

Group 70 Final Report

YIFAN PEI 20303063

MIN WU 20302508

1. Introduction

As we all know, we usually describe obesity in the unit of a country since the food structure is significantly different among different countries while remains similar of the same country(for example, the Irish people get 28.5256% of daily protein intake from animal product, and 21.4791% from vegetable intake and so on while the Nepalese get only 8.5788% in animal and 3.4866% in vegetable, the food structure of the 2 countries differs so much and the obesity rate of them is 3.8%(Nepal) and 26.9%(Ireland), this is a big contrast !).

We collected the percentage of different food as the nutrition intake from 170 countries, does the percentage of different kinds of food as the protein intake or fat intake or calories intake really matter? Why we have the same nutrition intake while the obesity of countries vary so much? And which food structure is most healthy with the lowest obesity? And the idea of the project is to find the suitable food structure to change our diet habit and decrease the obesity and improve the life quality. This really matters because only by changing diet habit can we live in a healthier lifestyle. So, the input is the food proportion of different nutrition intake of 170 countries, the output is the 3 obesity prediction models of the proportion of different food intake.




 Protein_Supply_Quantity_Data	2020/10/26 19:32	Microsoft Excel ...	43 KB
 Food_Supply_kcal_Data	2020/10/26 19:32	Microsoft Excel ...	43 KB
 Fat_Supply_Quantity_Data	2020/10/26 19:32	Microsoft Excel ...	42 KB

Fig 1 the input datasets

Dataset	Best model	Parameters
Food_Supply_kcal_Data	Random Forrest Regression Model	K: 8, Poly:2, Trees: 8
Fat_Supply_Quantity_Data	Random Forrest Regression Model	K: 20, Poly:3, Trees: 20
Protein Supply Quantity	Random Forrest Regression Model	K: 20, Poly:3, Trees: 8

Fig 2 the output prediction models and its parameters

2. Dataset and Features

We got the data from the US open data portal.

We will use 3 CSV file datasets including fat, protein, calories supply percentage from different foods, e.g. in dataset Fat_Supply_Quantity_Data.CSV, the Afghanistani fat intake from Alcoholic Beverages(0%), Animal Products(21.6397%),Animal fats(6.2224%) and ...the obesity rate of Afghanistan is 4.5%. Since there are 23 kinds of food in a row, so there are 23 features of each data point and these kinds of food that consists of one row. Also, because we collect the food proportions from each country as a row and there are 170 countries in total so the number of data points is 170.

Country	Animal Products	Animal fats	Cereals	Eggs	Fish, Seafo	Fruits - Exc	Meat	Miscellane	Milk - Excl	Vegetal Pr	Vegetable	Vegetables	Obesity
Afghanista	21.6397	6.2224	8.0353	0.6859	0.0327	0.4246	6.1244	0.0163	8.2803	28.3684	17.0831	0.3593	4.5
Albania	32.0002	3.4172	2.6734	1.6448	0.1445	0.6418	8.7428	0.017	17.7576	17.9998	9.2443	0.6503	22.3
Algeria	14.4175	0.8972	4.2035	1.2171	0.2008	0.5772	3.8961	0.0439	8.0934	35.5857	27.3606	0.5145	26.6
Angola	15.3041	1.313	6.5545	0.1539	1.4155	0.3488	11.0268	0.0308	1.2309	34.701	22.4638	0.1231	6.8
Antigua an	27.7033	4.6686	3.2153	0.3872	1.5263	1.2177	14.3202	0.0898	6.6607	22.2995	14.4436	0.2469	19.1
Argentina	30.3572	3.3076	1.3316	1.5706	0.1664	0.2091	19.2693	0	5.8512	19.6449	17.3147	0.1878	28.5
Armenia	29.6642	6.2619	2.5068	1.6196	0.2218	0.5468	10.8165	0.0361	10.4709	20.3384	12.8127	0.8717	20.9
Australia	24.1099	4.603	0.9908	0.7017	0.4515	0.4028	11.6002	0.052	6.5196	25.8901	20.3612	0.2144	30.4

