Zhaocheng (Raymond) Gu

2101 Wisconsin Avenue NW, Apt. 232, Washington, DC 20007 · zg127@georgetown.edu · (240)885-9263

WRITING SAMPLE

The following document is my essay for Intro to Data Science at Georgetown. This essay uses survey data and data science techniques to predict whether a Chinese trusts the local government or not. This pdf document is written by using R Markdown. The eassay is my own work. Please do not distribute it without my permission.

Introduction

Political trust has been a hot topic in political science. This project tries to build a model to best predict whether an individual in China trusts the local government or not, based on the knowledge learned in this course. And the success of this project depends on whether a best model can be built. For the rest of this report, we will go through 5 parts, including problem statement and background, data and data processing, analysis, results, and discussion.

Problem Statement and Background

Institutional trust is the extent of trust in governmental institutions such as the legal system, the parliament, and the police (Mishler & Rose, 2001). Better institutional trust can greatly promote social development through enhancing the political participation, regulatory compliance, and entrepreneurship performance (Ding, Au & Chiang 2015; Kwon, Heflin & Ruef 2013; Nunkoo & Smith 2013; Sohn & Kwon 2016).

Since the "reform and opening", China has changed from a totalitarian state to an authoritarian one and the society is getting freer and more diverse than before, which means Chinese people now are able to have different views toward the government and can even express them to some extent. As an authoritarian state, China has been successful in "winning" public support and trust from Chinese through "excellent" economic performances(Soest & Grauvogel, 2017) and propaganda.

And there are a number of articles that talk about factors that influence political trust in China. Manion(2006) have emphasized the importance of electoral democracy to the trust in local authorities in rural China. Chen and Shi(2001) pointed out the negative correlation between media exposure and political trust in China. Zhao and Hu(2017) suggested that improving satisfaction with the quality of public services is important to enhance trust in the city government and promoting national democracy is important to enhance trust in the central government. Also, they found that "younger citizens with higher education and higher income have less trust in central government"(Zhao & Hu, 2017). Recently, some scholars also examined the relationship between internet use and political trust in China but the results are mixed(Zhou et al., 2019; Lu et al., 2020)

Overall, the researches so far on political trust in China have adopted a "macro"-perspective and have focused on finding certain relationships between factors and people's attitudes towards the government. Those researchers have not tried to predict that attitude for individuals in China. However, this project will be based on a "micro"-perspective, which is to use a series of factors that may have impacts on political trust in China and machine learning techniques to build a best model to predict whether an individual trusts the government or not.

Data

In this project, we are going to use an existing dataset from CGSS(Chinese General Social Survey) that is administrated by Renmin University of China and is an very comprehensive survey covering demography, economy, society, and politics. And the 2010 wave of that survey is where our dataset has been drawn from, since that wave is the most comprehensive one in recent years and can provide sufficient information needed for our project. Each value of the CGSS dataset is from the response to each question by those respondents, which means the observation of the dataset as well as our analysis is obviously each respondent.

The dependent variable used here is the subjective level of political trust in local governments. The subjective level of political trust in governments is a quite straightforward variable in terms of our research question. And the local governments in China are usually the direct public services providers and are connected with local residents more tightly, which thus are a more suitable one to explore here compared to the central government.

This project has also used 12 other variables to predict the individuals' attitudes toward their local governments, including economic, political, demographic, and geographic factors. Except that family income and age are numeric variables and are measured by RMB(Ren Min Bi) and years respectively, all other variables are categorical ones.

There are five demographic variables, including age, gender, ethnicity, education attainment, and religious status. For age, we may expect that a younger people in China will be less likely to trust the local government, since he or she may be more radical than those elders. The gender can also affect an individual's attitude toward the local government, since politics are always considered to be male issues rather than female ones in China especially in those rural areas. Due to the fact that Han people are the dominant group in China and the most powerful positions are occupied by Han people in the local government as well as the central one, being a ethnic minority can also have influence on the attitude toward the local government. It is doubtless to include the education attainment in this analysis, since it is highly related to a individual's social status and the perspective that he or her think about the government. For the religious status, it can also be influential since the communist party advocates atheism.

