DATA SCIENCE CAPSTONE PROJECT

1. INTRODUCTION

1.1 BACKGROUND

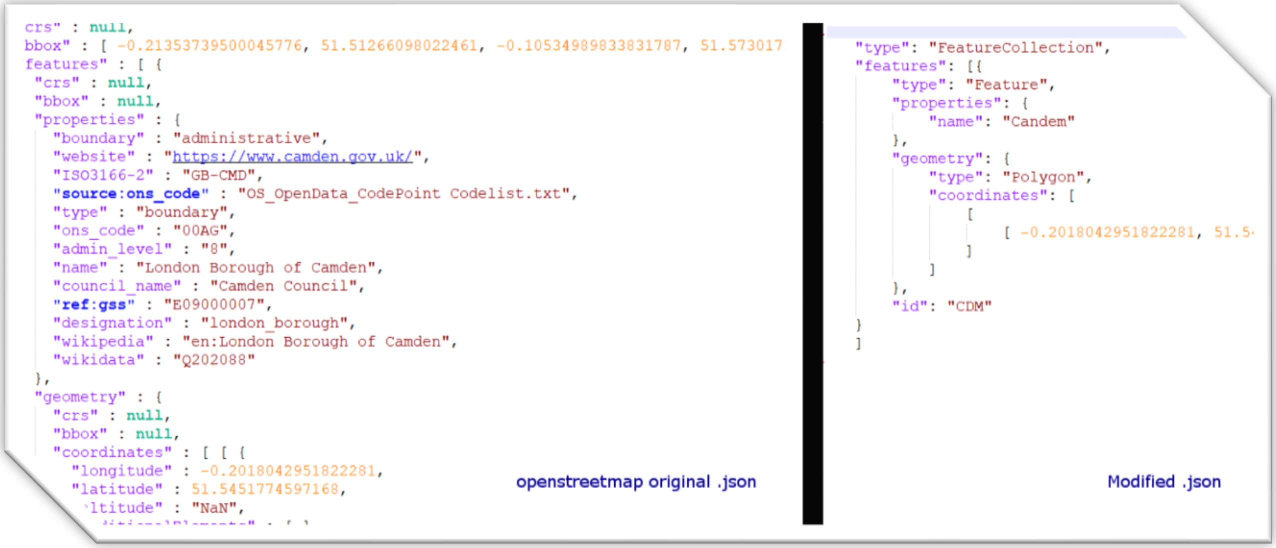
For this capstone project I decided to study a problem similar to the ones we already saw in this data science course. I decided to focus on a particular industry in the city of London with the intention of identifying the best area where to open a new venue. The field I want to analyse for this project is the fitness and wellness centers industry. Fitness centers have a market value of 4.5 billion pounds and employs almost 50000 people across the UK. Business analysis on this market have estimate that this industry is expected to grow due to a general increase in health awareness. Therefore, it is reasonable that new investors in this sector and already established brands desire to know which area is the best for investments. I will try to explain to the stakeholders of a fitness company which are the best locations for growing their business.

1.2 BUNSINESS PROBLEM

In order to identify the best location where to open a new fitness centers in London we have to define which are the most important factors that can influence the industry under investigation. Fitness and welless centers is a type of industry that rely on people willing to dedicate part of their time, effort and money to their physical health/aesthetic. Therefore, it is reasonable that demographic data such as, age, income, population density etc. might gives us some clues and help us resolve this problem. Another important data we should consider is the amount of gyms/fitness centers and their location. Moreover, for the purpose of this project, it is advantageous to find a way to subdivide the vast London expanse in smaller areas in a way that is intuitive and direct. This will facilitate our work and ultimately will help us explain the data the our shareholder

2. DATA

Completing this project means, first of all, finding data related to the problem of choice. I will use fundamentally three sources of information: fourthsquare for venues locations, openstreetmap.org for borough coordinates and the ONS of England (Office for National Statistics) for demographic and economic data. Openstreetmap.org is a valuable opensource collection of the world maps. I will use this resource to find and download all London borough borders coordinates point by point. The .json file provided by openstreetmap.org is going to be little dense of information so I will code a small python script on my local machine to transform all these .json file to a more readable and essential version. Lastly, I will push them to my github repository so they can be easily accessed from my notebook.



```
crs" : null,
bbox" : [ -0.21353739500045776, 51.51266098022461, -0.10534989833831787, 51.573017
features" : [ {
 "crs" : null,
 "bbox" : null,
 "properties" : {
   "boundary" : "administrative",
   "website" : "https://www.camden.gov.uk/",
   "ISO3166-2" : "GB-CMD",
   "source:ons_code" : "OS_OpenData_CodePoint Codelist.txt",
   "type" : "boundary",
   "ons_code" : "00AG",
   "admin_level" : "8",
   "name" : "London Borough of Camden",
   "council_name" : "Camden Council",
   "ref:gss" : "E09000007",
   "designation" : "london_borough",
   "wikipedia" : "en:London Borough of Camden",
   "wikidata" : "Q202088"
 },
 "geometry" : {
   "crs" : null,
   "bbox" : null,
   "coordinates" : [ [ {
     "longitude" : -0.2018042951822281,
     "latitude" : 51.5451774597168,
     ltitude" : "NaN",
```

```
"type": "FeatureCollection",
"features": [{
    "type": "Feature",
    "properties": {
        "name": "Candem"
    },
    "geometry": {
        "type": "Polygon",
        "coordinates": [
            [
                [ -0.2018042951822281, 51.5
            ]
        ]
    },
    "id": "CDM"
}
]
```

