DATA SENSUM

Can we Predict Area Average Income with Foursquare GeoCoding Venues Data?

Fabio Fulci

Dec, 2020

Table of Content

(1.) Introduction

1.1 Background

Foursquare City Guide, commonly known as **Foursquare** (https://foursquare.com/), is a local search-and-discovery mobile app developed by Foursquare Labs Inc. The app provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history.

Foursquare grants access via API to many data coming from his platform regarding all the venues registered. By Analysing the data in the platform it is possible to gather a lot of information regarding a specific city, town or neighbourhood. Therefore we will investigate in this case study, how can we use Foursquare valuable data to make predictions.

For instance by analysing the venues in a certain location it is possible to understand the area better: is it a Popular Touristic site, is it a residential area etc...

1.2 Problem

In this project we are going to verify if it is possible to accurately estimate the average income of an area by analyzing FOURSQUARE Venues and other publically available Geocoding Data (Latitude, Longitude, Altitude).

In Particular we are going to focus on Italy and we will try to build a model in order to predict the wealth of an Italian town and we are going to use as target variable the average taxable income per-capita of each ‘Comune’ (Town in Italian) registered by ISTAT.

The ‘Comune’ is the smallest administrative territorial entity in the Italian administrative system and can vary widely in size from few people up to more than 3 Million People for the biggest Comune which is Rome

1.3 Interest

For business or a no-profit organization can be useful to estimate the average income of a certain town or area. For Instance, It could be helpful for a non-profit organization that has to decide where to hold a Fundraising event or for a restaurant owner to decide where to open a new venue.

The average income of an area is a very useful information to make business decisions and understand better the territory. Unfortunately in many countries it is not possible to get updated data on average income and in many cases it is not available at the granular level needed. For instance you have the average income for the counties but not for each town, or you have it for the city and not for each neighbourhood. Therefore it is interesting to investigate if it is possible to estimate accurately the average income by looking at data freely available through the Foursquare API.

Moreover, this study allows us to find interesting patterns, for instance which kind of venues are more frequently found and are strongly correlated with the wealth of an area.

(2.) Data acquisition and cleaning

2.1 Data sources

1. **ISTAT** (Italian Government Census Data Collection Agency) : https://www.istat.it/en/

    a. Average Taxable income in Italy 2018 (last data publically available) for every town ('Comune' in Italian)
    b. Population Resident for each Town ('Comune' in Italian) in Italy
    c. Average Altitude from sea level for each Town ('Comune' in Italian) in italy

2. **LOCATIONIQ**.com API (https://locationiq.com/):

    a. Latitude and Longitude of each Town in Italy (Forward Geocoding)

3. **FOURSQUARE**.com API (https://foursquare.com/):

    a. List of Venues in a radius (SEARCH ENDPOINT)
    b. List of Venues Recommended in a radius (EXPLORE ENDPOINT)

2.2 Data cleaning

As the first step all data downloaded from ISTAT were combined in a single table using the Town unique code as key for the merge. All towns with a population lower than 10.000 people were dropped from the table to avoid building a model too influenced by the noise of small cities with very few venues registered on Foursquare. These gave us a total of 1213 towns covering almost the 70% of the Italian population

About 8 towns missing the Altitude data, so this data was added as the average altitude of the other towns in the same county.

The second step was to retrieve the Latitude and Longitude for each town. We got the Geographical Coordinated through LOCATIONIQ API. By using LOCATIONIQ Forward Geocoding EndPoint and query a string with "<Town Name>, <County Code>, <Region>" we got as a response the geographical center of town Latitude and Longitude.

The third step was to use the geographical coordinates of each town to interrogate Foursquare server. Firstly we used the "explore" endpoint in order to get for each town the number of expensive venues (price tag of 3 or 4) in a radius of 15 km of the town. Lastly we used the "search" endpoint to get all venues in a radius of 15 km of each town and from the API response we have extracted the total number of venues for each category.

At the end we found more than 600 venue categories and for each town we got the total count of venues per category. Finally we got the total number of venues (sum of all categories) for each town and we dropped from the table all categories where the total counts for all towns was less than 15.

2.3 Feature engineering and Selection

Now with the clean dataset we have tried to extract new features that could be highly correlated with the target variable. First of all we have extracted the Total population for each town living in a radius of 15 Km. in simple terms the sum of the population of all towns nearby with a distance inferior to km. This feature was useful in order to calculate the Per-Capita rate of the expensive venues in the area. In fact by simply dividing the Number of venues with the Town Population we would have got a misleading rate, because we had in fact extracted from Foursquare all venues in radius od 15 km from the town.

Moreover we have created a new binay Feature in order to label "suburbs" towns, i.e. small towns close to big ones. We have labeled as suburbs all towns with a population less than 5% of the total population in a radius of 15 km.

As last step we dropped from the table all non-numeric features (Town Name, Region Name, County Name etc...)

