BEST PLACE TO LIVE IN UK

Data Diggers

Yifang Zhou
University Of Southampton
Foundations of Data Science
yz5f17@soton.ac.uk

Richa Ranjan
University Of Southampton
Foundations of Data Science
rr2n17@soton.ac.uk

Anton Okhotnikov
University Of Southampton
Foundations of Data Science
ao2u17@soton.ac.uk

Shihuai Wang
University Of Southampton
Foundations of Data Science
sw1g17@soton.ac.uk

Daniel Pelnar
University Of Southampton
Foundations of Data Science
dp4n17@soton.ac.uk

ABSTRACT

The driving force for choosing this topic was the fact that all the members of this team have newly moved to UK. It was an interesting insight for us to be able to analyze the liveability of a city. This project aims at finding the most preferred city based on chosen factors that most people find important when moving to a new city. There are two main tasks that we perform. First, the factors influencing the liveability quotient of a city are analyzed, with the given weights and similar factors are grouped. Secondly, through an interactive application, the most liveable cities are recommended to the users based on their preferences.
Application link: https://github.com/zyfzjsc988/data_diggers

INTRODUCTION

Moving to a new city is always challenging! UK has always been a preferred choice for a wide population in all respects, may it be for higher studies, or work, or relocating in general. We, as a group of students who have moved to Southampton recently, came up with an analysis that helps people to decide which is the most liveable city in the United Kingdom. In order to find the most liveable city in the UK, we first needed to define a model with all the factors that the majority of people find important. After going through various researches and conducting an online survey, we came up with twelve factors. Then we found data for these factors. We managed to get about 84 UK cities as the number of observations. The database of our choice was MongoDB where most of the cleaning and merging took place, using Python, and also, manually. After performing analytical operations, we group similar factors. Another product of our work is an interactive online application

which takes inputs like nationality, age, factor preferences of the users, and suggests the best and the worst cities to live based on those inputs.

MODEL

0.1 DATA COLLECTION

For various factors that determine the liveability quotient of a city, related datasets are required. Various government websites provide data for these factors with respect to the cities in UK. The data has been collected from the following sources:

- UK Government website
- Kaggle.com
- National Health Survey

and some more. These datasets provided the data for various cities in UK. But to understand the weights of each factors affecting the liveability, as well as to ensure that the factors collected are realistic, we additionally relied on the following two sources:

- **The Happiness Ranking dataset:** This is the dataset from the UK National Statistics website, which provides the happiness levels of people living in UK. The data explains people's happiness level, by personal characteristics.
- **Survey:** An online survey was conducted where people were asked to rank each factor from 1 to 10 (10 being the most important), based on their preferences when moving to a new city. The factors are: Flat/House prices, Population, Road Traffic, GVA per worker, Unemployment rate, Noise level at night, Total Jobs available, Weather, Number of schools, Entertainment, Quality of higher education, Number of hospitals, Connectivity to other cities. Being aware of ambiguity and vagueness of some of these factors, we provided a short explanation for GVA per worker and Entertainment in order to help them make a more informed decision. Additionally, Age, Gender, Home Region and Employment Status were collected as well. This survey was conducted with a purpose of finding what factors do people look for, when they move to a new city. To this, we collected 123 responses are here are the results of this survey:

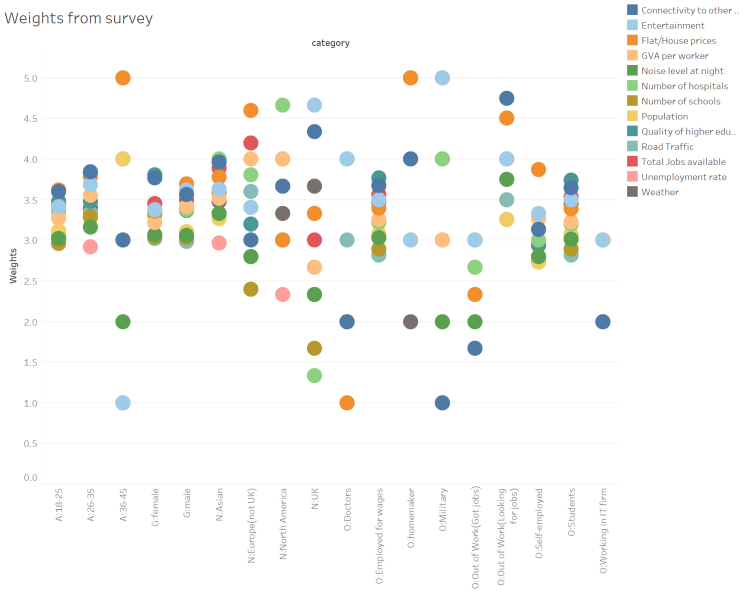

Figure 1: Responses from the Survey

0.2 DATA CLEANSING

Following are the cleansing processes we followed:

- **Consistency:** For the datasets collected from various sources, consistency was a major concern. To tackle this problem, we compared all the datasets to confirm if data were available for all the cities, across all factors. The cities which had no data across all the factors were omitted. As a result, the final count of cities ready for processing was narrowed down to 84.
- **Structure:** The datasets collected were all in different formats, like Excel, CSV, JSON, XML. So, the data structures were different, and also, there were additional wild card characters. So for basic sanitization, we used Python codes to clean all these datasets.
- **Missing Values:** Some of the city data had missing values for some factors. As we had to

perform mathematical calculations, we used average values for that field, or zero, in some cases, as required by the calculations.

0.3 DATA NORMALIZATION

Normalizing the data was necessary to have a uniform scale for our calculations. For the datasets obtained after performing general formatting/cleaning, the normalization process was carried out. The normalization techniques used were Z-score and Min-Max. In some cases, the cities had no numerical data. For instance, when we considered Medical facilities as one of the important factors, the dataset to work with was 'The Number of Hospitals each city has'. There were no numerical values, so we grouped them on the basis of similar cities and then normalized the data. We used the **Z-score** method for normalizing these datasets. Z-score was our choice of technique because we had some outliers. The outliers would inflate the results from Min-Max normalization method. Thus, we proceeded with Z-score.

0.4 DATA MANAGEMENT

The final datasets have been transformed into a different format such as a table-like structure, or a JSON file. Mainly, we ended up with two file formats, either an excel file, or a JSON file. Excel file was mostly the ones with numerical data that can be quantified. In case of JSON files, we had data that resulted from normalizing datasets like Hospitals, Pubs etc., as they had no numerical values. For consistent storage on the database, we used MongoDB, and stored all these files in the form of collections.

0.5 DATA ANALYSIS

This section is the main part of the project. These are analyses that have been done to show the importance of each factor, which is the analysis of relationship between happiness levels and factors (such as weather conditions and others).

The analysis of relationship between happiness and all the factors

In this analysis problem, the question is what are the most important factors that affect citiesâĂŹ livability and happiness level. The solving process could be abstracted as the feature selection[**?**] process in which all factors that needed be chosen are independent variables and the livability which is measured by the happiness rating of cities is independent variable. The function that was chosen at first for feature selection was *LASSO*[**?**]. However, we used Random Forest feature selection at last.

The process of experiments

0.5.1 Normalization. Before the training process, the feature matrix needs to be normalized by z-score normalization function as mentioned before.

0.5.2 Feature selection. There are lots of models to select features - *LASSO* is the most common one. We first tested *LASSO* and found out that it is not good enough for our project. Finally, we changed to the random forest regression model.

LASSO

$$Y = c_0 + c_1x_1 + c_2x_2 + c_3x_3 + \ldots + c_nx_n + \lambda \sum_{i=1}^{n} |x_i|$$

Because the training dataset is small in our project, we use cross validation to decrease the error of experiment. The evaluation function is mean sum of square (MSE) and we have tried to pick up the best λ (from 0.0001 to 0.1) whose result has the minimum MSE, to do the feature selection.
The result of using *LASSO*:

- The result of this model is swinging. Every testing might have different best λ and the difference between them is quit large.
- Most of the time, the minimum MSE comes from the result with all features.

Analysis of result:
Based on the result described above, we can confirm that this model is not suitable for the feature selection of our project because in most cases all the features are selected. This is not the result we expect. The reason might be relative to data set itself and the model we create. For example,

- The difference of the data in some features might be small. Some features have been measured according to different regions rather than cities. We deal with this problem by putting the same data to all the cities from same region. Therefore, some rows of training data have no much difference which might cause worse fitting.
- Another reason is that the relationship between independent variance and dependent variance is not linear. When using *LASSO*, we use the linear regression which cannot work with non-linear problem.