Figure 1. Comparison between the original json file and the simplified version

The demographic data I'm going to use will be gathered from the Office for National Statistics (ONS) of England which is an accredited source. I will consider various aspects of the population and I will consider the data per borough. In this way every area will have its own characteristics and will be easy to link them to the fitness industry for comparison. Firstly, the basic information that comes to my mind is to consider the total population per borough. Higher population is reasonably correlated to a higher number of facilities. We could argue that not all people that lives in a borough might want to go to the gym closest to home which might be especially true for those individuals that live close to the borough borders. However, this particular situation should not undermine the usefulness of this data since it is most likely a phenomenon that happens equally across all boroughs hence it's a constant, invariant to our problem.
The data of the total population per borough on its own its not enough. Another interesting parameter to consider is the average age per borough. We can expect that a younger population is more prone to engage in physical activities. Which means that a lower average age indicates a well suited borough for our business. Moreover, the average age data will also integrate productively with the total population data. For example, we could have a borough with a high number of people but with a high average age which will make this borough a little less appealing for our purposes. These two data can work together in defining more correctly boroughs of interest. Another parameter that we could consider is the average income per borough. A borough with a wealthy population is indicative of a more expensive area which can be a double edge sword. Properties and rents will be more expensive meaning that the investment required is going to be higher. However, it is reasonable to think that a wealthier population can afford to pay for more services, therefore making fitness centers a better investment. The population density (population per square km) is another parameter I will consider for this project. This data can be useful in further differentiate the areas. Most of London borough have similar sizes besides some bigger ones that are further towards the outskirts. Population density gives us the relation between population and borough extension. This ratio can be synonym of venues availability and possibly lower costs.
Lastly I will consider fourthsquare data. Fourthsquare is a reliable source of location data from which we can obtain very valuable information. I will make use of this platform to obtain various data from the business venues of interest. Fourthsquare allows us to search for venue type in a given area and can gives us also other venue insights such as: unique ids, geo-coordinates, names, photos, ratings etc. Solving my problem will require the use of only some of them. First of all, latitude and longitude are fundamental for the scope of my project. Mapping the location of all the venues of interest in London will give me one of the most important information; venue density and distribution. I will leverage this data by intertwining it with all the information gathered. Unique id will be useful to differentiate between venues with identical names which will be useful when dealing with bunsiness chains (e.g. Virgin Active, PureGym, The Gym, etc.).

3. METHODOLOGY

3.1. DATA COLLECTION

All the data gathered has been cleaned and organized to fit our needs and combined in one table. The "City of London" borough has been removed from the data due to its small dimensions and residents count. The population data considered is the ONS projection for 2021. The average age per borough was not directly available. However, it was possible to access how many people of any age were present in any given borough. To obtain the average age per borough I summed the age of all the people in every borough and divided it by the borough population. The average weekly income and population density per borough were easily accessible.
Getting all fitness centers location in London with fourthsquare has proven a little troublesome. Firstly I tried to get the venues by taking a radius from the center of every borough, however, I soon realized that this method was nor practical nor would give reliable results. The problems with this method were diverse. For example, the shapes of the boroughs are very complex and definitely not circular. Many areas were left out and there was high venue redundancy. The radius method would have never been able to cover all the London area effectively. Fortunately, fourthsquare allows to search venues without having to use circular areas. In fact it is possible to explore venues also in a quadrilateral area given the north-east and south-west points. Ultimately, I defined a rectangular area that would incorporate all London and subsequently subdivided it in smaller squares. This method solved all our previous problems. In fact, the borough shapes inconvenience has been bypassed, no areas have been left out and venue redundancy has been greatly reduced.

