CE 263 – Scalable Spatial Analytics

Final Project
Happiness Model
Fall 2015

Team members:
Harry Durbin

Professor:
Alexei Pozdnoukhov

Contents

Tables

Figures

1 Introduction

This section describes why the project was selected, goals, and the dataset used to create the model.

1.1 Why does this matter?

According to the Dalai Lama XIV, "the very purpose of our life is to seek happiness." Many great thinkers and philosophers throughout history have expressed similar views. More debatable is the question of how to go about seeking happiness. Regardless of how it is attained, it is important on an individual level, and also a societal level. If a society is happy, it could be assumed people would be less likely to commit crimes and cause other harm to each other, happy employees would likely work more effectively, and happy couples would likely raise happy children.

While it is clear that happiness is important for each of us and our communities, it is less clear how individual happiness and societal happiness influence each other. Perhaps happiness is partially infectious among a society—people will become more likely to increase happiness if they are immersed in other people with a good mood. If so, this would have an effect of creating localization of happiness levels geographically. Another possibility is happiness levels could be influenced by other external factors. This too may create localization of happiness levels geographically. Cultural values could also play a role, if people are taught to always maintain a positive attitude—to say one is happy irrespective of how unhappy one actually is. Alternatively, it could be possible that happiness level is more internally dependent, not relying on any type of external circumstances. If this is the cause, happiness levels across the world should be more or less random.

1.2 Project Goals

The purpose of this project is to:

- Evaluate happiness levels in countries across the world to see the spatial dependency.
- Create a model to try to predict happiness levels for individuals and averaged for countries.
- Find the features that correlate most with happiness levels.

1.3 Dataset

For this project, a dataset provided by World Values Survey was used. The dataset includes a collection of questionnaire results for over 86 thousand persons located throughout 60 countries, which was collected between 2010 to 2014 sixth wave (World Values Survey Association, 20150418) . The questionnaire asked hundreds of questions, one of which was, how happy do you consider yourself currently?

- 1 – very happy
- 2 – rather happy
- 3 – not very happy
- 4 – not at all happy

A possible issue with this data is that the happiness levels are not scientifically assigned—they are purely subjective as to how the respondent feels at that moment.

2 Approach and Methods

The following figure shows the methodology followed:

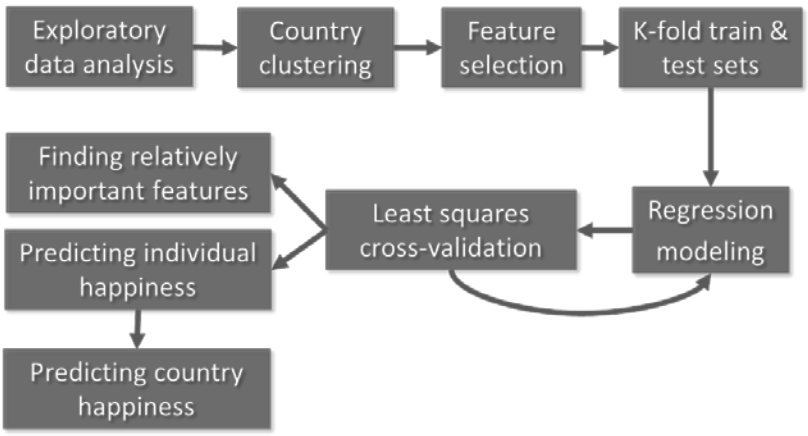

Figure 2-1: Approach Methodology

The particular steps shown in the figure above are discussed in more detail below.

2.1 Exploratory Data Analysis (EDA)

The purpose of the exploratory analysis was to get a sense of the distribution of surveys collected throughout each country. It was unknown whether a majority surveys were collected in one or two countries. Also, it is useful to see the typical happiness levels of countries across the world. A breakdown of happiness levels for each country is shown in the figure below.