For economic factors, total family income in 2009 and self-perceived economic status locally are chosen, which cannot only measure the absolutely economic conditions but also the relative ones that the individual thinks their families have.

Four other variables(province, CCP membership, household type, and internet utilization) have also been added to the analysis. Firstly, different provinces provide different public services based on their financial conditions and this thus may influence public trust in themselves. Also different resident in different regions of China may have different attitudes toward the government due to different cultural and historic factors. Secondly, whether an individual is a CCP(Chinese Communist Party) member can matter due to the party-state system. Thirdly, the household type of an individual determine where he or she can get public services and how he or she can get access to them. Due to the "dual economy" in China, people with different household types get different kinds of public resources provided by the local government, which can be an important factor that influence their attitudes toward the local government. Fourthly, individuals' internet utilization may matter as well since the new media based on the internet is really hard to control and can disseminate the information "different" from the official one.

After choosing appropriate variables, we need to clean the data. In order to make this project a classification, the subjective level of political trust needs to be transfer to whether a respondent trust the local government or not. Categories of ethnicity, religious status, and internet utilization should be grouped into only two, since the amount of observations in particular one group is larger than the total amount of observations in all other groups(Han versus not Han, not religious versus religious, and not regular versus regular). Also, Categories of some other variables including education, household type, region, and economic status need to be rearranged, due to small amounts of observations in some categories and the purpose to show "typical" attributes of these variables. Finally, since the data is from a survey, we need to transfer some "weird" values like -999999 into missing values. For the process above, I just used some basic functions like mutate, select, tibble, and left_join in tidyverse to achieve it. Here are the detailed descriptions of all the variables used in this project below.

Then, we should split the whole dataset, 80% of which is used for training models while the rest of it is used to test models. And characters in those categorical variables should be transferred into factors in R, in order to make it easy to create dummy variables for them while processing.

Table 1: Descriptions of variables

Variables	Descriptions
trust_or_not	Whether the respondent thinks the local government is trustworthy or not in 2010
age	The respondent's age in 2010
sex	The respondent's gender
ethnicity	Whether the respondent belongs to Han or not
educ	The education attainment of the respondent until 2010: primary school or below, middle, and some college or above
religion	Whether the respondent is religious or not in 2010
immigrate	Whether the respodent is an immigrant from another place in 2010: not immigrant, immigrant, and other
household_type	Which household registration type the respondent has in 2010: agricultural, urban, and others
region	Which region the respondent live in in 2010: east, northeast, center, and west
family_inc	The respondent's family income in 2009
econ_status	The respondent's self-perceived economic status of their family locally in 2010: below average, average, and above average
CCP	Whether the respondent is a CCP member or not in 2010
internet_use	How often the respondent used the internet in 2009: not regular and regular

skim_type	skim_variable	n_missing	complete_rate	factor.ordered	factor.n_unique	factor.top_counts	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.hist
factor	trust_or_not	53	0.9943767	FALSE	2	yes: 6086, no: 3286	NA	NA	NA	NA	NA	NA	NA	NA
factor	sex	0	1.0000000	FALSE	2	fem: 4908, mal: 4517	NA	NA	NA	NA	NA	NA	NA	NA
factor	ethnicity	17	0.9981963	FALSE	2	Han: 8519, Not: 889	NA	NA	NA	NA	NA	NA	NA	NA
factor	educ	10	0.9989390	FALSE	3	mid: 4554, pri: 3400, som: 1461	NA	NA	NA	NA	NA	NA	NA	NA
factor	religion	4	0.9995756	FALSE	2	not: 8205, rel: 1216	NA	NA	NA	NA	NA	NA	NA	NA
factor	immigrate	36	0.9961804	FALSE	3	not: 8464, imm: 901, oth: 24	NA	NA	NA	NA	NA	NA	NA	NA
factor	household_type	5	0.9994695	FALSE	3	agr: 4801, urb: 4150, oth: 469	NA	NA	NA	NA	NA	NA	NA	NA
factor	region	0	1.0000000	FALSE	4	eas: 3486, wes: 2450, cen: 2291, nor: 1198	NA	NA	NA	NA	NA	NA	NA	NA
factor	econ_status	23	0.9975597	FALSE	3	ave: 4721, bel: 3832, abo: 849	NA	NA	NA	NA	NA	NA	NA	NA
factor	CCP	14	0.9985146	FALSE	2	no: 8258, yes: 1153	NA	NA	NA	NA	NA	NA	NA	NA
factor	internet_use	49	0.9948011	FALSE	2	not: 7443, reg: 1933	NA	NA	NA	NA	NA	NA	NA	NA
numeric	age	3	0.9996817	NA	NA	NA	47.32201	15.70885	17	36	46	59	96	
numeric	family_inc	1162	0.8767109	NA	NA	NA	41178.04454	81076.77603	0	12000	25000	46800	2800000	