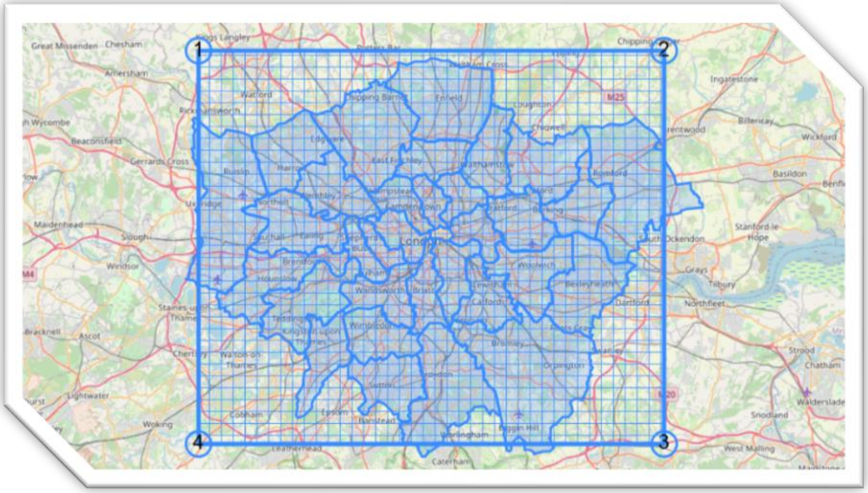


Figure 2. Map generated with folium showing all London boroughs and the reference square.

Fourthsquare was used to access geo-coordinates, unique ids and names of fitness centers in every square. This data has been gathered in one dataset. Subsequently every venue was given a borough based on their location. Gyms located outside any borough boundary have been removed. It is important to keep in mind that the foruthsquare method implemented will search for any venue that is fitness related. This is due to the fact that it was not possible to make a more selective venue exploration. Sometimes, in the fourthsquare database, swimming pools, track and field, yoga centers, etc. are marked as gyms/fitness centers.

3.2. EXPLORATORY DATA ANALYSIS

To understand which are the most valuable data we can use for our prediction we should verify that the data collected is actually meaningful.

3.2.1. Relation between venue number and total population.

Lets observe the relation between total population and number of gyms per borough. We expected to see an increase of venues following an increase population. A simple linear regression applied to the data shows the opposite (Figure 3A). If we try to analyze the data with a more complex method such as polynomial regression we can fit the data a little better. However, the fitting is still very imprecise. This relation shows that the data is very scattered without a strong dependance between number of gyms and total population per borough. Although it's interesting to see that this

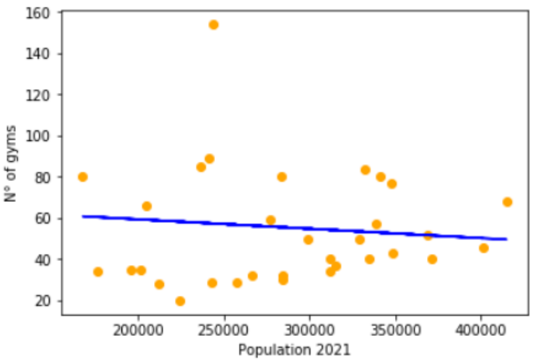


Figure 3A. Scatter plot of number of gyms and population – Simple linear regression

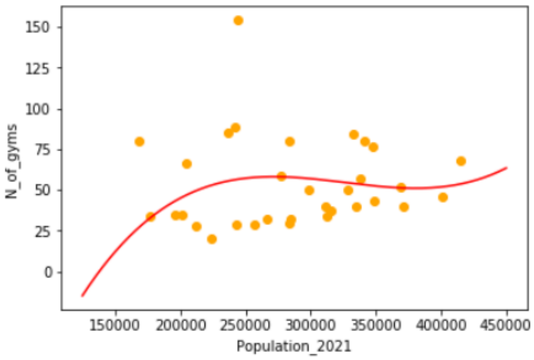


Figure 3B. Scatter plot of number of gyms and population - Polynomial regression

relation is not meaningful this also means that we should probably avoid to use this data in our predictive model.

3.2.2. Relation between venue number and population density.

In our assumptions we though reasonable that a higher population density would be correlated to a higher number of venues. As predicted, we can see that the relation between these two data shows an increase of venue numbers with the increase of population density. Oddly enough, while the previous exploratory analysis on the relation between total population and venue number was unsatisfying, the linear dependance in this case is much better. From this exploratory analysis we deduce that the population density is a more meaningful data then the total population be borough.

