THE BATTLE OF NEIGHBOURHOODS: FINDING A PLACE FOR YOUR GYM

week 4 – capstone project for IBM Data Science Professional Certificate provided by Coursera.

Gabriele Favia

Predicting where to place your business has a huge impact on its success

Given a city and a possible type of shop/business/POI, is it possible to estimate which one of the vacant premises is the most suitable for a new opening?

Objective: estimating which one of the neighborhoods composing the target city is the most suitable to host the next gym.

Data acquisition

- Data relative to the places for rent in Bari, acquired from a popular Italian website (https://www.casa.it/) making parametric https calls and treating the results using webscrapping.

- Information about the position of Bari neighborhood provided by Wikipedia using webscrapping techniques from https://it.wikipedia.org/wiki/Quartieri_di_Bari

- Map of the boroughs by converting the shapefile from the official website of the Bari Municipality http://opendata.comune.bari.it/dataset/circoscrizioni

- Getting the names of the Italian city with population rate similar to the one of Bari (between 180k and 420k), scrapping them from wikipedia https://en.wikipedia.org/wiki/List_of_cities_in_Italy#Cities

Building the dataset

Extraction of the venues around the business candidates found in Bari.

Extraction of the gyms found of the cities between 180k and 420k.

Extraction of the venues near the gyms of the other cities, giving a LIMIT parameter and a maximum RADIUS to the query.

Data cleaning and preparation

- Deleting all the gyms with a rating below a set threshold (7.0)

- One hot encoding for both the datasets

- Features candidates features and city-gyms features are intersected.

Supervised models: dealing with different techniques 1/2

Logistic Regression

Classification report with liblinear solver:

	precision	recall	f1-score	support
No	0.50	0.70	0.58	10
Yes	0.25	0.12	0.17	8
micro avg	0.44	0.44	0.44	18
macro avg	0.38	0.41	0.38	18
weighted avg	0.39	0.44	0.40	18

Support Vector Machines

Classification report with sigmoid kernel:

	precision	recall	f1-score	support
No	0.50	0.80	0.62	10
Yes	0.00	0.00	0.00	8
micro avg	0.44	0.44	0.44	18
macro avg	0.25	0.40	0.31	18
weighted avg	0.28	0.44	0.34	18

KNN

	precision	recall	f1-score	support
No	0.62	0.80	0.70	10
Yes	0.60	0.38	0.46	8
micro avg	0.61	0.61	0.61	18
macro avg	0.61	0.59	0.58	18
weighted avg	0.61	0.61	0.59	18

- ✓ Train ds size: 75%, test ds size: 25%

Supervised models: dealing with different techniques 2/2

Decision Tree

	precision	recall	f1-score	support
No	0.64	0.70	0.67	10
Yes	0.57	0.50	0.53	8
micro avg	0.61	0.61	0.61	18
macro avg	0.60	0.60	0.60	18
weighted avg	0.61	0.61	0.61	18

