IS THERE ANY LINK BETWEEN ECONOMIC FREEDOM AND THE HAPPINESS INDEX ACROSS COUNTRIES?

EC48W FINAL PROJECT

Çağatay Doruk Balcı
Mert Bakcacı
Fevzi Cem İz
Ece Taşan
Fatih Yavuz

Spring 2019

ABSTRACT

In this research, we tried to find a relationship between economic freedom and happiness index of countries and to construct a model which predicts the level of happiness conditional on economic freedom variables. We used three different methods of Machine Learning (Logistic Regression, K-Nearest Neighbors and Random Forest) to be able to find the best fitting model which would disclose this relationship. For several reasons (), we concluded that Logistic Regression method gives the most powerful and valid results. We hope that this work will contribute to the literature on economics of happiness.

INTRODUCTION

Happiness has been the main purpose of life throughout the history of humanity. From ancient philosophers to modern scholars, mankind has been seeking ways to define the term and trying to find means to achieve happiness on personal and universal levels. Until recent decades, researches on the area was mostly dominated by philosophers and sociologists, which leads to subjective interpretations of happiness thus a sound conclusion was never reached. With the developments in the technology, availability of mass data and the improvements in data interpretation techniques like ML, we are more confident about finding a more provable and less controversial answers to questions such as "What is the source of happiness?" and "Do economic activities have any impact on happiness?". There is an increasing number of researches to find relationship between happiness and numerous factors. This relationship especially attracts the attention of politicians and policy makers as the change of the average level of happiness in a country determines their chance of being re-elected. In this project, we have tried to find an answer to the question "Can we talk about a link between economic freedom and happiness across countries?".

DATA DESCRIPTION

We used two data sets: The first one is the "Economic Freedom of the World: 2018 Annual Report", the index published in Economic Freedom of the World by the Fraser Institute measuring the link between the effects of the policies and institutions of countries and economic freedom. Our second data set is "The World Happiness Report" which scores happiness across 155 countries according to economic production, social support, life expectancy, freedom, absence of corruption, and generosity.

Economic Freedom Data Set:

It uses five main variables:

- Size of Government: Higher government spending, taxation, and the size of government-controlled enterprises cause substitution of government decision-making for individual choice. Thus, economic freedom is reduced.
- Legal System and Property Rights: Protection of persons and their rightfully acquired property is a central element of both economic freedom and civil society. Indeed, it is the most important function of government.

- Sound Money: Inflation erodes the value of rightfully earned wages and savings. Sound money is thus essential to protect property rights. When inflation is not only high but also volatile, it becomes difficult for individuals to plan for the future and thus use economic freedom effectively.
- Freedom to Trade Internationally: Freedom to exchange—in its broadest sense, buying, selling, making contracts, and so on—is essential to economic freedom, which is reduced when freedom to exchange does not include businesses and individuals in other nations.
- Regulation: Governments not only use a number of tools to limit the right to exchange internationally, they may also develop onerous regulations that limit the right to exchange, gain credit, hire or work for whom you wish, or freely operate your business.

The World Happiness Data Set

The first World Happiness Report dates to 2012, hence it is relatively a new concept. With an increasing interest in happiness indicators by policymakers, governments and international organizations and civil society to be able to assess the progress of nations and come up with effective policies, the report continues to gain importance. The variables that the report uses to assess the countries' happiness levels are GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust and Government Corruption.

DATA CLEANING and Exploratory Data Analysis

Freedom dataset

We started by importing the packages: *pandas, numpy, matplotlib and seaborn.*
We defined the variable "freedom" as Economic Freedom of the World dataset which includes 5 main variables and their components.