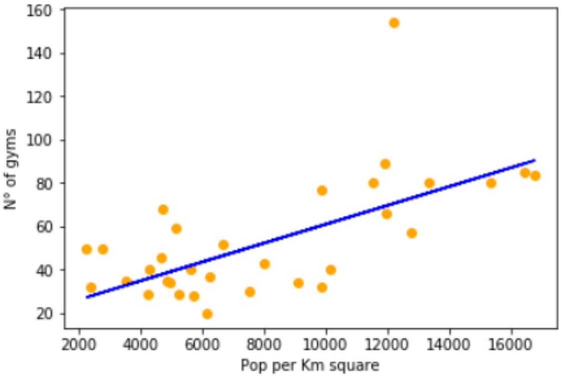


Figure 4. Scatter plot of number of gyms and population density

3.2.3. Relation between venue number and average income.

Surprisingly, the correlation between average income and venue numbers is robust (Figure5A-B). The higher the weekly salary the higher is the number of gyms in that borough. This means that our assumption that wealthier population can afford to pay for more services is valid. At the extremities there are: the lowest income borough (under 600 pounds weekly) with the lowest number of gyms and the higher income (more then 900 pounds weekly) with a number of gyms in the top 25%. In broughs where the income is in between these values the number of gyms tend to increase with an increasing income. From this data we can take that it is more advantageous to invest in boroughs were the venues prices are higher and the population richer. In figures 5A and 5B we can see the fitting of the data with a simple linear regression and polynomial regression respectively. The relation between the data do not show a perfect linear dependence but as the polynomial regression suggest it is characterized better by an exponential behaviour.

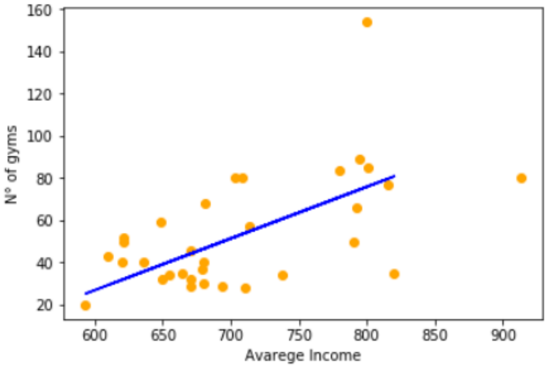


Figure 5A. Scatter plot of number of gyms and average weekly income – Simple linear regression.

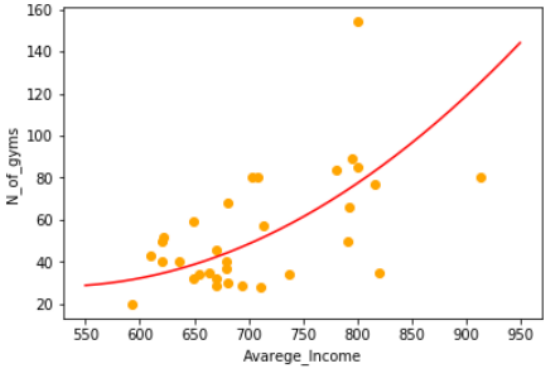


Figure 5B. Scatter plot of number of gyms and average weekly income - Polynomial regression.

3.2.4. Relation between venue number and average age.

To build our model we supposed that a younger population should be more willing to engage in physical activities then older ones. From this simple exploratory analysis on the relation between number of gyms and average age per borough it is possible to notice that our assumption is generally true. Boroughs were the population is older witness a decline of number of fitness centers. Figure 6A shows the data fitting with a simple linear regression showing a progressive decline as population ages. Figure 6B shows a more precise fitting with a polynomial regression that shows a rapid decrease of venues numbers past the average age of 36 years.

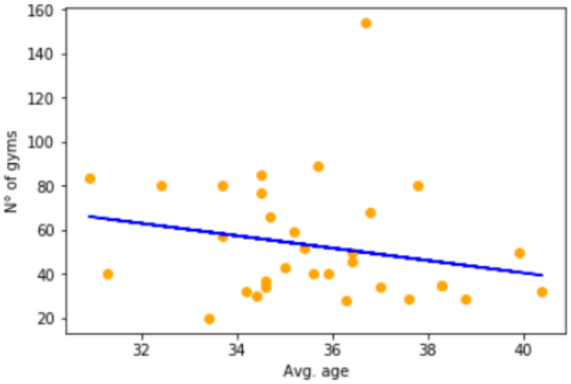


Figure 6A. Scatter plot of numbers of gyms and average age - Linear simple regression

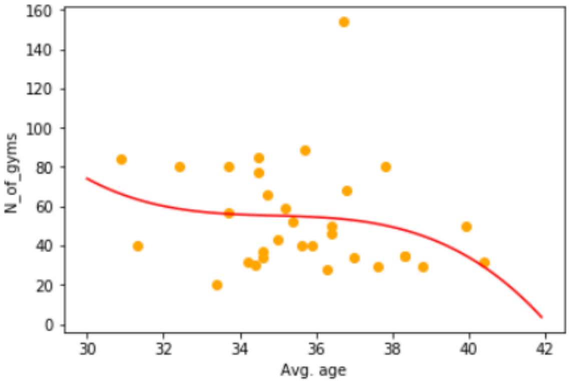


Figure 6B. Scatter plot of numbers of gyms and average age - Linear polynomial regression

3.2.5. Relation between venue number and borough extensions.

Taking a look at the relation between borough extensions and number of gyms we find an interesting but reasonable situation. Borough with higher extensions tend to have less gyms then the smaller ones. The smaller boroughs are located towards the center and more populated London areas. As we move farther away from the crowded center towards the outskirt we find bigger boroughs where the venue numbers are getting smaller. This means that the borough extension is inversely proportional to the number of gyms. Actually, one might argue that to consider this data is almost equivalent to considering the population density. However, taking a look at data distributions the relations between these two data and the number of venues is not inverted and identical. In fact a more thorough analysis with a polynomial regression shows a parabolic behaviour.

