Anndi Russell and Anshul Dabas

Professor Cathy Durso

10 June 2020

An Exploration into World-View and Happiness Using Extreme Gradient Boosting

Executive Summary

Our research explores whether we can predict the happiness score of a country a person lives in based on their answer to a series of survey questions relating to their perception and opinion of certain aspects of their society. We used Extreme Gradient Boosting with the "xgboost" R package to help answer our question. The model we produced using xgboost was not a good fit for our data. We were unable to predict the happiness score based on these survey responses with acceptable accuracy; we discuss the likely reason for this in the conclusion. We demonstrate many skills throughout this project including data cleaning and exploration, feature creation, model building, and model interpretation.

Data source and definitions

Our research involved two data sets. The first is from the Pew Research Center's Global Attitudes and Trends survey conducted in Spring 2018. The second is data from the World Happiness Report. This report has been published over multiple years; we chose the year 2018 to most closely match with our Pew data.

The Pew survey was a survey given via phone to approximately 30,000 respondents in 27 different countries. The survey first asked questions relating to a person's view of the country they live in, then asked more questions about a respondent's opinion about global politics and relationships between countries. For this research, we were only concerned with the first set of questions. We've listed the questions below. For the sake of brevity we have not listed the response options here as response options often can be inferred intuitively from the question itself (see Appendix A for a full list of questions with their responses).

- *Thinking about our economic situation, how would you describe the current economic situation in (survey country) - is it very good, somewhat good, somewhat bad, or very bad?*
- *When children today in (survey country) grow up, do you think they will be better off, or worse off financially than their parents?*
- *How satisfied are you with the way democracy is working in our country – very satisfied, somewhat satisfied, not too satisfied, or not at all satisfied?*
- *Compared with 20 years ago, do you think the financial situation of average people in (survey country) is better, worse, or do you think there has been no change?*
- *Thinking about the ethnic, religious, and racial makeup of (survey country), over the past 20 years do you think (survey country) has become more diverse, less diverse, or do you think there has been no change?*
    - *Follow-up for those who did not respond "I don't know" or "refused": Do you think this is a*

good thing or a bad thing for (survey country)?

- *Over the past 20 years, do you think equality between men and women in (survey country) has increased, decreased, or do you think there has been no change?*
    - *Do you think this is a good thing or a bad thing for (survey country)?*
- *Compared to 20 years ago, do you think religion has a more important role in (survey country), a less important role, or do you think there has been no change?*
    - *Do you think this is a good thing or a bad thing for (survey country)?*
- *Over the past 20 years, do you think family ties in (survey country) have become stronger, weaker, or do you think there has been no change?*
    - *Do you think this is a good thing or a bad thing for (survey country)?*

The World Happiness Report offers a wide variety of information relating to what makes people happy. For our analysis, we were concerned with the "happiness score" produced for each country. Researchers used data from the Gallup World Poll to compile this score. The Gallup World Poll asked respondents to think of a ladder, with the best life for them being a 10 and the worst being a 0. Respondents are then asked to rank their current life on that 0 to 10 scale. Researchers use survey weights to make this representative of the country and calculate a mean "happiness score" for every country. This is the number we use in our research.

We combined these two datasets by pulling the "happiness score" for a country and attaching it to all respondents from that country. It's important to emphasize again that this is not a happiness score for each individual person; rather, it is a happiness score for the country in which they live.

Exploratory Data Analysis

Our data set has 30109 rows. This is sufficiently large for the method we are using. An evaluation of missing values revealed that the only missing values we have are in the follow-up questions; this is expected, as people who answered "Don't know" or "Refused" to the first question were not asked the follow-up. However, we did discover that 61 people in Mexico were mistakenly asked at least one follow-up question. To correct this, we simply replaced these responses with "NA" values, as was consistent with how the rest of the data was reported.

The number of responses was approximately equal across all countries, except for India which had about twice as many responses as other countries. We left this in, as it would not negatively affect our model and we wanted to have as much data to train as possible.

It was important that we saw a variety of response distributions between countries. If all countries tended to have very similar responses, this would not be an interesting investigation. To check this, we plotted the density of responses for all countries as seen in Figure 1.

