EMPLOYING MACHINE LEARNING ALGORITHM TO DECIPHER AND FORECAST SUBJECTIVE WELL BEING

Mohammed Riyaan[1], Mohammed Saif[2], Syed Mohtashim Ahmed[3], Khutaija Abid[4]
[1,2,3]BE Students, IT Department, Lords Institute of Engineering and Technology
[4]Associate Professor, IT Department, Lords Institute of Engineering and Technology
khutaija@lords.ac.in

Abstract— In later a long time, machine learning has ended up an amazingly prevalent point within the innovation space. A critical number of inquiries about are attempting to embrace this innovation to move forward the proficiency of government administrations. We utilize machine learning to analyze the features related to the joy record and to form a few expectations. Utilizing the gotten dataset of the overview that collects a irregular number of Chinese people's bliss file and inquires important questions, we offer a few calculations to analyze the relationship between people's joy file and the answers to the issues they have replied. We will make expectations within the modern dataset roughly steady with the genuine values by utilizing the comes about. In our think about, we pick up a few vital highlights like salary, instruction, and wellbeing. This paper gives a few novel viewpoints to progress benefit for e- administration.

Keywords-*Machine learning; Happiness Index*

I. INTRODUCTION

Individuals living around the world are continuously yearning for joy, which is an unceasing truth. To way better show the intangible bliss in information, analysts have made the term joy list, which is utilized to degree people's sentiments and encounters of their survival and improvement, too meaning people's sense of joy. From a arrangement of successive numbers, able to degree people's joy normally and outwardly, making it more helpful to explore people's evaluation of their joy. That's what the bliss record can bring to us [1].

So why is it basic to explore the bliss list? What can it do for our society? Agreeing to a few specialists, the joy record may be a critical list to degree social agreement and an inescapable prerequisite of the scientific outlook on advancement [2]. Since bliss may be a searing point in 1 people’s day by day lives, it is significant to have a profound knowledge into that for the government to form more advantageous approaches and keen advancements for individuals. When those changes from the inquire about on bliss record gotten to be a reality, individuals will ended up more joyful, and the society will gotten to be more agreeable, shaping a ethical circle .Many factors affect people’s happiness, such as their income,living conditions, education level, health, and so on. And for different people, their understandings and interpretations of happiness are distinct. For example, an ordinary citizen A maysay that happiness means a higher salary and less overtime at night. Another professional manager B might think the happiestthing is that he will make the company a world brand.

The meaning of the happiness index is far more than happiness itself. It also includes the external living environmentand self-development conditions of the people. For example, SARS happened in 2002 once made people's happiness index drop. The successful launch of Shenzhou-5 in 2003 raised the happiness index again; people living in a city with low air pollution are relatively happier than those living in urban areaswith high air pollution. Those make the happiness index an intriguing topic to be investigated.

In many past studies on the happiness index, researchers mainly focus on the factors influencing the happiness index andthe significance of happiness index for the entire society[3,4]. Inour research, we not only analyze the relationship between the happiness index of some people and some factors but also predict the happiness index of other people through the analysisby utilizing machine learning, including using some models likeLinear Regression, Decision Tree, Random Forest, Gradient Boosting. We first preprocess the data, then apply several models, and then analyze, and finally come to a conclusion.

II. DATA AND PREPROCESSING

We obtain the dataset from a survey of 10968 randomChinese adults about their happiness index and some information in 2015 and divide it into two separate parts. One ofthem, including the happiness index of 8000 people, is our training set, which is used to analyze the features of the data. Theother part, including the rest of the data, is the testing set, whichwill be the part that we verify the results of machine learning andpredict the happiness index.

After separating the dataset, we first have a general view ofthe data. All of the operations done in our research are on the Jupyter Notebook. We import several packages and utilize the code to get a complete dataset:

TABLE I. The Complete Dataset From Jupyter Notebook A

	id	**happiness**	**survey_type**	**province**	**city**	**country**	**survey_time**	**gender**	**birth**	**nationality**	**…**	**neighbor_familiarity**	**public_service_1**	**public_service_2**
0	1	4	1	12	32	59	2015/8/4 14:18	1	1959	1	…	4	50	60
1	2	4	2	18	52	85	2015/7/21 15:04	1	1992	1	…	3	90	70
2	3	4	2	29	83	126	2015/7/21 13:24	2	1967	1	…	4	90	80
3	4	5	2	10	28	51	2015/7/25	2	1943	1	…	3	100	90

