ECON-386

Dr. Steven Levkoff

Data Regression Final Project

05/21/2020

**Estimating Happiness Score for Countries**

**Group 7**

Ernesto Truqui, Aruna Gossai, Tony Woods, Lauren Hidalgo, Conrad Smits, Yifan Liu

**Table of Contents**

|  |  |
| --- | --- |
|  | **Page Noº** |
| **Section 1: Introduction** |  |
| * Executive Summary | *2* |
| * Variables of Interest | *2* |
| * Dependant Variable | *3* |
| * Regions | *4* |
| * Data Collection Process | *4* |
| * Structure of Dataset | *4* |
| **Section 2: Data Cleanup** |  |
| * Breakdown of Process | *5* |
| * Issues with Multicollinearity | *5* |
| **Section 3: Regression Model Proposals** |  |
| * Models 1-3 | *6, 7, 8* |
| **Section 4: Regression Model Validation** |  |
| * Root Mean Square Error | *9* |
| **Section 5: Classification Model Proposals** |  |
| * Models 4-6 | *9, 10* |
| **Section 6: Classification Model Validation** |  |
| * Confusion Matrix and Accuracy | *11, 12, 13, 14* |
| **Section 7: Discussion of Results and Conclusion** |  |
| * Regression Model Discussion & Conclusion | *14* |
| * Classification Model Discussion & Conclusion | *14* |
| * Final Discussion & Conclusion | *15* |
| **Appendix & Bibliography** | *16-19* |

**Section 1: Introduction**

**Executive summary**

In this project, we aim to create two models to examine the relationship between a country's happiness score and a set of independent variables. Using the World Happiness Report of 2017, and other sources of hard data, we are looking to estimate the regression between a country’s given happiness score, and other independent factors that we have identified as variables of interest. Along with a regression, our goal is to solve a qualitative problem with the collected data sets. Therefore, we will be producing both a regression model as well as a classification model.

The happiness score was derived from the 2017 World Happiness Report, with a score for the years 2006 to 2015, and the independent variables or variables of interest (detailed below) were taken from Gapminder.org. The original dataset consists of inputs from 156 countries, which were divided into ten separate geographical regions, as well as six independent variables of the ten variables of interest, and a dummy variable for the regions’ categories. Using the raw data from those 156 countries, each matching with a given happiness score given in the report, our next step was to identify our variables of interest from Gapminder.org.

**Variables of Interest**

* **Democracy Score:** An index created using the data from the Economist Intelligence Unit to express the quality of democracies as a number between 0 and 100. The data is based on 60 different aspects of societies that are relevant to democracy, i.e. universal suffrage for all adults, voter participation, perception of human rights protection, and freedom to form organizations and parties.
* **Gini Coefficient:** A measure of income inequality. The index number ranges from zero to one. A zero meaning that there is no income inequality within the nation. A gini coefficient of .8 or .9 would mean that all of .8 percent of all the income in the country would be going towards a very small percentage of the population.
* **Child mortality:** This measures the number of deaths of children under 5 years of age, per 1000 live births.
* **Refugee share of Population**: The share of refugees as a percentage of the total population of the country of residence.
* **Percentage of Population with basic sanitation:** The percentage of people using at least basic sanitation service.
* **Mean Years of Schooling for Women 25-34:** The mean years a woman spends in school between the ages of 25 and 34 years.
* **Mean Years of Schooling for Men 25-34:** The mean years a man spends in school between the ages of 25 and 34 years.
* **Children and Elderly per 1000 Adults:** Measures the total dependency ratio in a country, which is the percentage of the population composed of children under 14 and adults over 65 years.
* **Population Density per Km²**: The average number of people per square kilometer.
* **Labor Force Participation 15+:** The percentage of people over 15 years old active in the labor force of the total population.