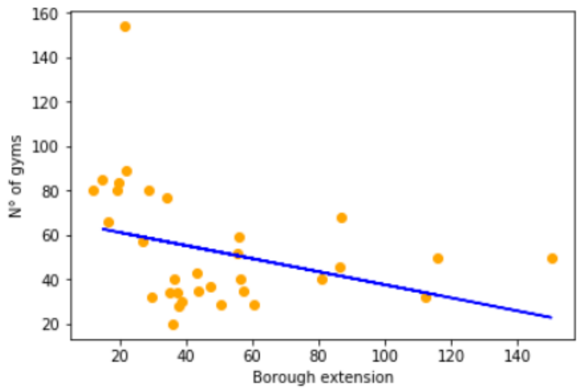


Figure 7A. Scatter plot of numbers of gyms and borough extensions - Linear simple regression

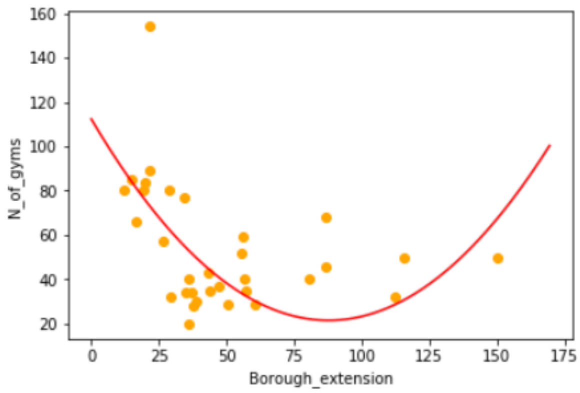


Figure 7B. Scatter plot of numbers of gyms and borough extensions - Polynomial regression

3.2.6. Relation between number of venues and boroughs.

It is useful to make use of a visual aid to understand the actual distribution of fitness/centers in the city of London. For this purpose I made a choropleth map of the London area (Figure 8). I want to address the Westminster borough and its oddly high number of venues. In this borough there are 154 fitness related venues. This data looked a little suspicious and made me second guess the method I implemented with fourthsquare to search the venues. After a thorough review of the code I could not find any error. However, as previously said, it must be kept in mind that the implemented fourthsquare method will gather all venues that are fitness related, not exclusively gyms but also track fields, sometimes swimming pools, yoga centers, etc. All those venues that had the keyword gyms and fitness center in the foruthsquare platform. I tried to be more selective but I could find a solution.

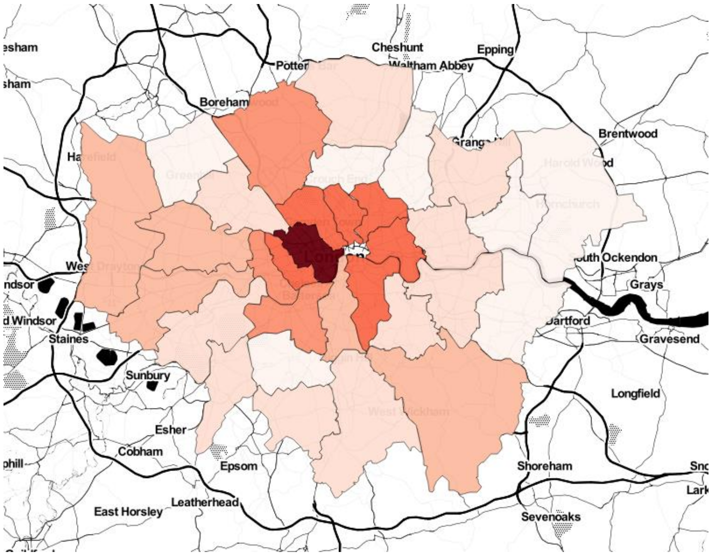


Figure 8. Choropleth map showing venue density in each borough

4. RESULTS

4.1. PREDICTIVE MODELLING

I will apply two difference predictive modellings using only the most reliable data. Firstly, I will use the multiple linear regression model. Multiple linear regression (MLR), is a statistical method that is able to predict an outcome taking into account multiple variables. The multiple linear regression model tries to find a linear relation between the independent and dependent variables. However, we should not expect a good result from this model since the relation between the dependent variables and the independent variables must be linear which, in our situation, is not really the case. Subsequently I will apply the Random Forest Regression. This method performs both regression and classification tasks. The Random Forest use multiple decision trees where each tree is trained with a different data sample. The final goal of this method is to make a prediction combining many decision trees instead of one. I chose the Random Forest Regression model fundamentally for two main reasons. Firstly, the relation in my data between the independent and dependent variables is not linear. Secondly, the choice of subdividing the analyzed area in borough was practical but left me with only 32 entries in my dataset, which leave little room for model training. Having a model that can be train with many data samples build by replacement will partially bypass my lack of entries. It is reasonable to expect a better result using the Random Forest Regression compared with the MLR method. As previously said, using multiple linear regression is probably a poor choice, however it will be interesting to see if this is actually the case and to see the differences with a better model.

