Prediction Of Life Quality

Tark Ramazan BASOGLU
Department of Computer Engineering
Hacettepe University, Ankara, TURKEY
tarik.basoglu@hacettepe.edu.tr

Emre DOGAN
Department of Computer Engineering
Hacettepe University, Ankara, TURKEY
emre.dogan@hacettepe.edu.tr

Abstract

In this study, we mention about the usage of using a machine learning approach to specify life qualities of cities instead of public research. We create an assorted dataset that contains statistical and physical features. To do that, we utilize from MAPZEN. We expect to predict the scores on MOVEHUB with high accuracy.

1. Introduction

Nowadays, we can easily see that cities differ considerably from each other in terms of their physical and social characteristics and that difference is highly influential in human life. We are making great efforts to determine the effects of these differences on human life and to make cities more livable and to change this imbalance positively.

In this situation, we are faced with a notion named quality of life.

"Quality of life (QOL) is the general well-being of individuals and societies, outlining negative and positive features of life. It observes life satisfaction, including everything from physical health, family, education, employment, wealth, religious beliefs, finance and the environment."[2]

By this definition, there are various social and physical criteria that influence the quality of life. The number of researchs and studies carried out in this area is increasing day by day. While life quality information for large cities is easily accessible, it is not possible to find reliable results for cities that are not big enough.

In this project, we purpose to achieve higher efficiency in shorter time and reduce the burden on a human in such researches. Rather the laborious and time-consuming processes of public researches we also aim to provide a new, flexible and developable method by making use of machine learning experiences. Thus, we get a chance to detect the life qualities for any cities in the world. At the same time, we are expecting to be able to observe which physical factors effects the life quality with which rates.

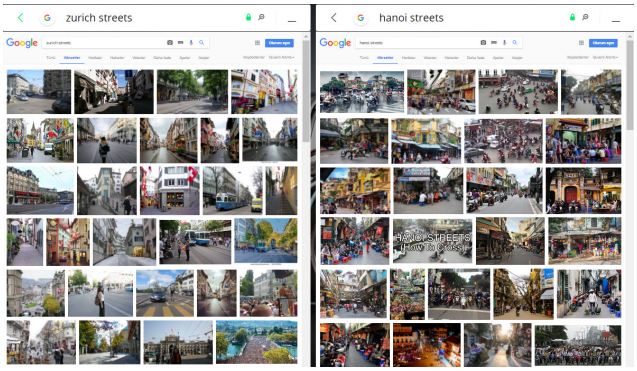

Figure 1. The reflection of the crowd difference between the Hanoi and Zrich on the street photos

MoVEHUB
There is a platform named MOVEHUB that helps you make informed decisions about where to move to around the world. And it has a city ranking list consists of over 200 cities. We utilized this list as the main target in the estimation results.

MAPZEN
Mapzen is an open and accessible mapping platform that is focused on the core components of geo platforms, including search, rendering, navigation, and data.

2. Related Work

There are numberless researchs done to measure life quality in cities every year. In this researches generally, lots of criteria are considered to obtain correct results. Such researches have been carried out in the form of public opinion polls up to now.
MOVEHUB: MOVEHUB is similar research that includes

quality of life score in over 200 cities. It also has a purchase power, health care, pollution, move hub rating and crime rating on different scopes. We want to predict that life quality scores from this platform.

NUMBEO: Also It has a Quality of Life ranking that contains much features aspect of life. The estimation score is calculated by using an empirical formula that consists of similar features as a parameter with MOVEHUB like purchasing power, pollution, cost of living. It can be said that it contains more comprehensive statistics.

Movehub City Rankings: There is a kernel to predict MOVEHUB rating score with other parameters on MOVEHUB done by Jonathan Bouchet on KAGGLE[4]. In this study, linear regression has been used as a kernel function. Quality of Life and City Competitiveness[5] is research about life quality criteria and they help to identify key metrics.

Predicting Sentiment about Places of Living: In this thesis, a ranking list of cities of MERCER is taken using. As a result of the quality of living survey 230 cities of the world are ranked on the list. This study is based on binary classification by using microblogs on Twitter to rank the cities. A classifier with a strong baseline for predicting sentiment about places of living is trained using logistic regression model.It would be correct to say that our work is a combination of other works. We purpose to get life quality scores like Movehub City Rangkings but we try to collect own dataset and use more varied features.