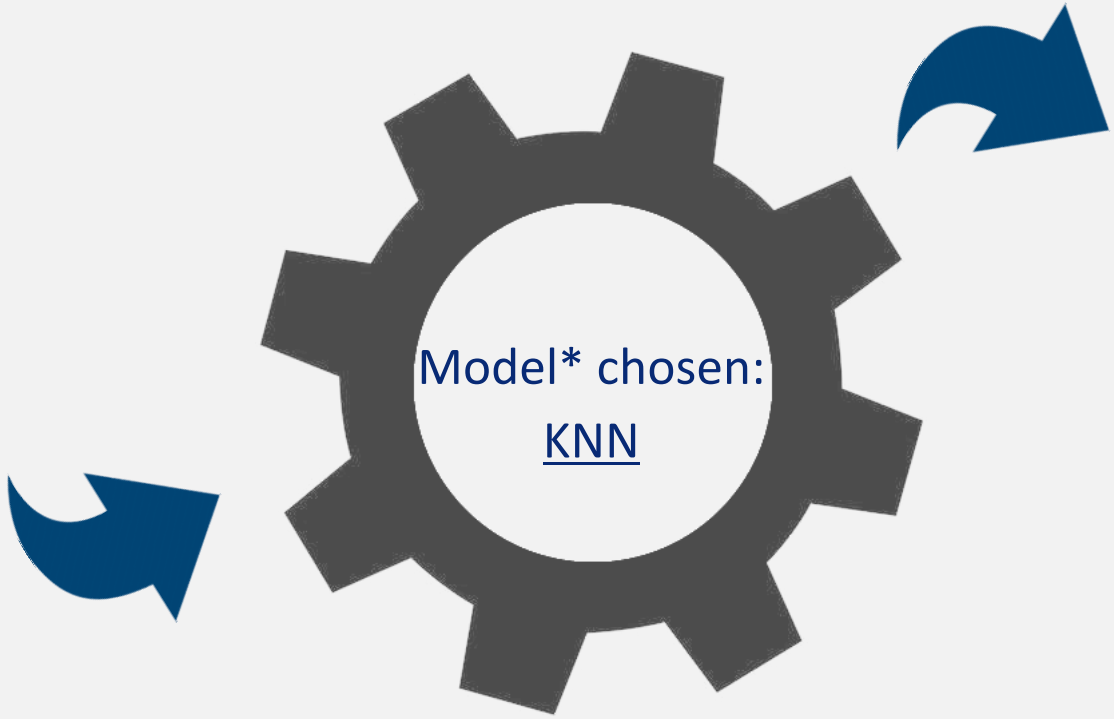
Naïve Bayes

	precision	recall	f1-score	support
No	0.43	0.30	0.35	10
Yes	0.36	0.50	0.42	8
micro avg	0.39	0.39	0.39	18
macro avg	0.40	0.40	0.39	18
weighted avg	0.40	0.39	0.38	18

- ✓ Train ds size: 75%, test ds size: 25%

Results 1/2

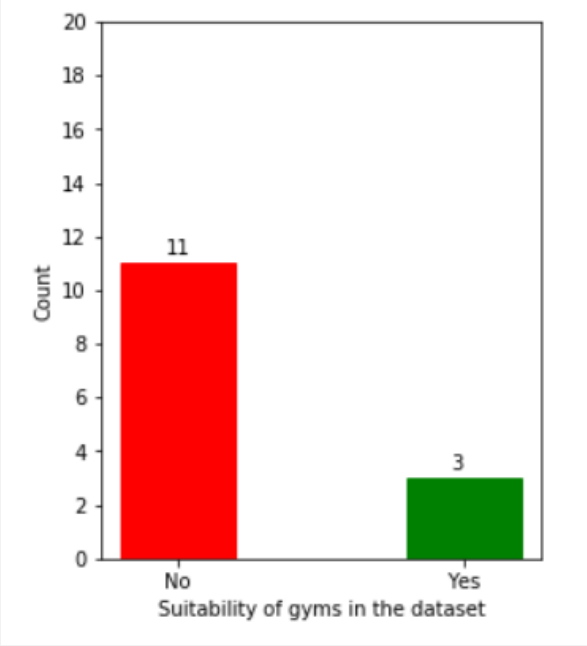
Dataset of the candidates for business places in Bari



Coordinates of "Suitable" places for a business cold start

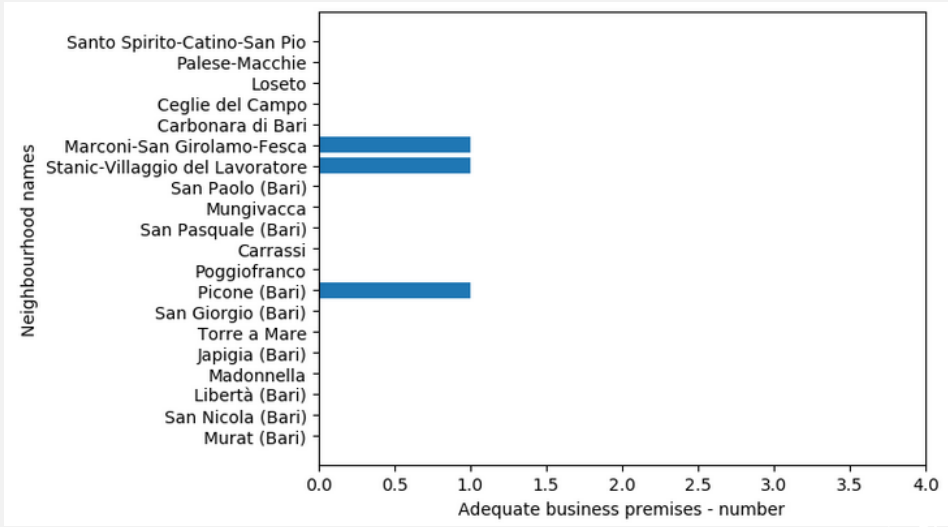
*Retrained with the whole "gyms → venues " dataset