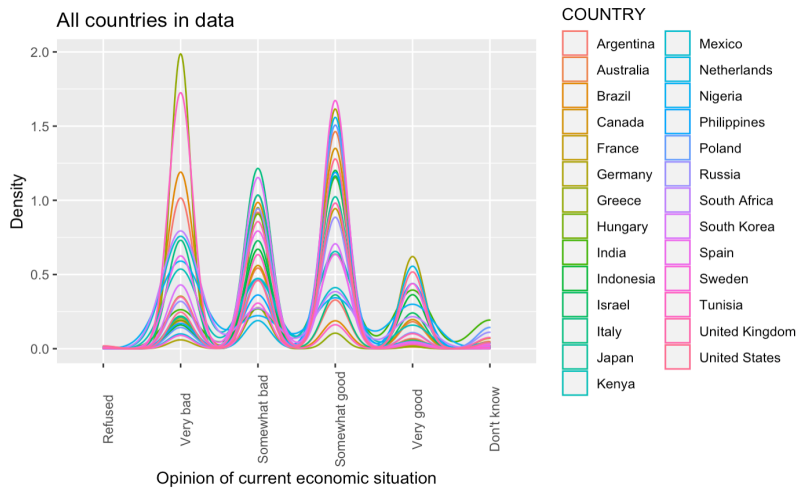


Figure 1: Density of responses to economic situation question by country.

Additionally, we selected one country that had a high happiness score (Netherlands), one with a medium happiness score (Poland), and one with a low happiness score (Tunisia), and compared their distributions of responses to these questions. The Netherlands would not be representative of all happy countries here (same with Poland and Tunisia), but it did allow us to further explore what differences we may find between countries of varying happiness scores. We did this same exploration for each of the 4 stand-alone questions in our data.

One important step we took with our data was to create a new feature column composed of responses to a question and its follow up (using the paste functionality). The follow-up question does not hold value for us on its own; if a person responded “Good thing”, the meaning can only be captured when combined with their response to the question before it. This resulted in 20 paired responses, with a sample density plot for the question regarding diversity in Figure 2.

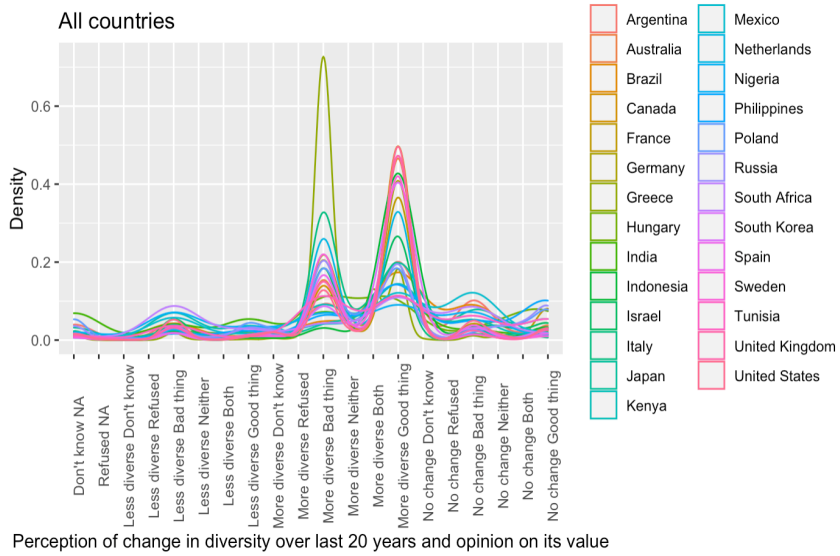


Figure 2: Density of responses to diversity question plus follow-up by country.

Again in this plot, as in Figure 1, we see a spread of responses between countries. We explored this same distribution between a high happiness, medium, and low happiness country as we did above to get a cursory glance at what differences our model may be able to pick up. We did this same exploration for all paired questions, in addition to exploring responses to just the first of the paired questions on their own.

Xgboost Method Explained

Now that we have explored the data, a reminder of our research question is important. We are going to use survey responses to predict the happiness score of the country a person lives in. Using Xgboost, we will build a model so we can predict the happiness score for future survey respondents, and we will also be able to explore which variables are the best predictors of our outcome.

