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 the Intro to Data Science course at Georgetown. This essay uses survey data and data science techniques to predict whether a Chinese trusts the local government or not. In addition to model building, this essay involves reviewing relevant literature, data cleaning and processing, descriptive analysis, data visualization, and explanations of the logic for each model. Figures and tables are attached in the appendix at the end of formal writing. This pdf document is written by using R Markdown. The essay is my own work. Please do not distribute it without my permission. The folder can also be accessed on my GitHub.

```
library(haven)
require(tidyverse)
require(caret)
require(recipes)
require(rsample)
require(rattle)
require(yardstick)
require(skimr)
require(ranger)
require(vip)
require(pdp)
library(kableExtra)
library(grid)
library(gridExtra)
require(kernlab)
## Import data
cgss <- read_dta("Data/cgss2010_14.dta")
```

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 the best model can be built. For the rest of this report, we will go through 5 parts, including the 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 by enhancing political participation, regulatory compliance, and entrepreneurship performance (Ding et al., 2015; Kwon et al., 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 about 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 the Chinese through "excellent" economic performances and propaganda (Soest & Grauvogel, 2017).

And there are a number of articles that talk about factors that influence political trust in China. Manion (2006) 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 (Lu et al., 2020; Zhou et al., 2019)

Overall, the research so far on political trust in China has adopted a "micro"-perspective and has focused on finding certain relationships between factors and people's attitudes about the government. Those researchers have not tried to predict that attitude for individuals in China. However, this project will be based on a "macro"-perspective, which is to use a series of factors that may have impacts on political trust in China and machine learning techniques to build the 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) which is administrated by the Renmin University of China and is a 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. There are 11783 observations in the original dataset. More information can be found on the official website.

```
## Select variables needed
raw_dat <-
  cgss %>%
  select(trust = d303,
         internet = a285,
         birth = a3a,
         gender = a2,
         personal_inc1 = a8a,
         family_inc1 = a62,
         province = s41,
         educ1 = a7a,
         party = a10,
         ethni = a4,
         status = a64,
         immig = a21,
         reli = a5,
         household = a18)
## Clean data
raw_dat <-
  raw_dat %>%
  mutate(trust_or_not = ifelse(trust>3, "yes",
                               ifelse(trust<0, NA,
                                      "no"))) %>%
  mutate(year = replace(raw_dat$birth, raw_dat$birth<0,
                        NA)) %>%
  mutate(age = 2010-year) %>%
  mutate(sex = ifelse(gender==1, "male", "female")) %>%
```

```
  mutate(family_inc = replace(raw_dat$family_inc1,raw_dat$family_inc1>9999990,
                              NA)) %>%
  mutate(educ = ifelse(educ1==1|educ1==2|educ1==3, "primary school or below",
                       ifelse(educ1==-3|educ1==14, NA,
  ifelse(educ1==9|educ1==10|educ1==11|educ1==12|educ1==13, "some college or above",
         "middle")))) %>%
    mutate(CCP=ifelse(party==1, "yes",
                              ifelse(party>1, "no",
                                     NA))) %>%
    mutate(ethnicity = ifelse(ethni==1, "Han",
                    ifelse(ethni==-3, NA, "Not Han"))) %>%
     mutate(religion = ifelse(reli==1, "not religious",
                    ifelse(reli==-3, NA, "religious"))) %>%
  mutate(immigrate = ifelse(immig==3, "immigrant",
                    ifelse(immig==-3, NA,
                    ifelse(immig==4, "others",
                           "not immigrant")))) %>%
  mutate(household_type = ifelse(household==1, "agricultural",
                    ifelse(household==2, "urban",
                    ifelse(household<0, NA, "others")))) %>%
  mutate(internet_use = ifelse(internet<0, NA,
                    ifelse(internet==4|internet==5,
                           "regular", "not regular"))) %>%
    mutate(econ_status = ifelse(status<0, NA,
                    ifelse(status==1|status==2, "below average",
                    ifelse(status==3, "average", "above average")))) %>%
  mutate(family_inc=as.numeric(family_inc))
## Continue using join function to clean data
province_key <- tibble(province=c(1,2,3,4,5,6,7,8,9,10,11,
                                  12,13,14,15,16,17,18,19,
                                  20,21,22,23,24,25,26,27,28,
                                  29,30,31),
                       prov_chi=c("  ",
                                  "  ",
                                  "    ",
                                  "  ",
                                  "  ",
                                  "  ",
                                  "  ",
                                  "    ",
                                  "  ",
                                  "  ",