The dataset has 3726 rows and 36 columns. Summary statistics:

```
freedom.describe()
```

	year	ECONOMIC FREEDOM	rank	quartile	1a_government_consumption	1b_transfers	1c_gov_enterprises	1d_top_marg_tax_rate	1_size_gover
count	3726.000000	3003.000000	3003.000000	3003.000000	3137.000000	2766.000000	3080.000000	2679.000000	3079.0
mean	2001.347826	6.519640	68.307026	2.497835	5.862426	7.672901	5.737987	5.813177	6.2
std	12.735125	1.133638	41.343417	1.118963	2.270241	2.138957	3.242377	2.654083	1.4
min	1970.000000	1.970000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.6
25%	1995.000000	5.855000	33.000000	1.000000	4.450000	6.207809	4.000000	4.000000	5.2
50%	2005.000000	6.680000	66.000000	3.000000	6.082353	8.432251	7.000000	6.000000	6.3
75%	2011.000000	7.350000	102.000000	3.000000	7.571360	9.482289	8.000000	8.000000	7.2
max	2016.000000	9.190000	162.000000	4.000000	10.000000	10.000000	10.000000	10.000000	9.9

8 rows × 34 columns

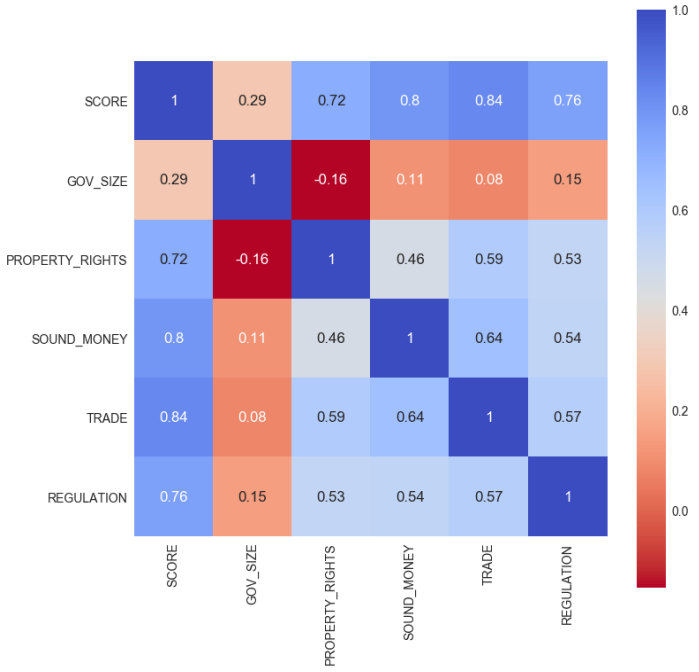
Then we checked for the number of null values in each column. Then we dropped the columns with more than 1242 null values (which is more than ⅓ of the total number of values). We renamed variable names. We changed the null variables with the median value in the columns left.

Yet, some values are still missing. The cause may stem from the country did not have this information for that year causing impossible to take median from. We can overcome this problem by taking the median of the respective quartile since countries of same quartile are similars.

Heatmap of economic freedom variables

As the correlation matrix of all components of main variables is too complex to interpret, we are going to use only the heatmap of main features.

Using only the five main features, we can see that most of them are strong correlated to the final score. Except Government Size: it doesn't have a strong relationship with the economic freedom score.



Happiness dataset:

Summary of the data:

```
In [49]: happy16.describe()
```

Out[49]:

	Happiness Rank	Happiness Score	Lower Confidence Interval	Upper Confidence Interval	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
count	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000
mean	78.980892	5.382185	5.282395	5.481975	0.953880	0.793621	0.557619	0.370994	0.137624	0.242635	2.325807
std	45.466030	1.141674	1.148043	1.136493	0.412595	0.266706	0.229349	0.145507	0.111038	0.133756	0.542220
min	1.000000	2.905000	2.732000	3.078000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.817890
25%	40.000000	4.404000	4.327000	4.465000	0.670240	0.641840	0.382910	0.257480	0.061260	0.154570	2.031710
50%	79.000000	5.314000	5.237000	5.419000	1.027800	0.841420	0.596590	0.397470	0.105470	0.222450	2.290740
75%	118.000000	6.269000	6.154000	6.434000	1.279640	1.021520	0.729930	0.484530	0.175540	0.311850	2.664650
max	157.000000	7.526000	7.460000	7.669000	1.824270	1.183260	0.952770	0.608480	0.505210	0.819710	3.837720

Then, we checked the missing values. There are no missing values. We changed the column names.