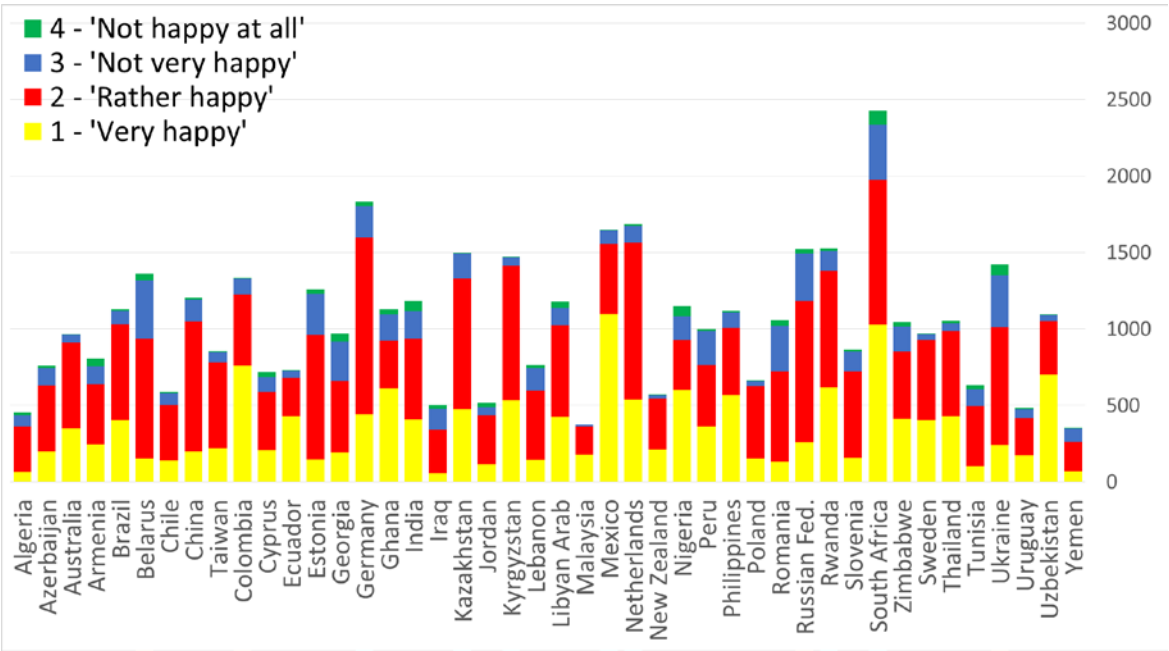

Figure 2-2: Exploratory Analysis of Survey Results

As shown in the figure, a huge majority of people in countries all of the world consider themselves either “very happy” (1) or “rather happy (2).

Interestingly, the country with the highest percentage of very happy people was Uzbekistan, followed closely by Mexico. On the other hand, Iraq appears to be the country with the largest percentage “not very happy” (3), which is understandable due to the past and continuing conflict.

2.2 Clustering

As previously noted, the cultural values could factor in to how people respond about their happiness levels. For these reason, I wanted to cluster countries into geographic regions with the assumption that countries located close to each other would be more likely to share similar cultural values. The cluster group number is then used as an additional feature when training the model. Initially, I was thinking to create a cluster for each continent, but I decided to increase it to 10 cluster groups. The quantity of 10 is an arbitrary number that was selected for obtaining clusters using MiniBatch KMeans. The following cluster groups are presented below:

- **Group 0:** ['Mexico']
- **Group 1:** ['China' 'Hong Kong' 'Japan' 'Malaysia' 'North Korea' 'Philippines', 'Singapore' 'Taiwan' 'Thailand']
- **Group 2:** ['Armenia' 'Azerbaijan' 'Bahrain' 'Cyprus' 'Egypt' 'Georgia' 'Iraq' 'Jordan' 'Kuwait' 'Lebanon' 'Palestinian Territory' 'Qatar' 'Turkey' 'Yemen']
- **Group 3:** ['Rwanda' 'South Africa' 'Zimbabwe']
- **Group 4:** ['India' 'Kazakhstan' 'Kyrgyzstan' 'Pakistan' 'Russian Federation' 'Uzbekistan']
- **Group 5:** ['Argentina' 'Chile' 'Uruguay']
- **Group 6:** ['Algeria' 'Ghana' 'Libyan Arab Jamahiriya' 'Morocco' 'Nigeria' 'Spain' 'Tunisia']
- **Group 7:** ['Belarus' 'Estonia' 'Germany' 'Netherlands' 'Poland' 'Romania' 'Slovenia' 'Sweden' 'Ukraine']
- **Group 8:** ['Australia' 'New Zealand']
- **Group 9:** ['Brazil' 'Colombia' 'Ecuador' 'Peru' 'Trinidad and Tobago']