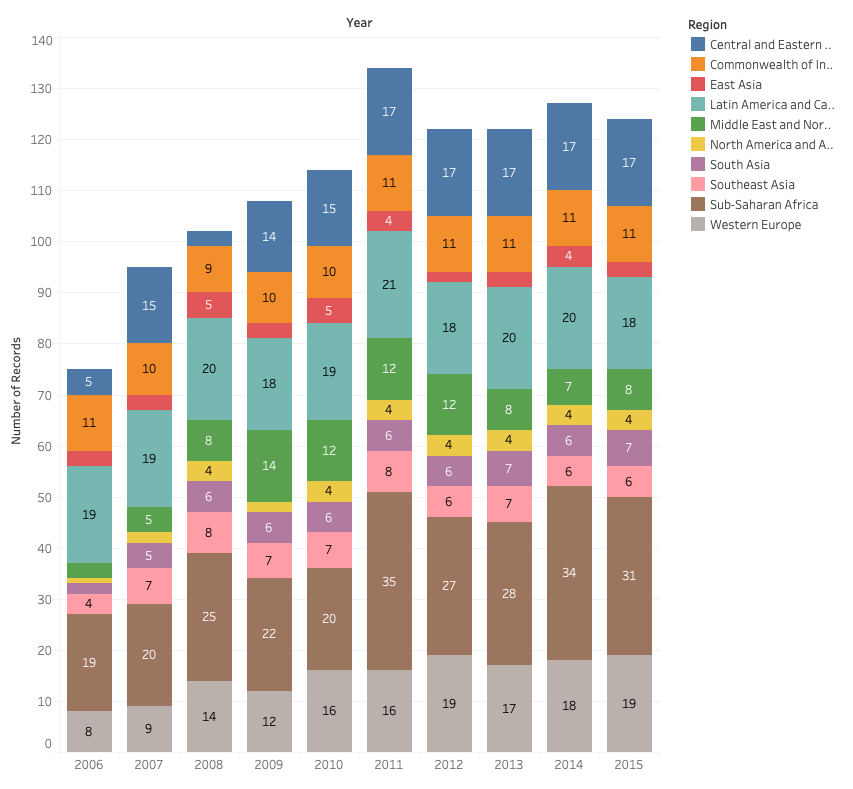
**Dependent Variable**

* **Happiness Score:** The score is based on multiple independent variables that ranks 156 countries by how happy their citizens perceive themselves to be.

**Regions**

We have included a dummy indicator for regions, including Western Europe, Central and Eastern Europe, Commonwealth of Independent States, Southeast Asia, South Asia, East Asia, Latin America and Caribbean, North America and ANZ, Middle East and North Africa, and Sub-Saharan Africa.

By doing this, we will be able to input variables into the model to retrieve an output of which region the record likely belongs to.



**Data Collection Process**

The data we collected is a mix of hard and soft data from the websites. An example of the hard data is the happiness score from the World Happiness Report, because it was obtained through scientific observation and measurement. An example of the soft data we observed were the independent variables, as these include many census data, which is often based one subjective responses from the nation’s population.

We exported the data collected from the World Happiness Report into an Excel file where all the variables for the corresponding scores were attached. However, we could not use the independent variables from the World Happiness Report to create our model. If we were to have used them, then we would have essentially just been looking for the formula that was used to generate the happiness scores for each country, as opposed to examining the relationship between the variables.

Therefore, we decided to find our own credible independent variables for the 156 original countries. We found this information on Gapminder.org. Gapminder is a Swedish non-profit organization that promotes the achievement of the United Nations Millennium Development Goals and sustainable development through the sharing and understanding of socio-economic, political, and environmental information.

From there, we extracted the variables of interest into a seperate excel sheet and commenced the process of cleaning the data.

**Structure of the Data Set**

We are using supervised learning in our data to best predict our dependent variable given all the independent variables we have. The reason it is supervised learning is due to the fact that both our x and y variables are known. Our data is structured as cross-section data by averaging across the time dimension for each country's metrics.

We have two datasets: “df” and “df\_dum”. The latter has a column region for the dummy variables, whereas the former does not. We have done this in order to perform both quantitative and qualitative analysis on the data.