Merging two datasets:

We needed a common denominator for the two datasets to continue with our analysis. Our common link was the country variables, so we merged both of them. While merging them we used the 2016 data, as it was the most recent year.

```
In [55]: merged = pd.merge(freedom16,happy16)
```

```
In [56]: merged.head()
```

Out[56]:

	YEAR	ISO_CODE	COUNTRY	SCORE	RANK	QUARTILE	GOV_CONSUMPTION	TRANSFERS	GOV_ENTERPRISES	TOP_MARG_TAX_RATE	...	Happiness Score
0	2016	ALB	Albania	7.54	34.0	1	8.232353	7.509902	8.0	8.0	...	4.655
1	2016	DZA	Algeria	4.99	159.0	4	2.150000	7.817129	0.0	4.5	...	6.355
2	2016	AGO	Angola	5.17	155.0	4	7.600000	8.886739	0.0	9.5	...	3.866
3	2016	ARG	Argentina	4.84	160.0	4	5.335294	6.048930	6.0	4.0	...	6.650
4	2016	ARM	Armenia	7.57	29.0	1	7.264706	7.748532	8.0	5.0	...	4.360

5 rows × 41 columns

```
In [57]: merged.columns
```

Out[57]:

```
Index(['YEAR', 'ISO_CODE', 'COUNTRY', 'SCORE', 'RANK', 'QUARTILE',
       'GOV_CONSUMPTION', 'TRANSFERS', 'GOV_ENTERPRISES', 'TOP_MARG_TAX_RATE',
       'GOV_SIZE', 'IMPARTIAL_COURTS', 'PROTEC_PROP_RIGHTS', 'MILITARY_INTERF',
       'INTEGRITY_LEGAL_SYST', 'GENDER_ADJUSTMENT', 'PROPERTY_RIGHTS',
       'MONEY_GROWTH', 'STD_INFLATION', 'INFLATION', 'FOREIGN_CURRENCY',
       'SOUND_MONEY', 'TARIFFS', 'BLACK_MARKET', 'CONTROL_MOVEMENT', 'TRADE',
       'CREDIT_MARKET_REG', 'LABOR_MARKET_REG', 'REGULATION', 'Region',
       'Happiness Rank', 'Happiness Score', 'Lower Confidence Interval',
       'Upper Confidence Interval', 'Economy (GDP per Capita)', 'Family',
       'Health (Life Expectancy)', 'Freedom', 'Trust (Government Corruption)',
       'Generosity', 'Dystopia Residual'],
      dtype='object')
```

The majuscule variable names are from freedom dataset and minuscules from happiness.

We needed a way to narrow down the set of variables we could do future selection from. For this we took the correlations of every variable with the happiness score. Then we added any variable with more than 0.3 to our new dataframe, high_corr.

```
In [72]: main_2.corr()["Happiness Score"].sort_values(ascending=True).head(4)

Out[72]: TRANSFERS           -0.528596
         GOV_CONSUMPTION     -0.503665
         GOV_SIZE            -0.291076
         TOP_MARG_TAX_RATE   -0.199507
         Name: Happiness Score, dtype: float64

In [112]: main_2.corr()["Happiness Score"].sort_values(ascending=False).head(15)

Out[112]: Happiness Score         1.000000
          PROPERTY_RIGHTS         0.634884
          PROTEC_PROP_RIGHTS      0.591973
          MILITARY_INTERF         0.554763
          INTEGRITY_LEGAL_SYST    0.495799
          TRADE                   0.493676
          IMPARTIAL_COURTS        0.424768
          FOREIGN_CURRENCY        0.421851
          GOV_ENTERPRISES         0.405362
          SOUND_MONEY             0.402104
          CONTROL_MOVEMENT        0.395403
          TARIFFS                 0.377800
          REGULATION              0.351171
          GENDER_ADJUSTMENT       0.329706
          CREDIT_MARKET_REG       0.282291
          Name: Happiness Score, dtype: float64

In [113]: high_corr_vars = ["Happiness Score","TRANSFERS","GOV_CONSUMPTION","PROPERTY_RIGHTS","PROTEC_PROP_RIGHTS","MILITARY_INTERF","INTE
                ,"TRADE","IMPARTIAL_COURTS","FOREIGN_CURRENCY","GOV_ENTERPRISES"]
          high_corr_vars += ["SOUND_MONEY","CONTROL_MOVEMENT","TARIFFS","REGULATION","GENDER_ADJUSTMENT","GOV_SIZE"]

In [135]: high_corr = main_2.loc[:,high_corr_vars]

In [140]: print(np.mean(high_corr["Happiness Score"]))
          sns.distplot(high_corr["Happiness Score"],bins=20)

          5.414888888888891
```