							17:33							
4	5	4	1	7	18	36	2015/8/10 9:50	2	1994	1	…	2	50	50
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
7995	7996	2	2	29	82	124	2015/7/21 19:36	1	1981	1	…	3	40	50
7996	7997	3	1	12	32	61	2015/7/31 16:00	2	1945	1	…	4	80	80
7997	7998	4	1	16	46	78	2015/8/1 17:48	2	1967	1	…	4	75	70
7998	7999	3	1	1	1	8	2015/9/22 18:52	2	1978	1	…	2	56	67
7999	8000	4	1	1	1	3	2015/9/28 20:22	2	1991	1	…	3	80	80

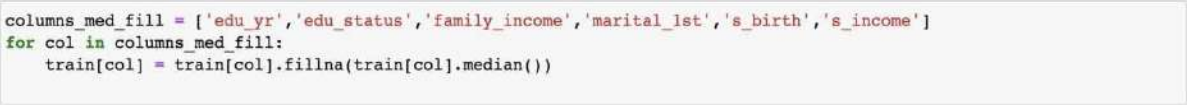
8000 rows h 140 columns

II. The Complete Dataset From Jupyter Notebook B

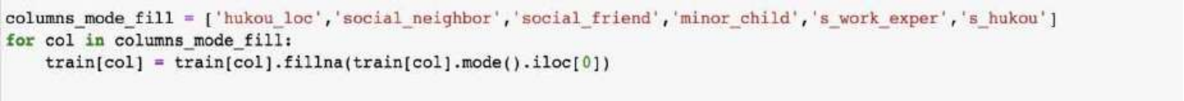
larity	**public_service_1**	**public_service_2**	**public_service_3**	**public_service_4**	**public_service_5**	**public_service_6**	**public_service_7**	**public_service_8**	**public_service_9**
4	50	60	50	50	30.0	30	50	50	50
3	90	70	70	80	85.0	70	90	60	60
4	90	80	75	79	80.0	90	90	90	75
3	100	90	70	80	80.0	90	90	80	80
2	50	50	5	5	50.0	5	5	5	5

			0	0		0	0	0	0
…	…	…	…	…	…	…	…	…	…
3	40	50	50	50	40.0	50	50	60	50
4	80	80	80	80	80.0	60	60	80	80
4	75	70	70	80	80.0	70	75	70	75
2	56	67	70	60	70.0	60	70	80	70
3	80	80	80	80	80.0	80	80	80	80

We can see a table with 8000 rows and 140 columns, whichreveals the number of samples and the available quantity of features. However, this dataset cannot be utilized to analyze therelationship between those features and happiness index since there is some invalid data and meaningless data which will obstruct our analysis. As a result, we need to deal with the databriefly.We first delete the columns with more than half of the data missing because they have no practical reference value. After that, we look for the column with missing data and use two different approaches to fill the null data. One is to use the median to fill the missing data, and the other is to fix the data with mode.



```
columns_med_fill = ['edu_yr','edu_status','family_income','marital_1st','s_birth','s_income']
for col in columns_med_fill:
    train[col] = train[col].fillna(train[col].median())
```



```
columns_mode_fill = ['hukou_loc','social_neighbor','social_friend','minor_child','s_work_exper','s_hukou']
for col in columns_mode_fill:
    train[col] = train[col].fillna(train[col].mode().iloc[0])
```

The codes above are the two approaches. Besides filling thenull data, we also have to revise those incorrect data. When carefully looking at the form, there are some strange numbers like -8 in the column of the happiness index, which we need tomodify in order to make the analysis reliable. Just as we fill thenull data, we also use mode and median to replace those erroneous data. We finally acquire a new table with 8000 rowsand 129 columns at the end of those preprocessing steps. Thosedata left can be used in our further research.Before starting to analyze the data, we can first look at several bar plots about the relationship between some features and the numerical value of the happiness index from 1 to 5.

