Data Analysis

This section is the main part of project. There are two analyses have been done to present the importance of each factors, which are the analysis of relationship between happiness and factors (such as weather condition) and the analysis of survey.

**The analysis of relationship between happiness and factors**

In this analysis problem, the question that which are the most important factors to affect cities’ livability has been asked. The solving process could be abstracted as the feature selection process in which all factors that needed be chose are independent variables and the livability whose measurement is the happiness rating of cities is independent variable. The function to do feature selections is LASSO which we tried firstly and Random Forest we used finally.

The process of experiments:

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

Feature selection:

There are lots of model to select features, LASSO is the most common one. We first test LASSO and find LASSO is not good enough for our project. Finally, we change to the random forest regression model.

LASSO:

Equation.

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 lambda (from 0.0001 to 0.1) whose result has minimum MSE, to do the feature selection.

The result of using LASSO:

1. The result of this model is swinging. Every testing might have different best lambda and the difference of them are quit large.
2. In the most of 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 the most of time 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.

1. The difference of the data in some features might be small. Some features have been measured according to different shires rather than cities. We deal with this problem by putting the same data to all the cities from same shire. Therefore, some rows of training data have no much difference which might cause worse fitting.
2. 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 come out of LASSO is also not our expected. For example, it can only output the weight for each factor which are not showing the importance ranking for all the factors. It is because the two of these factors might have correlationship.

This supposition has been confirmed by measuring the pearson correlation coefficient between two factors. For example, some factors have strong positive correlationship with population.

|  |
| --- |
| the population and FrostDay\_perYear pearsonr correlation coefficient is 0.073343, p-value is 0.507316  the population and GVA pearsonr correlation coefficient is -0.234764, p-value is 0.031591  the population and Rainfall\_perMon pearsonr correlation coefficient is 0.165111, p-value is 0.133379  the population and SummerDay\_ave\_temperature pearsonr correlation coefficient is -0.284169, p-value is 0.008801  the population and SummerNight\_ave\_temperature pearsonr correlation coefficient is -0.261417, p-value is 0.016308  the population and Sunshine\_perMon pearsonr correlation coefficient is -0.340634, p-value is 0.001521  the population and WinterDay\_ave\_temperature pearsonr correlation coefficient is -0.345143, p-value is 0.001303  the population and WinterNight\_ave\_temperature pearsonr correlation coefficient is -0.173028, p-value is 0.115495  the population and hospitals\_number pearsonr correlation coefficient is 0.726583, p-value is 0.000000  the population and house\_price pearsonr correlation coefficient is -0.284419, p-value is 0.008740  the population and number\_of\_universities pearsonr correlation coefficient is 0.662404, p-value is 0.000000  the population and population pearsonr correlation coefficient is 1.000000, p-value is 0.000000  the population and pubs\_number pearsonr correlation coefficient is 0.546639, p-value is 0.000000  the population and road\_traffic\_2015 pearsonr correlation coefficient is 0.510170, p-value is 0.000001  the population and road\_traffic\_2016 pearsonr correlation coefficient is 0.498474, p-value is 0.000001  the population and school\_number pearsonr correlation coefficient is 0.810281, p-value is 0.000000  the population and stations\_number pearsonr correlation coefficient is 0.686357, p-value is 0.000000  the population and total\_jobs pearsonr correlation coefficient is 0.985825, p-value is 0.000000  the population and traffic\_noise pearsonr correlation coefficient is 0.778560, p-value is 0.000000  the population and unemployment pearsonr correlation coefficient is 0.375785, p-value is 0.000427  the population and hospitals\_number\_per\_person pearsonr correlation coefficient is -0.198825, p-value is 0.069816  the population and pubs\_number\_per\_person pearsonr correlation coefficient is -0.203885, p-value is 0.062853  the population and stations\_number\_per\_person pearsonr correlation coefficient is -0.170241, p-value is 0.121568  the population and school\_number\_per\_person pearsonr correlation coefficient is -0.374345, p-value is 0.000452  the population and total\_jobs\_per\_person pearsonr correlation coefficient is -0.218554, p-value is 0.045795  the population and number\_of\_universities\_per\_person pearsonr correlation coefficient is 0.108700, p-value is 0.324996 |

In summary, LASSO could not be the good methods to do the feature selection in our project and we should care about the relationship between two factors.

Random Forest:

There are three reasons why random forest is more suitable for our feature selection.

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

Brief introduction of the calculate process

Therefore, random forest is most suitable and has been chose to be the model in our project.

|  |
| --- |
| The rank of importance: less to most  ['SummerDay\_ave\_temperature',  'SummerNight\_ave\_temperature',  'Sunshine\_perMon',  'road\_traffic\_2015',  'traffic\_noise',  'school\_number\_per\_person',  'FrostDay\_perYear',  'pubs\_number',  'total\_jobs\_per\_person',  'number\_of\_universities',  'total\_jobs',  'Rainfall\_perMon',  'stations\_number\_per\_person',  'house\_price',  'GVA',  'number\_of\_universities\_per\_person',  'WinterNight\_ave\_temperature',  'population',  'WinterDay\_ave\_temperature',  'pubs\_number\_per\_person',  'unemployment',  'school\_number',  'hospitals\_number\_per\_person',  'hospitals\_number',  'stations\_number',  'road\_traffic\_2016']  Out[53]:  {'FrostDay\_perYear': 2.8603382127892094e-05,  'GVA': 0.012105026105635082,  'Rainfall\_perMon': 0.004313505494751562,  'SummerDay\_ave\_temperature': 0.0,  'SummerNight\_ave\_temperature': 0.0,  'Sunshine\_perMon': 0.0,  'WinterDay\_ave\_temperature': 0.032802557726910396,  'WinterNight\_ave\_temperature': 0.023259177597378482,  '\_id': ObjectId('5a4e02a57c4c7b10887dac8d'),  'hospitals\_number': 0.13849379181378044,  'hospitals\_number\_per\_person': 0.11223550170873713,  'house\_price': 0.009815594922222743,  'number\_of\_universities': 0.0020814155846021884,  'number\_of\_universities\_per\_person': 0.015763443160266254,  'population': 0.02749113337050372,  'pubs\_number': 4.621981803086575e-05,  'pubs\_number\_per\_person': 0.052369125668876734,  'road\_traffic\_2015': 0.0,  'road\_traffic\_2016': 0.1910347576824795,  'school\_number': 0.10026176707881813,  'school\_number\_per\_person': 0.0,  'stations\_number': 0.17258035819124543,  'stations\_number\_per\_person': 0.006260976034285652,  'total\_jobs': 0.004009982598814446,  'total\_jobs\_per\_person': 0.00030544077285242147,  'traffic\_noise': 0.0,  'unemployment': 0.09474162128768097} |