Fig 3 part of the Fat_Supply_Quantity_Data CSV file

Country	Alcoholic E	Animal Pr	Animal fat	Aquatic Pr	Cereals - E	Eggs	Fish, Seafo	Fruits - Exc	Meat	Milk - Excl	Vegetal Pr	Vegetable	Vegetables	Miscellaneous	Obesity
Ireland	0.7787	28.5256	0.3115	0	14.0504	1.4017	3.2849	0.8543	13.5407	9.6847	21.4791	0.0425	1.9681	0.0897	26.9
Israel	0.11	29.564	0.0204	0.0041	12.7832	1.2755	2.6773	0.8476	16.5363	8.3374	20.436	0.0326	1.7889	0.0896	26.7
Italy	0.1827	26.7179	0.1405	0	16.4832	1.7191	3.7754	0.8853	12.5299	8.0191	23.2845	0.0141	1.9907	0.1499	22.9
Jamaica	0.103	24.6534	0.048	0.0137	16.5134	0.3912	4.8593	1.0913	12.6081	5.7515	25.3466	0.0343	1.6747	0.3706	24.4
Japan	0.2541	27.7588	0.0058	0.0635	12.4213	3.5687	9.5744	0.3292	9.8112	4.1058	22.2383	0.0289	1.9923	0.0635	4.4
Jordan	0.0072	17.4589	0.0289	0	24.224	0.7235	1.1287	0.615	10.361	4.7826	32.5447	0.0145	2.0838	0.1592	33.4

Fig 4 part of the Protein_Supply_Quantity_Data CSV file

Regarding the amount of each data point, we use “Polynomial Features” to fit the training data. The unit of the datasets is percentage, which means the real value of them is in the range of [0,1], so there is no need to do data normalization.

Also, there is some data which is missing (as shown “NA” in obesity value cells), at first we decided to write a python file that detects the invalid data first and deletes the rows containing the “NA”, remaining the valid data to train. However, we finally found the performance of prediction models we got was not as good as we just filled with the mean value of that column. So we use a function named “get_features” using pandas to replace the “NA” of mean value as the data preprocessing.

```
# give a constant mean value when there is a "NA" of value
def get_features(file):
    obesity = file['Obesity']
    features = file.iloc[:, 1:24]

    # print(len(obesity.dropna())/len(obesity)) 98%, we do not need to drop those
    # features with N/A

    obesity.fillna(obesity.mean(), inplace=True)
    return features, obesity
```

3. Methods

In our project, we are using 4 kinds of machine learning models: Linear Regression Model, Ridge Model and Random Forrest Regression Model and the baseline model is Dummy Regressor.

1) Linear Regression Model

More generally, a linear model makes a prediction by simply computing a weighted sum of the input features, plus a constant called the bias term (also called the intercept term), as shown in

equation: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$, \hat{y} is the predicted value, n is the number of features, x_i is the i th feature value, θ_j is the j th model parameter (including the bias term θ_0 and the feature weights $\theta_1, \theta_2, \dots, \theta_n$), and we would choose the best parameters to minimize the cost function (MSE). Gradient Descent is a very generic optimization algorithm capable of finding optimal solutions to a wide range of problems. The general idea of Gradient Descent is to tweak parameters iteratively in order to minimize a cost function. Concretely, we start by filling θ with random values (this is called random initialization), and then we improve it gradually, taking one baby step at a time, each step attempting to decrease the cost function (MSE), until the algorithm converges to a minimum. An important parameter in Gradient Descent is the size of the steps, determined by the learning rate hyperparameter.

2) Ridge Model

For a linear model, regularization is typically achieved by constraining the weights of the model. So we will apply Ridge Regression model as adding the L2 penalty to train the data. Ridge

Regression is a regularized version of Linear Regression: a regularization term equal to $\alpha \sum_{i=1}^n \theta_i^2$ is added to the cost function. This forces the learning algorithm to not only fit the data but also keep the model weights as small as possible. The hyperparameter α (which equals to $1/2C$) controls how much you want to regularize the model. If $\alpha = 0$ ($C = +\infty$) then Ridge Regression is just Linear Regression. If α is very large (C is very small close to 0), then all weights end up very close to zero and the result is a flat line going through the data's mean. To choose the best value of C , we use cross-validation to plot the MSE with the range of C values.

3) Random Forrest Regression Model

The Random Forest algorithm introduces extra randomness when growing trees; instead of searching for the very best feature when splitting a node, it searches for the best feature among a random subset of features. This results in a greater tree diversity, which (once again) trades a higher bias for a lower variance, generally yielding an overall better model. Decision trees algorithm in machine learning is a kind of decision support tool which uses trees' structure. It is also a very effective method in many other areas like operational research. Basically, trees decision tries to split feature space into independent and undividable unit and calculate conditional probability for the unit in tree. Random Forrest algorithm based on decision tree and basically, it is an application of decision trees multiply and make decision (including classification and prediction) based on the statistics results of each decision tree.

4) Baseline Model: Dummy Regressor

DummyRegressor is a regressor that makes predictions using simple rules. This regressor is useful as a simple baseline to compare with other (real) regressors. In our project the score of it is -0.068 and the performance of it is worse than any of the models above, which means all the models above are better than random predictions.