I have used the skim function to show the distribution of each variable below. There are obviously missing values in the dataset since the data is from a survey and respondents may not be able to answer some questions. And we can use KNN to "fill" those missing values based on corresponding non-missing values from their k nearest neighbors. As the data summary shows from below, there is an imbalance of classes for the dependent variable, which can be addressed by setting different sampling method in the cross validation. And we also need to get logarithms of family income and then rescale both two numeric variables, since the distribution of family income is right skewed and the units of age and family income are not the same. Lastly, we need to create dummy variables for those independent variables that are categorical to built our models.

Analysis

With the processed data, we can now build our models. I have tried almost all methods learned in this course, including logistic regression, k-nearest neighbors, random forest, classification trees, and support vector machines models, since the aim of this project is to find a model with the best performances.

The logistic model originates from the regression model but transforms its estimates such that the predicted probabilities can only take values between 0 and 1. Although the logistic model allows the relationships between the Xs and Y to be non-linear, it represents a "s-curve" on a probability space, which is a strong assumption and may not be the case in reality. But this model is quite easy to interpret its coefficients for each Xs and is thus often used in empirical researches in social sciences since it originates from the OLS model.

The logic of k-nearest neighbors model is to find k nearest neighbors of a certain observation in the entire training data and then classify it based on the "majority vote". This algorithm is easy to implement but must ensure that all numeric variables have been rescaled so that the distance among observations will not be biased by different ranges of different variables. And KNN may have poor performance in high dimensions.

Classification trees model is also easy to understand. For a classification tree, we start from a variable and choose a certain value to break, so that the observations with different categories can fall into corresponding areas following the breaking point. Then we make similar splits based on the existing splits as best as we can for several times and a model can be built. Another thing to point out for this model is that the number of splits can influence the result of the model, which is that less splits can result in under-fitting and more splits can result in over-fitting.

The random forest method is based on the classification trees method. What a random forest method does is to grow many classification “trees” and chooses an “average” one of them. Since the random forest involves many independent predictions and is “random”, it always performs better than some other algorithms with a relatively higher accuracy and a smaller possibility to be over-fitting.

Finally, what the support vector machine methods are going to do is to try to draw a boundary to separate observations with different categories as best as we can. And there are three different ways to “draw” such a boundary. The first one is to “draw” a linear boundary and a second one is to “draw” a curve(polynomial) one, while the third one is to draw a “circle”(radical boundary). Also, we can set how many wrong classifications the model can make to adjust this model.