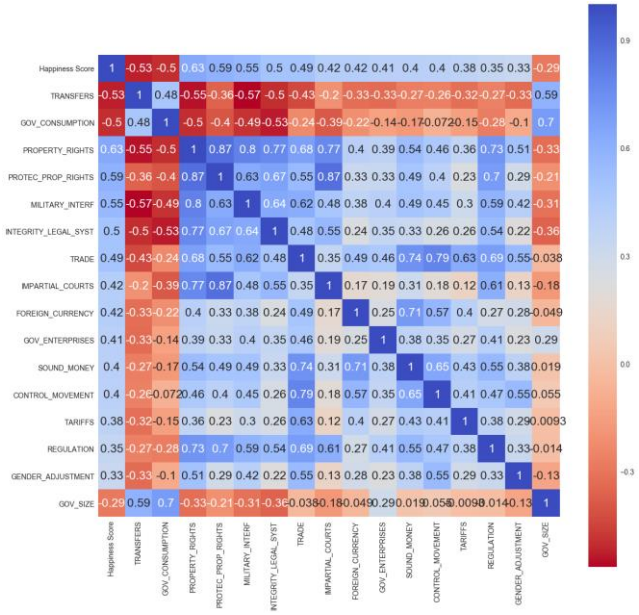
Since our happiness score was not a binary variable, we could use it as the dependent variable for our linear regression. However, we also wanted to use classification methods in our study as well. So we first look at the distribution graph of the happiness score variable.

As we can see above the peak is at 5, and the mean score of the Happiness Score is 5.41. In order to create our binary variable for happiness, we created a new variable “Happy”. If a country’s happiness score is above 5, its value is 1. Else, it’s zero.

Correlation matrix of variables:

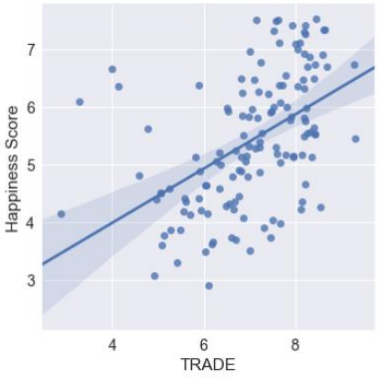
```
In [116]: plt.figure(figsize=(20,20))
          sns.heatmap(high_corr.corr(),  square=True, annot=True, cmap='coolwarm_r')

Out[116]: <matplotlib.axes._subplots.AxesSubplot at 0x29ebeecfb70>
```



Property rights is highly correlated with protec_property rights, military interference, integrity_legal, impartial_ courts.

We were concerned about possibility of multicollinearity in linear regression model and thought about dropping protec_property_rights military_interference and integrity_legal, since their correlation with property_rights variable was more 0.7. However, since this is a machine learning project and predictive power is more important than interpretation of the variables (as in an econometrics project), we decided to keep them.



We can see from the graphs above that Happiness Score has a positive relationship with both property rights and trade variables.