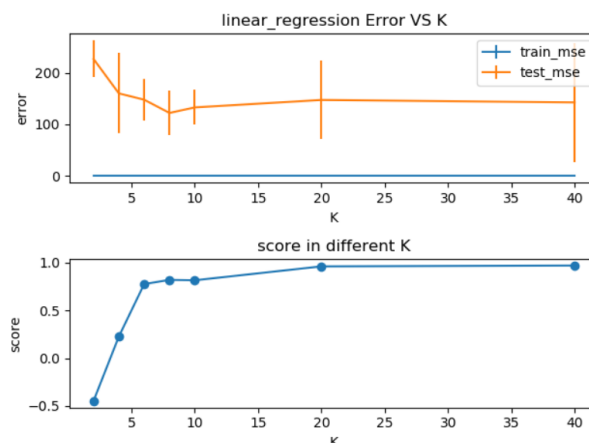
4. Experiments/Results/Discussion

In the experiments stage, selecting hyperparameters and training models based on the best hyperparameters. I listed all the hyperparameters I need select with cross validation method.

- 1) Linear Regression Model: Degree of Polynomial features, K in KFold;
- 2) Ridge Model: C, Max Iteration, degree of polynomial features, K in KFold;
- 3) Random Forrest Regressor, Number of Trees, criterion, degree of polynomial features, K in KFold.

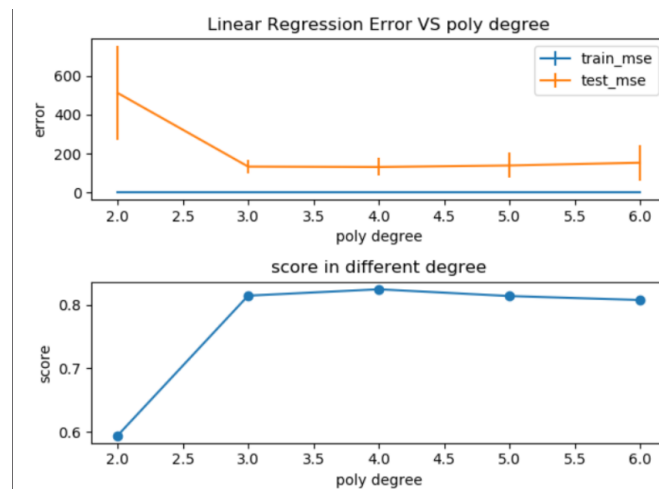
Basically, I need to select 10 hyperparameters in all three kinds of models in each dataset.

I will take an example in the first data set: "fat_supply_quantity_data". The first is linear regression model, I use default value in degree of polynomial features to select the best K in KFold.



According to the result of cross validation experiment, I found that when degree = 8 the test error is the smallest, but the score is not the biggest, so made a trade-off here, selected 8 as the best K in KFold.

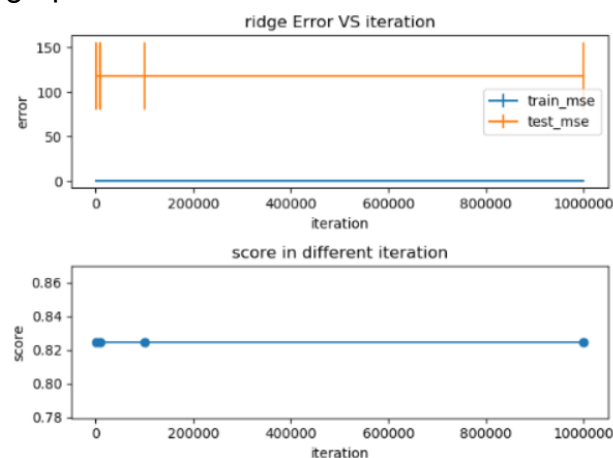
Then based on the best K value, I started to search the best degree of polynomial features, the result graph is as below.



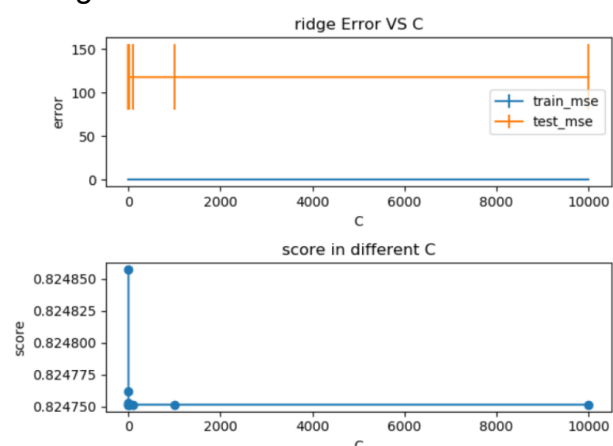
According to the graph below, I chose the best degree as 3.