We have decided to not get rid of outliers in the original datasets. If desired, these may later be replaced with NA values, or left as is.

**Section 2: Data Cleanup**

**Breakdown of Process**

Below are the steps we took in cleaning up our data. Please refer to the Rmarkdown file for the corresponding lines of code:

1. Line 10: Loaded data from Github into Rmarkdown.
2. Line 30-32: Created a key for the regions, each region corresponds with a unique number in order to be able to aggregate it.
3. Line 34: Changed the region data from categorical to numeric in order to be able use it in the model.
4. Line 36-37: Eliminated countries with no observations (Kosovo, Taiwan and Sudan) and no democracy values (South Sudan, Georgia, Belize).
5. Line 39-40: Converted the dataset from pooled cross-sectional to cross sectional by taking the mean of each column by country and aggregating this.
6. Line 42-43: Added a column for the region name titled “rname” in order to incorporate it into the dataset.
7. Line 45: Created dummy variables from the region name and assigned this to a new dataset titled “df\_dum”.
8. Line 47: Tested for multicollinearity using a covariance matrix, excluding regions and year from the matrix (see section titled “Issues with Multicollinearity” for details).
9. Line 49-56: Dropped problem variables found in multicollinearity test, which were: “men\_edu”; “sanitation”; “child\_mortality”; & “elder\_child”. Dropped these from both the “df” dataset as well as the “df\_dum” dataset.
10. Line 58-66: Dropped unnecessary variables, which were: “df$i..id”; “df$region”; “df\_dum$year”; “df\_dum$rname”; and “df\_dum$rname\_West\_EU” – got rid of this single dummy variable to prevent multicollinearity.
11. Line 69-70: Created binary “very happy” variables to use in classification models.

**Issues with Multicollinearity**

As clear in the “Variables of Interest” section, we originally began with ten independent variables. However, after running a correlation matrix, we saw high correlation coefficients of above .5 between several variables, which were “Child Mortality”, “Percentage of Population with basic sanitation”, “Mean Years of Schooling for Men 25-34” and “Children and Elderly per 1000 Adults”.

Although multicollinearity poses more of a problem in economics than in data analytics, we wanted a low threshold for multicollinearity in order to ensure relationships between independent variables did not affect our models in any way. When deciding what variables to omit, we looked at which variables were more significant than others and took that into consideration. For example, we eliminated men’s education as opposed to women’s education since women’s education had more variability than men’s, and men’s was more static. Therefore we kept women’s as it would create a better model to guess the happiness score from.

**Section 3: Regression Model Proposals**

In this section, three group members have created a regression model using the tidy dataset. The tables below outline the important information and diagnostics for each model to allow for easy comparison.

* *Please note that all models have been created using the same training/testing split of 70/30, and use the same Set Seed of “123”, to allow for proper comparison between the models.*
* *Please see the Appendix for a full display of each models’ coefficients as well as confidence intervals.*