4.1.1. Multiple linear regression

The multiple linear regression will be build starting from tree of the more meaningful data. This data where chosen based of the exploratory analysis and from testing. The data considered in this method are: population density, average income and borough extensions.

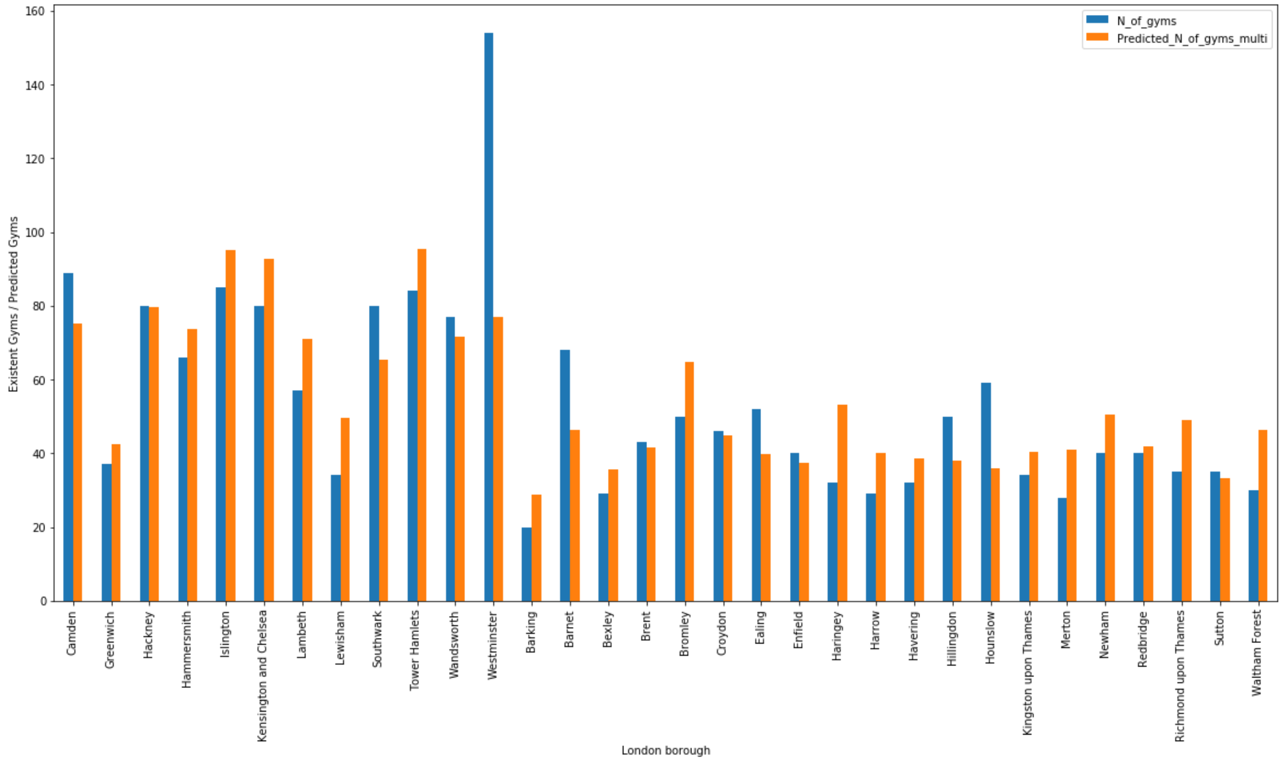


Figure 9. Predicted number of gyms with MLR method

This method has proven to be very inconsistent. In the barchart above the variance score is 74% which could be said to be decent. However, the variance score differ widely depending on the data entries used to train the model. This result is not replicable. The model is failing to make a reliable prediction. This is due, as said before, to a dataset too small to train the model and the missing linear dependence between independent and dependent variables.

4.1.2. Random Forest Regression

This model will be build starting from the same data used also in the multiple linear regression model. The number of estimators (number of classification trees considered) set to 60 and a max depth (maximum depth of a tree) set to default.