The result of *LASSO* is also not what we expected. For example, it can only output the weight for each factor which is not showing the importance ranking for all the factors. It is because the two of these factors might be correlated.
This supposition has been confirmed by measuring the *Pearson's Correlation Coefficient* between two factors. For example, some factors have strong positive correlation with population, seen in **Table.1**.

Factors	Pearson correlation coefficient
Number of Schools	0.81
Number of Hospitals	0.73
Number of Stations	0.69
Number of Universities	0.66

Table 1: The correlation coefficient between factors to population

In summary, *LASSO* could not be classified as a good method to do the feature selection in our model and we should care about the relationship between two factors.

Random Forest[?][?]
There are three reasons why the Random forest is more suitable for our feature selection.

- The first reason is that it is not limited to dealing with linear problems, it is also suitable for non-linear problems.
- The result of random forest is a set of scores for factors, these scores represent how important of this factor is among all factors, to reach the maximum accuracy.
- It ignores the correlation between two factors when we calculate the importance score.

Therefore, *Random Forest* might be more suitable and has been chosen to be the algorithm for our model.

0.6 APPLICATION

After performing the analysis, a web application was built, where individual users can interact with the application, and based on the inputs (age, gender etc.) entered by the user, the most preferable city would be recommended. To do this, we used the **Flask** framework for Python server part and HTML and CSS for static client part. The structure of the application can be described as:

(1) **User Input:** Users can interact with the application by filling in the form their details like age, gender, factors' preferences etc.. Based on the user's inputs server performs calculations.
(2) **Calculations:** The algorithm for recommendations is based on the analysis carried out by our model and represents a weighted sum of factors input by the user.
(3) **Results:** The results rendered by the application includes three most suitable cities and the three worst cities to live in, based on the user's inputs.
(4) **Survey result-based recommendations:** Second part of recommendations comes up from the survey results calculated before. For example, if a user enters a certain age, gender, location, etc., the application gets the factor's

importance from the survey report and calculates new recommendations of cities for user according to input. This part of recommendations represents the main trend for people of similar gender, age, location, etc. that came from the survey.

the application results look like what is shown in figure 2:

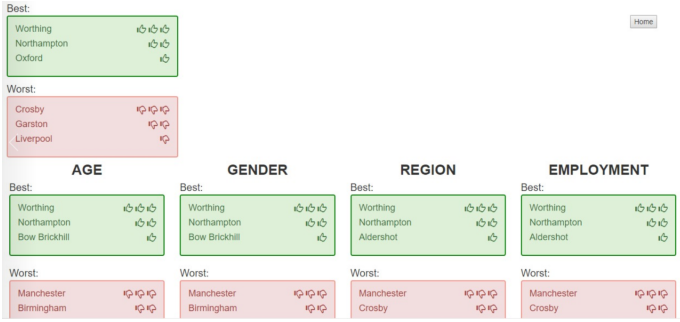

Figure 2: Responses from the Survey

CONCLUSION / RESULTS

As mentioned under the Application section, the result obtained will be two different recommendations, one based on our analysis, and the other one based on similar parameters from the survey conducted before. Using **Random forest**, we estimated all 26 coefficients of which 13 are the main factors and 13 are sub-factors of those main factors. Figure 3 provides a list of all 26 factors and sub-factors. For each sub-factor, the main factor is indicated in brackets.

Also, the factors and their weights can be observed as a result of the Random Forest algorithm, in the figure 4. The result from the Random Forest yields that the six most important factors and sub-factors are road traffic for 2016, stations number, hospitals number, hospitals number per person, schools number and unemployment. The factors and the results are discussed below:

- **Weather:** This result was very surprising because it does not include the factors 'Weather' or any of the five sub-factors. In our opinion,