**# Key for interpreting Significance codes:**  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model #1** | | | **Comments** | |
| *Team member* | Lauren Hidalgo | |  | |
| *Regression equation* | M2 = lm(happiness ~ .) | | A mid-level complex linear model that transforms the variables refugeen, women’s education and population density to their values to their log values to try to capture their less linear relationship observed in the data. | |
| *Statistical significance*  *Codes: 0 ‘\*\*\*’ ; 0.001 ‘\*\*’ ; 0.01 ‘\*’* | \*\*\* | Intercept, r\_name\_Cen\_E\_EU, womens\_edu,  r\_name\_SS\_AF | Asian regions seem to be very significant overall. Women's education and democracy are the most significant independent variables. The intercept being significant implied that there is a correlation to be observed in the model. | |
| \*\* | democracy, r\_name\_E asia, r\_name\_S asia, r\_name\_common |
| \* | r\_name\_SE asia |
| *Confidence intervals of est. model parameters* | *Tight intervals:* democracy, gini, refugee, labour  *Large intervals:* Intercept, rname\_S\_ASIA, rname\_E\_ASIA, rname\_NA\_OC  See table 1 in Appendix | | *Tight intervals:* variables with a spread <0.1 between lower and upper bounds. More likely to return the same values for different inputs.  *Large intervals:* variables with a spread >1.5 between lower and upper bounds. Produce more volatility in outputs and error. | |
| *Goodness of fit (Adjusted R squared)* | 70.52% | | 70.52% of variation is explained by only those independent variables that affect the dependent variable. | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model #2** | | | **Comments** | |
| *Team member* | Aruna | |  | |
| *Regression Code or Estimated regression or variables used?* | Happiness = 3.79 + .02\*democracy - .01\*gini - .09\*refugee + .60\*women\_edu - .03\*pop\_den + .01\*labour - .87\*Cen\_E\_EU - .74\*Common - 1.41\*E\_Asia + .09\*LA\_Carib - .00\*NAF + .43\*NA\_OC - .2\*S\_ASIA - .39\*SE\_ASIA - .87\*SS\_AF | | 5 fold cross validation with ridge penalty | |
| *Statistical significance*  *Codes: 0 ‘\*\*\*’ ; 0.001 ‘\*\*’ ; 0.01 ‘\*’* | \*\*\* |  | N/A standard errors are not meaningful for penalized regression | |
| \*\* |  |
| \* |  |
| *Confidence intervals of est. model parameters* |  | | N/A standard errors are not meaningful for penalized regression | |
| *Goodness of fit (Adjusted R squared)* | 75.42% | |  | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model #3** | | | **Comments** | |
| *Team member* | Yifan | | Trying to fit a linear model with non-dummy variables and has better in-sample and out-sample prediction. Start with the full model with 6 variables, next construct all the possible reduced models with 5 variables, then test all in- and out-sample errors for all the reduced models. Finally pick out one reduced model from the full model and the all the reduced model based on the average performance of in- and out-sample errors. | |
| *Regression equation* | happiness=2.377+0.019∙democracy-0.002∙refugee+0.136∙women\_edu+0.00008∙pop\_den+0.010∙labour+ε | |  | |
| *Statistical significance*  *Codes: 0 ‘\*\*\*’ ; 0.001 ‘\*\*’ ; 0.01 ‘\*’* | \*\*\* | Intercept, democracy, women\_edu | Intercept, democracy and women\_edu are statistically significant at 0.1% level.  Democracy is positively correlated with happiness scores. For one unit increase in democracy, the happiness score is expected to increase by 0.019.  Women’s education is positively correlated with happiness scores. For one year increase in women’s education, the happiness score is expected to increase by 0.136.    Other variables are not significant even at a 10% level | |
| \*\* |  |
| \* |  |
| *Confidence intervals of est. model parameters* | See table 3 in Appendix | |  | |
| *Goodness of fit (Adjusted R squared)* | 57.79% | | 57.79% variation in happiness is explained by the model | |

**Section 4: Regression Model Validation**

**Root Mean Square Error**

We used the Root Mean Square Error (RMSE) to measure the average deviation in the actual data relative to the prediction. The RMSE measures the expected/predicted value conditional on the independent variables in the model. The RMSE IN measures the in-sample error, whereas the RMSE out measures the out of sample error of the model. Lower RMSE values indicate a better fit, and the measures below are rounded to 4 decimal places.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MODEL** | **RMSE IN** | **RMSE OUT** |
| **1 - Lauren** | LM with all variables | 0.5360 | 0.5912 |
| **2 - Aruna** | Ridge regression with 5 fold CV | .5494 | .5855 |
| **3 - Yifan** | Linear regression model with 5 variables (democracy, refugee, women\_edu, pop-den,labour) | 0.6935 | 0.7025 |

**Section 5: Classification Model Proposals**

In this section, three group members have created a classification model using the tidy dataset. The tables below outline the important information and diagnostics for each model, so that they can be easily compared in order to be able to select the best model.