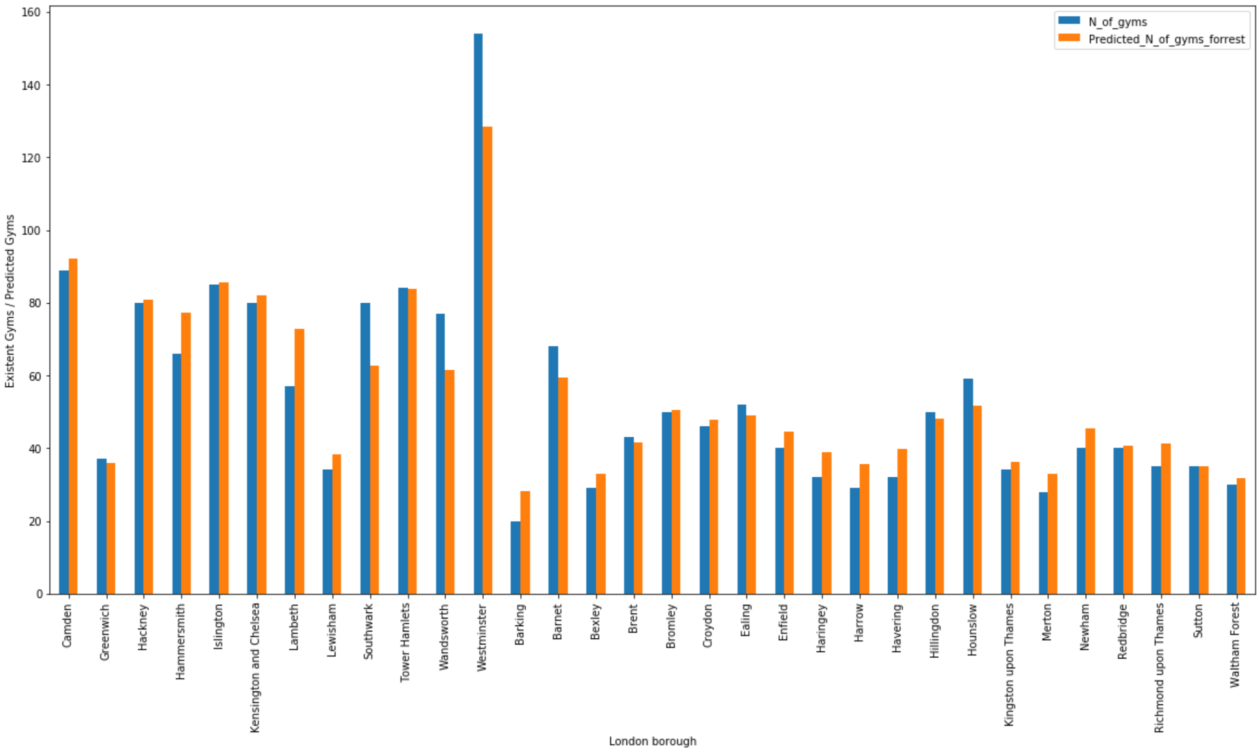


Figure 10. Predicted number of gyms with the Random forest regression method

This model is much more solid and reliable. The variance is always above 90%. In contrary to the MLR method this model gives replicable results. The barchart above shows the number of gyms that should be present in each borough taking into account the data of population density, average income and borough extension. The chart below (Figure 11) provides a visual understanding of this model prediction. In figure 11 is showed the variation between the actual number of gyms and the gyms predicted by the model. The green values indicate that in that specific borough there should be more fitness centers, therefore indicating an area suitable for investors. The red values indicate borough with a oversaturation of fitness centers thus making it a less appealing area for new investors. Figure 12 is showing a choropleth map of the London area subdivided by borough where the variation data has been used to locate the best borough for our business. The darker the color the more appealing the borough is. The best three boroughs for our purpose predicted by this model, in descending order are: Lambeth, Hammersmith and Barking.