Results

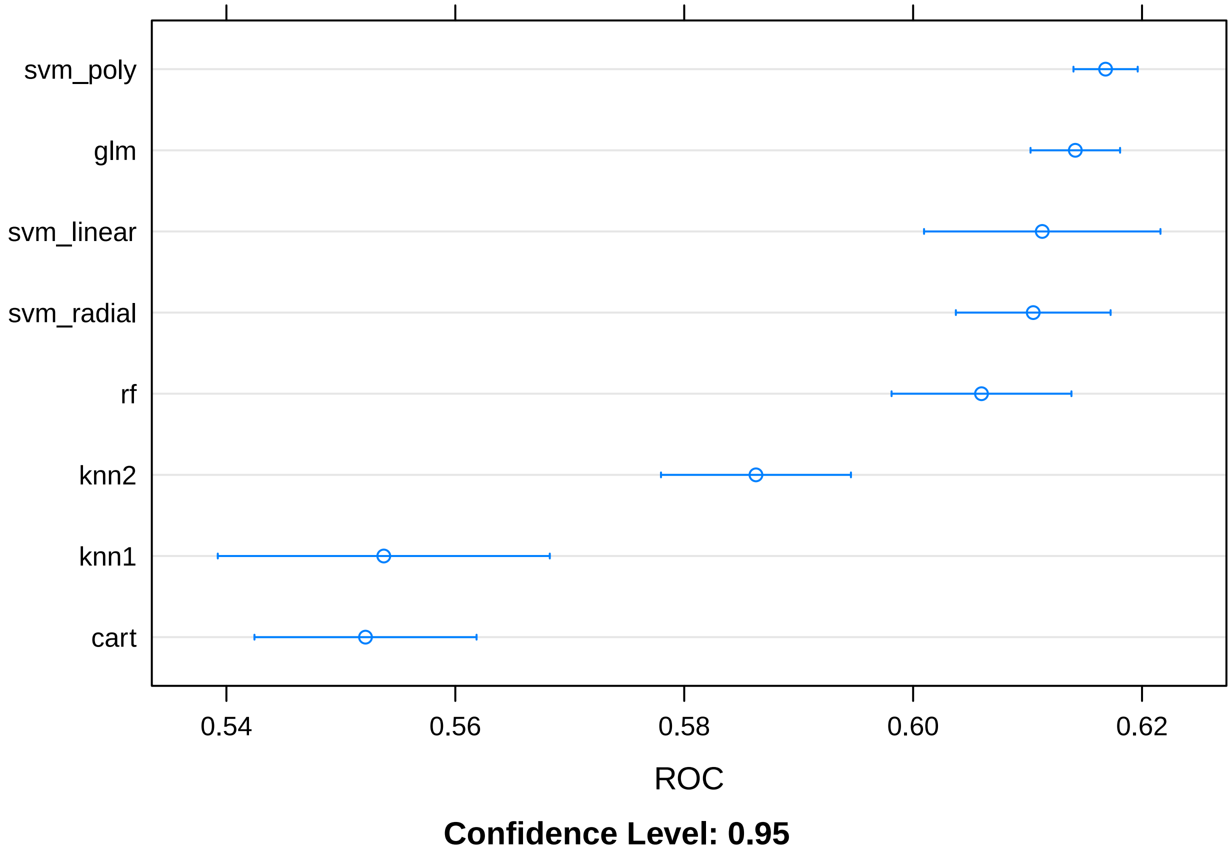
As we can see from Figure 1, all models we have built have ROCs larger than 0.5, which means those models are useful to predict the individuals’ attitudes toward the local government in China through using the variables mentioned above to some extent. Specifically, while two KNN models and the classification trees model perform the worst, other five models(random forests, linear boundary, radial boundary, logistic model, and polynomial boundary) all have ROCs larger than 0.6. The best two models(logistic model and polynomial boundary) seem to even have ROCs larger than 0.61.

From the comparison, the support vector machine model with the polynomial boundary performs the best, while the logistic model performs slightly worse than that. However, a warning message saying that the support vector machine model with the polynomial boundary ”fails to search lines and returns NAs. Due to this warning, we may think the comparison is biased and we can turn to the logistic model since it only performs slightly worse than the support vector machine model.

Table 2: Performance on the testing data

.metric	.estimator	.estimate
roc_auc	binary	0.3889198
accuracy	binary	0.5441052

Figure 1: Comparison among different models



However, when I try to apply the logistic model to the testing data, it performs far more worse than it did on the training data. Its ROC is only about 0.36 that is far more below than 0.6 and it only makes right predictions for about half of the testing data.