2.3 Feature Selection

The questionnaire used in the survey is extremely comprehensive, covers a wide range of topics, and has approximately 250 questions. The survey is included as an Appendix for reference. After reviewing the questionnaire, several features of interest were selected. In addition to the survey questions, three additional features were added to the data based on the country (latitude and longitude of country center, and the country cluster group as discussed above). All of the features used are listed below:

- Country (V2)
  - What country are you located? (list of country codes provided)
- Happiness (V10)
  - Would you say you are:
    1 very happy, 2 rather happy, 3 not very happy, 4 not at all happy

- Health (V11)
  - How would you describe your state of health these days?
    1 very good, 2 good, 3 fair, 4 poor
- Marital Status (V57)
  - Are you: 1 married, 2 living together, 3 divorced, 4 separated, 5 widowed, 6 single
- Children (V58)
  - How many children do you have?
    0 through 8 (8 if 8+)
- Purpose (V143)
  - Do you think about the meaning and purpose of life?
    1 often, 2 sometimes, 3 rarely, 4 never
- Prayer (V147)
  - How often do you pray?
  - 1 several times per day, 2 once per day, 3 several times per week, 4 only at services, 5 only holy day, 6 once per year, 7 less, 8 never
- Employment (V229)
  - Are you employed now? How many hours per week?
    YES: 1 full time, 2 part time, 3 self-employed
    NO: 4 retired, 5 housewife, 6 student, 7 unemployed, 8 other
- Intellectual Work (V231)
  - Are your work tasks manual or intellectual?
    1 = mostly manual, 10 = mostly intellectual
- Creative Work (V232)
  - Are your work tasks routine or creative?
    1 = mostly routine, 10 = mostly creative
- Social Class (V238)
  - Would you describe yourself in:
    1 upper class, 2 upper middle class, 3 lower middle class, 4 working class, 5 lower
- Sex (V240)
  - Gender: 1 = male , 2 = female
- Age (V242)
  - How old are you? 00-99
- Education (V248)
  - Highest educational level attained?
    1 = no formal education, 9 = university level w/ degree
- Town (V253)
  - Size of town: 1 = under 2,000 , 8 = 500,000 and more

- Longitude (of the country centroid)
- Latitude (of the country centroid)
- Cluster – cluster group determined by MiniBatch KMeans

2.4 Model Training

As a first step, the raw data was cleaned to exclude all data that had missing information in the features noted above, and for happiness levels. Then, the raw data was divided into a train and test dataset using k-folds with the number of folds set to 10. An array of features was created for the training set.

After data was properly configured, various models were created:

- Decision tree
- Nearest neighbor
- Random Forest

The models were initialized by setting initial parameters and then the models were fit to the training data features and happiness levels for each model.

2.5 Predictions

After training, each model was used to predict happiness levels for the individuals in the test set. The happiness levels were rounded to the nearest integer for cross validation with the actual happiness levels in the test set. The error was determined by root mean squared error (RMSE) and the average RMSE was determined for all of the test set predictions. This process was repeated iteratively using various model types and parameters.

After selecting the best model, the training and test data was combined to make predictions average country happiness for each country. The process of running predictions using the training dataset is not ideal as the model could still be over-fit to the training data and therefore not represent true accuracy. Nonetheless, this was performed to allow comparison of predicted and actual average country happiness. Another RMSE cross validation was performed, this time using the country average happiness levels.