3. The Approach

3.1. Collecting Dataset

In this section, the specification of a dataset that we created, the process of collecting data from different resources, the analysis of data and the problems we met during the project are mentioned

We aim to predict the quality of life scores of cities on MOVEHUB. MOVEHUB is a company that makes researches about cities and it has a research that includes different metrics, cities, and scores. Quality of life is one of these metrics. Dataset has been downloaded from KAGGLE with CSV format.

There are also different researches about the cities quality of life, like NUMBEO and MERCER.Additionally, life quality scores are different for different researches. In the figure, you can see different researches and lists. We chose MOVEHUB because its dataset is downloadable and it included more cities than others. In MOVEHUB 216 cities from different countries and regions are evaluated scientifically with using different sources like NUMBEO, CIA WORLD FACTBOOK, census from several governments and WHO.

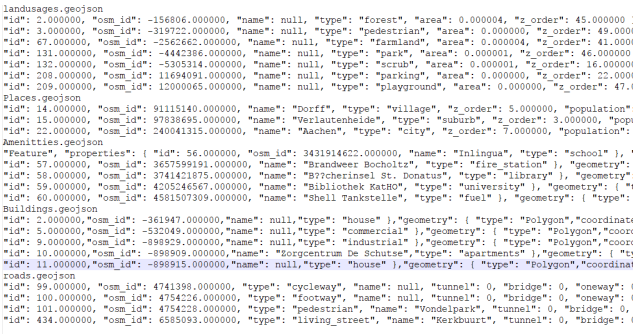

Figure 2. Contents of geojson files.

Our second data set is MAPZEN.It includes datasets that have information about map components for a certain city. Every city, even every town has their own datasets. Also, you can create your specific dataset for a specific area.We used this dataset to make inferences about cities.

Data for 116 cities were ready to download directly. Other cities were created manually and it took 30-60 minutes for each city. Mapzen allowed us to create up to five cities. t decreased waste of time.

Datasets included geojson files that were generated for different purpose and usage. Geojson is a file format that is an open mapping standard that uses text to describe geographical features, locations, and attributes, based on JSON. Full specification can be found on their website. In the Figure 2 you can see the file content. Features are determined by analyses and checking coordinates and types with Google earth. The number of features was decreased 200 to 85. You can see features that we used, on the table.

As shown in the table, ratios were used to compare cities. Generally, capitation features were used. We notice that there are some differences between the population we found and the real population. A new dataset was created for the population. We downloaded this dataset from worldpopulationreview.com. Missing ones were collected from Wikipedia.

During the process we encountered some problems:

- There are different cities named same. If there is a link on movehub, it is easy to choose correct one otherwise the more popular was chosen.
- Some areas had been logged as 0 because they were too small.We did not count them.
- Data has a lot of noise.
- Files are encoded 'utf8'.It was converted to ASCII format with PowerShell script.

- Some cities have huge files. We can not open these files directly. Cities were used after deleting unnecessary features from their huge files.
- We encounter stop iteration error with some landusages.geojson files. It can not be solved.

Finally, Train and Validation datasets were created but city number decreased to 150.If we didnt use buildings and landusages files, we could use all of the cities. But half of the features are in these two files. Furthermore, prediction problem is converted to classification problem with a different class number.A feature table is shown end of documenet.

3.2. Modelling

After the dataset has been made workable, there are a few important points to need to be aware of before deciding the modeling. Firstly, it should be kept in mind that the dataset has many features that do not affect the result at the same ratio. It is quite troublesome to calculate them correctly before the process.

Our goal is to estimate the life qualities by comparing the physical characteristics of cities, so our problem falls into the category of regression problems. Our main focal point is the regression solutions but we have also tried solutions to this problem by transforming our problem into a classification model by labeling it according to the scores of the dataset. In fact, we were hoping to create regression and classification models using deep learning but it was not possible to produce an efficient result since the number of cities in our dataset was not enough to train the model.
We choose three methods for both regression and classification:

- Support Vector Machine
- Decision Tree
- Random Forest

We have used the scikit-learn library to implement these methods on python.

