Modeling Willingness to Vote

By Carlos Gustavos Salas Flores, Aryan Poonacha, Jincheng Wu, Biniam Garomsa

https://github.com/cs582/DukeDatathon2021

I. Introduction

We investigate the likelihood for an individual to vote in corrupt and less-corrupt Asian democracies and overall, as measured by survey results and corruption indexes, using demographic data as predictors. We create 2 models, one for corrupt and the other for less corrupt countries, and generalized them to both waves. We tested a final model using all countries from both waves to predict likelihood to vote.

II. Literature Review

Previous studies on the correlation between corruption and voting are mainly focused on how corruption changes the preference or choices of voting, but not the willingness of voting. Nicholas Charron and Andreas Bågenholm (2015) have done this research in European countries, and Eric Chang (2019) has similar work focusing on Asian democracies and is also utilizing the Asian Barometer Survey dataset. Our work attempts to predict the general willingness of voting based on the demographic features of the voters, and compare between groups of corrupt and less corrupt Asian democracies to find out if corruption of the governments affects willingness to vote. With that being said, our work is closely related to those previous ones, and we have also attempted to migrate some conclusions from previous works to help explain our models and results.

III. Initial Exploratory Data Analysis

Initially, to explore the large dataset, we divided the data columns into 3 main 'anchor' categories for consideration: countries, individual demographics (Age, Gender, etc.), and survey variables (question response data). We made visualizations of specific demographic datapoints to view their suitability and distribution as predictors. We also looked at a large correlation matrix of all the variables together, and then subset down to only include large correlation pairs, and considered possible causal hypotheses and prediction mechanisms that might explain some of them with specific visualizations (appendix 1).

After considering a few possibilities, we decided to focus on specifically using demographics to predict likelihood of having voted in the previous election (willingness to vote), and comparing two such models: one for corrupt democracies and the other for less corrupt democracies.

We then compared the predictors to each other and to the willingness to vote outcome variable with more visualizations, and to easily identify with visuals clustered areas where specific demographics were more likely to vote or not vote (appendix 2).

IV. Methodology

1. Choosing Variables: To measure likelihood to vote, we use the survey question, "Have you voted in the last election?" as the independent variable we want to predict (q027 in wave 1, q38 in wave 2). To find predictors, we created a large correlation matrix and subset down to only those with at least some significant correlations, which essentially complemented the list of suggested important variables in the Github page, and was primarily demographic data: age, gender, income, etc.
2. Cleaning Dataset: We excluded survey responses of 'chose not to answer', 'do not understand the question', etc. for the specific survey question q38/q027 response. We projected the dataset to

only include the predictor and predicted variables and removed non-democratic countries that don't hold elections from the dataset. Other miscellaneous cleaning (remove NA values, etc.)

3. Defining Corruption: We initially use the survey responses (specifically, 2 questions that asked respondents to rank corruption at the municipal and national levels respectively, Q117 & Q118) to rank the countries in a 'custom' corruption index, and then compare that ranking to a general measure of corruption, the Corruption Perceptions Index (https://www.transparency.org/en/cpi/2008) It is worth mentioning that during analysis, we found discrepancies between survey responses and corruption indices. Many countries that were considered highly corrupt by survey responders were ranked low on corruption indexes, and vice versa. This is further discussed in the analysis.
4. Initial hypotheses: Individuals in corrupt countries are less likely to respond that they have voted in the last election than individuals in non-corrupt countries.
5. Creating And Testing Models:

We initially only worked on Wave 2 data, and then generalized to both waves in the last step.

Step 1: We first fit and tested SVM models for each country (in wave 2) with the appropriate test and training data splits. We then split the data into 2 groups: corrupt and less corrupt, and then ran separate models to predict willingness to vote for both. We iterated through different hyperparameters and kinds of regularization to ensure the models were fitting well in all the above contexts for test data.

Step 1.5: We also created dataframe subsets for each country and tested our models on them to ensure that they would generalize well to all countries' data, and there's no country-specific variation that might skew willingness based on country characteristics not captured by the predictor demographic data.

Step 2: We then merged the wave 1 and wave 2 datasets and applied the same data cleaning pipeline, and ran the 2 models on the corrupt and less corrupt splits.

Step 3: Finally, we used the entire merged wave 1 and wave 2 dataset to create a general model to predict willingness to vote using only demographic controls.

V. Models And Analysis

Dependent Control X variables: Gender, Marital Status, Level of Education, Age, Income

Predicted/Independent Y variables: Probability of having voted in last election

The answers to question "Have you voted in the last election?", is either yes or no, and would make the data linearly separable. So we hypothesized that a linear SVM model would best fit it, and also compared it with and experimented with logistic regression models.