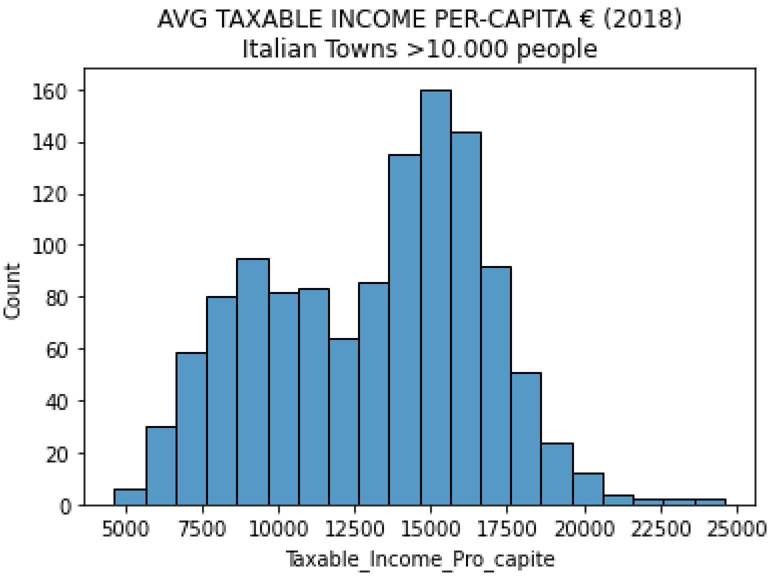
(3.) Exploratory Data Analysis

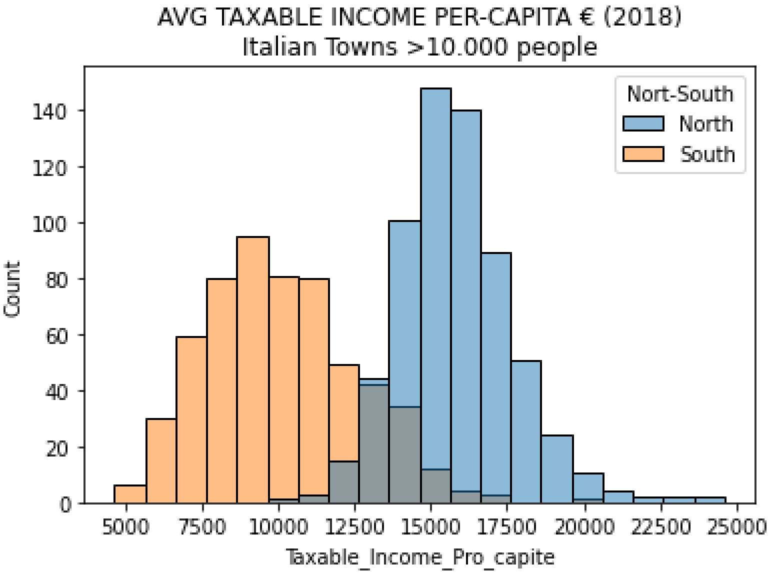
3.1 Calculation of target variable

The Target Variable was the Average Taxable Income per-capita of the town. It was calculated taking the Total Taxable Income in 2018 of the town taken from ISTAT divided by the Population resident in the town, also a data taken from ISTAT. We should point that the data is almost 3 years old (the latest available) while the Foursquare Data is from Dec, 2020. The timing difference can certainly affect negatively the outcome but we don't expect that the variance of the target variable inside the group has not changed a lot in the past 2 years.

We should notice that this data is the best proxy of the "wealth" of the town we could find, however it has some issues. In fact being a census data of the "Taxable" income it is a good proxy of the wealth of the town if the rate of Tax Evasion they don't differ too much from one town to the others. As we have noticed for instance there are few towns near the Swiss Border with a strangely low Taxable Income, and this is due to the high rate of residents there who work in Switzerland and so have an Income that is not Taxable in Italy.

As it is possible to see in the Histogram on the right, the target variable has a bi-modal distribution. This particular distribution is due to the strong divide between the richer North and the poorer South of the country. The different income distribution of the two parties

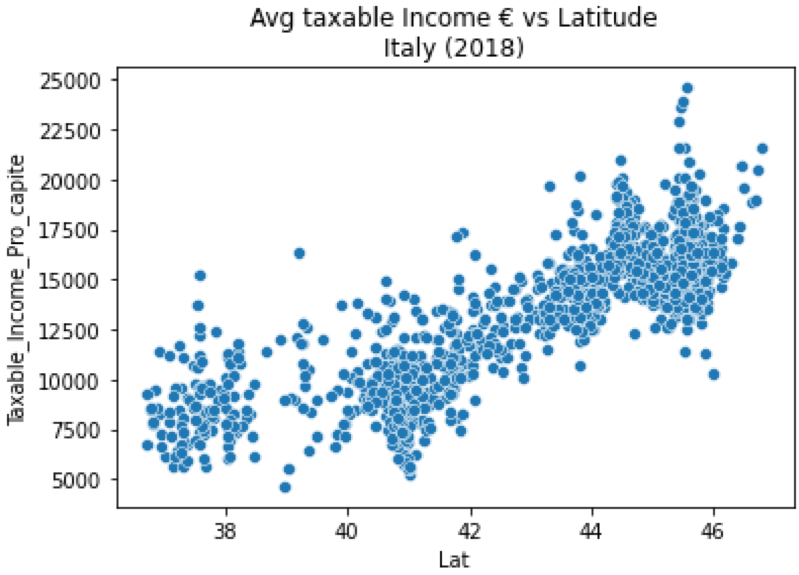




of the country is even more evident in the second Histogram.