MODELS

A. *Linear Model*

Linear Model is a basic model in machine learning. The basicformula of this model is:

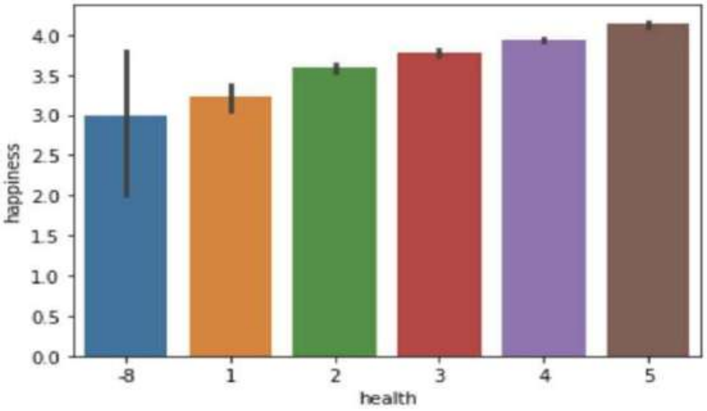


Figure 1. The numerical value of the happiness about health

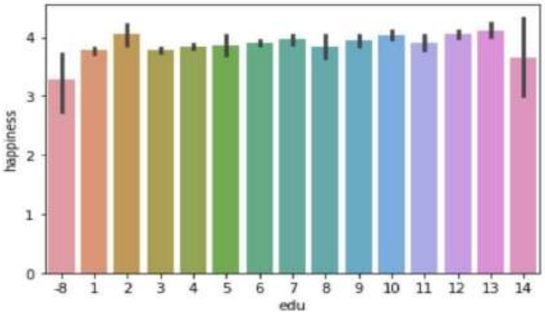


Figure 2. The numerical value of the happiness about edu

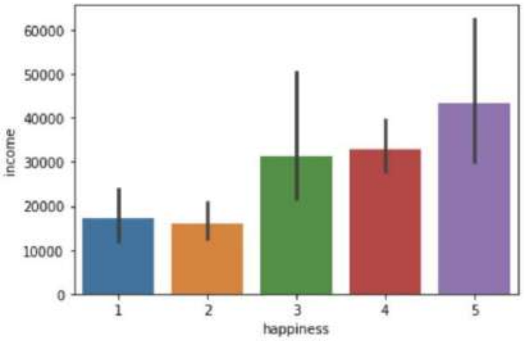


Figure 3. The numerical value of the income about happiness

In those graphs, we can generally know that after ignoring the wrong data, these features are roughly proportional to the happiness index, which means when people are healthier, richer,and receiving more education, they are assumed to be happier. These provide us a rough view of what the analyses will be.

$$Y = XB + U \qquad (1)$$

In this equation, Y is a matrix with a series of multivariablemeasurements, X is an observation matrix of independent variables, which can be a design matrix, B is a matrix containingparameters that are usually estimated, and U is a matrix containing errors.

When we try to quantify the results of linear regression, weneed the parameter R^2. It is the ratio of the sum of squares of regression to the sum of squares of total deviations, representingthe proportion in the sum of squares of total deviations that canbe explained by the sum of squares of regression.

R^2 is between 0 and 1. The larger it is, the more accurate themodel is, and the more significant the effect of the regression is.It is generally considered that the goodness of fit of the model with this value over 0.8 is relatively high.

We also need the Lasso Regression to help us find some features properly. It can reduce some unimportant feature parameters to 0, which helps us select the features which are significant.

B. Decision Tree Model

The Decision Tree is a method of machine learning. It is a tree structure. Each internal node represents a judgment on an attribute. Each branch represents the output of a judgment result,and finally, each leaf node represents a classification result. Thismodel has several benefits: it is flexible in operation, easy to understand, and able to deal with missing values. Here is a sketchmap of the Decision Tree:

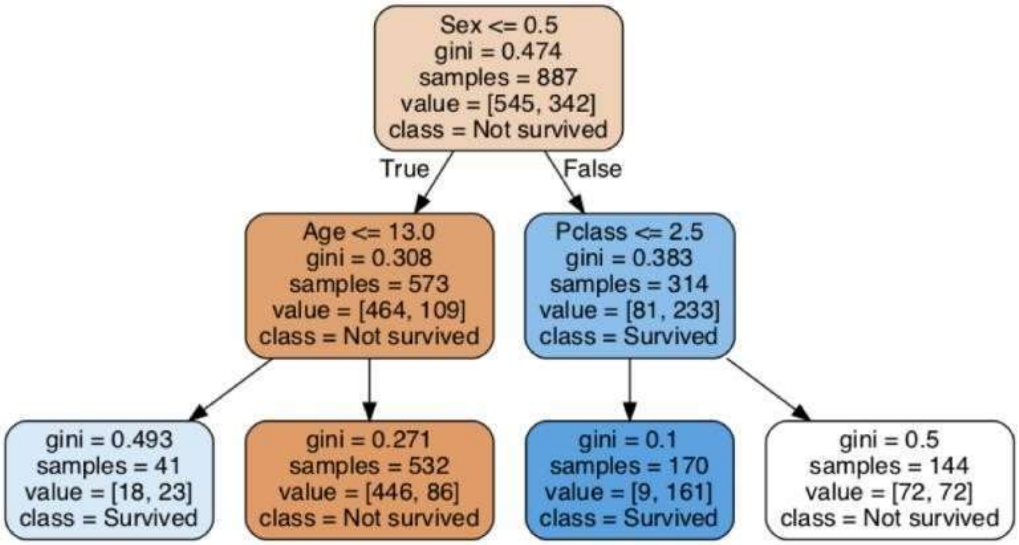
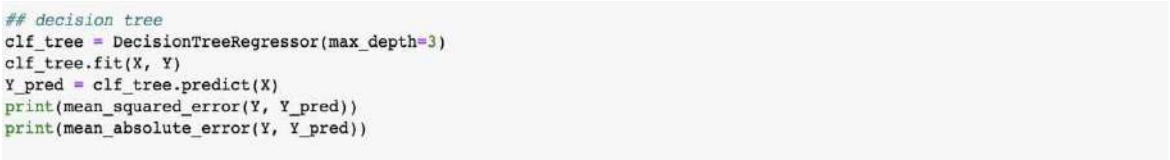


Figure 4. The Decision Tree

There are two different strategies when using the Decision Tree Model. One is regression, and the other is classification. Inour study, we are focusing on the regression. We have the codebelow:



```
## decision tree
clf_tree = DecisionTreeRegressor(max_depth=3)
clf_tree.fit(X, Y)
Y_pred = clf_tree.predict(X)
print(mean_squared_error(Y, Y_pred))
print(mean_absolute_error(Y, Y_pred))
```

In theory, we can continuously segment trees to increase accuracy. But for new data, it is easy to overfit. In this code, wehave a max depth, which means when the depth reaches it, the splitting is stopped, and the operation ends. It should be noted that this parameter should not be too large to avoid overfitting.From this code, we can test the model by seeing the result of themean squared error and the mean absolute error.

C. Random Forest and Gradient Boosting

These two models are based on the algorithm of the DecisionTree. We are now introducing two new concepts: bagging and boosting. In bagging, we select partial subsets from all data to train and model each time. Then we evaluate multiple models toget the final pattern. Bagging can not only reduce the variance of the ultimate model but also reduce the probability of overfitting. We can see how it works in the graph below.