3 Results

This section presents the results of the approach.

3.1 Features Correlated with Happiness

One of the objectives of this project was to find which features correlate most with happiness levels. To determine this, after creating and training decision tree and random forest models, the relatively important features were extracted. Results are presented in the figure below:

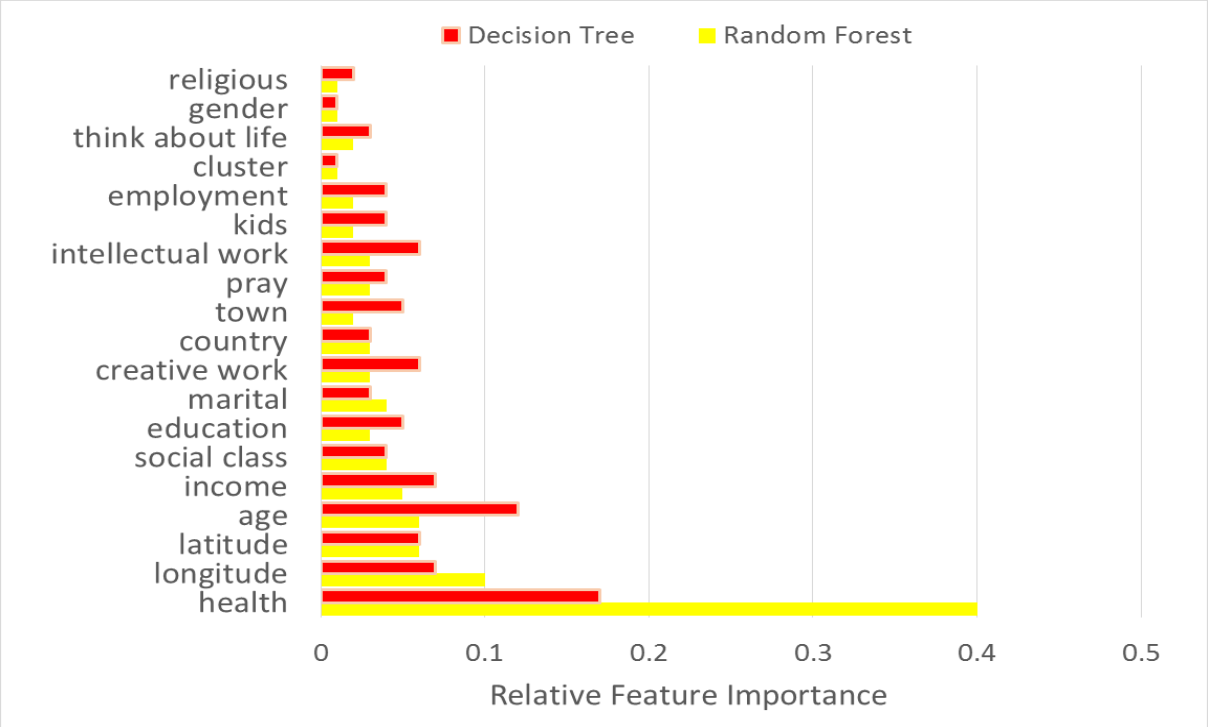

Figure 2-1: Relatively Important Features

As one would expect, health has the strongest correlation with happiness for both models, and was an especially important feature in the random forest model. This indicates there is a type of mind and body connection, in which mental and physical health may influence each other. Interestingly, longitude was also found to be an important feature. Latitude was expected to be more important as it influences the amount of sunlight and warmth a country has, which could logically influence happiness levels. It is not clear why longitude has such importance and may be a coincidence of countries lining up with similar happiness levels. One surprise of a feature that was expected to possibly have more importance is a person's religious standing, however, this was not an important feature in the model.

3.2 Individual Happiness Level Predictions

While feature importance was determine using random forest and decision trees, the model that provided the lowest RMSE was k-nearest neighbors, using 10 neighbors. This model yielded an RMSE of 0.5 for the test data set. The predictions were not as accurate as hoped as 0.5 is relatively significant as a majority of people had happiness levels of 1 or 2. The cause of the large error may be partially due to

the fact that these numbers are integers, so a difference between 1 vs 2 creates a seemingly large variation.