**# Key for interpreting Significance codes:**  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model #4** | | | **Comments** | |
| *Team member* | Ernesto | |  | |
| *Regression equation* | veryhappy = 0.1563(democracy) + 0.1114(gini) + 0.0883(refugee) + 0.7373(women education) + 0.0020(Pop density) + 0.1269(labor) | | Logistic regression model | |
| *Statistical significance*  *Codes: 0 ‘\*\*\*’ ; 0.001 ‘\*\*’ ; 0.01 ‘\*’* | \*\*\* | - | These were the variables that proved significant at the 10% level, or lower | |
| \*\* | Intercept, democracy |
| \* | - |
| . | Gini, Women education |
| *Confidence intervals of est. model parameters* | See Table 4 in Appendix | |  | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model #5** | | | **Comments** | |
| *Team member* | Conrad | |  | |
| *Regression equation* | Very happy = -12.8534(democracy) + 4.9528(gini) + -1.4453(refugee) + -9.6857(women\_edu) + 2.1092(pop\_den) + -3.4277(labour) + 12.0365(rname\_Cen\_E\_EU) + 6.2600(rname\_Common) + 2.4696(X.rname\_E\_ASIA) + 4.4783(rname\_LA\_Carib) + 3.3727(rname\_ME\_NAF) + -15.2491(rname\_NA\_OC) + 0.2930(rname\_S\_ASIA) + 4.1151(rname\_SE\_ASIA) + 4.4002(rname\_SS\_AF) | | Support vector machine classification with 5 fold cross validation using the radial basis function | |
| *Statistical significance*  *Codes: 0 ‘\*\*\*’ ; 0.001 ‘\*\*’ ; 0.01 ‘\*’* | \*\*\* | - | These were the variables that proved significant at the 5% level, or lower | |
| \*\* | - |
| \* | - |
| *Confidence intervals of est. model parameters* |  | | NA | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model #6** | | | **Comments** | |
| *Team member* | Tony | |  | |
| *Regression equation* | veryhappy = -19.5898 + 0.0747(democracy) + 0.1007(gini) - 0.0171(refugee) + 1.0944(women\_edu) + 0.0011(pop\_den) - 0.0596(labour) - 21.8802(Cen\_E\_EU) - 20.2720(Common) - 19.1994(E\_Asia) + 1.1474(LA\_Carib) - 0.7377(ME\_NAF) + 18.9084(NA\_OC) - 9.0266(S\_Asia) - 3.3392(SE\_Asia) - 18.7207(SS\_AF) | | Logistic regression that has no variable restrictions and adds regional dummy variables | |
| *Statistical significance*  *Codes: 0 ‘\*\*\*’ ; 0.001 ‘\*\*’ ; 0.01 ‘\*’* | \*\*\* |  | These were the variables that proved significant at the 5% level, or lower | |
| \*\* |  |
| \* |  |
| *Confidence intervals of est. model parameters* | See Table 6 in Appendix | |  | |

**Section 6: Classification Model Validation**

The classification task we set out to complete was to predict whether a country was very happy or not. And the threshold we set for a country to qualify as very happy was a 6.5 or above on the happiness score, which ranges from 1-10.