                                  "  ",
                                  "  ",
                                  "    ",
                                  "     ",
                                  "  ",
                                   "  ",
                                  "  ",
                                   "  ",
                                   "  ",
                                   "  ",
                                   "  ",
```

```
                                    "  ",
                                    "  ",
                                    "  ",
                                    "   ",
                                     "  ",
                                     "  ",
                                     "  ",
                                     "  ",
                                    "  ",
                                    "   "),
                        prov_eng=c("Shanghai",
                                   "Yunnan",
                                   "Nei Mongol",
                                   "Beijing",
                                   "Jilin",
                                   "Sichuan",
                                   "Tianjin",
                                   "Ningxia Hui",
                                   "Anhui",
                                   "Shandong",
                                   "Shanxi",
                                   "Guangdong",
                                   "Guangxi",
                                   "Xinjiang Uygur",
                                   "Jiangsu",
                                   "Jiangxi",
                                   "Hebei",
                                   "Henan",
                                   "Zhejiang",
                                   "Hainan",
                                   "Hubei",
                                   "Hunan",
                                   "Gansu",
                                   "Fujian",
                                   "Xizang",
                                   "Guizhou",
                                   "Liaoning",
                                   "Chongqing",
                                   "Shaanxi",
                                   "Qinghai",
                                   "Heilongjiang"))
raw_dat <- left_join(raw_dat, province_key, by = "province") ## Regional variable
raw_dat <-
  raw_dat %>%
  mutate(region =
           ifelse(prov_eng=="Heilongjiang"|prov_eng=="Jilin"|prov_eng=="Liaoning", "northeast",
                  ifelse(prov_eng=="Shanxi"|prov_eng=="Henan"|prov_eng=="Hubei"|prov_eng=="Anhui"|prov_
                         "center",
                         ifelse(prov_eng=="Hebei"|prov_eng=="Tianjin"|prov_eng=="Beijing"|prov_eng=="Sha
## Cleaned data
dat <-
  raw_dat %>%
  select(trust_or_not, age, sex, ethnicity, educ, religion, immigrate, household_type, region, family_i
```

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 about 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, educational attainment, and religious status. For age, we may expect that younger people in China will be less likely to trust the local government, since they may be more radical than those elders. Gender can also affect an individual's attitude about 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 an ethnic minority can also influence the attitude about the local government. It is doubtless to include educational attainment in this analysis, since it is highly related to an individual's social status and the perspective that s/he thinks about the government. Religious status can also be influential since the communist party advocates atheism.

For economic factors, total family income in 2009 and self-perceived economic status are chosen, which can not only measure the absolute 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. First, different provinces provide different public services based on their financial conditions and this thus may influence public trust in themselves. Also, different residents in different regions of China may have different attitudes about the government due to different cultural and historic factors. Second, whether an individual is a CCP(Chinese Communist Party) member can matter due to the party-state system. Third, the household type of an individual determines 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 influences their attitudes about the local government. Fourth, individuals' internet utilization may matter as well since the new media based on the internet is really hard to control and can disseminate "different" information 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 transferred to whether a respondent trusts the local government or not. Categories of ethnicity, religious status, and internet utilization should be grouped into only two since the number 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.

```
## Organized as a table
variables <- tibble(Variables=c("trust_or_not", "age", "sex", "ethnicity", "educ", "religion", "immigra
                    Descriptions=c("Whether the respondent thinks the local government is trustworthy or
variables %>%
  kbl(caption  = "Descriptions of variables") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

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

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.