Xgboost (extreme gradient boosting), uses decision trees to build a model for our data. The main principle of xgboost is that it fits a new model on the residuals from the previous model and combines these models together. It does this continuously until the model performance is not improved by fitting more models (using root mean square error as an evaluation metric).

Xgboost will first build a naive model (F0), then calculate the residuals of that model. It will build a model predicting the residuals (h1) then combine them together into a new model (F1):

F1(x)=F0(x)+h1(x)

It then calculates the residuals on F1, builds a model (h2) to predict the residuals, then combines these together (F2):

F2(x)=F1(x)+h2(x)

This continues until there is no improvement on model performance by increasing the complexity, or until some stopping criterion for the number of trees built.

The full “TreeBoost” algorithm is defined as follows:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}_{R_{jm}}(x), \quad \gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma).$$

Citation: Wikipedia

Data Requirements for xgboost

Xgboost requires a particular data input format. To prepare for this, we first dropped any columns that were unnecessary (survey number, country, follow-up questions), then did a 70/30 train/test split. Since all of our predictors were categorical, we used one-hot-encoding with the sparse.model.matrix() function in R. This turns every survey response option into its own column, with a 1 indicating an answer, and a “.” otherwise. This sparse matrix only includes the predictor variables, approximately 120 of them (xgboost can handle large sets like this, and we

have 30,000 rows in our data so that number of variables is not a problem) (Fig. 3). The outcome variable (happiness score) is put into its own list. These are together put into xgboost's "DMatrix" (Fig. 4).

```
Create sparse matrix of just predictors
```{r}
sparse_matrix_train <- Matrix::sparse.model.matrix(happiness_score ~ ., data = dat_train_df,drop.unused.levels = FALSE)[,-1]
sparse_matrix_test <- Matrix::sparse.model.matrix(happiness_score ~ ., data = dat_test_df,drop.unused.levels = FALSE)[,-1]
```
```

```
XGBoost input
```{r}
dat_train <- xgb.DMatrix(data = sparse_matrix_train,label = output_train)
dat_test <- xgb.DMatrix(data = sparse_matrix_test,label = output_test)
```
```

Figures 3 and 4: Sparse matrix creation and xgboost DMatrix creation

Application of xgboost

We used the following libraries to apply Xgboost modeling: library(xgboost), library(haven), library(car), library(SHAPforxgboost), library(Seurat).

First, we prepared a list of parameters to be passed while fitting the Xgboost model on our training dataset. The outcome variable was a continuous variable (the happiness score) so we performed a regression using Xgboost (these hyperparameters are: booster = 'gbtree' and objective = 'reg::linear'). Further, we used subsample and colsample_bytree hyperparameters to deal with overfitting for our model. Both of these hyperparameters randomly sample the data and variables from the training set for different iterations. We used 'rmse' root mean square error as the evaluation metric (to evaluate regression performance).

```
```{r}
param_trees <- list(booster = "gbtree"
                , objective = "reg:linear"
                , subsample = 0.7
                , max_depth = 5
                , colsample_bytree = 0.7
                , eta = 0.037
                , eval_metric = 'rmse'
                , base_score = 0.012
                , min_child_weight = 100)
```
```

Figure 5: Parameters for xgboost

After setting up the required parameters appropriately, we performed cross-validation on our dataset by using xgboost internal cross-validation method xgb.cv with relevant hyperparameters as follows:

```
Run xv
```{r}
target <- output_train
foldsCV <- createFolds(target, k=7, list=TRUE, returnTrain=FALSE)
xgb_cv <- xgb.cv(data=dat_train,
                 params=param_trees,
                nrounds=100,
                prediction=TRUE,
                maximize=FALSE,
                folds=foldsCV,
                gamma=0,
                early_stopping_rounds = 30,
                print_every_n = 5)
```
```

Figure 6: Cross validation for xgboost

After performing cross-validation on the training set, we found the best number of rounds to be 100; this is what we used as the maximum number of trees (nrounds hyperparameter) in our final model.