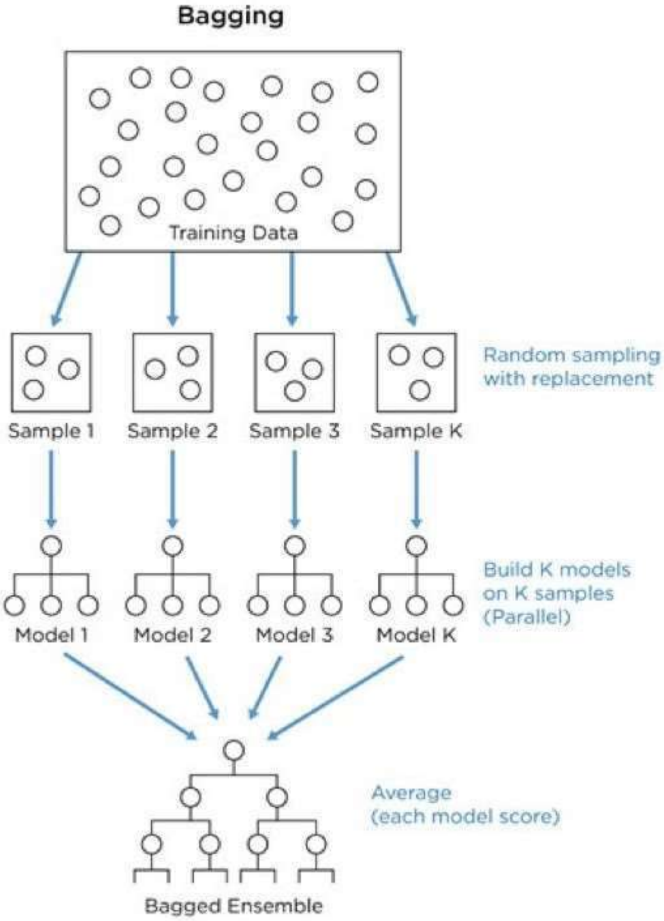

Bagging
Training Data
Random sampling with replacement
Sample 1
Sample 2
Sample 3
Sample K
Build K models on K samples (Parallel)
Model 1
Model 2
Model 3
Model K
Average (each model score)
Bagged Ensemble

Figure 5. A sketch map of how the bagging works

There is another idea which is called boosting. When utilizing boosting, we will build a series of models. Then we increase the weight of each data point that was previously miscalculated by the model. Finally, we put all models together,which can help us reduce bias. Here is a sketch map of how thisidea works.

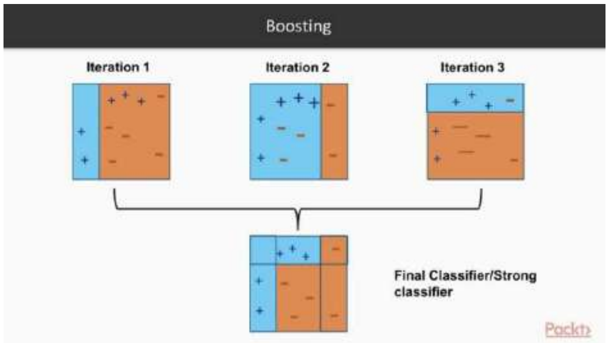
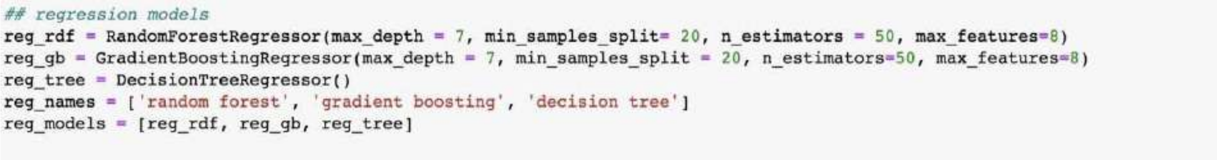


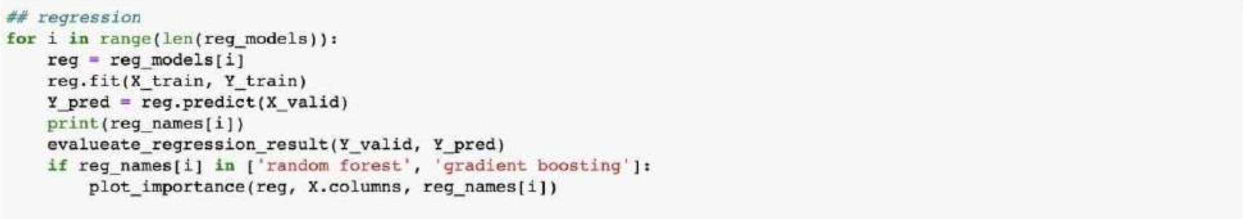
Figure 6. A sketch map of how the boosting works

In this graph, we try to split the square so that the partition'stwo parts contain the same sign. We first have Iteration 1, and itsuccessfully separates two plus signs on the left side. However,the division on the right side is not particularly clear. As a result,we need Iteration 2, which pays attention to the right part. We need Iteration 3 for the same reason. After combining those threemodels, we can get a strong classifier.