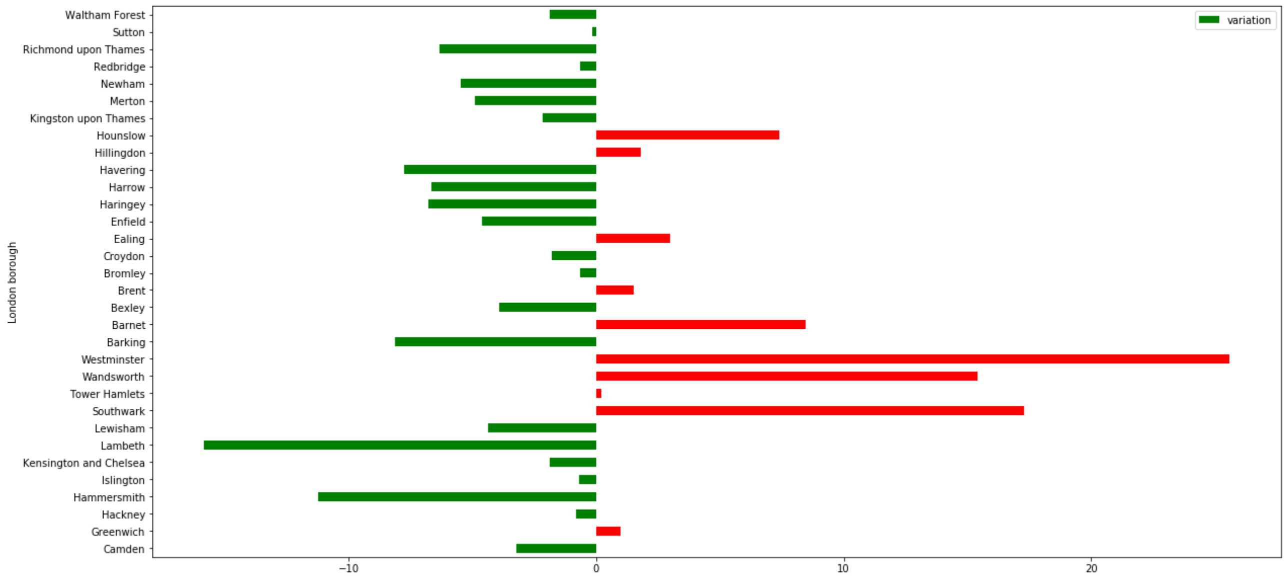


Figure 11. Variation chart between the actual number of gyms and the predicted number by the Random forest regression model

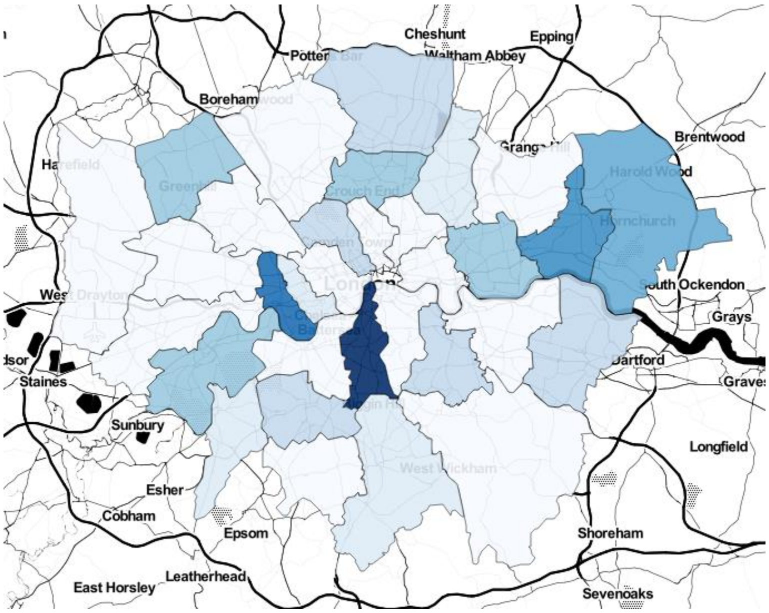


Figure 12.Choropleth map of the variation data.

5. DISCUSSION & RECOMMENDATIONS

Solving this problem has proven more difficult then expected for various reasons. Some of the data considered for this project were not as meaningful as I had assumed. The exploratory data analysis showed that the total population and the average age data was weakly related with the number of venues. However, the rest of the data was more useful. Another data related problem was due to the choice of working with the London boroughs as area of interest. This was a practical choice because it was an intuitive way to subdivide London, it was easy to find demographic and economic data for such areas and allowed an easy visualization of the results. However, on the other hand, it limited our dataset size greatly making the training of predictive models very difficult and sometimes impossible. Moreover, exploring all the venues in the vast London expanse was challenging and required a series of tries and observations before coming up with a working method. The fourthsquare api was not as venue selective as I would have liked. I could not find a way to search for venues that where purely fitness centers. Instead, every venue that responded to the keyword gym/fitness-center has been taken into account. It is not uncommon to see swimming pools, tennis courts and yoga centers into the list of explored venues. For a more thorough study of this problem I recommend consider how the demographic and economic data variates in time. In this way it is possible to get an evaluation on the borough growth and would increase the data pool considerably. Other data that might be meaningful to consider is venue popularity and rating.

6. CONCLUSION

In this report we analyzed London demographic and economic data to make a prediction on which is the best location for investment for the fitness industry. We considered data such as: population density, average income and average age, etc. The predictive models used were Multiple Linear Regression and Random Forest Regression. The MLR approach didn’t give us reliable results since there is not a sufficient linear dependence between the considered independent a dependent variables. The Random Forest approach gave much more meaningful results with a variance score of 90%. This model predicted that the best boroughs for investments in the fitness industry are Lambeth, Hammersmith and Barking while in boroughs such as Westminster and Southwark the fitness market is very competitive due to excessive amount of fitness centers.