Step 1 results: after iterating through a few variations and different hyperparameters, both models were able to achieve high accuracy with a linear SVM: ~78%* accuracy on test data for the less corrupt countries model and ~86%* accuracy on test data for the corrupt countries model.

Step 1.5 results: all countries' individual dataframes did well with the model, with high accuracy scores that reflected the models for each of the 2 groups in step 1 (70-90%*).

Step 2 results: After merging wave 1 and wave 2 and applying the same corrupt/less corrupt split as in step 1, we create 2 SVM models with 79%* (non-corrupt) and 81%* (corrupt) accuracy, respectively.

Step 3 results (Final Combined Model): When combining the datasets and using all countries together, the linear classifier had an 84%* accuracy score to predict willingness to vote on the test set. Thus, the model was successfully able to generalize to the entire dataset, and separate models for corrupt and less corrupt controls don't generalize as well; as such, corruption as a predicting variable for willingness to vote is not significant. We can affirm this by plotting responses to questions on corruption and responses on willingness to vote (see Appendix 4).

*: all the accuracy percentages mentioned are with models produced at the time, exact percentage on reproduction can slightly vary.

We have identified three potential sources of error. One is the variance in demographic information from the dataset. Another plausible explanation is that the assumption of voters in corrupted countries being less likely to vote is inaccurate. Eric Chang (2019) brings forth an argument that goes against our initial assumption. He points out that "institutionalized corruption promotes greater electoral tolerance of corrupt politicians in Asian democracies" (Chang 2019, 307). This suggests that voters living in countries with history of corruption may already have had corruption perceived as an integrated part of the functioning mechanism of the political system, and thus are more tolerant of corrupt politicians and are still willing to vote for them.

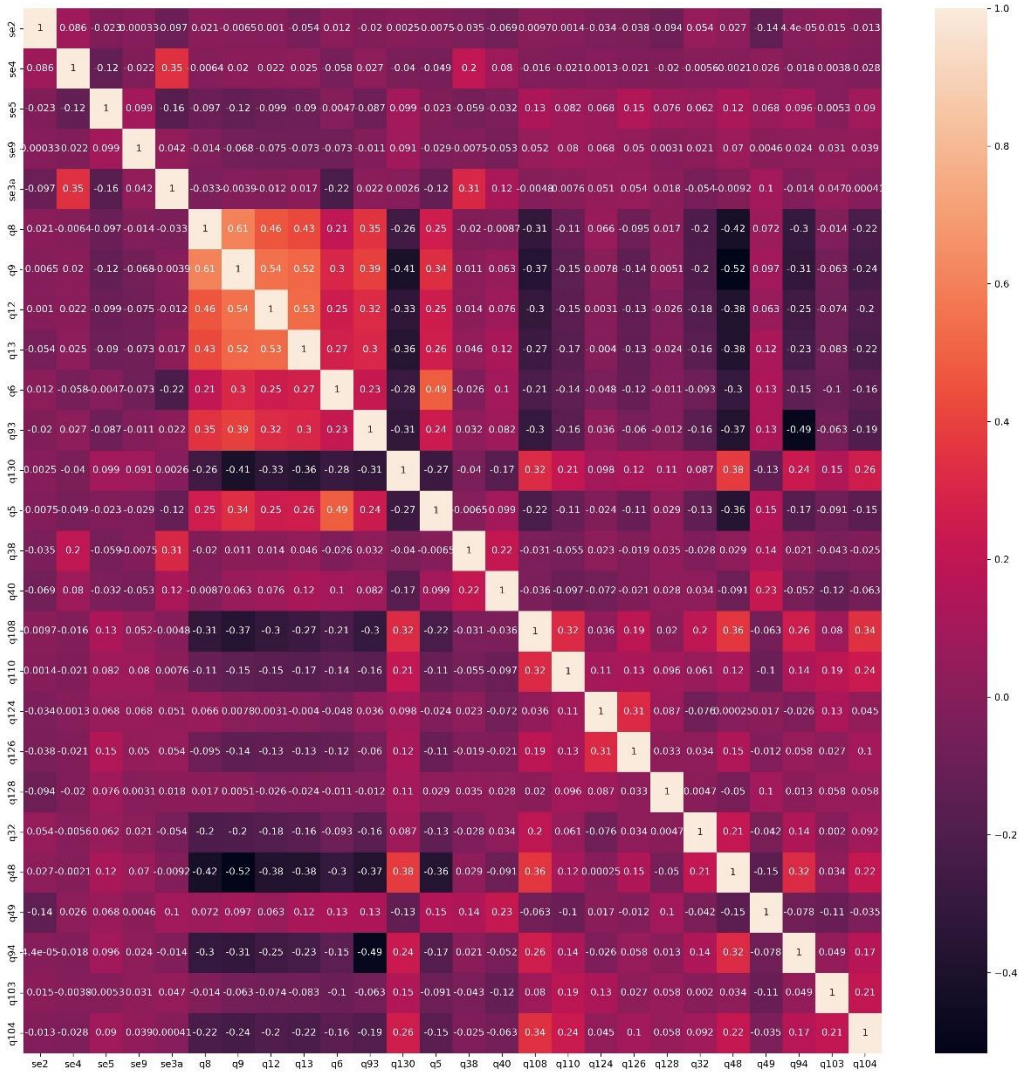
Another dimension that may interfere with the correlation between corruption and voting is the party system. Charron and Bågenholm (2016) discover that when voters have limited party options, political ideologies would prevail and drive voters to vote for their preferred parties regardless of corruption scandals. When the number of effective parties in the system is large, the effects of corruption on voting behavior would be more significant because people have more choices that they can choose from. While this is also an analysis focusing on the specific voting behavior but not willingness to vote, the logic and conclusion still can be migrated to our research and help point out a potential source of variance. Among the countries in our research, some have one party dominant system (e.g., Singapore and Japan), some have single party system (e.g., Vietnam), and some have multiple party coalition governments (e.g., Thailand). We also suspect party systems resulted in the discrepancies of perception of corruption and the fact. Responders from countries with multi-party systems tend to underestimate corruption, while responders under single party and one-party dominant system tend to overestimate corruption. It is beyond the scope of this research to quantitatively analyze the influence of the party systems and integrate this variable into our model, but we do recognize the this as an influencing factor that is worthwhile of further investigation.

