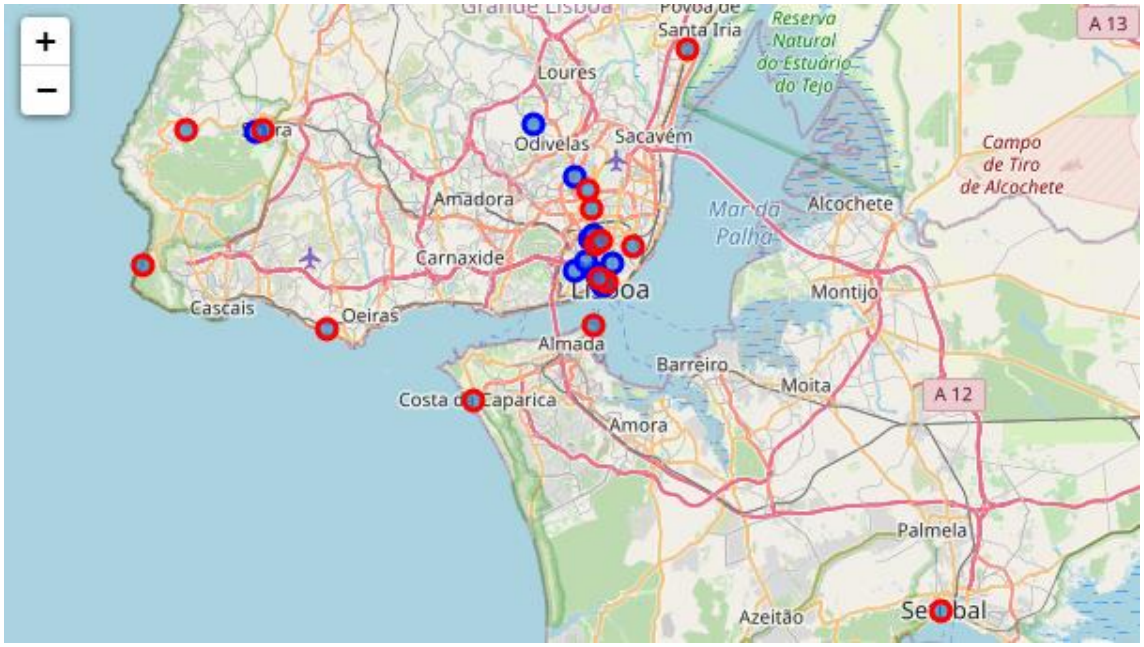
Having the Adresses and the respective location data from the restaurants, it will be necessary to properly clean it and prepare it for posterior analysis.

Afterwards, we will be able to compare the data obtained from foursquare and try to identify trends among the similar restaurantes and differences between the two different kinds.

Moreover, we will be able to build machine learning models, such as decision trees, that will help us determine wether or not a certain location is fit to have a new restaurant.

## Results

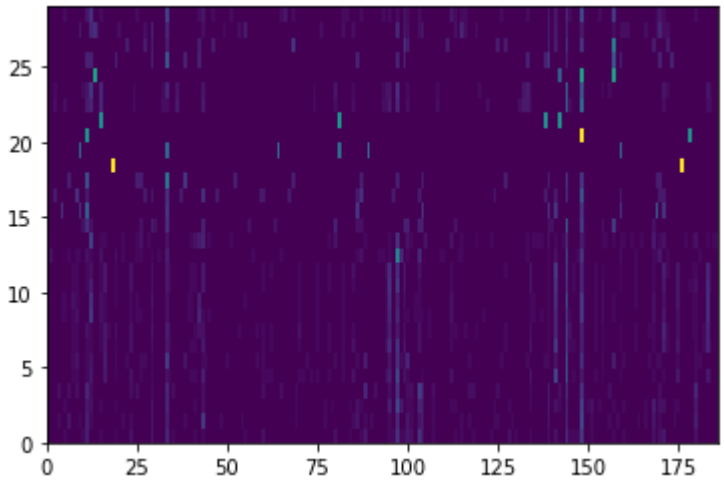
Map with original points from the different restaurants selected. In blue we can see the locations of the restaurants with a high rating, whereas in red we have the restaurants with a low score.



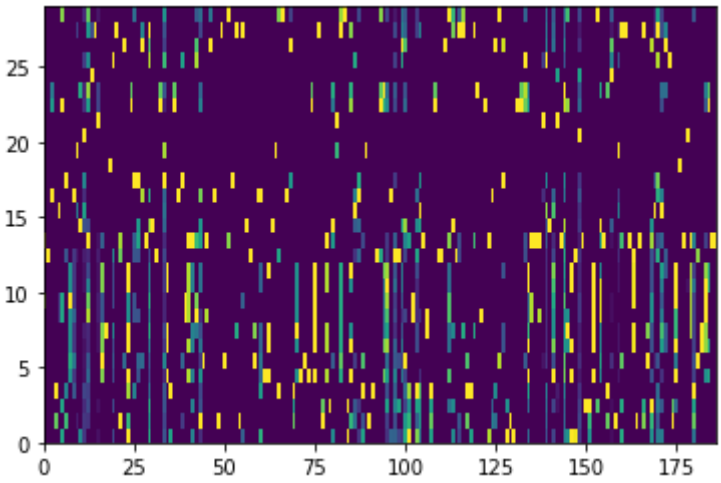

From these coordinates it was possible to extract the different venues location data.

From this data we applied one hot encoding, which allowed to transform categorical data into binary one, by turning each new category into a new feature.

These results can be seen in the following heat map (each row corresponds to a restaurant and the columns correspond to the features). It is possible to say that each row/restaurant is represented by a histogram. This histogram encodes the quantity represented by each feature in the restaurants' surroundings.



However the information obtained from the foursquare is not standardized. Therefore it was necessary to scale the data obtained. In order to have a feel for the impact of the scaling we can look at the new heatmap which originated from the scaling performed.




After cleaning the data it was necessary to split the data into a training set and test set. In order to do this we used the train_test_split provided by sklearn and used a test size of 0.33 (so 33% of the data was used in the test set and 66% in the training set).

Finally it was possible to apply our classification algorithms.

The ones applied were Decision Trees and SVM. In the end the accuracy obtained was 90% for both algorithms.

## Discussion

As it is possible to see, we were able to obtain an accuracy of 90% using both Decision trees and SVM. This shows how promising this new method to classify new locations for restaurants.

However it is important to keep in mind that the current dataset was quite small and in order to get more realistic results a much larger one should be used. Such low quantity of data entries and high amount of features may lead to overfitting, that due to the low amount of data is not visible.

Moreover, adding another features besides the location data would also be ideal (these features could include the average cost per person, type of cuisine...)

## Conclusion

Overall this project was quite useful. It allowed to put into practice everything learned along all the project's progress.

Regarding the results obtained they seem good, but due to the low quantity of data, no real conclusion can be made. At most its usefullness falls on the essential train acquired to derive those results.