'SummerDay_average_temperature', (Weather)
'SummerNight_average_temperature', (Weather)
'Sunshine_per_Month', (Weather)
'road_traffic_2015', (Road Traffic)
'traffic_noise', (Noise level at night)
'school_number_per_person', (number of schools)
'FrostDay_perYear', (Weather)
'pubs_number', (Entertainment)
'total_jobs_per_person', (Total jobs)
'number_of_universities', (Quality of higher education)
'total_jobs',
'Rainfall_per_Month', (Weather)
'stations_number_per_person', (Connectivity to other cities)
'house_price',
'GVA',
'number_of_universities_per_person', (number of schools)
'WinterNight_average_temperature', (Weather)
'population',
'WinterDay_average_temperature',
'pubs_number_per_person', (Entertainment)
'unemployment',
'school_number',
'hospitals_number_per_person', (number of hospitals)
'hospitals_number',
'stations_number', (Connectivity to other cities)
'road_traffic_2016'

Figure 3: Factors list

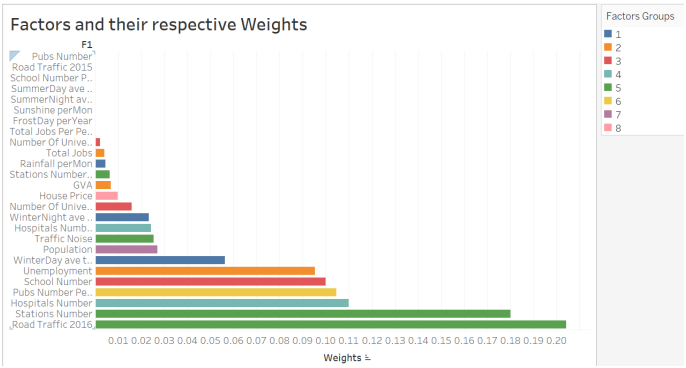

Figure 4: Factors and their Weights

it might be due to the fact that for all of 84 cities in our sample, there are only 16 places in the UK where weather related factors of our interest are measured.

- **Road Traffic:** The first very important thing to mention here is that Road traffic for 2016 were found to have the highest magnitude in our Random Forest Analysis, but at the same

time Road Traffic for 2015 was found to have one of the lowest magnitudes. This paradox is difficult to explain. Our suggestion is that since more and more people become an owner of a car each year, the intuition is that from 2015 to 2016, the amount of new drivers (and as a result of higher road traffic) increased so dramatically that the factor became highly significant. It would be very interesting to find a data for year 2017 to see how this trend is progressing and whether our intuition is correct.

- **Number of Stations:** Number of Station is a proxy for Connectivity to other cities and unfortunately there is no relevant study that would investigate connectivity to other cities and the amount of happiness (well-being). Nevertheless, as in the previous case, a common sense and intuition can be used here to support our findings that indeed number of stations is correlated with happiness of the people living in that particular city.
- **Hospitals number and hospitals number per person:** The random forest analyses concluded them both important with 3rd and 4th highest magnitude of all 26 factors. "Number of hospital" ended up more important by a decent margin (0,026). There might be a lot of reasons for it but one of it may be that people most likely value more âĂIJhigher-qualityâĂİ hospitals over the quantity of hospitals available in a city.

LIMITATIONS AND FUTURE WORK

Some of the limitations observed in our model are as follows:

(1) **Limited Datasets:** Most of the datasets are concentrated during the years 2015 and 2016. Since this is only 1-2 years ago, the data collected on our factors should still be representative to a good extent of what is happing in 2017/2018. However, the interactive application can become outdated in the next 10 years as the factors can change much more due to the slowly and ever changing human culture and what things(factors) people value.

(2) **Small Data Sample:** There were insufficient data for all factors across all cities. Therefore, the list had to be narrowed down to 84 cities only. This might cause a problem with statistical inferences as we do not have consistent estimators asymptotically and testing (F and T test) is out of question as well. Also, the survey conducted had 123 responses in total, which is not a very significant value in terms of population.

(3) **Weight of Factors not quantifiable:** While building the application, the user's input requires them to rank factors based on their preferences, but some of the factors cannot be quantified. For example, the "weather" can be ranked as 5(the most important factor) by the user, but it cannot be justified as to what kind of weather is preferred. There are some more examples like that in the application.

(4) **No Rules for defining Weights:** There is no standard rule for defining weights of the factors. As a result, we created our own ranking and used it to perform the analysis on those factors.

REFERENCES