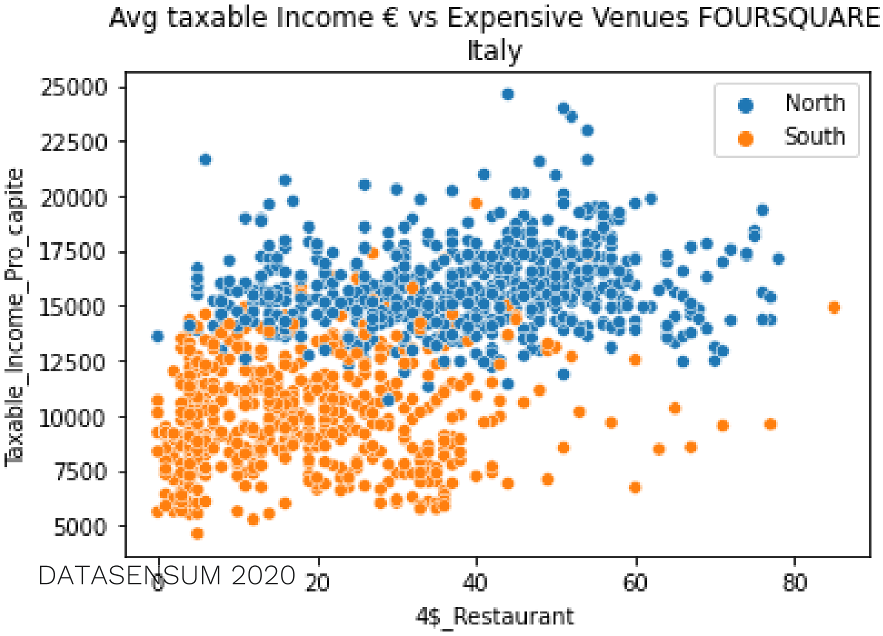
3.2 Relationship between Avg Income and Latitude

As it is possible to see from the scatter plot below, there exists a strong positive correlation between the target variable and the Latitude of the town. Indeed the Northern part of Italy tends to be much richer than the south.
The Latitude is indeed the variable that has the highest correlation (0.82) with the Average Taxable income of Italian towns.



3.3 Relationship between Avg Income and FOURSQUARE Expensive Venues

The strongest correlated variable extracted from FOURSQUARE database is the Number of Expensive Venues recommended in a radius of 15 Km from the town.