Random Forest is the model applying bagging to the Decision Tree, and Gradient Boosting is the one applying boosting to the Decision Tree. For Random Forest, we randomlyselect a part of the data at a time and put it back, and choose some features randomly each time. And for Gradient Boosting,every time we train the model to fit the gradient error before fitting. In order to prevent overfitting, we only select several dataand some features each time. We have the code below to utilizethose two models.



```
## regression models
reg_rdf = RandomForestRegressor(max_depth = 7, min_samples_split= 20, n_estimators = 50, max_features=8)
reg_gb = GradientBoostingRegressor(max_depth = 7, min_samples_split = 20, n_estimators=50, max_features=8)
reg_tree = DecisionTreeRegressor()
reg_names = ['random forest', 'gradient boosting', 'decision tree']
reg_models = [reg_rdf, reg_gb, reg_tree]
```



```
## regression
for i in range(len(reg_models)):
    reg = reg_models[i]
    reg.fit(X_train, Y_train)
    Y_pred = reg.predict(X_valid)
    print(reg_names[i])
    evalueate_regression_result(Y_valid, Y_pred)
    if reg_names[i] in ['random forest', 'gradient boosting']:
        plot_importance(reg, X.columns, reg_names[i])
```

In these two models, besides the parameter max depth, whichhas been shown in the code of the Decision Tree, we also get some new parameters like the min samples split, n estimator, max features. Those can help to make the calculation more detailed.

RESULT ANALYSES

A. Linear Regression Model

We utilize the code we obtain the following results:

Table Iii. The Result Of Linear Regression Model

Dep. Varible:	Happiness	**R-squared:**	0.258
Model:	OLS	**Adj.R-squared:**	0.246
Method:	Least Squares	**F-statistic:**	21.70
Date:	Wed, 09 Sep 2020	**Prob (F-statistic):**	0.00
Time:	23:09:46	**Log-Likelihood:**	-8552.9
No. Observations:	8000	**AIC:**	1.736e+04
Df Residuals:	7873	**BIC:**	1.825e+04
Df Model:	126		
Convariance Type:	Nonrobust		

In this model, R^2 is just 0.258, which is much lower than theexpected 0.8 or more. This shows that our fitting result is not very ideal, suggesting that many features may not be related tothe happiness index. As a result, we need to get some improvement for our Linear Regression Model. We need to findout the features that are closely related to the happiness index and use these features to fit the model. We apply the Lasso Regression and update our features selected to make linear regression, and we get the following results:

Table IV. The Result of Lasso Regression Model A

Dep. Varible:	Happiness	**R-squared:**	0.906
Model:	OLS	**Adj.R-squared:**	0.906
Method:	Least Squares	**F-statistic:**	2.567e+04
Date:	Wed, 09 Sep 2020	**Prob (F-statistic):**	0.00
Time:	23:25:49	**Log-Likelihood:**	-12894.
No. Observations:	8000	**AIC:**	2.579e+04
Df Residuals:	7997	**BIC:**	2.581e+04
Df Model:	3		
Convariance Type:	Nonrobust		

Table V. The Result of Lasso Regression Model B

	coef	**std err**	**t**	**p>|t|**	**[0.025**	**0.975]**
income	2.289e-09	5.84e-08	0.039	0.969	-1.12e-07	1.17e-07

edu	0.0738	0.004	16.923	0.000	0.065	0.082
health	0.9014	0.007	134.180	0.000	0.888	0.915

Table VI. The Result of Lasso Regression Model C

Omnibus:	974.478	**Durbin-Watson:**	1.900
Prob (Omnibus):	0.000	**Jarque-Bera (JB):**	4442.637
Skew:	0.518	**Prob (JB):**	0.00
Kurtosls:	6.501	**Cond. No.:**	1.34e+05

We add a precondition to the selection of feature X, askingthat the feature coefficient must be greater than 0, and the LassoRegression helps us eliminate collinearity. As a result, we gain a new R^2 which is 0.906, a pretty convincing number, suggesting that the model is well fitting. What's more, from thetable we receive from the code, we can clearly see that there arethree features which are closely related to the happiness index: income, education, and health. We will also get similar results inthe following study.

B. Decision Tree, Random Forest, and Gradient Boosting

The reason for why these models is put together for analysisis that they have certain commonness. As a result, we can compare them to figure out which can best predict the results. We can first have a look at the mean square error and the meanabsolute error of these three models.

TABLE VII.The Error Table

	Decision Tree	Random Forest	Gradient Boosting
Mean Square Error	0.935	0.609	0.621
Mean Absolute Error	0.672	0.549	0.557

From the table above, we can discover that for the DecisionTree Model, when having the same max depth, the results of thatare far more inaccurate than that of the other two models. The mean square error is even close to 1, suggesting that the prediction is somewhat quite different than the actual data. Theother two models have similar mean square error and mean absolute error, indicating that they have similar accuracy in predicting the results.