```
set.seed(2004)
dat <-
  dat %>%
  mutate(
   trust_or_not = factor(trust_or_not, levels = c("no", "yes")),

   sex = factor(sex,level=c("male",'female')),

    ethnicity = factor(ethnicity,levels= c('Han',
                                'Not Han')),

    educ = factor(educ,levels= c('primary school or below',
                                "middle",
                                "some college or above")),

   religion = factor(religion,levels= c("not religious",
                                "religious")),

   immigrate = factor(immigrate,levels= c('not immigrant',
                                'immigrant',
                                "others")),

   household_type = factor(household_type,levels= c('agricultural',
                                'urban',
                                "others")),

   region = factor(region,levels= c('east',
                                'northeast',
                                "center",
                                "west")),

    econ_status = factor(econ_status,levels= c('below average',
                                'average',
                                "above average")),
```

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	

```
    CCP = factor(CCP,levels= c('no',
                               'yes')),

    internet_use = factor(internet_use,levels= c('not regular',
                               'regular'))
   )
## Split data
set.seed(123)
splits <- initial_split(dat,prop = .8,strata = trust_or_not)
train_data <- training(splits) # Use 80% of the data as training data
test_data <- testing(splits)
```

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. But we can use KNN to "fill" those missing values based on corresponding non-missing values from their k nearest neighbors. Also, there is an imbalance of classes for the dependent variable, which can be addressed by setting a 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 build our models. Tables 1 to 3 show the definitions of variables used and descriptive statistics for both the training and the testing datasets.

```
## Check the training data
kable(skim(train_data))%>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

```
# Data processing for the training data
rcp <-
  recipe(trust_or_not ~ .,train_data) %>%
  step_impute_knn(all_numeric()) %>%
  step_impute_knn(all_nominal()) %>%
  step_dummy(all_nominal(),-trust_or_not) %>%
  step_log(family_inc, offset = 1) %>%
  step_range(all_numeric()) %>%
  prep()
# Apply the recipe to the training and test data
train_data2 <- bake(rcp,train_data)
test_data2 <- bake(rcp,test_data)
# Cross validation
set.seed(1988)
folds <- createFolds(train_data2$trust_or_not, k = 5)
control_conditions <-
  trainControl(method='cv',
               summaryFunction = twoClassSummary,
               classProbs = TRUE,
```

```
                sampling = "up",
                index = folds

  )
```

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 machine 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 an "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 research in social sciences since it originates from the OLS model.

The logic of the 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.

The 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 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 fewer 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 choose 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 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, the second one is to "draw" a polynomial curve, and 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.

```
## Logistic Regression
mod_logit <-
  train(trust_or_not ~ .,
        data=train_data2,
        method = "glm",
        metric = "ROC",
        trControl = control_conditions
  )
## KNN
mod_knn <-
  train(trust_or_not ~ .,
        data=train_data2,
        method = "knn",
        metric = "ROC",
        trControl = control_conditions
```

```
  )
## Second KNN
mod_knn2 <-
  train(trust_or_not ~ .,
        data=train_data2,
        method = "knn",
        metric = "ROC",
        tuneGrid = expand.grid(k = c(1,3,5,7,10,50)),
        trControl = control_conditions
  )
## CART
mod_cart <-
  train(trust_or_not ~ .,
        data=train_data2,
        method = "rpart",
        metric = "ROC",
        trControl = control_conditions
  )
## Random Forest
mod_rf <-
  train(trust_or_not ~ .,
        data=train_data2,
        method = "ranger",
        metric = "ROC",
        trControl = control_conditions
  )
# Linear Boundary
mod_svm_linear <-
  train(trust_or_not ~ . ,
        data=train_data2,
        method = "svmLinear",
        metric = "ROC",
        tuneGrid = expand.grid(C = c(.5,1)),
        trControl = control_conditions
  )
# Polynomial Boundary
set.seed(1234)
mod_svm_poly <-
  train(trust_or_not ~ .,
        data=train_data2,
        method = "svmPoly",
        metric = "ROC",
        trControl = control_conditions
  )
# Radial Boundary
mod_svm_radial <-
  train(trust_or_not ~ .,
        data=train_data2,
         method = "svmRadial",
        metric = "ROC",
        trControl = control_conditions
  )
```