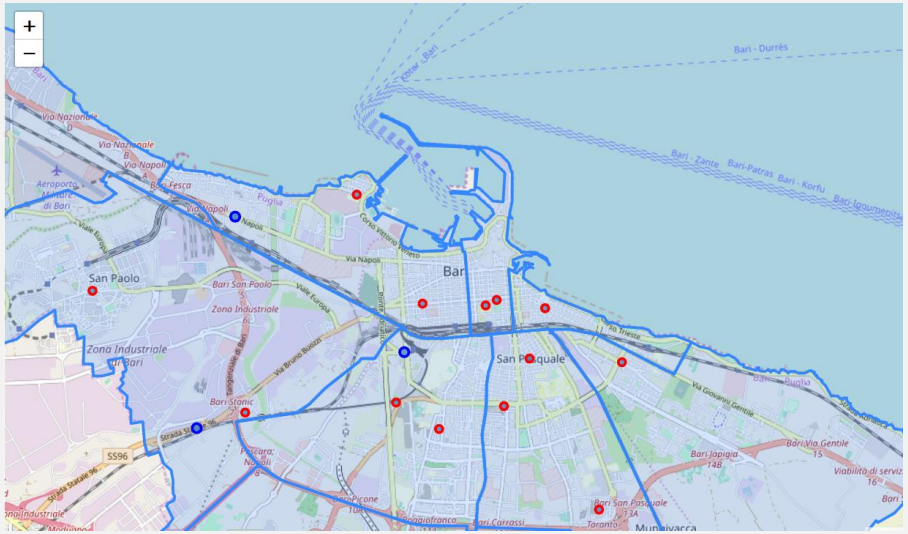
Results 2/2



Suitable and non-suitable candidates in Bari

Number of candidates per neighborhood





Map of Bari:

- Polygons: borough areas
- Red: neighborhoods
- Blue: places “Suitable”

Conclusion, limitations and future directions 1/2

- Three neighbourhoods won "the battle" : "Marconi-San Girolamo-Fesca", "Stanic-Villaggio del Lavoratore" and "Picone".

- The rating of a gym can be discriminated only by the surrounding businesses? That is of course not true but this assumption can be taken in account for future development as a subset of feature to analyze.

- The whole research of venues is limited by the Foursquare API number of calls per day, therefore the LIMIT parameter of venues around each gym has been limited to 100 entries.

- The parameter RADIUS used for the search of venues is set to 250 meters, this is considered the maximum limit of distance in which a person is willing to walk before or after the gym session. No information in the social sciences literature has been found about it.

Conclusion, limitations and future directions 2/2

- The cities to train with have population between 180000 and 420000 units: this choice has been made looking at the calls limit in Foursquare API and because the model wanted to take in account the behavior of an average city inhabitant in Italy.
- The host used to search for free location in Bari is casa.it, scrapped from a single webpage. An extensive system should be able to use API calls or at least being able to scrap multiple websites being aware that the possible clones of announces should be deleted in the phase of data preparation.
- The boroughs of Bari are following the old grouping system: that's because the new one (with "Municipi") tends to group too many neighborhoods together, making the differentiation of zones too much simplistic.
- The threshold rating imposed to 7.0 is the result of average quality perception of a point of interest.

THANK YOU FOR YOUR ATTENTION