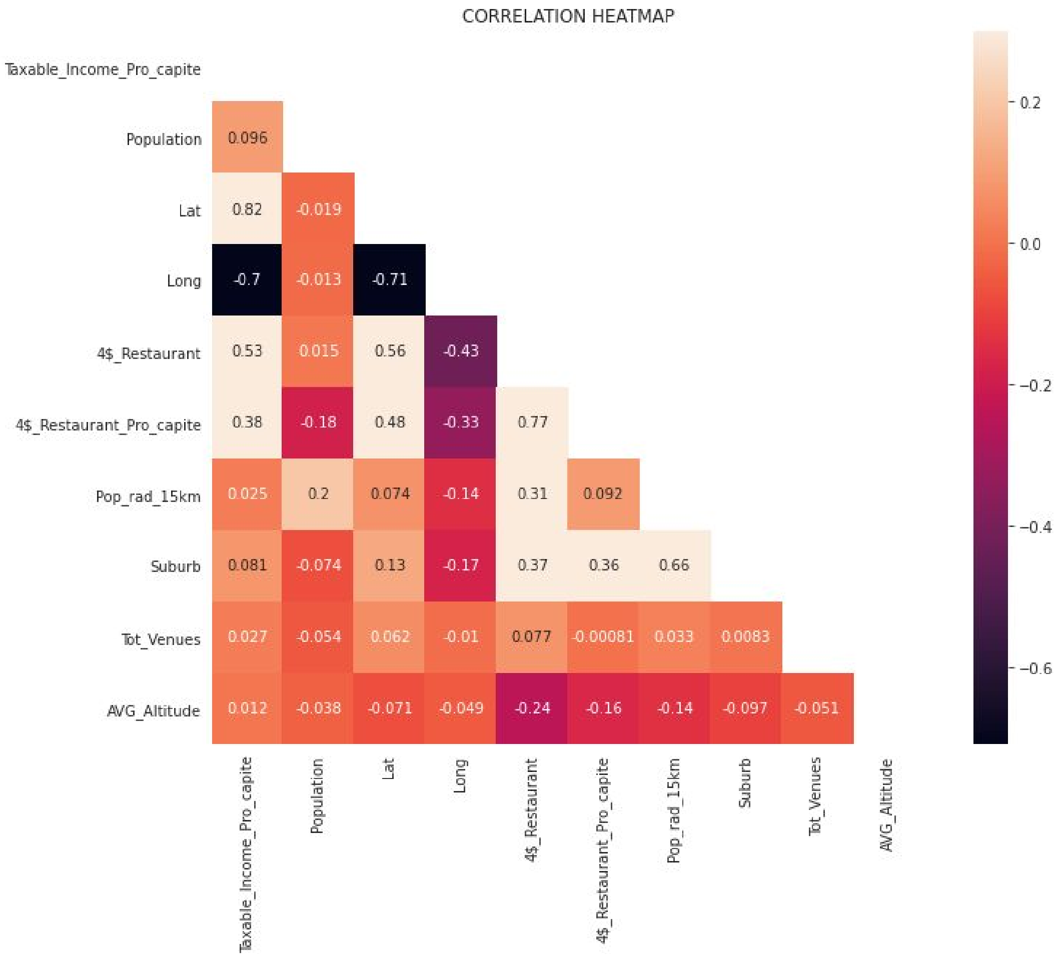


As you can see from the graph on the left the correlation is due primarily to the fact that the Towns in the North have an average Number of recommended expensive venues listed on Foursquare that is much higher.
In fact if we take only the Northern towns the correlation between the Number of Venues and the target variable falls to 0.

Unfortunately the amount of venues listed on Foursquare is very limited and represent just a small portion of the real venues existing. For instance in Rome the biggest city in Italy we found only 27 expensive venues which show a huge lack of cover in that area of Foursquare Services.

3.4 Correlation between Avg Income and other variables (Heatmap)

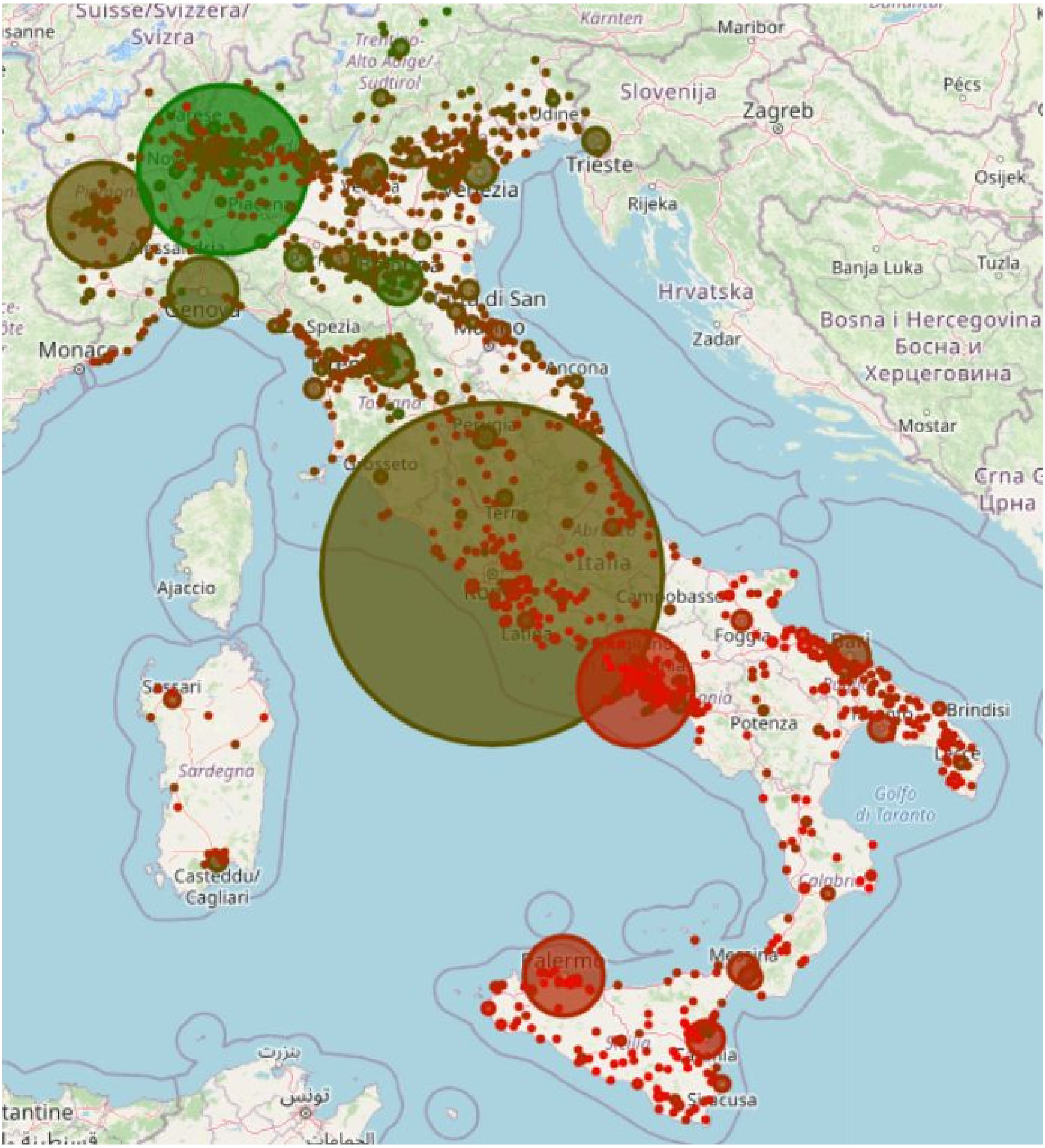
Analysing the heatmap of the correlation is evident that beside the Latitude there aren't any variables strongly correlated to the target variable.
The Longitude is indeed strongly correlated too, but it depends too from the latitude, in fact because of the shape of the Italian Peninsula the southern cities tend to be also more eastern.
Moreover contrary to what we expected the Total number of Venues have no correlation with the target variable. Finally the rate of Expensive Venues per Population (4$_Resturant_Per_Capita) appears to be less correlated to the target variable compared to the total number of Expensive Venues (4$_Restaurant). This doesn't make much sense because we would expect that the ratio to be more relevant than the total number. In simple terms if a town has an high ratio of expensive venues per resident we would expect it to be a rich town, however if the town has a relatively high number of expensive venues but a low ratio of expensive venues per person (because it is a big city), we would expect it to be necessarily a rich town. The lack of cover of Foursquare evidently is to blame for this counterintuitive outcome