The follow table presents the happiness levels for a completely random group of 20 individuals. It is includes to provide a representative sample of the individual happiness results.

Table 3-1. Happiness Levels (Actual and Predicted) for 20 Random Individuals

Predicted Happiness Level	Actual Happiness Level	RMSE	RMSE (%)	Correct Prediction
2	2	0	0%	Yes
3	3	0	0%	Yes
2	3	1	50%	No
2	1	1	50%	No
2	2	0	0%	Yes
2	2	0	0%	Yes
2	1	1	50%	No
2	3	1	50%	No
2	2	0	0%	Yes
2	2	0	0%	Yes
2	2	0	0%	Yes
3	2	1	33%	No
2	2	0	0%	Yes
2	1	1	50%	No
2	1	1	50%	No
2	2	0	0%	Yes
3	1	2	67%	No
2	2	0	0%	Yes
2	3	1	50%	No
2	1	1	50%	No
2.15	**1.90**	**0.55**	**25%**	**50%**

3.1 Average Country Happiness Level Predictions

After predicting the happiness level of each individual in the survey, the average happiness level of each country was calculated by totaling the happiness levels and dividing by the number of people surveyed in each country. The predictions were compared to the actual average country happiness levels and results are presented in the following table:

Table 3-2. Average Country Happiness Levels (Actual and Predicted)

Country	Actual	Predicted	RMSE	RMSE (%)
Uzbekistan	1.39	1.80	0.41	29%
Mexico	1.40	1.75	0.35	25%
Ecuador	1.48	1.73	0.25	17%
Colombia	1.52	1.76	0.24	16%
Malaysia	1.56	1.75	0.19	12%
Philippines	1.60	1.70	0.10	6%
Sweden	1.63	1.64	0.01	1%
Ghana	1.67	1.76	0.09	5%
Thailand	1.67	1.83	0.16	10%
Kyrgyzstan	1.68	1.76	0.08	5%
New Zealand	1.68	1.78	0.10	6%
Australia	1.69	2.07	0.38	22%
Rwanda	1.70	1.66	0.04	2%
Nigeria	1.72	1.55	0.17	10%
Brazil	1.74	1.98	0.24	14%
Netherlands	1.76	1.89	0.13	7%
Uruguay	1.79	1.63	0.16	9%
Kazakhstan	1.80	1.84	0.04	2%
South Africa	1.80	2.07	0.27	15%
Libyan Arab	1.81	2.01	0.20	11%
Zimbabwe	1.82	1.87	0.05	3%
Poland	1.82	1.77	0.05	3%
Taiwan	1.84	1.87	0.03	2%
Peru	1.88	1.38	0.50	27%
Germany	1.90	1.91	0.01	1%
Chile	1.92	2.23	0.31	16%
India	1.92	2.24	0.32	17%
Azerbaijan	1.92	2.03	0.11	6%
Cyprus	1.94	1.90	0.04	2%
Armenia	1.97	2.07	0.10	5%
China	1.97	2.08	0.11	6%
Jordan	1.98	1.90	0.08	4%
Slovenia	2.00	1.98	0.02	1%
Lebanon	2.05	2.02	0.03	1%
Yemen	2.06	1.61	0.45	22%
Russian Federation	2.07	1.77	0.30	14%
Algeria	2.09	2.12	0.03	1%
Tunisia	2.10	1.73	0.37	18%
Estonia	2.14	1.52	0.62	29%
Ukraine	2.17	1.78	0.39	18%
Georgia	2.18	1.88	0.30	14%
Romania	2.23	1.76	0.47	21%
Belarus	2.23	1.93	0.30	13%
Iraq	2.26	2.15	0.11	5%
Average	**1.85**	**1.85**	**0.20**	**11%**

As shown in the table, the model was fairly accurate with an average RMSE of 11% for the country predictions. Several country predictions had an RMSE of only a few percentage points, and Iraq was predicted to be the least happy country, matching actual levels. On the other hand, the results had some large discrepancies as Estonia was predicted to be the happiest country, yet according to the actual survey data is one of the least happy countries.