Support Vector Machines

Support Vector Classification

For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\zeta_i \tag{1}$$

subject to the constraints:

$$y_i(w^T\phi(x_i)+b) \geq 1-\zeta_i \tag{2}$$

and

$$\zeta_i \geq 0, i = 1\ldots N \tag{3}$$

Support Vector Regression

For this type of SVM the error function is:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\zeta_i + C\sum_{i=1}^{N}\zeta_i^{\cdot} \tag{4}$$

which we minimize subject to:

$$w^T\phi(x_i)+b-y_i \leq \epsilon+\zeta_i^{\cdot} \tag{5}$$

$$y-w^T\phi(x_i)-b_i \leq \epsilon+\zeta_i \tag{6}$$

$$\zeta_i\zeta_i^{\cdot} \geq 0, i = 1\ldots N \tag{7}$$

Where C is the capacity constant, w is the vector of coefficients, b is a constant,ζ_i represents parameters for handling nonseperable data. The index i labels the N training cases. Note that $y \in \pm 1$ represents the class labels and xi represents the independent variables. ϕ is a kernel function that is used to transform the data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid overfitting.

Kernel Functions

$$K(\mathbf{X_i},\mathbf{X_j}) = \begin{Bmatrix} \mathbf{X_i}\cdot\mathbf{X_j} & \text{Linear} \\ (\gamma\mathbf{X_i}\cdot\mathbf{X_j}+\mathrm{C})^{\mathrm{d}} & \text{Polynomial} \\ \exp(-\gamma|\mathbf{X_i}-\mathbf{X_j}|^2) & \text{RBF} \\ \tanh(\gamma\mathbf{X_i}\cdot\mathbf{X_j}+\mathrm{C}) & \text{Sigmoid} \end{Bmatrix}$$

Where $K(X_i, X_j) = \phi(X_i \cdot \phi X_j)$ that is the kernel function, represents a dot product of input data points mapped into the higher dimensional feature space by transformation ϕ.
Gamma is an adjustable parameter of certain kernel functions.The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

Decision Trees

Given a binary categorization, C, and a set of examples, S, for which the proportion of examples categorized as positive by C is p+ and the proportion of examples categorized as negative by C is p-, then the entropy of S is:

$$Entropy(S) = -p_+ \log_2(p_+) - p_- log_2(p_- \quad (8)$$

Given an arbitrary categorization, C into categories $c_1 \ldots C_n$, and a set of examples, S, for which the proportion of examples in c_i is p_i, then the entropy of S is:

$$Entropy(S) = \sum_{i=1}^{n} -p_i \log_2(p_i) \quad (9)$$

The information gain of attribute A, relative to a collection of examples, S, is calculated as

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{S_v}{S} Entropy(S_v) \quad (10)$$

The information gain of an attribute can be seen as the expected reduction in entropy caused by knowing the value of attribute A.In a corresponding regression tree, the standard deviation is used to make that decision in place of information gain.

$$S = \sqrt{\frac{\sum(x-\mu)^2}{n}} \quad (11)$$

Random Forest

The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as: $g(x) = f_0(x) + (f_1(x) + \ldots$ where the final model g is the sum of simple base models ϕ. This broad technique of using multiple models to obtain better predictive performance is called model ensembling. In random forests, all the base models are constructed independently using different subsamples of the data.

4. Experimental Results

4.1. Regression Solutions

Support Vector Regression

Especially, in the sigmoid and RBF graphs, we can see that the results are pressed into a small-sized interval. The changes in the features don't affect the value of results correctly, the results changing with much smaller impact. The model cannot produce a result lower than 50.

