12/6/2021

Money can buy happiness afterall*!

to some extent.

PAYAL MOHAPATRA & AKASH PANDEY

Akash Pandey – 3334115 | Payal Mohapatra - 3334086

PROJECT OBJECTIVE

Happiness is a subjective emotion and means very different to everyone. However, it is without a doubt that it is one of the most sought-after things in the world. In this project, we attempt to objectively analyze the causality between happiness and country-wide metrics like GDP, social security, life expectancy at birth, etc.

RESOURCES AND DATA PREPARATION

We used the official datasets provided by World Happiness Report, which are pre-processed (parsed into columns and updated with missing data points) in Kaggle. We have simplified the labels for indexing purposes and combined the data from 2015 and 2016 as our training dataset. We have used the model trained on these years to predict the happiness scores of data from 2017 and assess the model performance.

```
df_2015 = pd.read_csv("Dataset/2015.csv")
df_2016 = pd.read_csv("Dataset/2016.csv")

## Pre-process data :: Append the years and generate a complete dataset
h_cols = ['Country', 'Region', 'GDP', 'Family', 'Life', 'Freedom', 'Generosity', 'Trust']
def prep_frame(df_year, year):
    df = pd.DataFrame()
    # Work around to load 2015, 2016, 2017 data into one common column
    target_cols = []
    for c in h_cols:
        target_cols.extend([x for x in df_year.columns if c in x])
    df[h_cols] = df_year[target_cols]
    df['Happiness Score'] = df_year[[x for x in df_year.columns if 'Score' in x]]
    # Append year and assign to multi-index
    df['Year'] = year
    df = df.set_index(['Country', 'Year'])
    return df
df = prep_frame(df_2015, 2015)
df = df.append(prep_frame(df_2016, 2016), sort=False)
```

METHODOLOGY

VISUAL ANALYSIS

We performed a quick visual analysis to assess if there is any direct relation between the individual features and the happiness scores. For majority of the features across the years we found some sort of a linear trend. Below is a figure showing the trend between the six features and happiness scores in the year 2015.

Fig.1. Trends between all the six reported features and happiness scores for the year 2015. The horizontal axis depicts happiness sore and the vertical axis the respective feature score.

MODEL DEVELOPMENT

This is a regression problem. We are modeling the relationship between the metrics reported and the happiness score so that we can predict an 'unseen' country's happiness score based on these metrics.

For our development we have used scikit-learn libraries. Based on the visual analysis, we have chosen Linear Regression (LR) and Random Forest (RF) approaches to build our model. We experimented with various methodologies to arrive at the best-performing model. We considered fitting a model to the entire dataset with and without feature reduction and models fitting to a particular geographic region and then using these individual models to predict in the entire input space. We use root mean squared error (RMSE) between the actual and predicted outputs to judge the model quality. More details in the upcoming section.

I. SINGLE MODEL APPROACH

A. Without Feature Reduction

This is a vanilla approach where we use the two models to fit the training data and use a validation set to predict the accuracy. In this case the although the Random Forest regressor and linear regressor have very close validation RMSE , 0.53 and 0.55 respectively. We chose the Random Forest regressor and found that its test accuracy (test RMSE = 0.54) is superior to the test accuracy of Linear Regressor (test RMSE = 0.67).

B. With Feature Reduction

For feature reduction, we perform two steps.
Step 1 : Perform cross correlation between input features and output and pick the top 4 .
Step 2: Perform cross correlation between selected input features from Step 1 and remove the redundant one is any. Our threshold was 0.8.

```
#################################################################################
## Use kbest features function for regression to get 4 best features for the whole data
features = np.asarray(df.columns)
select_features = SelectKBest(f_regression, k=4)
X_new = select_features.fit_transform(X_train, Y_train)
filter = select_features.get_support(indices = True)
print("Best Selected features:")
print(features[filter[0]])
print(features[filter[1]])
print(features[filter[2]])
print(features[filter[3]])
print('\n')

df_kbest = pd.DataFrame()
df_kbest = df[[str(features[filter[0]]), str(features[filter[1]]), str(features[filter[2]]), str(features[filter[3]]), 'Happiness Score']]
print('\n ******************Data Frame with 4 best features********************** \n')
print(df_kbest)

spearman_cormatrix= df_kbest.corr(method='spearman')
print('\n ******************Spearman Correlation********************** \n')
print(spearman_cormatrix)

plt.figure(1)
sns.heatmap(spearman_cormatrix, vmin=-1, vmax=1, center=0, cmap="viridis", annot=True)
plt.show()

### Drop Life since GDP and Life highly correlate (0.8)

df_kbest_corr = pd.DataFrame()
df_kbest_corr = df_kbest.drop(['Life'], axis='columns')
print('\n ******************Data Frame with 3 best features********************** \n')
print(df_kbest_corr)
#################################################################################################################
```