Then for the second kind of model, Ridge model. Firstly, using the same method to get the best K in KFold and degree of polynomial features, secondly, I started to search the best C and max_iteration. The reasons why I want to adjust the max_iteration but not using the default value is that I considered there may exist problems when the amount of whole data is insufficient and lead under-fitting.

To select C, I get the result graph as below.

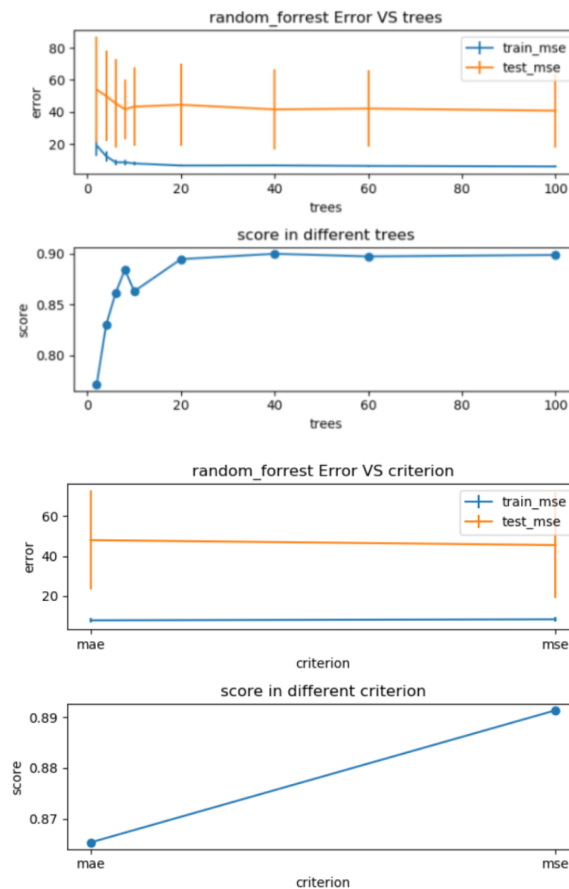


I found that in this graph I can not see there was obvious differences in test and train error due to the changes of iteration. So as a conclusion, I will use default value of max_iteration. Then it comes to C, I choose a wide range of if.



I found that with the changes of C, there is not obvious differences in model error (optimization range is in $1e4$), I also adjusted the range of C values, but the result remained. So, I choose the default value of C as the best C value.

Then is Random Forrest Regression model, I need to select “number of trees” and “criterion” besides K and degree.



Here I want to mention that “criterion” means the rule of measuring loss.

Basically, the primary metrics are: MSE error(the cost function)and model score(the accuracy).

For baseline model, I chose “Dummy Regressor” from sklearn, which was used in three data and the results from “fat_supply_quantity” is worse than the two other data sets.

Conclusion: After a series of experiments, we found that the in the three data sets and, the performance in “food kcal”is the best, so basically to deal with obesity question, we may focus on the heat in the food, try to eat low heat food may contribute more than other methods.

For overfitting: I don’t think we have overfit our training set, because we choose the simplest model while keeping the MSE low and high score to avoid overfitting, and as we can see in the plots above, each plot we don’t see any more than 1 curves which means at the point of the curve comes the best model parameter.

For baseline model, which is at random predictions, as we can see from the final score(-0.068) which is lower than any models above, and the MSE of the baseline model is also the biggest which means any models performs much better than just random predictions.

5. Summary

In this project, 4 algorithms are used in three datasets trying to explore the most important factor leading to obesity. As is shown in the all plots in last section, the best algorithm is random forest algorithm, and I think one of the reasons that why forest algorithm is better because it can solve high-dimensional question better and my datasets have more features than usual and compared with other algorithms, its performance keeps the best. And the scores of Random Forrest Regressor of the 3 datasets are as follows:

Dataset	Score(Random Forest)	Dummy Regressor
Fat_Supply_Quantity_Data	0.34	-0.068
Food_Supply_kcal_Data	0.48	-0.068
Protein_Supply_Quantity_Data	0.27	-0.068

As we can see from the scores above, the proportion of different foods with calories intake matters most, which means in order to decrease the obesity, changing the food structure of calories intake has more impacts.

This is interesting because as a common sense, fat in the food is the biggest contributor to gain weight, but as the result shows even you eat a high fat diet you may not get as fat as the one who eats in low fat but with high calories.

6. Contributions

Yifan Pei: do the core coding, including data pre-processing, model establishment and training, hyperparameters' experiments and selection.

Min Wu: find the datasets and do some research to write the proposal and the final report.

7. Include a github link (or similar) to the code written as part of the project.

https://github.com/pyfppp/ML_Group70_Project.git