VI. Conclusions

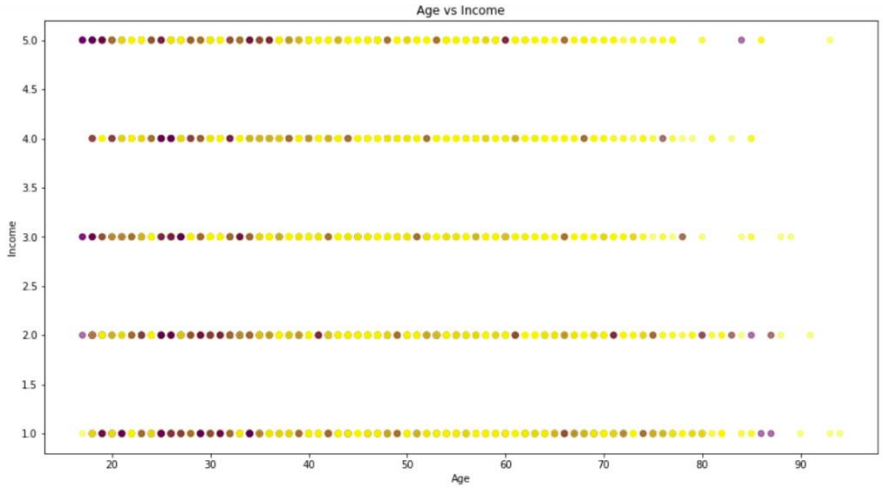
1. From our final model, we clearly see an increase in accuracy with the combined dataset; clearly, the splitting into corrupt and less corrupt groups doesn't influence willingness to vote as much as general demographics. A combination of general demographic characteristics is the best general predictor of willingness to vote overall, which our final model does with 84% accuracy.
2. From existing literature, 'obviousness' of corruption in a given society and confidence in government machinery that exists alongside transparent corruption and a multi-party or dual-party system that gives more choices can also influence willingness to vote; these should be avenues for further investigation.