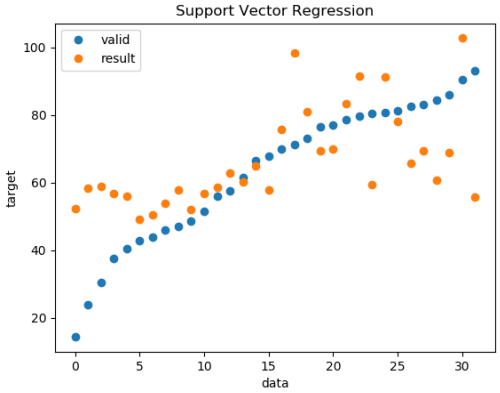

Figure 3. Linear kernel C=5

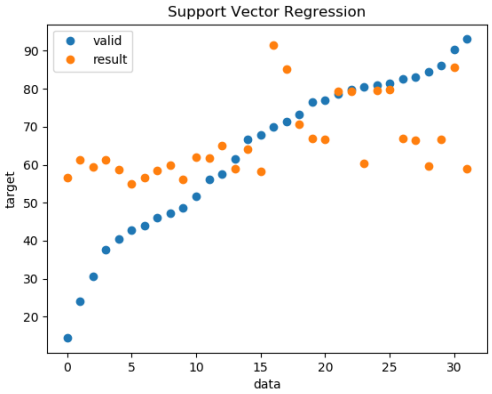

Figure 4. Sigmoid kernel C=100

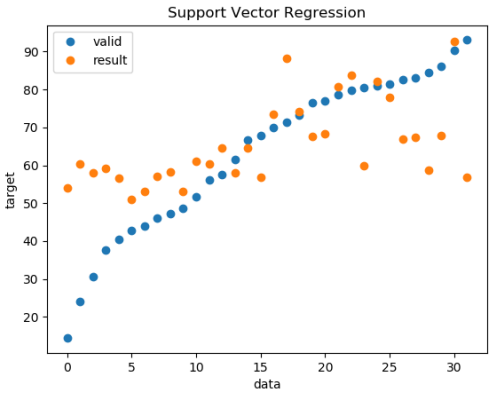

Figure 5. Radial Basis kernel C=100

Decision Tree Regression

Figure 6. max depth = 6

Random Forest Regression

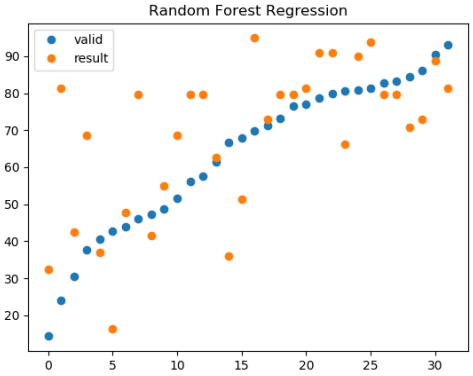

Figure 7. max depth = 6,bootstrap=false

Figure 8. max depth = 6,bootstrap=true

In this two graph, we can easily see the positive effects of the using larger dataset obtained by utilizing bootstrap method on the results. So, the variance of the results is decreasing and the results are becoming more meaningful

4.2. Classification Solutions

Support Vector Classification

Class size=2	Rbf	sigmoid	linear
C:5	0.65625	0.65625	0.875
C:10	0.65625	0.65625	0.84375
C:25	0.9375	0.65625	0.84375
C:50	0.90625	0.90625	0.8125
C:100	0.90625	0.90625	0.8125

Class size=2	Rbf	sigmoid	linear
C:5	0.5	0.5625	0.59375
C:10	0.625	0.5	0.625
C:25	0.625	0.65625	0.625
C:50	0.625	0.625	0.46875
C:100	0.625	0.625	0.375

Class size=10	Rbf	sigmoid	linear
C:5	0.15625	0.15625	0.1875
C:10	0.34375	0.15625	0.125
C:25	0.34375	0.34375	0.09375
C:50	0.34375	0.375	0.09375
C:100	0.25	0.34375	0.09375

Decision Tree Classification

Table 1. Class size:2

max depth	accuracy	maxdepth	accuracy
1	0.8125	2	0.78125
3	0.78125	4	0.75
5	0.6875	6	0.71875
7	0.71875	8	0.71875
9	0.71875	none	0.71875

Table 2. Class size:4

max depth	accuracy	maxdepth	accuracy
1	0.59375	2	0.59375
3	0.46875	4	0.46875
5	0.5	6	0.375
7	0.46875	8	0.40625
9	0.40625	none	0.40625

Table 3. Class size:10

max depth	accuracy	maxdepth	accuracy
1	0.34375	2	0.375
3	0.25	4	0.25
5	0.28125	6	0.25
7	0.25	8	0.25
9	0.28125	none	0.3125

Random Forest Classification

[h] Bs:Bootstrap

Table 4. Class size:2

max depth	Bsfalse	Bstrue	maxdepth	Bsfalse	Bstrue
1	0.8125	0.7812	6	0.7812	0.90625
2	0.8125	0.8437	7	0.8437	0.9375
3	0.8437	0.875	8	0.8125	0.9375
4	0.75	0.9375	9	0.8437	0.9375
5	0.8125	0.8125	none	0.8437	0.9375

