Predicting the Perception of Democracy:
Discrepancy Between the Reality and Anticipation of the Level of Democracy

Introduction

Asian Barometer Survey(ABS) aspires to strengthen intellectual and institutional capacity for research on democracy by generating and disseminating scientifically reliable and comparable data (Hu Fu Center for East Asia Democratic Studies, National Taiwan University, n.d.). To accomplish this mission, ABS needs to obtain data that can provide more valuable insights and promote use cases that can showcase the insightfulness of the data. The data sets that ABS provides are very extensive in that they include not only the demographic data but also the individuals' perceptions of society. However, the shortcomings also come from their extensiveness, which is that the data cannot be easily interpreted.

In this project, we predict the discrepancies in the level of a country's democracy, which is the difference in the respondents' current perceived and the desired level of democracy, with latent features discovered by factor analysis. Specific descriptions about the calculation of the discrepancy variable will be shown in the following EDA section.

Furthermore, in order to give guidance about which information is more insightful in the context of a given object of interest, the target variable, we investigate the variables that considerably contribute to the prediction and suggest a direction of survey development. This will allow ABS to design and manage its big data more strategically.

Prior Work/Literature Review

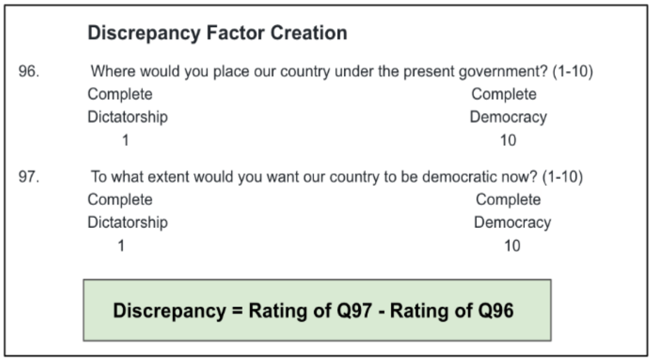
Several past pieces of research have analyzed the satisfaction level of democracy and its relatedness to other socio-economic factors. Global survey-based research done by Richard Wike and Shannon Schumacher (2020), found out that people who were dissatisfied with democracy were also dissatisfied with the current economy and elected officials and had low expectations about the future of their children. However, this study only looks at fragments of data because the satisfaction level of democracy is interpreted with a single dependent factor at a time. In other words, it fails to capture the complete dynamics of socio-political perceptions.

Hence, in this study, we will use big data from ABS and build a machine learning model to see the full picture and quantify the dynamics of socio-political perceptions. Here, we introduce the “discrepancy” variable as our target variable and try to predict it with the survey responses. The significance of the discrepancy variable lies in the fact that it captures the direction (+ or -) of differences in the desired level and the current level of democracy, while satisfaction is just a scalar metric. In addition, the perception about the current level of democracy alone should not be used since the response may be tainted by political factors influencing one’s perception of the present. For example, a rightist person under a left party may rate the current level of democracy low when really, he or she is just dissatisfied with the ruling party. Therefore, the discrepancy variable is more suitable to employ as a factor to diagnose the sentiment about democracy.

Exploratory Data Analysis

The original data set contains over 12,000 survey responses from eight Asian countries. There are 217 columns, each presenting answers to survey questions. We excluded columns that had too many categorical variables or were not interpretable and were irrelevant. Also, we excluded columns with 40% or more null values. We then plotted a correlation matrix shown in (Exhibit 1) with the highest correlation values. These questions were the ones not asked in specific countries, which obviously leads to a high correlation in null values. After data cleaning and manipulation, we were left with 98 columns and 6875 rows.

As shown in Figure 1, our new target variable discrepancy is a simple subtraction of the current level of democracy (question 96) from the desired level of democracy (question 97). Two questions are responded to on a scale of 1(complete dictatorship) to 10(complete democracy). The distribution of the discrepancy variable is approximately normal as shown in Figure 2.