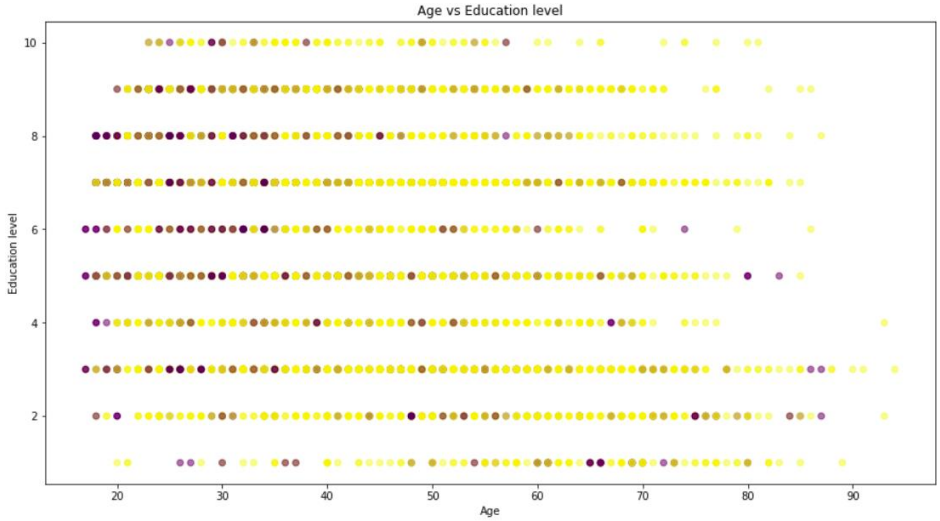
VII. Appendix

1. Correlation Matrix of variables



2. Demographic data plots showing uniform distribution (purple means voted, yellow means didn't vote; lower overlap of purple points look brown), and concentration of voters vs. non-voters for certain demographics (richer, older, etc.)





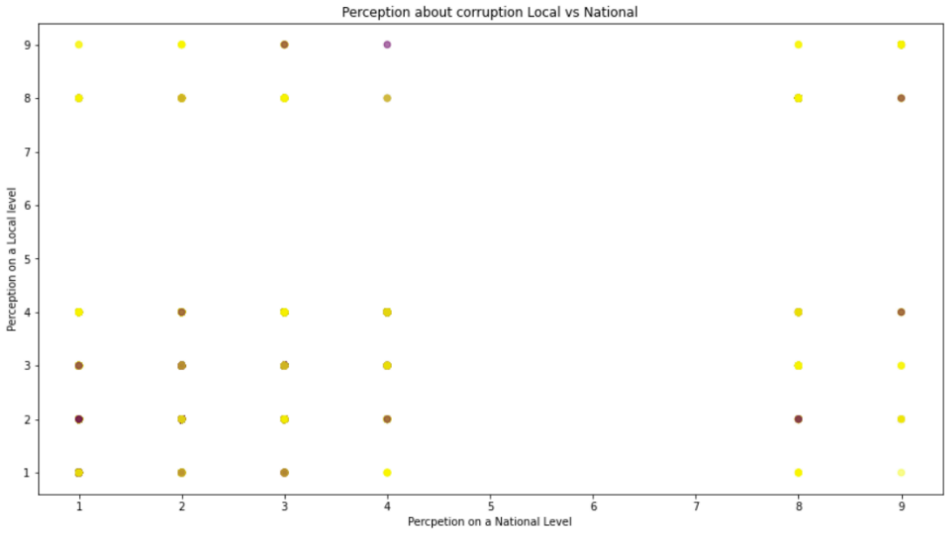
3. Ranking of Countries by (left to right) surveyed perception of corruption at national and municipal level, and CPI 2008 rankings:

(Least corrupt to most corrupt going down. 'x' marks a country whose rank in CPI is significantly different from the surveyed ranking.) Same country across cols = same color

Q117 – Perception of Corruption at a Local Level (wave 2)	Q118 – Perception of Corruption at a National Level (wave 2)	Corruption Perceptions Index (2008)
Mongolia	Taiwan	Singapore (Rank 4) x
Taiwan	Philippines	Japan (Rank 18)
Philippines	Malaysia	Taiwan (Rank 39)
Indonesia	Mongolia	South Korea (Rank 40)
South Korea	South Korea	Malaysia (Rank 47)
Malaysia	Japan	Thailand (Rank 80)
Japan	Indonesia	Mongolia (Rank 102) x

Thailand	Thailand	Vietnam (Rank 121)
Vietnam	Vietnam	Indonesia (Rank 126) x
Singapore	(Singapore – No Information)	Philippines (Rank 141) x

4. Plotting responses on questions about corruption and responses on question on willingness to vote:



VIII. References

Chang, Eric C. "Corruption Predictability and Corruption Voting in Asian Democracies." *Public Choice* 184, no. 3-4 (2019): 307–26. https://doi.org/10.1007/s11127-019-00760-x.

Charron, Nicholas, and Andreas Bågenholm. "Ideology, Party Systems and Corruption Voting in European Democracies." *Electoral Studies* 41 (2016): 35–49. https://doi.org/10.1016/j.electstud.2015.11.022.

"2008 - CPI." Transparency.org. Accessed November 7, 2021.

https://www.transparency.org/en/cpi/2008.