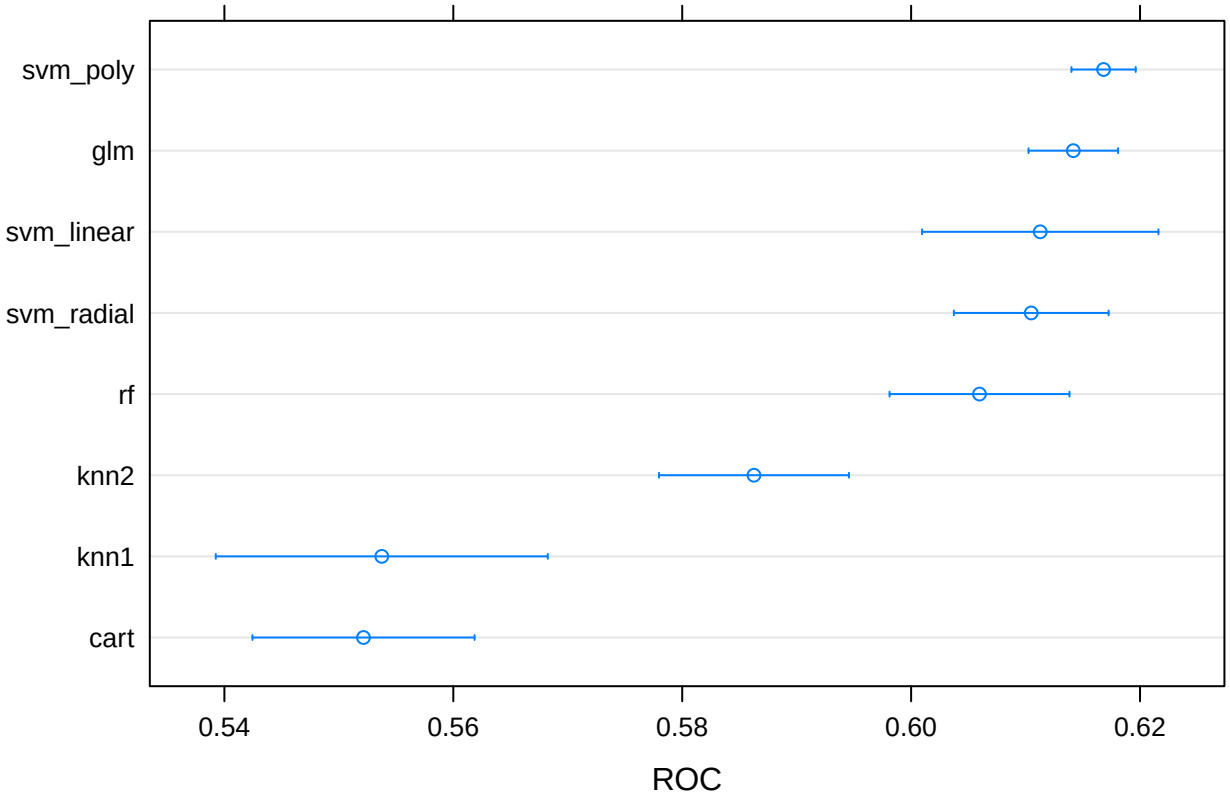
Results

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 about the local government in China by using the variables mentioned above to some extent. Specifically, while two KNN models and the classification trees model perform the worst, the 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. Therefore, our final model is the support vector machine model with the polynomial boundary.

```
## Models comparison
mod_list <-
  list(
    glm=mod_logit,
    knn1 = mod_knn,
    knn2 = mod_knn2,
    cart = mod_cart,
    rf = mod_rf,
    svm_linear = mod_svm_linear,
    svm_poly = mod_svm_poly,
    svm_radial = mod_svm_radial
  )
dotplot(resamples(mod_list),metric = "ROC", main = "Figure 1: Comparison among different models")
```