3.5 Avg Income and Population Geographic Distribution

The map below shows clearly the income divide between North and South. Each circle represents a different town, while the circle size represents the town population. Finally the color or the circle range from red (lowest avg taxable income) to green (highest avg taxable income)
With the link below is possible to access to the interactive map online with more information

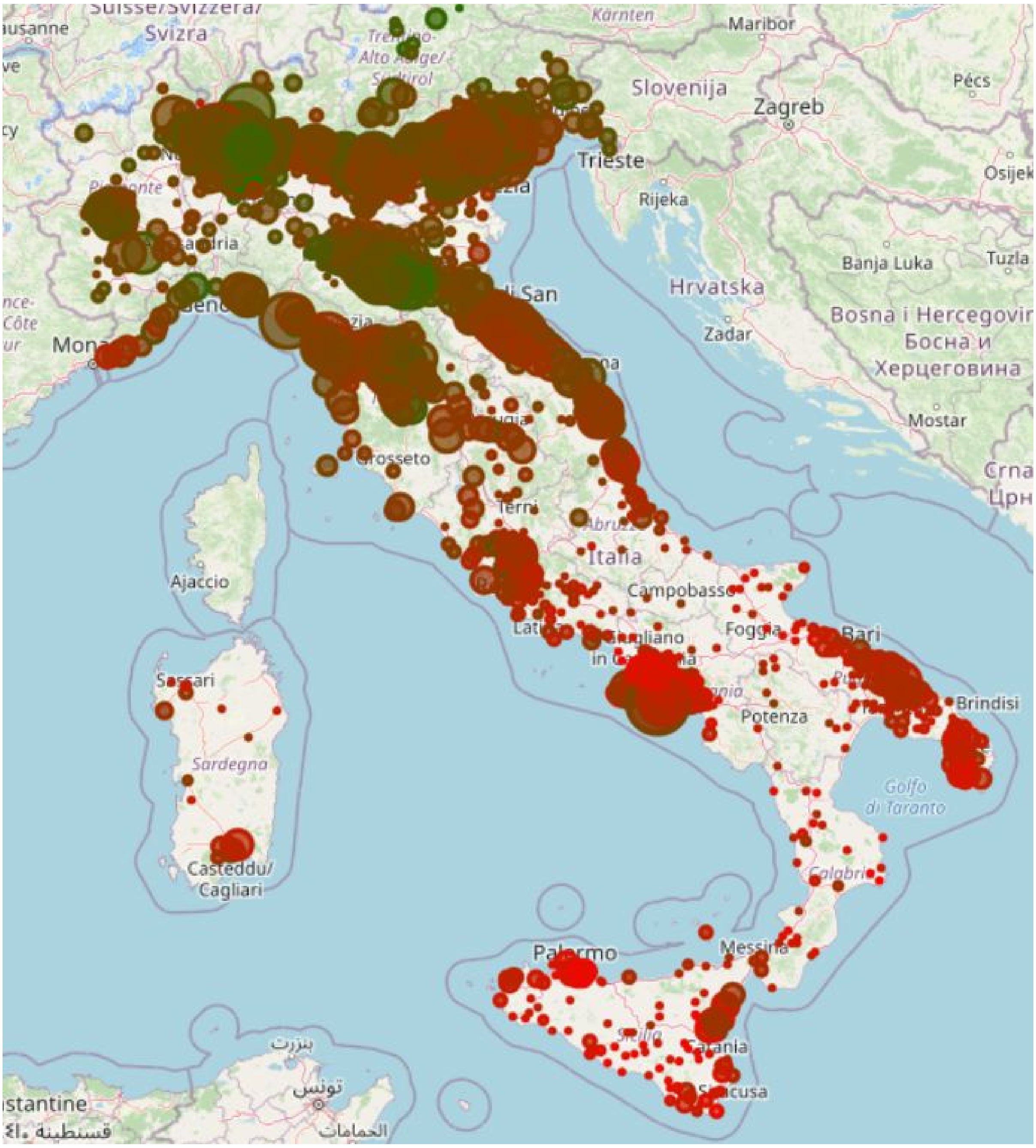
link interactive map



3.6 Avg Income and Foursquare Expensive Venues Geographic Distribution

In the map below instead is it possible to see the distribution of the expensive FOURSQUARE venues. In this map the size of the circle represents the number of expensive Foursquare venues found while the range of color from red to green represent the avg taxable income as in the map before. With this map we can also notice that there is a lack of coverage for the Foursquare venues in Italy, in fact many big cities (Rome, Turin, Palerm, Milan) have a very low number of venues compared to their population size.

link interactive map



(4.) Predictive Modeling

4.1 Model evaluation Metrics

As metrics to evaluate the regression model we have chosen to use R-Squared in order to measure the model performance, and R-square Adjusted in order to test the optimal number of features to us has input of the model. In fact when we add more independent variables or predictors to a regression model, it tends to increase the R-squared value, which could lead to overfitting and can return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much is determined by the addition of independent variables.

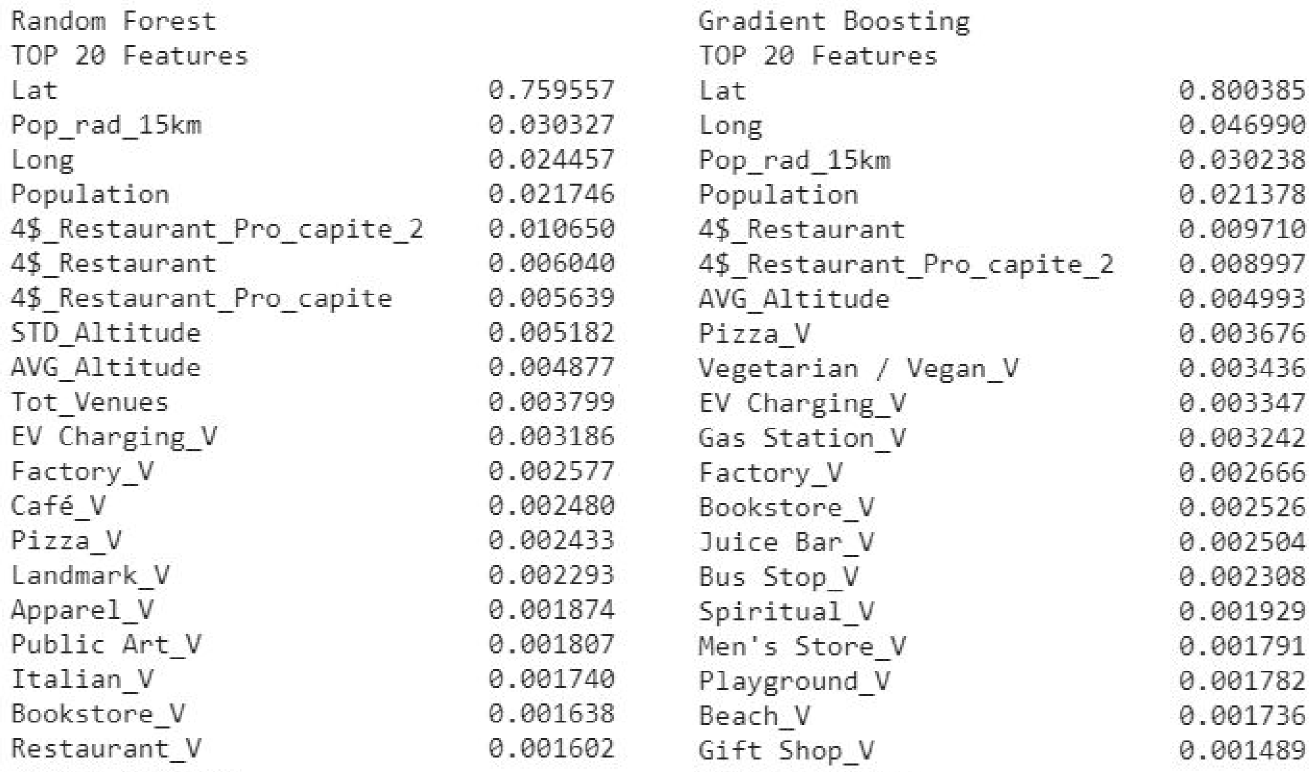
The Adjusted R-squared is calculated with the following formula, where k is the number of features and n is the number of samples:

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1} \right]$$

4.2 Performance Models Compared

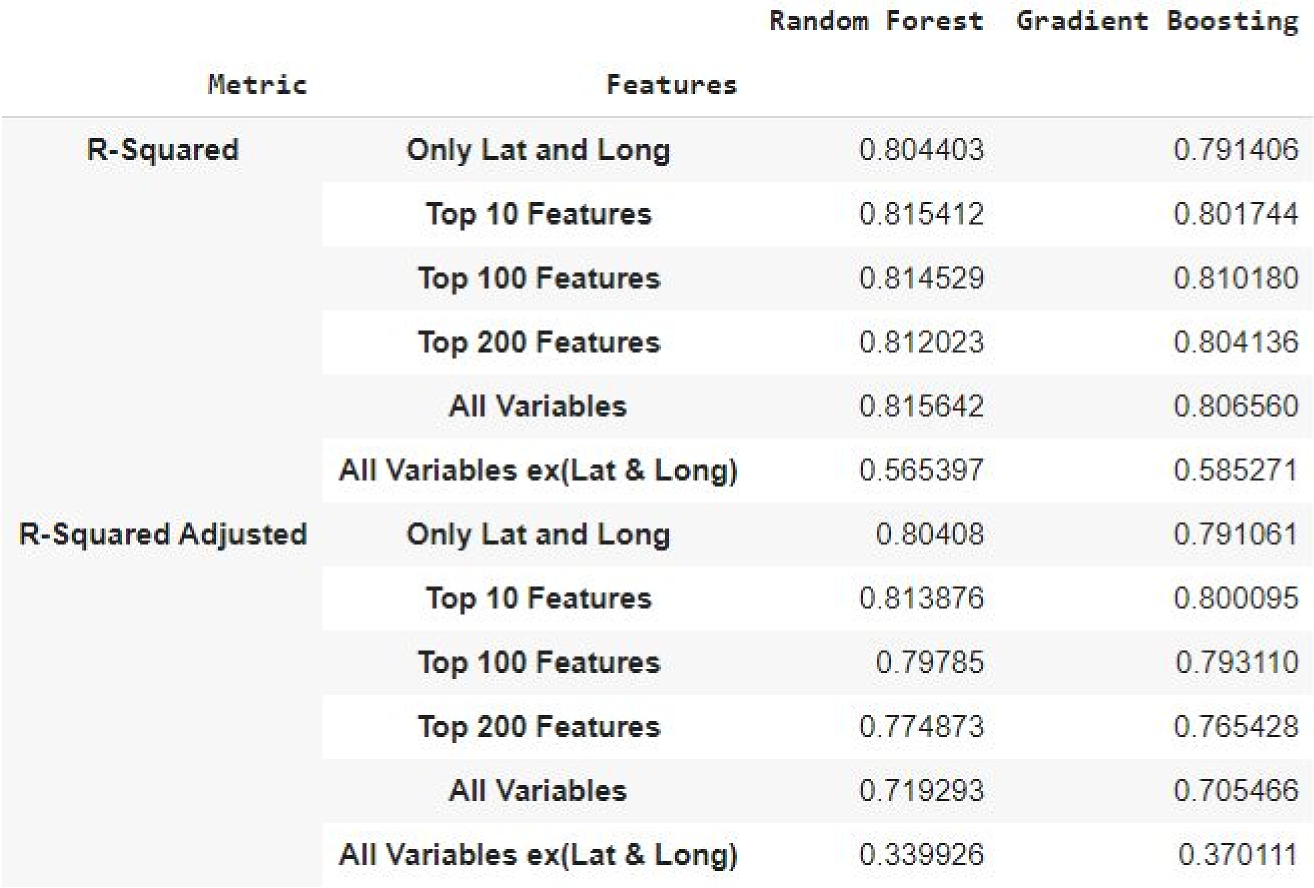
For building the model we have decided to use two ensemble algorithms, which are among the best performers for this type of regression problems. The Machine Learning algorithms tested were Random Forest Regressor and Gradient Boosting Regressor.