Discrepancy Factor Creation

96. Where would you place our country under the present government? (1-10)

Complete Dictatorship 1	Complete Democracy 10

97. To what extent would you want our country to be democratic now? (1-10)

Complete Dictatorship 1	Complete Democracy 10

Discrepancy = Rating of Q97 - Rating of Q96

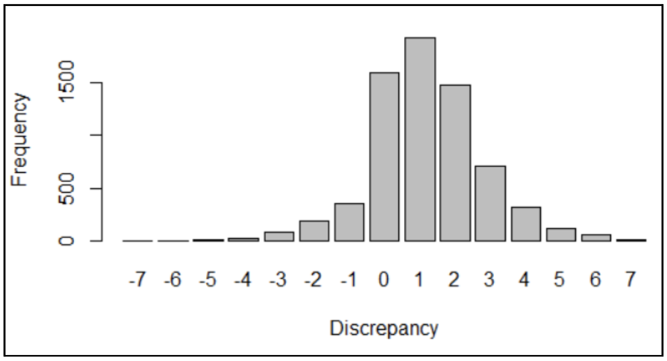


Figure 1. Discrepancy Factor Creation

Figure 2. Distribution of the Discrepancy

We determined the number of features via the Figure 3 shown below. According to the graph generated via the elbow method, the optimal number of factors is 17. These factors can be characterized by the questions to which they show a significant correlation and thus can be interpreted as a latent feature. For example, feature MR11 shows high correlations with questions 7 through 14, which were a subset of questions regarding trust in institutions and media. Hence, we have created a subset of the field of questions regarding institutional trust. Moreover, the most insightful finding was that we were able to create cross-sectional question groups. Survey questions in different categories were grouped under latent features, which were not interpretable before factor analysis. Feature MR2 for instance, incorporated sentiments about three different fields of questions: trust government, authoritarian values, and approve the government. The rest of the features and their characteristics are specified in Exhibit 2.

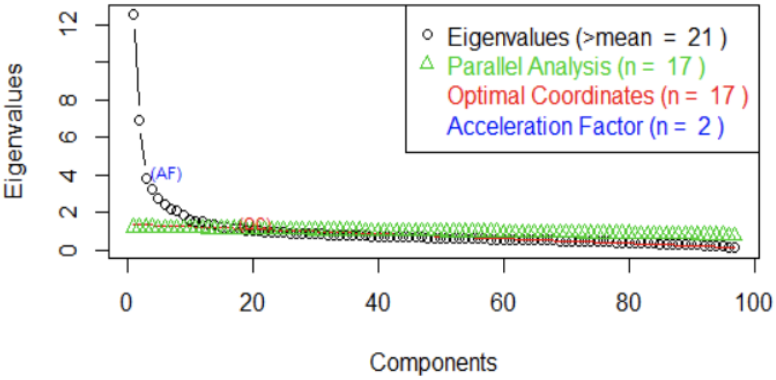


Figure 3. Optimal Number of Factor Selection via Elbow Method

Proposed Methodology

We randomly assigned 75% of the dataset as training data and 25% as test data to evaluate our model performance based on Out-of-Sample Root Mean Squared Error. For predictive modeling, we first generated a basic multivariate linear model and obtained a Root Mean Squared Error of 0.958. We then ran a random forest, which is a decision-tree-based ensemble model well known for its high prediction accuracy. From this model, we obtained a Root Mean Squared Error of 0.927. In addition, we deployed Lasso with minimum lambda for variable selection and built a Post-Lasso Random Forest model based on those selected variables to investigate whether variable selection helps avoid overfitting the data and to compare the pre-and-post variable selection model performance. From the Post-Lasso Linear model and Random Forest model, we obtained Root Mean Squared Error of 1.059 and 1.197 for each. We concluded that the Pre-Lasso Random Forest model showed the best model performance and thus decided to pursue this model as our final model.