Still, we want more insights from the “best” model we got from this project, although it performs badly on the testing data. From Figure 2, we see that for the logistic model, the 5 most important variables are whether a respondent is from the northeastern part of China, respondent’s age, whether a respondent thinks your family’s economic status is equal to the average level, whether a respondent is a Han people, and whether a respondent is from the central part of China.

Then, we can try to examine the marginal effects of those variables and Figure 3 has shown those to us. Since the “best” model is the logistic model and there is no quadratic term or interaction term in the model, the partial dependency lines for those variables are all linear even for the only numeric variable. Two geographic variables play important roles in the model and have different effects on an individual’s attitude toward the local government. Being in the northeastern part of China is likely to decrease the possibility for an individual to trust the local government, while being in the central part of China has a completely different effect on this. This can be due to the qualities of public services provided by different local governments or even the different kinds of culture in different regions of China. Then, the effect of self-perceived economic status being equal to the average level can decrease the possibility for an individual to trust the local government. This can be due to the fact that the rich are satisfied with their conditions as well as the government and the poor are too tired to care about politics, while only the middle class is left to be dissatisfied with the government. The negative effect of being a ethnic minority is easy to understand, since Han people are the dominant group in China and the heads of local governments are always Han people. The effect of age is kind of hard to understand, since what we usually think is that if an individual is younger, he or she may be less likely to trust the government due to the higher education attainment and the tendency for a young people to be more likely to be cynical and radical. However, the plot below just shows the opposite.