All the metrics are the results were calculated as mean of a 10 fold cross-validation test. We proceed with the following steps: (1) we fitted the model with all features in order to calculated the feature importance with the built-in Sklearn method `feature_importances_;` `(2)`we sorted all features by importance; (3) we fitted the model to calculate the metrics with different amount of features as input [top 2, top 10, top 100, top 200, All]. Below you can see the top 20 features for the 2 models:



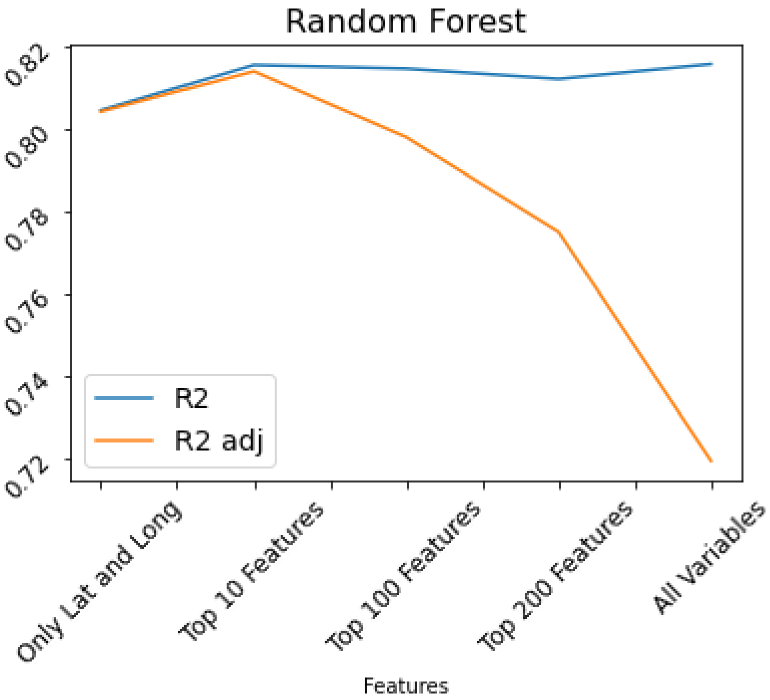
```
Random Forest                                Gradient Boosting
TOP 20 Features                              TOP 20 Features
Lat                              0.759557    Lat                              0.800385
Pop_rad_15km                     0.030327    Long                             0.046990
Long                             0.024457    Pop_rad_15km                     0.030238
Population                       0.021746    Population                       0.021378
4$_Restaurant_Pro_capite_2       0.010650    4$_Restaurant                    0.009710
4$_Restaurant                    0.006040    4$_Restaurant_Pro_capite_2       0.008997
4$_Restaurant_Pro_capite         0.005639    AVG_Altitude                     0.004993
STD_Altitude                     0.005182    Pizza_V                          0.003676
AVG_Altitude                     0.004877    Vegetarian / Vegan_V             0.003436
Tot_Venues                       0.003799    EV Charging_V                    0.003347
EV Charging_V                    0.003186    Gas Station_V                    0.003242
Factory_V                        0.002577    Factory_V                        0.002666
Café_V                           0.002480    Bookstore_V                      0.002526
Pizza_V                          0.002433    Juice Bar_V                      0.002504
Landmark_V                       0.002293    Bus Stop_V                       0.002308
Apparel_V                        0.001874    Spiritual_V                      0.001929
Public Art_V                     0.001807    Men's Store_V                    0.001791
Italian_V                        0.001740    Playground_V                     0.001782
Bookstore_V                      0.001638    Beach_V                          0.001736
Restaurant_V                     0.001602    Gift Shop_V                      0.001489
```

In the table below is shown the performance of the 2 models for different number of features:



		Random Forest	Gradient Boosting
Metric	Features		
R-Squared	Only Lat and Long	0.804403	0.791406
	Top 10 Features	0.815412	0.801744
	Top 100 Features	0.814529	0.810180
	Top 200 Features	0.812023	0.804136
	All Variables	0.815642	0.806560
	All Variables ex(Lat & Long)	0.565397	0.585271
R-Squared Adjusted	Only Lat and Long	0.80408	0.791061
	Top 10 Features	0.813876	0.800095
	Top 100 Features	0.79785	0.793110
	Top 200 Features	0.774873	0.765428
	All Variables	0.719293	0.705466
	All Variables ex(Lat & Long)	0.339926	0.370111

As it is possible to see the Random Forest Regressor is the one that is performing best. Moreover we get the best R2 Adjusted fitting the model with the top 10 features only. If we fit the model with all the variables we would get an higher R-squared but a much lower R-Squared adjusted.
The graph below show the relationship between R-Squared and R-Squared adjusted



(5.) Conclusions

The study has shown that it is possible to predict the average income with a pretty good accuracy, in fact the final model selected has a RSME of 1.465€, which is pretty good considering that the target variable range between 5.000€ and 30.000€.
Moreover we have learned that the best predictor for income per capita in Italy is the Latitude by far of an Italian all the rest of variables add little to the accuracy of the model.
The data of the Foursquare venues in general don't appear to be enough to predict accurately the avg income of an Italian town. In fact if we fit the model only with Foursquare data we can reach at best an RSME of 2.244€. However we cannot deny the fact that Foursquare Venues data could be a very good predictor if the Foursquare coverage of the Italian territory would improve.