Table 5. Class size:4

max depth	Bsfalse	Bstrue	maxdepth	Bsfalse	Bstrue
1	0.5625	0.5625	6	0.5625	0.625
2	0.6562	0.5937	7	0.5625	0.53125
3	0.6562	0.6875	8	0.625	0.59375
4	0.6562	0.625	9	0.625	0.625
5	0.5937	0.5937	none	0.625	0.625

Table 6. Class size:10

max depth	Bsfalse	Bstrue	maxdepth	Bsfalse	Bstrue
1	0.3437	0.2812	6	0.2187	0.34375
2	0.3125	0.2812	7	0.3125	0.375
3	0.1875	0.3437	8	0.2187	0.3437
4	0.3125	0.25	9	0.3125	0.25
5	0.1875	0.3125	none	0.3437	0.28125

As a sample result, we got a weights for RFC with Class size = 4 and max depth = 3. The ones with the highest weight respectively, $(parking, 0.121997)$, $(pier, 0.065002)$, $(garage, 0.054224)$, $(residental, 0.043953)$, $(footway, 0.037077)$, $(firestation, 0.037060)$, $(school, 0.035133)$, $(terrace, 0.034628)$, $(cinema, 0.032273)$, $(crossing, 0.026875)$, $(cylceway, 0.026056)$, $(college, 0.023532)$, $(train_station, 0.022929)$, $(hospital, 0.022152) and (fuel, 0.021980)$.

5. Conclusions

5.1. Summarizing

In this study, our purpose was to get an effective result of life quality estimation according to physical characteristics of cities. But we can easily see the negative impacts of the difficulties we have encountered when creating datasets on the results

The fact that the properties are obtained from statistically indeterminate data and the fact that the data set is not fully used(almost 1/3 of the data is not able to used, especially major cities and the small cities can keep more exceptions) might cause the weights of the properties to be determined incorrectly and thus the model cannot be constructed sufficiently accurately.
However, if we compare it with the last project in the related studies, it can be said that our work has produced a useful result if taken into account in the classification.

5.2. Future Work

In this Project, we tried to find score or class of cities from map data that consist of physical attributes. It can be combined with a different dataset that includes information about the economy, security, education. For example, It can be the countable number of school per person but it is more meaningful to know the number of teacher per student or book per student in the library. Also, MAPZEN includes coordinates and it was not used(except roads).Coordinates can be used to find average distances between a house and a school or it can be detected that place points are located equally everywhere in city or not.

In addition to this MAPZEN and MOVEHUB are different resources. It is not known if the MOVEHUB surveys cover all of the city or a part of the city. Creating data with this knowledge can be more suitable.

Also, the number of data is very few for machine learning.Data number can be increased by taking part of a city instead of the whole city. Data can be labeled for different purpose like predicting development status.

Feature Tables

Table 7. buildings.geojson

church	industrial	apartments	university
office	house	residential	school
palace	terrace	public	commercial
train station	shed	hospital	temple
retail	chapel	college	detached
garage	mosque		

Table 8. amenities.geojson

library	fuel	hospital	police
school	townhall	university	fire station

Table 9. landusages.geojson

zoo	pedestrian	sports centre	university
park	stadium	school	place of worship
fuel	parking	grass	pitch
footway	theatre	meadow	retail
library	playground	hospital	heath
college	cinema	pier	forest
residential	golf course	commercial	scrub
railway	farmyard	farmland	water areas

Table 10. amenities.geojson

city	village	suburb	town

Table 11. amenities.geojson

living street	pedestrian	steps	trunk link
path	rail	subway	cycleway
pier	light rail	funicular	raceway
tram			

Table 12. amenities.geojson

subway entrance	bus stop	station	crossing
helipad	tram stop		

References

[1] https://www.movehub.com/

[2] https://www.numbeo.com/

[3] https://www.kaggle.com/jonathanbouchet/movehub-rating-prediction

[4] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, Peter Kontschieder; The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4990-4999

[5] Robert J. Rogerson,Quality of Life and City Competitiveness,May 1 1999,Volume: 36 issue: 5-6, page(s): 969-985 Issue published: May 1, 1999

[6] https://mapzen.com/data/metro-extracts/

[7] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[8] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

[9] http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[10] http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[11] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[12] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.htm

[13] https://elib.uni-stuttgart.de/handle/11682/9297

[14] Artificial Intelligence: A Modern Approach (3rd Edition), Stuart Russell and Peter Norvig. Prentice Hall, 2009

[15] Machine Learning: A Probabilistic Perspective, Kevin Murphy, MIT Press, 2012