Figure 2: Variables importance

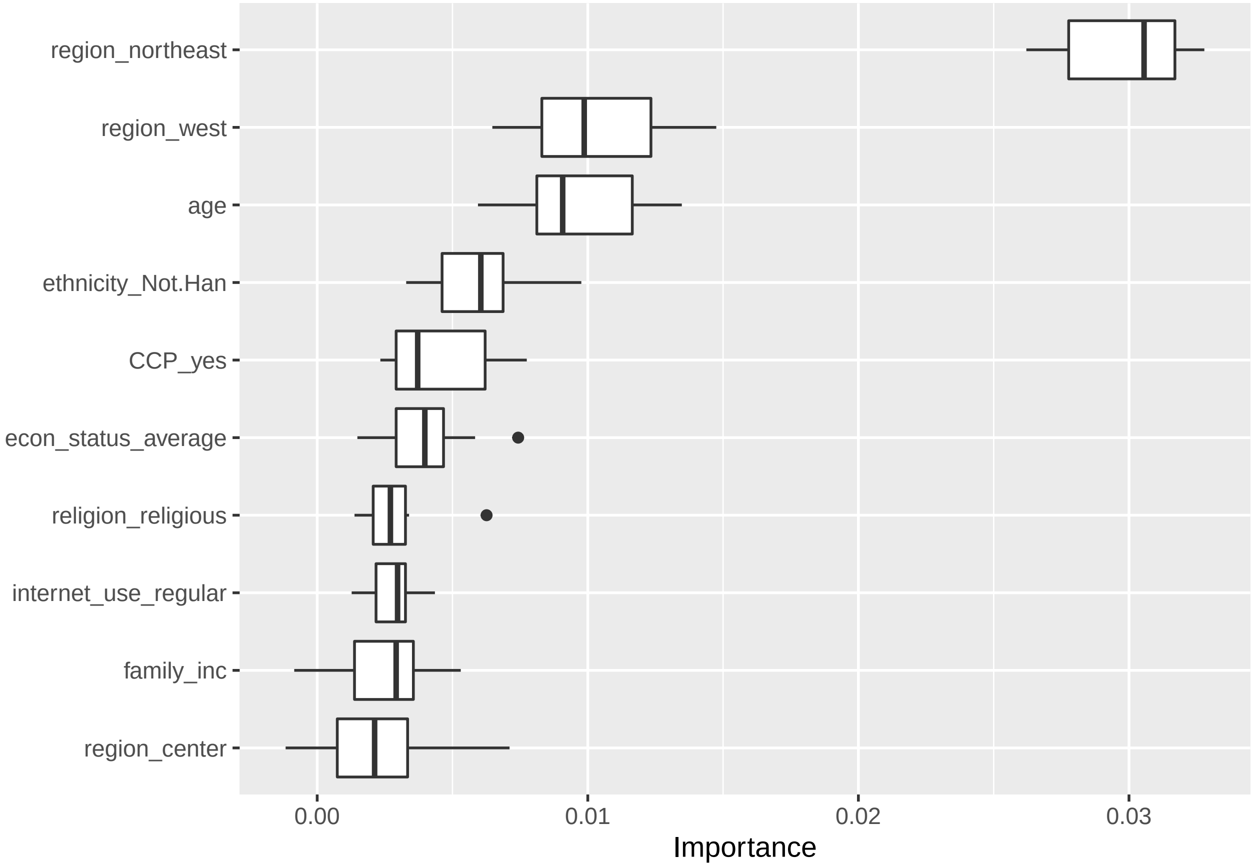
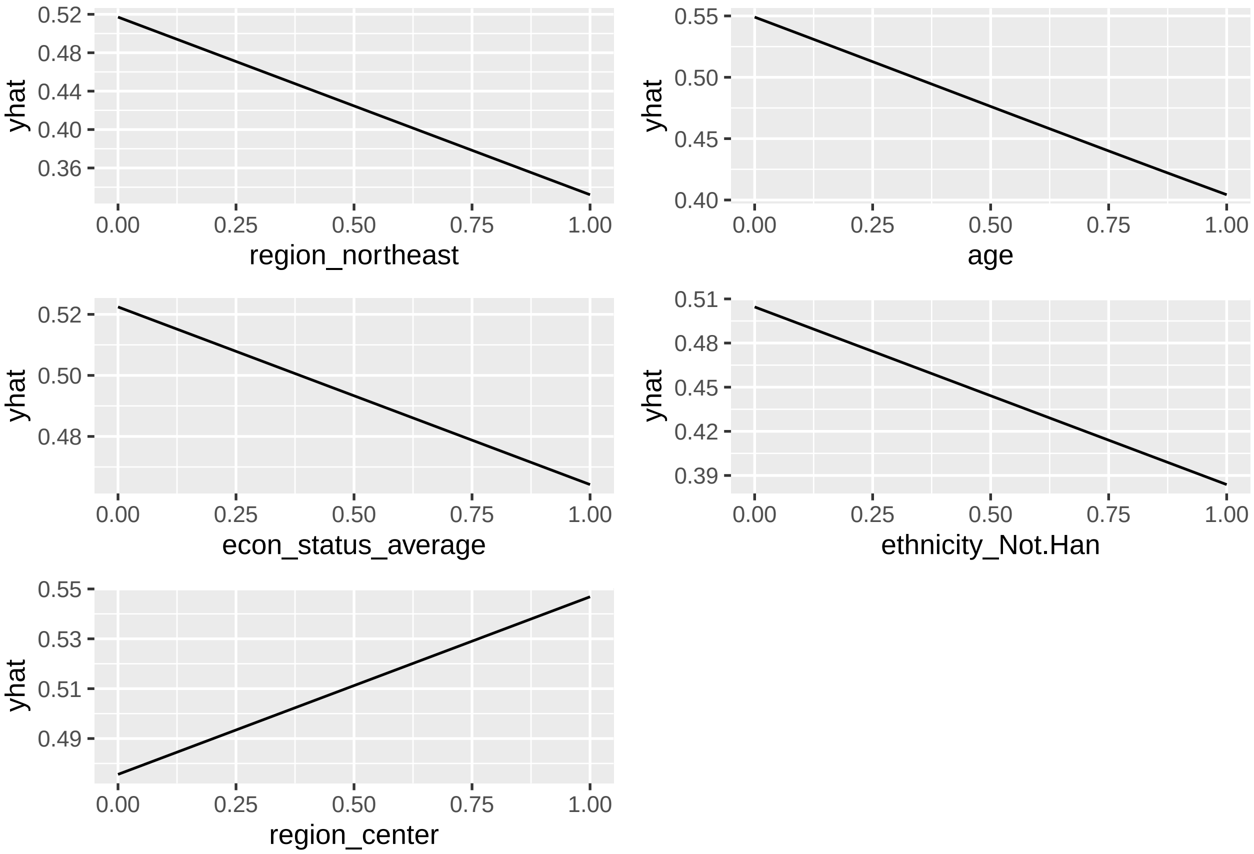