Figure 1: Comparison among different models



Confidence Level: 0.95

However, when I try to apply the polynomial boundary model to the testing data, it performs far worse than it did on the training data. From Table 4, we see that its ROC is only about 0.40 which is far below 0.6 and it only makes the right predictions for about half of the testing data.

```
# Logit
pred_probability <- predict(mod_svm_poly,newdata = test_data2,type="prob")
pred <- predict(mod_svm_poly,newdata = test_data2,type="raw")
performance <- tibble(truth = test_data2$trust_or_not,
                      prob = pred_probability$yes,
                      pred = pred)
bind_rows(
  performance %>% roc_auc(truth,prob),
  performance %>% accuracy(truth,pred)
) %>%
  kbl(caption  = "Performance on the testing data") %>%
  kable_classic(full_width = F, html_font = "Cambria")
```

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 polynomial boundary model, the 5 most important variables the model adopts to make predictions for the training dataset are whether a respondent is from the northeastern part of China, whether a respondent is from the western part of China, respondent's age, whether a respondent gets a middle school degree, and whether a respondent is a communist party member.

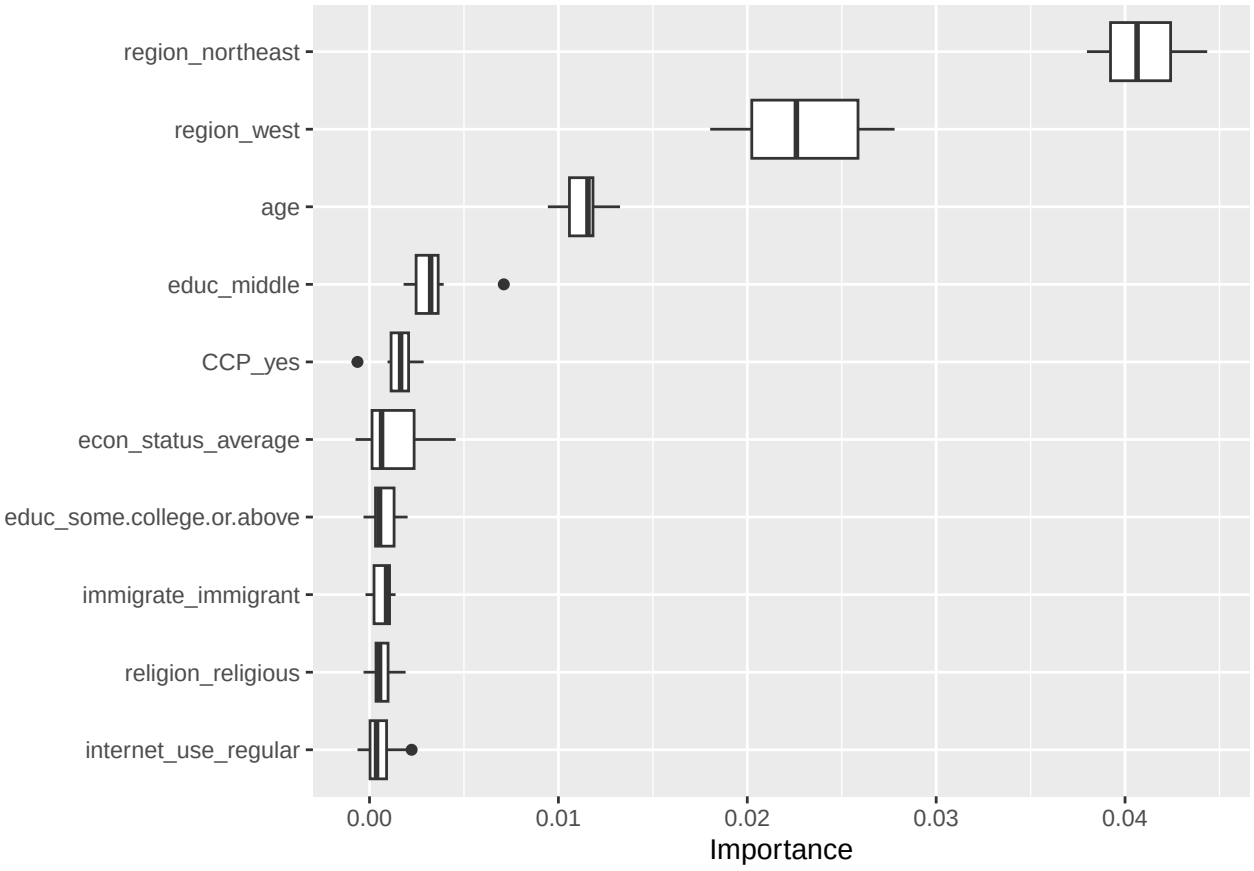
Table 2: Performance on the testing data