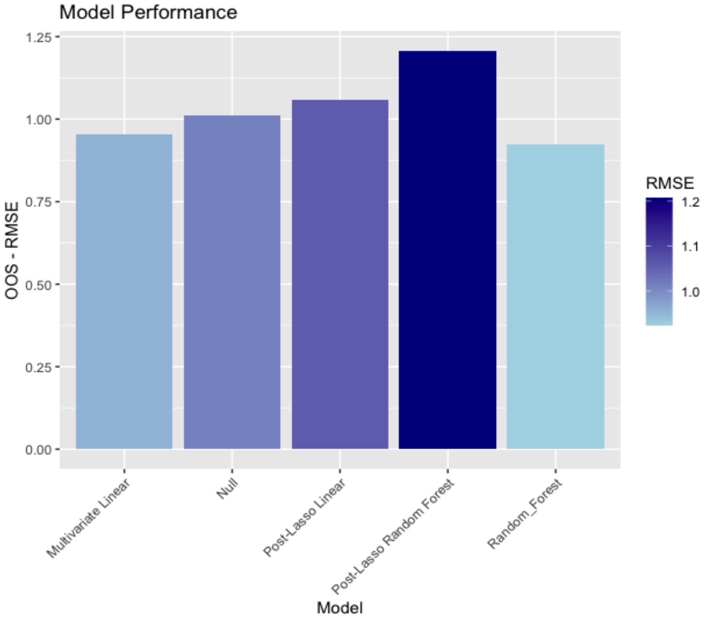


Figure 4. Out-of-Sample RMSE of Each Model

Analysis

To interpret the importance of each variable in our final random forest model, we analyzed the percent Increase in mean squared error, hereafter, %IncMSE. The %IncMSE is computed from permutating Out-of-Bag data. The mean squared error on out-of-bag data of each tree is recorded before and after permuting each variable. Then the differences between two mean squared error values are averaged over all trees and normalized by the mean squared error of the differences (*Package randomforest* 2018). This allows us to evaluate the importance of each variable in explaining the target variable, discrepancy. Table 1 is the values of %IncMSE for each input variable sorted in descending order.

Rank	Variable	%IncMSE
1	**MR5**	**31.5177001**
2	**MR3**	**30.4169487**
3	**MR4**	**24.8379523**
4	**q098**	**24.554054**
5	MR1	22.0888807
6	MR10	19.8284293
7	MR11	19.7816168
8	MR17	18.720374
9	MR2	18.6312169
10	MR9	17.6759455
11	MR14	17.4696935
12	MR8	17.345875
13	MR15	17.3247357
14	MR16	17.186262
15	q105	14.3884006
16	MR13	14.3593808
17	MR7	14.1863974
18	MR6	13.5435286
19	MR12	11.9521618
20	q010	11.8851639
21	q121	11.6222232
22	q123	10.7768927
23	q008	10.7366549
24	q009	9.98698487
25	q106	9.50523422
26	q007	8.94715284
27	q006	8.50084331
28	q128	7.25873329
29	q127	5.7442917
30	q027	5.62658169
31	q005	5.17860828

Table 1. Importance of Each Variables (% Increase in MSE)

MR3, MR5, q098, and MR4 were found to be most important in predicting the target variable. MR3 is a latent feature that represents the evaluation of the current regime compared to the most recent authoritarian rule on its degree of corruption, maintaining social orders and economic development. MR4 features the current regime evaluation on freedom and equality, compared to the most recent authoritarian rule, and MR5 features country, religion, and annual income level. Q098 asked about the individual's level of satisfaction with the way democracy works in one's country.