**Confusion Matrix and Accuracy**

We used a Confusion Matrix to test both in and out of sample errors, which provides a visual layout of the performance of the model. We also tested for accuracy to get another measure for out of sample error. An accuracy measure of >95% is usually considered to be a good criteria to strive for in building a model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model #4** | | | | | **Comments** | |
| *Team member* | Ernesto | | | |  | |
| *Regression equation* | veryhappy = 0.1563(democracy) + 0.1114(gini) + 0.0883(refugee) + 0.7373(women education) + 0.0020(Pop density) + 0.1269(labor) | | | | Logistic regression model | |
| Confusion matrix | |  |  |  | | --- | --- | --- | | OUT OF SAMPLE | | | |  | False | True | | False | 35 | 2 | | True | 5 | 3 |  |  |  |  | | --- | --- | --- | | IN SAMPLE | | | |  | False | True | | False | 81 | 5 | | True | 2 | 16 | | | | |  | |
|
|
| **Out of sample** | | | | | | |
| Accuracy | 84.44% | | | | 95% CI: (0.7054, 0.9351) | |
| Sensitivity | 87.50% | | | |  | |
| Specificity | 60.00% | | | |  | |
| Positive predicted value | 94.59% | | | |  | |
| Negative predicted value | 37.50% | | | |  | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model #5** | | | | | **Comments** | |
| *Team member* | Conrad | | | |  | |
| *Regression equation* | Very happy = -12.8534(democracy) + 4.9528(gini) + -1.4453(refugee) + -9.6857(women\_edu) + 2.1092(pop\_den) + -3.4277(labour) + 12.0365(rname\_Cen\_E\_EU) + 6.2600(rname\_Common) + 2.4696(X.rname\_E\_ASIA) + 4.4783(rname\_LA\_Carib) + 3.3727(rname\_ME\_NAF) + -15.2491(rname\_NA\_OC) + 0.2930(rname\_S\_ASIA) + 4.1151(rname\_SE\_ASIA) + 4.4002(rname\_SS\_AF) | | | | Support vector machine classification with 5 fold cross validation using the radial basis function | |
| Accuracy | 93.33% | | | | 95% CI: (81.73%,98.60%) | |
| Confusion Matrix | |  |  |  | | --- | --- | --- | | OUT OF SAMPLE | | | |  | False | True | | False | 40 | 3 | | True | 0 | 2 |  |  |  |  | | --- | --- | --- | | IN SAMPLE | | | |  | False | True | | False | 83 | 14 | | True | 0 | 7 | | | | |  | |
| **Out of sample** | | | | | | |
| Accuracy | 93.33% | | | | 95% CI: (0.8173, 0.9860) | |
| Sensitivity | 100% | | | |  | |
| Specificity | 40.00% | | | |  | |
| Positive predictive value | 93.02% | | | |  | |
| Negative predicted value | 100% | | | |  | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model #6** | | | | | **Comments** | |
| *Team member* | Tony | | | |  | |
| *Regression equation* | veryhappy = -19.5898 + 0.0747(democracy) + 0.1007(gini) - 0.0171(refugee) + 1.0944(women\_edu) + 0.0011(pop\_den) - 0.0596(labour) - 21.8802(Cen\_E\_EU) - 20.2720(Common) - 19.1994(E\_Asia) + 1.1474(LA\_Carib) - 0.7377(ME\_NAF) + 18.9084(NA\_OC) - 9.0266(S\_Asia) - 3.3392(SE\_Asia) - 18.7207(SS\_AF) | | | | Logistic regression that has no variable restrictions and adds regional dummy variables | |
| Confusion matrix | |  |  |  | | --- | --- | --- | | OUT OF SAMPLE | | | |  | False | True | | False | 34 | 2 | | True | 6 | 3 |  |  |  |  | | --- | --- | --- | | IN SAMPLE | | | |  | False | True | | False | 79 | 2 | | True | 4 | 19 | | | | |  | |
|
|
| **Out of sample** | | | | | | |
| Accuracy | 82.22% | | | | 95% CI:  .6795, .9200 | |
| Sensitivity | 85.00% | | | |  | |
| Specificity | 60.00% | | | |  | |
| Positive predicted value | 94.44% | | | |  | |
| Negative predicted value | 33.33% | | | |  | |

**Section 7: Discussion of Results and Conclusion**

**Regression Model Discussion & Conclusion**

The final model measuring the continuous happiness score is a penalized regression, we selected Model 2. To decide between a ridge or lasso penalty term we used a for loop. In the glment package, an alpha of 0 indicates a ridge penalty and an alpha of 1 indicates a lasso penalty. This package allowed us to test hybrid penalty terms as well. We tested the following alpha values: 0, .25, .5, .75, 1. Ultimately, the pure ridge penalty term, an alpha of 0, consistently reported the smallest out of sample RMSE. RMSE is the validation metric used for the models measuring happiness.