Linear Regression

```
In [123]: predictions = reg1.predict(X_test)

In [124]:
from sklearn import metrics
print("MSE: ", metrics.mean_squared_error(y_test,predictions))
print("R^2: ",metrics.r2_score(y_test,predictions))
```

```
MSE:  0.6191782766415055
R^2:  0.5026294144617753
```

*

	coefficients
TRANSFERS	-0.089920
GOV_CONSUMPTION	-0.169366
PROPERTY_RIGHTS	0.240651
PROTEC_PROP_RIGHTS	0.401891
MILITARY_INTERF	-0.033290
INTEGRITY_LEGAL_SYST	-0.044552
TRADE	-0.160979
IMPARTIAL_COURTS	-0.188164
FOREIGN_CURRENCY	0.026394
GOV_ENTERPRISES	0.048236
SOUND_MONEY	0.001927
CONTROL_MOVEMENT	0.103832
TARIFFS	0.153182
REGULATION	-0.296468
GENDER_ADJUSTMENT	0.603295
GOV_SIZE	0.134693

We used all the variables that were in the high_corr dataframe, and their coefficients can be seen above. We can see that the variables property_rights, protect_property_right, regulation and gender_adjustment have the most economical effect on our predictions. These are in line with our predictions since they were also the variables that correlated with happiness score the most.
As a result of linear regression, we get R-squared of 0.50 and MSE 0.61. This shows us that the variables we used in our model can explain 50% of the dependent variable's variance. We think this result is successful and shows that conditions for economic freedom has a strong relationship with that country's happiness.

CLASSIFICATION METHODS

Our Approach Towards the Classification Procedure

For our classification problems we had two problems we needed to solve to obtain the best results. First one was choosing the best set of variables and the second one was to pick the best (hyper)parameters for our models.

To solve these problems our approach was similar to grid search.
We had a list of highly correlated variables, sorted from highest correlation to the least. We wrote a loop to evaluate our fit with every set of variables, increasing the size of our set by one in every

iteration. In every iteration there was also an inner loop to find the best (hyper)parameter value. We will explain them in the following sections.

Logistic Regression:

```
features = []
param_grid = {'C':[.01,.03,.1,.3,1,3,10]}
for i in range(len(high_corr_vars)-1):
    features = high_corr_vars[0:i+1]

    X_train, X_test, y_train, y_test = train_test_split(high_corr[features],high_corr["Happy"], test_size=0.3, random_state=101)

    kf = KFold(len(X_train), n_folds=5)

    logreg=LogisticRegression()
    logreg.fit(X_train,y_train)

    gs_logreg = GridSearchCV(logreg, param_grid=param_grid, cv=kf)

    gs_logreg.fit(X_train, y_train)
```

We used the feature selection method explained previously. In the inner loop of the feature selection method we tried to optimize the "C" hyperparameter, trying different values ranging from 0.01 to 10. Our method automatically selected the "C" value providing the best accuracy score.

The best feature set and regularization parameter can be seen below.

```
Optimal Regularization Parameter C:10
['TRANSFERS', 'GOV_CONSUMPTION', 'PROPERTY_RIGHTS', 'PROTEC_PROP_RIGHTS', 'MILITARY_INTERF', 'INTEGRITY_LEGAL_SYST', 'TRADE',
'IMPARTIAL_COURTS', 'FOREIGN_CURRENCY']
Average accuracy score on cv (KFold) set: 0.851
Accuracy score on test set is: 0.829
```

We find optimal result as follows:

We found out that when we use variables named transfers, government consumption, property rights, protection of property rights, military interference, integrity of the legal system, trade, impartial courts and foreign currency, we reached the highest accuracy score on test set.

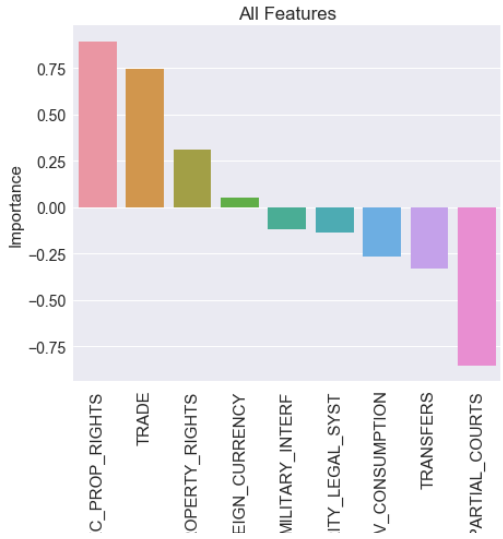
The average accuracy score on KFold set is 0.808 and accuracy score on test set is: 0.805.

```
preds=logreg.predict(X_test)
print(classification_report(y_test,preds))
```