Finally, we ran the Xgboost model on our training dataset using xgb.fit with the appropriate hyperparameters as follows:

```
nrounds <- xgb_cv$best_iteration
xgb.fit <- xgb.train(params = param_trees
                  , data = dat_train
                  , nrounds = nrounds
                  , verbose = 1
                  , print_every_n = 5
                 )
```

Figure 7: Fitting xgboost model

Additionally, we applied the predict method on the test dataset to evaluate the predictions using the above model.

Method Results and Analysis

We used the following two evaluation techniques to interpret or analyze the results produced by Xgboost model: Importance matrix plot and SHAP (SHaply Additive exPlanation) values plot for the features. The importance matrix is a table with the first column including the names of all the features actually used in the boosted trees, and the second column reports the 'importance' value of that feature. Wee had ~120 features in the importance plot, so for simplicity, we display in Figure 8 the top 10 features.

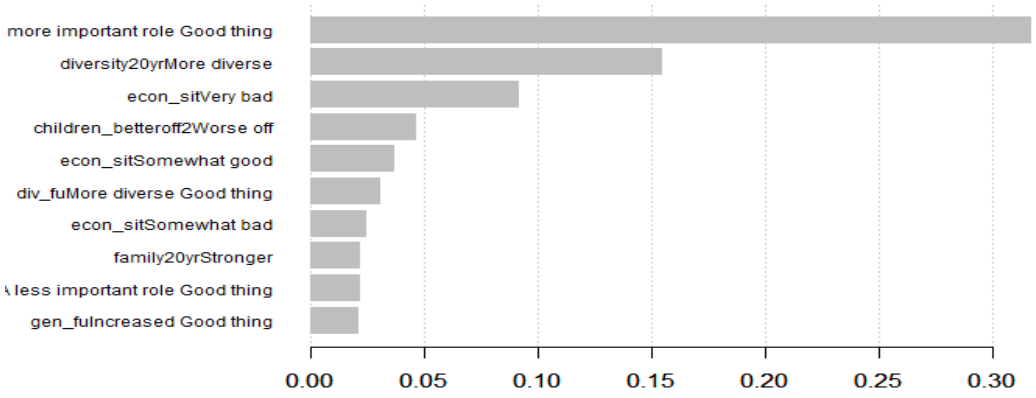
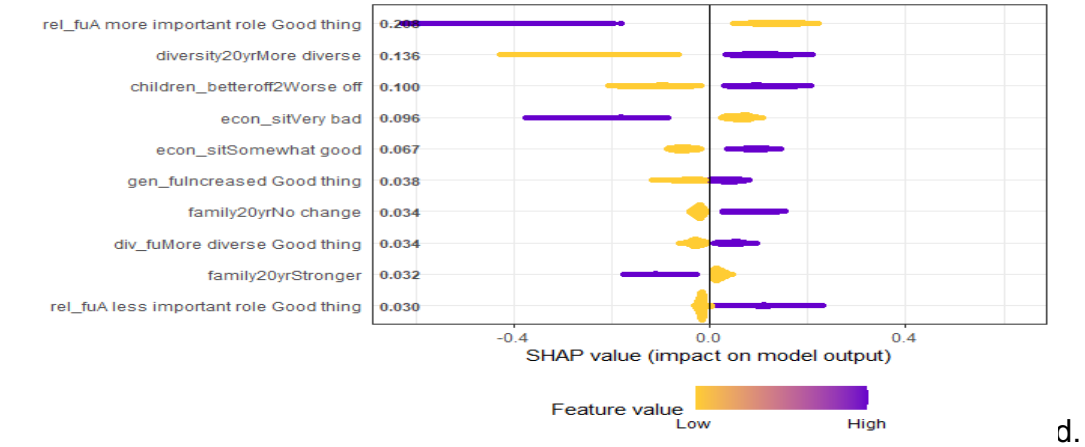


Figure 8: Importance matrix for top 10 predictors

According to the above graph, only three features had their importance values > 0.05. To be considered as a good predictor the importance value should be closer to 1, which we do not see in our research.

Next, we used SHAP (SHaply Additive exPlanation) values, which is similar to the importance matrix; it is a different way to calculate the most important predictors. It calculates the importance of a feature by comparing what a model predicts with and without the feature. Since the order in which a model sees features can affect its predictions, this is done in every possible order, so that the features are fairly compared.



d.