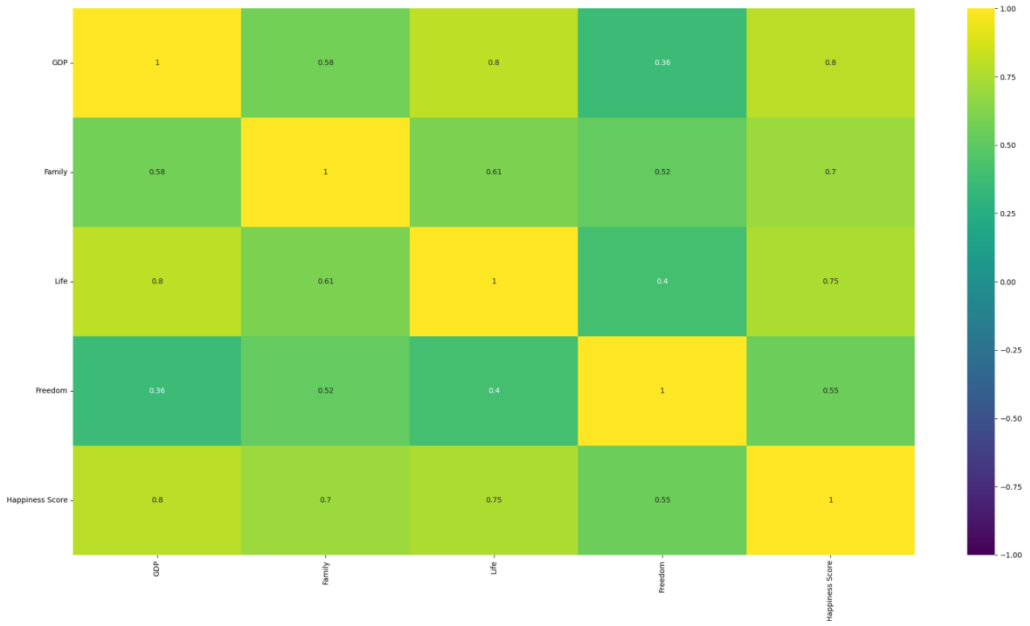

Fig.2. Heat Map showing correlation between the top 4 input features. In this case *Life* and *GDP* have a correlation of 0.8. And GDP has a higher correlation with the Happiness score, so we have chosen to drop the feature *Life* in this case.

The top 3 feature selected are – ***GDP, Family and Freedom***. Using feature reduced dataset we observe that both the regressors perfome equally well (LR with validation RMSE or 0.64 and RF with 0.69). Similar trend was observed in the test data too test RMSE of LR is 0.72 and RF is 0.78.

However, both RF and LR have lower RMSE score overall when trained without feature reduction. This brings us to our next approach of modelling based on regions and feature selction to improve accuracy.

II. REGIONAL MODELS BASED ENSEMBLE APPROACH

Motivation : We observed that countries in regions like Australia, North America and Western Europe always ranked above 40 amongst the 158 countries with a few exceptions, whereas countries in sub-saharan Africa ranked lower than 70. This probably means unique patterns of causality between happiness and the features regionally. This was our motivation to build regional models with feature reduction and ensemble them to make our prediction.

In this approach, we are building models for each regions based on the top four features that contribute to the happiness score.
Step 1 : Segment data into regional dataframes.
There are 10 geographical regions we can divide the data into.

Step 2 : Carry out feature reduction for each region. (Following similar approach as Fig.2.)
Table 1, shows the frequency at which the features occur as one of the top four correlated with happiness score in the 10 regions.

Table 1. Tabulation of the frequency at which each feature appears in top 4 contributors of happiness. For example, GDP ranked 1 for 8 regions, Trust ranked 4 in 5 regions and so on.

Features	**Ranked 1**	**Ranked 2**	**Ranked 3**	**Ranked 4**
GDP	8/10			
Family	1/10	8/10		
Life	1/10	1/10	5/10	
Freedom		1/10	4/10	2/10
Generosity			1/10	3/10
Trust				5/10

Table 1, highlights one of the most interesting observations in this analysis, .i.e., the only two regions that did not list GDP as the top or even one of the top 4 contributors of happiness are the countries with the highest GDP per capita. Shown in Fig.3. are the GDP per capita of all the regions for the year 2015 and 2016. North America and Australia & New Zealand have the highest GDP amongst the others and do not report GDP to be their top contributor to happiness unlike others. This lets us draw a more philosophical conclusion, that probably money can buy happiness but only till a certain extent. Beyond which it is Family, quality of Life and Freedom, are what that determine a country's happiness.