The selected model did not report standard errors, as they are not meaningful in penalized regressions. As a result, we could not calculate the variable significance or the confidence intervals for the parameter estimates. The final model also used transformed variables for population density, women's education, and refugee population share. The log of population density and refugee population share is used to reduce variance, across countries. When plotting women’s education and happiness, the log of women's education seemed to have a better fit for the happiness variables.

**Classification Model Discussion & Conclusion**

When we chose our classification model we wanted to be as accurate as possible in our main goal: predicting whether a country would be very happy. Therefore, we went with the support vector machines model, Model 5, because of a few key attributes: its out-of-sample accuracy was significantly higher than the logistic models at 93%, it had a much better Negative Prediction score than the logistic models, and it seems to provide the best base for future predictions on happiness levels . The function “svm” can be used as either regression or c-classification, but since our dependent variable is a factor we had to go with c-classification to use the svm function.

Next we had to set the parameters for our svm model. We went with a five-fold cross validation using the radial basis function. We went with a five-fold to assess the quality of the model using the accuracy rate since it’s classification. Then the training and testing sets were created with five different levels of gamma. This gave us our in and out of sample error for the training and testing set, respectively.

Because we used the support vector machine to build a classification model, Given our best results with a gamma of 0.010, those were then used for the confusion matrix, giving us a very high accuracy both for in and out of sample error.

**Final Discussion & Conclusion**

Throughout this project, we faced several challenges, particularly when it came to cleaning and preprocessing the data.

First, we lost a lot of observations from aggregating the data in an attempt to make sure every country had observations representing the mean of these values. Still, even after doing this there were countries with no data. This was due to a disconnect between the original countries used in the happiness dataset, which were not used across the US. For example, we couldn't find any data besides the happiness variable for Kosovo and Taiwacouldntn, and were forced to eliminate these countries.

Another issue that we faced was the subjectivity of dealing with outliers. It was challenging to decide when it was appropriate to eliminate a country; for example would it be appropriate to do so if there was only one missing value; or should we eliminate countries only with several missing values? In the end, we took the approach of “less is more” and attempted to eliminate the fewest amount of observations possible by replacing these with NA values.

NA values also made us drop records instead of inputting values, which was a choice we had to make that we felt worked best with our dataset. We also had to drop four variables due to the issues with multicollinearity which were discussed previously, which was essential to ensure that our model was actually just observing the correlations we wanted it to and properly measuring happiness.

Another issue that we found along the way was trying to balance the tradeoff between in sample fit and out of sample fit. Both chosen models assumed that in sample fit tends to be better than out of sample fit. This assumption is usually true, but might not be all the time. For example, for the chosen classification model, the accuracy out is actually better than the accuracy in for certain levels of Gamma.

To conclude, both models showed strong predictive power for happiness.

**APPENDIX**

Confidence Interval with Center

Table 1: MODEL 1

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Center | 2.50% | 97.50% |
| (Intercept) | 3.9227 | 2.3927 | 5.4527 |
| democracy | 0.0135 | 0.0037 | 0.0234 |
| gini | -0.0104 | -0.0369 | 0.0161 |
| log(refugee +1) | -0.0149 | -0.0369 | 0.0070 |
| log(women\_edu) | 0.6372 | 0.2866 | 0.9879 |
| log(pop\_den) | -0.0378 | -0.1506 | 0.0750 |
| labour | 0.0112 | -0.0029 | 0.0253 |
| rname\_Cen\_E\_EU | -1.2247 | -1.7024 | -0.7469 |
| rname\_Common | -1.1573 | -1.8458 | -0.4688 |
| `rname\_E\_ASIA | -1.8621 | -3.1371 | -0.5870 |
| rname\_LA\_Carib | -0.2647 | -0.9153 | 0.3858 |
| rname\_ME\_NAF | -0.3212 | -0.9814 | 0.3389 |
| rname\_NA\_OC | 0.2058 | -0.5847 | 0.9963 |
| rname\_S\_ASIA | -0.5467 | -1.4715 | 0.3781 |
| rname\_SE\_ASIA | -0.7909 | -1.5115 | -0.0703 |
| rname\_SS\_AF | -1.3260 | -2.0733 | -0.5788 |