After discussing the accuracy of their predictions, we then focus on how the Random Forest Model and the Gradient

Boosting Model are different in the features importance. It is used to record which feature segmentation is used for each tree and how much error is reduced in this segmentation. Here are two graphs showing the features importance in those two models:

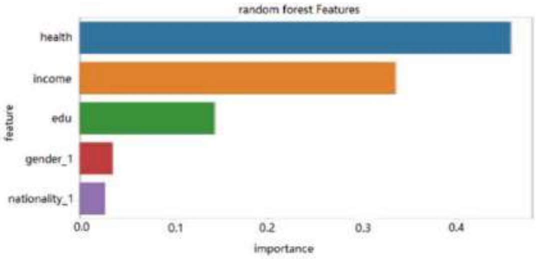


Figure 7. Random forest Features

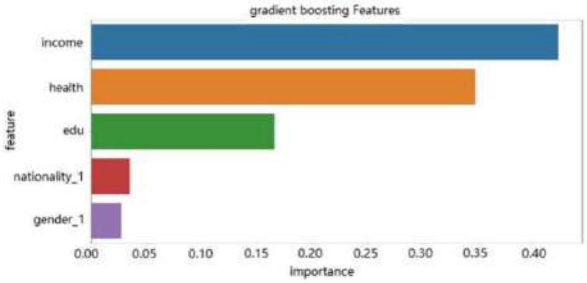


Figure 8. Gradient boosting Features

While in the Gradient Boosting Model, the first andsecond positions are adjusted. The others are almost the same. The reason for this difference may be due to the different emphasis of their algorithms. Although they are not perfectly equal, we can say that those features that are highly shared are important features related to the happiness index. In those two graphs, income, health, and education are the three main important features as we find in the Linear Regression Model. The difference is the priority of their importance. In the Random Forest Model, health is the most significant feature, income is the second important one.

IV. CONCLUSION

Our study utilizes the linear model and some tree models to analyze the features that are related to the happiness index and to make some predictions. From the analyses, we can see that themost related features are income, health, and education. In the Linear Regression Model, our pattern successfully fits the actualresult. In those tree models, we can clearly rank the features importance, showing which features are closely related to the happiness index. It is very difficult to design public policy for happiness and well-being without a strong understanding of the factors contributing to people ÿ̇s happiness and how theyinteract with policies. Our research can be beneficial to society since we know what features can affect people's happiness. Asa result, the government can put some effort in those parts. Thegovernment can improve our educational system to make morepeople get education; the government can improve the medicalinsurance system and medical facilities in order to keep peoplehealthy; the government can also provide some financial subsidies to some poor people to make them feel happy and satisfied. Those can improve our quality-of-life quality, makingus gain a sense of happiness.

E-government (also known as e-government) is the use of technical communications equipment, such as computers and theInternet, to provide public services to citizens and others in a country or region. In recent years, the use of E-government brings more and more attention in improving peoples' happiness which is a permanent part of E-governance[5]. Further, imaginative use of it can cross government, business and civil society divide and serve as an enabler of people-centered development. Using e-government strategies to identifying people's specific need and obtaining instant feedback is an effective way to improve citizens' happiness. We need to point out that our data have some limitations; our tree models have some shortcomings in prediction; we also only analyze the relationship between some features and happiness index. In thefuture research, we need to try to use more and more ideal models to fit the data.

REFERENCES

[1] Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve, eds. 2020. World Happiness Report 2020. New York: Sustainable Development Solutions Network Miao Yuanjiang.

[2] Blanchflower, David G., and Andrew J. Oswald. Happiness and the human development index: The paradox of Australia. No. w11416. National Bureau of Economic Research, 2005.

[3] Frey, Bruno S., and Alois Stutzer. "What can economists learn from happiness research?." Journal of Economic literature 40.2 (2002): 402- 435.

[4] Frey, Bruno S., and Alois Stutzer. Happiness and economics: How the economy and institutions affect human well-being. Princeton University Press, 2010.

[5] Silcock, Rachel. "What is e-government." Parliamentary affairs 54.1 (2001): 88-101.