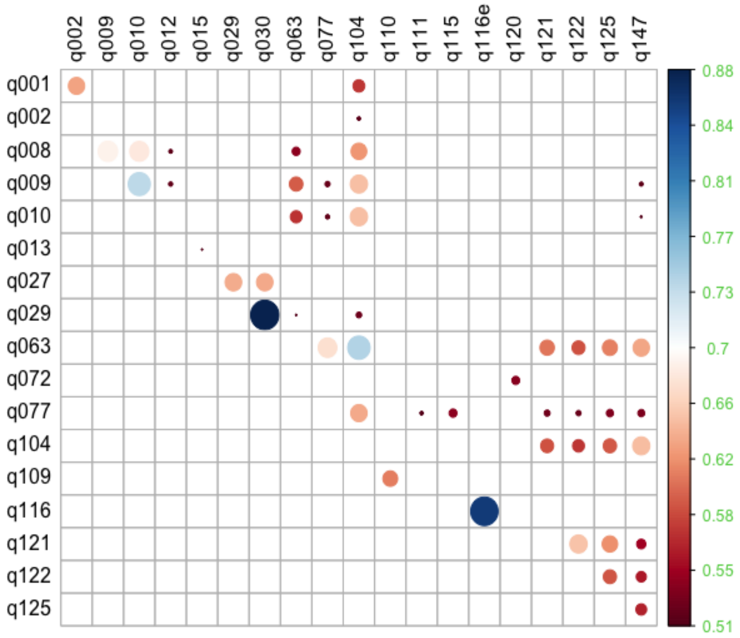
Conclusion

Our model's prediction accuracy was not as high as it ideally would be due to data structure and our limitation of resources. However, we were still able to generate a valuable insight in terms of the use of latent features as a dependent variable in predictive modeling. In our model, three out of four most important variables were latent features defined from factor analysis. Therefore we believe that our data mining structure holds great value in the fact that it identifies hidden important factors of prediction models. We believe that the following studies can refer to our approach to similar analysis.

In addition, some undiscovered features in ABS such as MR5 and perception of the degree of freedom of speech(question 105) showed higher importance in our predictive model. Discovering that there is differential importance among the questions, we are able to identify which area of interest ABS must focus on in order to generate a more insightful analysis. Hence, we suggest that ABS generates and collects more related questions in such areas. Meanwhile, current survey questions are proxies rather than a direct indicator of the subjects in question. ABS will be able to make its data set more versatile if it devises questions that can provide more direct measurements.

Appendix

(Exhibit 1) Correlation Matrix of Na



(Exhibit 2) Extracted features by factor analysis

Feature #	Feature Explanation
MR2	trust government & authoritarian values & approve the government ("keep it as it is" ideas)
MR11	trust in institutions
MR10	participation in election
MR3	**current regime evaluation** compared to most recent authoritarian rule, on corruption, maintaining social orders and economic development
MR4	current regime evaluation compared to most recent authoritarian rule, on freedom and equality
MR5	**country, religion, and annual income level**
MR14	political party association & democratic legitimacy and preference
MR8	subjective self-eval on social and economic status, and economic situation of the country
MR7	authoritarian values and obeying for the "State"
MR6	witnessing government corruption or bribe-taking
MR12	traditionalism (emphasis on personal relations and fate)
MR9	political participation - took social action on solving personal, family, or

	neighborhood problems, or problems with government officials and policies
MR1	democratic legitimacy and preference, partisanship, education demographics
MR13	citizen empowerment, system responsiveness and political support
MR16	evaluation of current regime and future economy
MR15	political participation - took other action on solving personal, family, or neighborhood problems, or problems with government officials and policies
MR17	trust in press and media

References

Hu Fu Center for East Asia Democratic Studies, National Taiwan University. (n.d.). *Program Objectives*. Asian Barometer Survey. Retrieved November 6, 2021, from http://asianbarometer.org/intro/program-objectives.

Richard, W., Shannon, S. (2020). *Democratic Rights Popular Globally But Commitment To Them Not Always Strong.* Pew Research Center https://www.pewresearch.org/global/2020/02/27/satisfaction-with-democracy/

Fortran original by Leo Breiman and Adele Cutler, & R port by Andy Liaw and Matthew Wiener. (2018, March 22). *Package randomforest*. CRAN. Retrieved November 7, 2021, from https://cran.r-project.org/web/packages/randomForest/index.html.