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 analysis are LASSO which we tried firstly and Random Forest we used finally.

Lasso:

1. The process of experiment:

Normalization:

Lasso:

1. The result of experiment:

The randomness of result

No selection

1. The analysis of experiment result:

Reason of failure:

Limitation of Lasso:

Random Forest:

Why use this?

The process of selection:

The result of feature selection by random forest:

**The analysis of survey**

Some differences why?

在训练模型之前，首先对特征矩阵进行normalization （此处插入 normalization部分）

Feature selection:

整个问题可以抽象成特征选择过程，特征选择的一般方法为Lasso:因此本实验首先使用了lasso来进行特征选择。

实验过程：由于训练样本量少，为保证实验的准确性，使用交叉验证来生成不同的训练集和测试集来训练，并且选择 λ 区间为（0.0001到0.1）中MSE最小的情况来作为最终的特征选择的结果。

实验结果：1. 在进行反复实验之后，发现实验结果并不稳定，λ取值存在较大波动，（可能需要图）。2. 多数情况下最小MSE存在于没有去特征的情况

分析：

若以上两种情况发生，可以评判该模型是失败的。因为，在多数情况下并没有得到特征选择的结果，这有可能有两种原因造成。

首先是训练数据集本身的问题，

由于某些数据集并不是对城市的数据统计，而是对郡如组合of几个城市的综合评定，因此当format这些信息到城市时，这几个城市都使用了同样的信息。比如，这就会导致某些条目数据相同，差异不大导致拟合效果变差。

另一个重要的原因是某些自变量和因变量关系可能是非线性的，而我们使用了线性的模型去拟合它。比如，因此Lasso作为线性模型并不适合解决这个问题。

Lasso产生的结果也不是我们所期望的。它并未直接得到各factor重要性比较，最终使用lasso生成的每个因素的权重并不能反映其对happiest的影响，这是由于factor之间可能是非独立的。

为了验证这个猜想，我们使用了pearson相关系数去测试。结果表明，某些因素之间存在很强的相关性。比如：

Correlation: pearson, find some have strong positive correlation.

|  |
| --- |
| 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 |

结论：因此 LASSO 并不适合作为特征选择的方案。

Random forest：

随机森林是决策树的升级版，它可以处理非线性数据，并且它在进行特征选择的时候它会对各特征的单独重要性评分，因此相比于lasso， 它更加适合本次实验。

选择过程：

结果：

|  |
| --- |
| 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} |

Compare with survey: find some difference. Why?