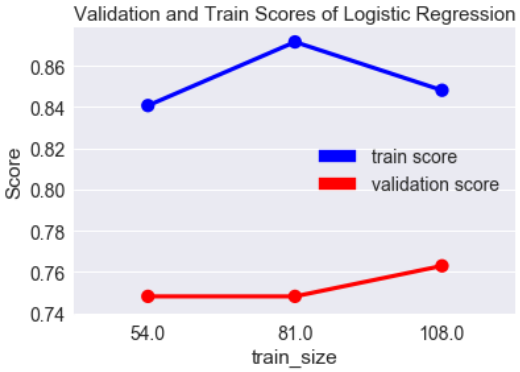
	precision	recall	f1-score	support
0	0.76	0.76	0.76	17
1	0.83	0.83	0.83	24
avg / total	0.80	0.80	0.80	41

Evaluation of the Logistic Regression Model

As can be seen, our precision (the ratio of correctly predicted positive observations to the total predicted positive observations) and recall (the ratio of correctly predicted positive observations to the all observations in actual true class) are both 0.80 which is pretty good. Thus, the f1 score which is the weighted average of the precision and recall, showing the overall performance of the model, is also 0.80.



Looking at the importance of the features determining the economic freedom, protection of property rights and trade are the most important ones positively affecting the happiness level while transfers and impartial courts have negative effects.



We can see in the right the validation and train scores of the logistic regression model. This shows us that if we had more data our model could provide better results. In this case, more data would mean more countries, which is not possible. However this is still a good indicator that our model is healthy and provides better results with more data.

K Nearest Neighbours:

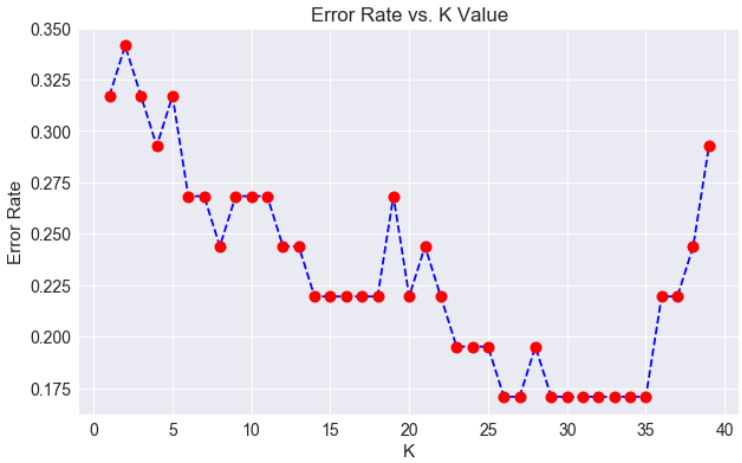
We chose K Nearest Neighbours method, which is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in, as our second method as it is a widely used method for classification. We will compare the power of this prediction with the logistic regression results later.

We used the same method for feature selection and this time in our inner loop, we tried to find the best K value, providing the best accuracy score.

```
Ideal K:25
['TRANSFERS', 'GOV_CONSUMPTION', 'PROPERTY_RIGHTS', 'PROTEC_PROP_RIGHTS']
Average accuracy score on cv (KFold) set: 0.777
Accuracy score on test set is: 0.805
```

The test with the highest accuracy score (0.805) uses transfers, property rights and the protection of property rights as variables. We constructed a KNN model with the above-mentioned variables and K values from 1 to 40.