Table 2: MODEL 2

|  |  |
| --- | --- |
| **Variable** | **Parameter Estimate** |
| *(Intercept)* | 3.8132 |
| *democracy* | 0.0159 |
| *gini* | -0.0145 |
| *log(refugee + 1)* | -0.0965 |
| *log(women\_edu)* | 0.6197 |
| *log(pop\_den)* | -0.0385 |
| *labour* | 0.0099 |
| *rname\_Cen\_E\_EU* | -0.96177 |
| *rname\_Common* | -0.8412 |
| *rname\_E\_ASIA* | -1.5464 |
| *rname\_LA\_Carib* | 0.02130 |
| *rname\_ME\_NAF* | -0.0593 |
| *rname\_NA\_OC* | 0.3737 |
| *rname\_S\_ASIA* | -0.2518 |
| *rname\_SE\_ASIA* | -0.4714 |
| *rname\_SS\_AF* | -0.9563 |

Table 3: MODEL 3

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Center** | **2.50%** | **97.50%** |
| *(Intercept)* | 2.377 | 1.265 | 3.490 |
| *democracy* | 0.020 | 0.011 | 0.028 |
| *refugee* | -0.002 | -0.028 | 0.023 |
| *women\_edu* | 0.136 | 0.089 | 0.184 |
| *pop\_den* | 0.00008 | -0.00011 | 0.00028 |
| *labour* | 0.010 | -0.005 | 0.025 |

Table 4: MODEL 4

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Center** | **2.50%** | **97.50%** |
| *(Intercept)* | -33.92248 | -59.0572 | -16.8889 |
| *democracy* | 0.15638 | 0.0613 | 0.2879 |
| *gini* | 0.1115 | -0.0104 | 0.2630 |
| *refugee* | 0.0883 | NA | 0.2654 |
| *women\_edu* | 0.7373 | 0.0733 | 1.6800 |
| *pop\_den* | 0.0020 | -0.0003 | 0.0070 |
| *labour* | 0.1269 | -0.0401 | 0.3050 |

Table 5: MODEL 5 – **NA**

Table 6: MODEL 6

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Center | 2.50% | 97.50% |
| (Intercept) | -19.589769081 | -52.94197 | 2.587098 |
| democracy | 0.074757796 | -.06644746 | 2.506416 |
| *gini* | 0.100728275 | -.2474129 | .5016639 |
| refugee | -0.017123206 | NA | .2777640 |
| women\_edu | 1.094451546 | .07137466 | 2.667193 |
| pop\_den | 0.001175746 | -.0004596843 | .009045811 |
| labour | -0.059635909 | -.3468480 | .1779024 |
| rname\_Cen\_E\_EU | -21.880216937 | NA | 571.1449 |
| rname\_Common | -20.272077045 | NA | 883.0397 |
| rname\_E\_ASIA | -19.199438628 | NA | 5762.091 |
| rname\_LA\_Carib | 1.147441617 | -4.128745 | 7.444101 |
| rname\_ME\_NAF | -0.737721600 | -8.745186 | 7.895000 |
| rname\_NA\_OC | 18.908459627 | -1.393054 | NA |
| rname\_S\_ASIA | -9.026673809 | NA | 1467.242 |
| rname\_SE\_ASIA | -3.339244970 | NA | 5.630653 |
| rname\_SS\_AF | -18.720799975 | NA | 326.1746 |

**Bibliography**

<https://www.gapminder.org/data/>

<https://worldhappiness.report/ed/2020/>

<https://en.wikipedia.org/wiki/Gapminder_Foundation>