Figure 3: Partial dependencies for the 5 most important variables



Discussion

The conclusion here is that the logistic model may be the "best" model to predict an individual's attitude toward the local government in China given my knowledge so far. The limitation of this project is that the model chosen performed not well on the testing data and the comparison among models may be biased.

As I have stated in the proposal and the beginning of this report, the aim of this project is to find a model with the best performance to predict an individual's attitude toward the local government in China based on the knowledge learned in this course. And the success depends on whether I can achieve this goal. It seems that I have partly achieved this goal, since I have successfully find a model with the best performance on the training data among a series of algorithms that I have learned in this course. However, this project is also unsuccessful due to two reasons. The first reason is definitely about the warning, which can make the comparison biased. I have reached out to TA and she has gave me very good suggestions. But I still fail to solve this problem due to the time constraint. The second one is about the prediction for the testing data. The model performed well on the training data but performed badly on the testing data. A good model should perform well both on the training data and on the testing one. And I am still wondering why this has happened. Therefore, I would say this project is partly successful but also is partly unsuccessful.

If given more time, I think I can improve this project in three ways. Firstly, I can have more time to address the warning through finding solutions online and connecting with TA or the professor. With more time given, the probability of addressing this issue can increase and the best model can be chosen with certainty. Secondly, I can continue exploring those variables and can get more insights from using methods such as checking individual conditional expectations of those important variables and their interaction relationships, which can contribute to the improvement of the existing model and may thus improve its performance on the testing data as well as the training one. Thirdly, I have added two more variables into the project after getting the feedback from the professor. If there is plenty of time, I can try to find more variables that may have impacts on individuals' attitudes toward the government in China through doing more literature review and searching for relevant information on the internet.

Reference

Chen, X. and Shi, T. (2001) Media effects on political confidence and trust in the People's Republic of China in the post-Tiananmen period. East Asia: An International Quarterly 19(3): 84–118.

Ding, Z., Au, K, and Chiang, F. (2015) Social trust and angel investors' decisions: A multilevel analysis across nations, Journal of business venturing, 30 (2):307-321.

Kwon, S. W., Heflin, C., and Ruef., M. (2013) Community social capital and entrepreneurship, American sociological review, 78 (6):980-1008.

Lu, H., Tong, P. & Zhu, R. (2020) Does Internet Use Affect Netizens' Trust in Government? Empirical Evidence from China. Social indicators research 149 (1): 167-185.

Manion, M. (2006) Democracy, community, trust: The impact of elections in rural China. Comparative Political Studies 39(3): 301–324

Mishler, W., & Rose, R. (2001) What are the origins of political trust? Testing institutional and cultural theories in post-communist societies, Comparative Political Studies,34(1):30–62.

Nunkoo, R., and Smith, S. L. (2013) Political economy of tourism: Trust in government actors, political support, and their determinants, Tourism management (1982), 36:120-132.

Sohn, K., and Kwon. I. (2016) Does trust promote entrepreneurship in a developing country?, Singapore economic review, 63 (5):1385-1403.

Soest, C. & Grauvogel, J. (2017) Identity, procedures and performance: how authoritarian regimes legitimize their rule, Contemporary Politics, 23:3, 287-305, DOI: 10.1080/13569775.2017.1304319

Zhao, D. and Hu, W. (2015) Determinants of public trust in government: Empirical evidence from urban China. International Review of Administrative Sciences 0(0): 1–20.

Zhou, D., Deng, W. & Wu, X. (2019) Impacts of Internet Use on Political Trust: New Evidence from China, Emerging Markets Finance and Trade, DOI: 10.1080/1540496X.2019.1644161