.metric	.estimator	.estimate
roc_auc	binary	0.3960896
accuracy	binary	0.5441052

Then, we can try to examine the marginal effects of those variables and Figure 3 has shown those to us. Since 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 about the local government. Being in the northeastern and the western parts of China is likely to decrease the possibility for an individual to trust the local government. This can be due to the quality of public services provided by the local government or even the local culture in different regions of China. However, the effects of the remaining three variables are hard to understand. 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 higher educational attainment and the tendency for young people to be more likely to be cynical and radical. However, the plot below just shows the opposite. Then, the effect of obtaining a middle school degree can increase the possibility for an individual to trust the local government compared to those who only finish primary school or below. One possible explanation is that the longer you have been in the mandatory education system administered by the government, the more likely you get influenced by governmental propaganda, which shows the feature of an authoritarian state. Also, the negative effect of being a party member is hard to understand, since we believe party members' loyalty is what the party regularly examines and tries to maintain. This can be explained by the fact that the party's disciplines have not been strictly followed at the grassroots level, indicating that the party's control has been weakened.

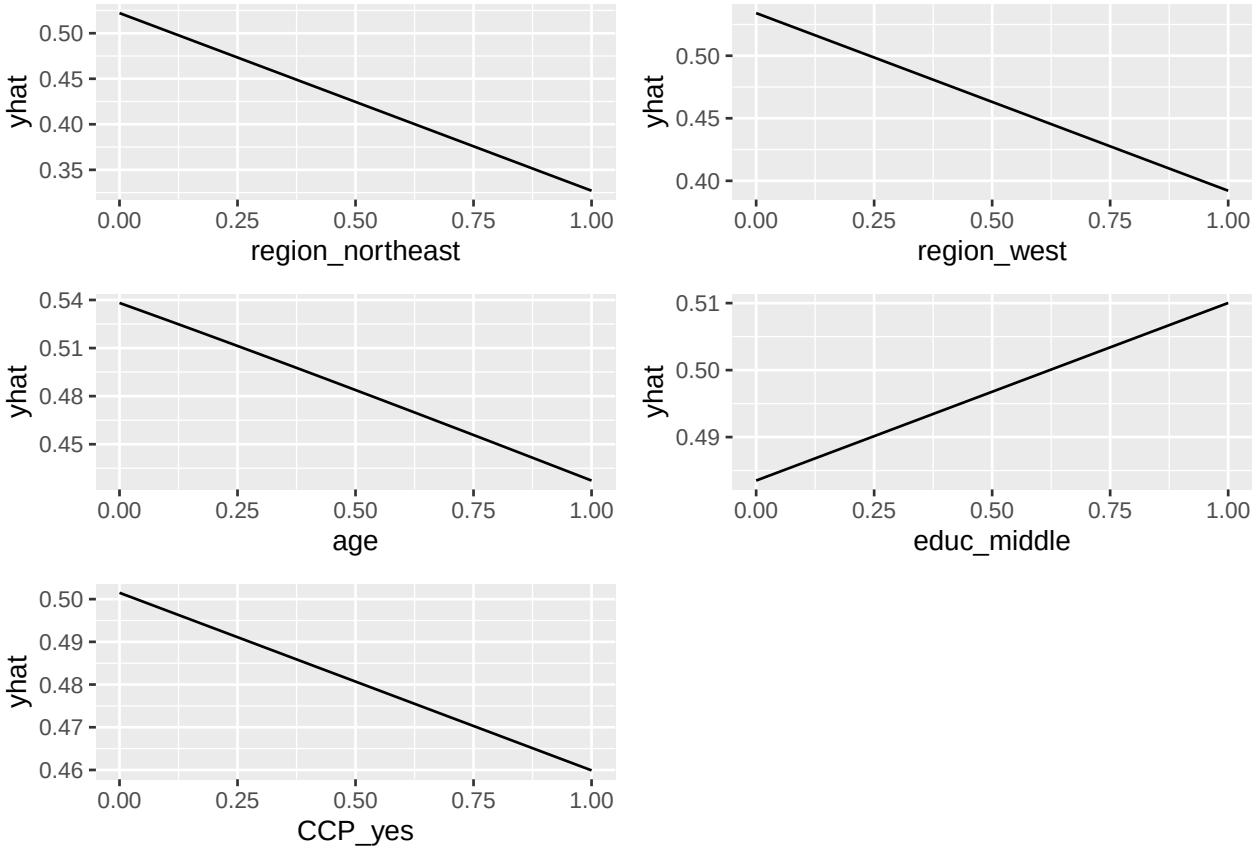
```
## Variables importance
vi_plot<-
  vip(mod_svm_poly,
      train = train_data2,
      method="permute",
      nsim = 10,
      geom = "boxplot",
      target = "trust_or_not",
      reference_class = "yes",
      metric = "accuracy",
      pred_wrapper = predict) +
  ggtitle("Figure 2: Variables importance")
vi_plot
```

Figure 2: Variables importance




```