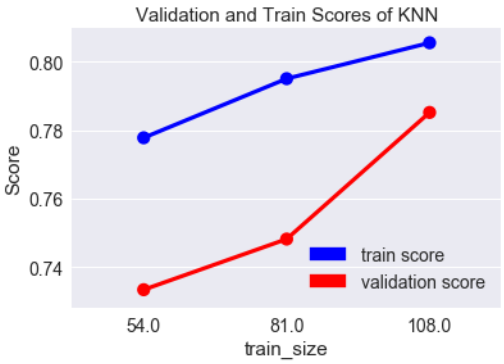
The process which we chose the best K value can be seen below. As can be seen from the graph, 25 provides the best error rate, thus it's our k value.



Evaluation of the K-Nearest Model:

	precision	recall	f1-score	support
0	0.92	0.61	0.73	18
1	0.76	0.96	0.85	23
avg / total	0.83	0.80	0.80	41

As can be seen, our precision rate is 0.83 and recall is 0.80 which are pretty good. Thus, the average f1 score is 0.80. Average accuracy score on cv (KFold) set is 0.777. The highest accuracy score on test set is 0.805.



In the left, we can see the train and validation scores of our KNN model, which just as the model before shows us that our model is healthy and provides better results with more data.

Random Forest Classifier:

Random forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. We obtained the highest accuracy scores with transfers, government consumption, property rights, protection of property rights, military interference, integrity of the legal system, trade, impartial courts and foreign currency variables.

Average accuracy score on cv (KFold) set: 0.746

Accuracy score on test set is: 0.707

Evaluation of the Random Forest Model

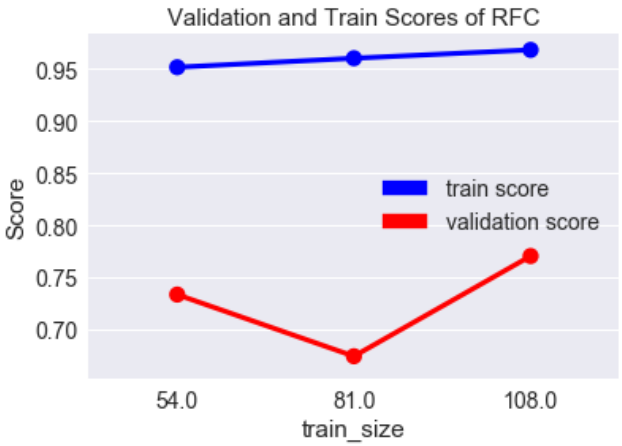
```
preds = rfc.predict(X_test)
print(classification_report(y_test,preds))
```

```
             precision    recall  f1-score   support

          0       0.62      0.62      0.62        16
          1       0.76      0.76      0.76        25

avg / total       0.71      0.71      0.71        41
```

As can be seen, our precision rate is 0.71 and recall is 0.71. Thus, the average f1 score is again 0.71, even though this result is considerably successful, it is below the results conceived from our previous models.



Validation and train scores show a converging trend here, which increases the trust in our model. However, we'd like the mention here that with different test sets, the shape of the graph changes and raises doubt about its health.

Comparison of the methods used

The average accuracy score on KFold set we obtained using logistic regression method is 0.808 while the average accuracy score KFold set of the KNN method was 0.777 and that of random Forest method was 0.746. It seems that there is not much difference between the scores of KNN and Random Forest models, however that of Logistic Regression model is better than others. We can say that the logistic regression method makes more accurate predictions compared to KNN and Random Forest method predictions.

Random forest classifier also provides some inconsistency in its validation scores, therefore is not as reliable as KNN and logistic regression.

Conclusion

In our research, we tried to predict a relationship between the determinants of the economic freedom and happiness index of countries. To assess the value of a deeper investigation on this relationship we made a linear regression analysis and found out that the explanatory power of those variables for the estimation of the happiness index were far from being negligible (R-squared: 0.50). This showed us that happiness index can be predicted to a good extent using those variables, we continued our research by trying to find the best fitting method for this prediction. We used three methods: Logistic Regression, K-Nearest Neighbours and Random Forest. Our main point of comparison was their average accuracy score on KFold test and their validation scores to test the soundness of the models. The lowest accuracy belonging to the Random Forest method, the winner of this comparison was the Logistic Regression Model with an average K-fold accuracy score 0.808 and a consistent validation score.