Figure 9: SHAP values for top 10 predictors

Many top values in the SHAP plot match up with the top values in the importance matrix, though they still are not reporting high numbers which further emphasizes that no feature is a great predictor for our data. Additionally, the R^2 value for the model was low (0.28), which reflects the low accuracy of the model.

Conclusion

We attempted to build a model that would predict the happiness level of a person's country of residence based on their answer to survey questions regarding their world-view. The model we built using the Xgboost package was low-performing. The low performance of the model likely is due to the large variation in how people within a certain country answered a survey question. We do think there is a link to be explored between world-view and happiness level, though a better model might be built on data at a higher level of granularity, such as happiness level for a specific person or region.

Citations:

https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

https://en.wikipedia.org/wiki/Gradient_boosting

https://blog.datascienceheroes.com/how-to-interpret-shap-values-in-r/

https://www.analyticsvidhya.com/blog/2016/01/xgboost-algorithm-easy-steps/

https://www.youtube.com/watch?v=3CC4N4z3GJc

Data:
https://www.pewresearch.org/global/2019/04/22/a-changing-world-global-views-on-diversity-gender-equality-family-life-and-the-importance-of-religion/

https://worldhappiness.report/ed/2018/

Appendix A:

Survey questions and responses

Full list of questions with responses and variable names ("DO NOT READ") indicates that the survey reader did not read those options aloud:

Thinking about our economic situation, how would you describe the current economic situation in (survey country) – is it very good, somewhat good, somewhat bad, or very bad?

1 Very good
2 Somewhat good
3 Somewhat bad
4 Very bad
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: econ_sit

When children today in (survey country) grow up, do you think they will be better off, or worse off financially than their parents?

1 Better off
2 Worse off
3 Same (DO NOT READ)
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: children_betteroff2

How satisfied are you with the way democracy is working in our country – very satisfied, somewhat satisfied, not too satisfied, or not at all satisfied?

1 Very satisfied
2 Somewhat satisfied
3 Not too satisfied
4 Not at all satisfied
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: satisfied_democracy

Compared with 20 years ago, do you think the financial situation of average people in (survey country) is better, worse, or do you think there has been no change?

1 Better
2 Worse
3 No change
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: financial20yr

Thinking about the ethnic, religious, and racial makeup of (survey country), over the past 20 years do you think (survey country) has become more diverse, less diverse, or do you think there has been no change?

1 More diverse
2 Less diverse
3 No change
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: diversity20yr

Follow-up (if not 8/9 response): Do you think this is a good thing or a bad thing for (survey country)?

1 Good thing
2 Bad thing
3 Both (DO NOT READ)
4 Neither (DO NOT READ)
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: diversity20yr_fu

Over the past 20 years, do you think equality between men and women in (survey country) has increased, decreased, or do you think there has been no change?

1 Increased
2 Decreased
3 No change
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: gender20yr

Follow-up (if not 8/9 response): Do you think this is a good thing or a bad thing for (survey country)?

1 Good thing
2 Bad thing
3 Both (DO NOT READ)
4 Neither (DO NOT READ)
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: gender20yr_fu

Compared to 20 years ago, do you think religion has a more important role in (survey country), a less important role, or do you think there has been no change?

1 A more important role
2 A less important role
3 No change
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: religion20yr

Follow-up (if not 8/9 response): Do you think this is a good thing or a bad thing for (survey country)?

1 Good thing
2 Bad thing
3 Both (DO NOT READ)
4 Neither (DO NOT READ)
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: religion20yr_fu

Over the past 20 years, do you think family ties in (survey country) have become stronger, weaker, or do you think there has been no change?

1 Stronger
2 Weaker
3 No change
8 Don't know (DO NOT READ)
9 Refused (DO NOT READ)
Ref: family20yr

Follow-up (if not 8/9 response): Do you think this is a good thing or a bad thing for (survey country)?

1 Good thing
2 Bad thing
3 Both (DO NOT READ)
4 Neither (DO NOT READ)
8 Don’t know (DO NOT READ)
9 Refused (DO NOT READ)

Ref: family20yr_fu