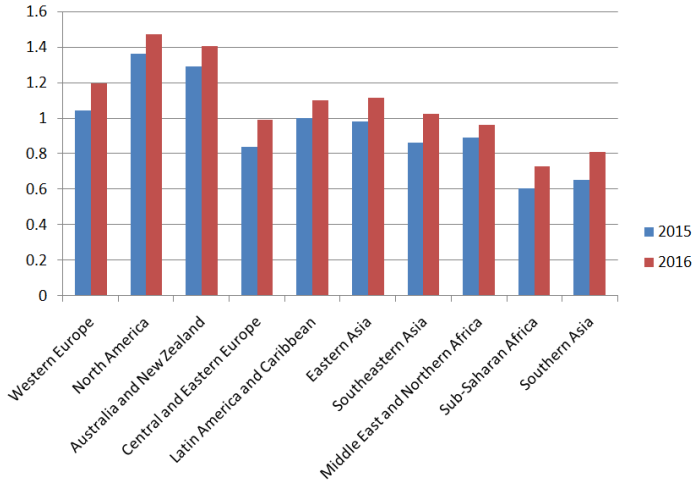

Fig.3. GDP per capita wrt regions

Step 3 : Select model with lowest RMSE for each region and save.

```
## Dump the model with lowest RMSE in a pickle file
    pkl_filename = "regional_model_{0}.pkl".format(i)
    if (RF_RMSE < LR_RMSE) :
        print('Choosing RF model for Region',i)
        with open(pkl_filename, 'wb') as file:
          pickle.dump(RF_model, file)
    else:
        ## Append LR to model
        print('Choosing LR model for Region',i)
        with open(pkl_filename, 'wb') as file:
          pickle.dump(LR_model, file)
```

Step 4 : Build an ensemble of the trained models to predict the happiness score of a country based on the region they belong to.

```
df_western_europe = df[df['Region'] == 'Western Europe']

df_north_america = df[df['Region'] == 'North America']
            .
            .
            .
df_south_asia = df[df['Region'] == 'Southern Asia']

for i in range (10) :
    if i == 0 :
        df_current = df_western_europe
        df_current = df_western_europe.drop(['Life','Generosity'], axis ='columns')
    elif i == 1:
        df_current = df_north_america
        df_current = df_north_america.drop(['GDP','Family'], axis = 'columns')
            .
            .
            .
            .
    elif i == 9:
        df_current = df_central_eastern_europe
        df_current = df_central_eastern_europe.drop(['Generosity','Life'], axis = 'columns')

    X_test = df_current.drop(['Happiness Score', 'Region'], axis='columns')
```

```
Y_test = df_current['Happiness Score']
pkl_filename = "regional_model_{0}.pkl".format(i)
print('Reading file', pkl_filename)
## Load from file
with open(pkl_filename, 'rb') as file:
    pickle_model = pickle.load(file)
# Calculate the accuracy score and predict target values
Y_predict =  pickle_model.predict(X_test)
test_RMSE = np.sqrt(sklearn.metrics.mean_squared_error(Y_predict,Y_test))
print('Test RMSE at iteration', i ,' is', test_RMSE)
test_RMSE_hist.append(test_RMSE)

test_RMSE_avg += test_RMSE
```

We observed that this model has the lowest test RMSE (0.39).

RESULTS AND SUMMARY

- The data was extracted from world happiness report from 2015-2017.
- Using 2015 and 2016 data as the training set, an optimized model was developed to predict the happiness score of countries. All developed models are tested on 2017 dataset.
- Model was trained using linear regression and random forest regression with and without feature reduction on the whole dataset as well as on the region-wise dataset.
- RMSE of all the models on the test dataset (2017) is shown in Fig.4. and it can be noted that Regional Ensemble model has outperformed all other models.

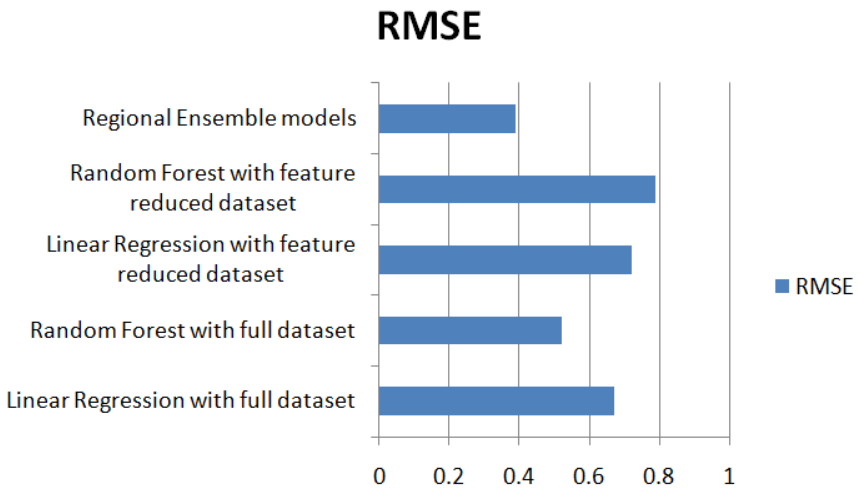

Fig.4. RMSE of all the experimented models on test data (2017 data).

Conclusion

Overall it can be concluded that GDP, Life expectancy, family and freedom are the strongest contributors to the happiness with below evidences:

- Referring to the scatter plot of Fig.1, it can be observed that generosity and trust does not show any trend with respect to the happiness score.
- Referring to table 1, it can noted that in fact GDP, life expectancy, family and freedom occur most frequently as the top ranked contributors to the happiness score.

All the source code can be accessed at https://github.com/payalmohapatra/The-Happiness-Project.git .
Video presentation of the work can be found at https://youtu.be/wja6_fjiVCo .