The results are displayed graphically on the map below to provide another view of the data. Many countries had predictions in the correct happiness level range, while many others were not. Viewing the data spatially, it is apparent that certain regions have higher happiness levels than others. For example, the northern part of Europe has higher happiness levels but transitions to lower levels through southeastern Europe, the Middle East, and Northern Africa. High happiness levels also appear in the northern part of the Americas (Mexico, Colombia, Ecuador), the Ocean regions (Australia, New Zealand, Indonesia), and Central Asia (Uzbekistan, Kyrgyzstan).

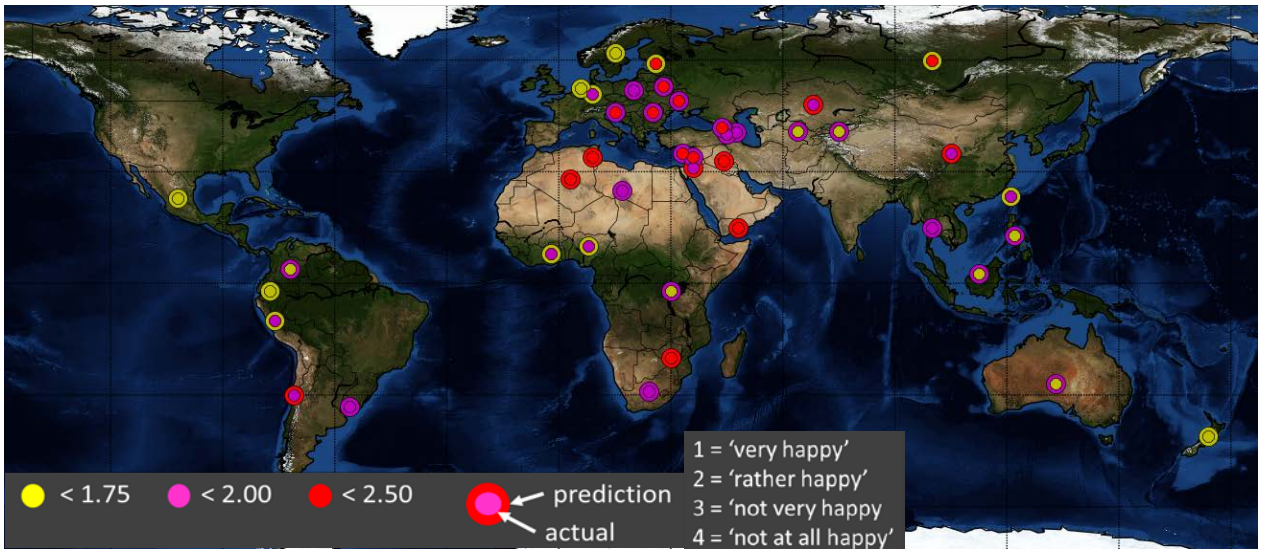

Figure 2-2: Average Country Happiness Levels (Actual and Predicted)

4 Conclusion

Key features that correlate with happiness were determined, and found happiness to be by far the strongest link, indicating a physical and mental health (mind / body) connection. While the model could make some correct predictions about individual and average country happiness, it was not impressively accurate. The model is disappointing to not have more accuracy; however, this may simply indicate that people are not significantly affected by external events. While the results indicate happiness may be heavily influenced by internal attitude, the results also indicate, at least to a certain extent, happiness is influenced by external factors that may be occurring directly to the person or to the country en masse. This geographic happiness link is demonstrated on the map as certain regions have higher happiness levels than others. Perhaps Aristotle was partially correct when he said, "happiness depends upon ourselves".

References

World Values Survey Association. (20150418). *World Values Survey Wave 6 2010-2014 OFFICIAL AGGREGATE.* Retrieved from http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp

Appendix A – Python Code Printout

Appendix B – Questionnaire