# PDP plot
region_northeast_pdp <- partial(mod_svm_poly, pred.var = "region_northeast", plot = TRUE,prob=T,
                    plot.engine = "ggplot2")
region_west_pdp <- partial(mod_svm_poly, pred.var = "region_west", plot = TRUE,prob=T,
                    plot.engine = "ggplot2")
age_pdp <- partial(mod_svm_poly, pred.var = "age", plot = TRUE,prob=T,
                     grid.resolution = 20,
                    plot.engine = "ggplot2")
educ_middle_pdp <- partial(mod_svm_poly, pred.var = "educ_middle", plot = TRUE,prob=T,
                                  plot.engine = "ggplot2")
CCP_yes_pdp <- partial(mod_svm_poly, pred.var = "CCP_yes", plot = TRUE,prob=T,
                    plot.engine = "ggplot2")
gridExtra::grid.arrange(region_northeast_pdp,region_west_pdp, age_pdp,educ_middle_pdp,CCP_yes_pdp,
                        top=textGrob("Figure 3: Partial dependencies for the 5 most important variables
```

Figure 3: Partial dependencies for the 5 most important variables



Discussion

The conclusion here is that the polynomial boundary model may be the "best" model to predict an individual's attitude about 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 dataset.

As I have stated in the proposal and at 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 about 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 found 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 the predictions made for the testing dataset. 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 partly unsuccessful.

If I was given more time, I think I can improve this project in two ways. First, 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. Second, I added two more variables to the project after getting feedback from the professor. If there is plenty of time, I can try to find more variables that may have impacts on individuals' attitudes about the government in China by doing more literature reviews and searching for relevant information on the internet.

References

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