Final Project: Predicting Political Attitudes and Political Dissidents in China Using Machine Learning and Survey Data

Georgia Institute of Technology
ISyE/Math 6783
Student Name: Zhongyun Zhang

Introduction

- Literature review: many studies have worked to examine political values, ideologies, and attitudes in China and have explained factors that tend to affect a person's political attitudes
  - Some of them explain that political attitudes in China are heavily affected by the people's evaluations of the economic conditions of the country and of the family (Zhang, 2015)
  - Besides, other studies argue that political attitudes are also affected by the cultures and values, such as Confucianism and collectivism (Chu, 2016; Shi, 2014)
  - Moreover, some other research finds that assessment of the country's distributive and redistributive mechanisms have worked to affect their attitudes toward the current regime (He and Su, 2018)
  - Furthermore, studies have also found that the people's perception of the country's protection of political rights are linked to their political attitudes (Wang and You, 2016)
- Research goals of this project
  - What the existing study lacks is a more systematic study that can identify the specific factors that affect the political attitudes in China more consistently
  - Through machine learning techniques, we can better identify the specific features that more consistently work to affect the people's political attitudes in China and predict a person's political attitudes when given these features.
  - Thus, based on the existing research on the political attitudes in China, it is meaningful to use machine learning models to further discover which factors tend to affect the people's political attitudes and their political support of the government and to build models that work to predict people's political attitudes.

Data (from Wave 4 of the Asian Barometer Survey (ABS IV) on mainland China)

Dependent Variables (Output Variables)

- Political Attitudes as Ordinal Variables
  - Support for open and competitive election
  - System capable of solving problems our country face
  - Proud of system
  - System deserves the people's support, rather living under current system than others,
  - Degree of systemic changes need
  - Support of multi-party system.
- Political Attitudes as Dichotomous Variables
  - people who disapprove or strongly disapprove the statement that only one political party should be allowed to stand for election and hold office are coded as 1 and -1 otherwise
    - 22.03% of the 4,068 respondents are coded as 1.
  - Regarding the degree of changes expected of the system of government, people who think the current system of government needs major change or should be replaced are coded as 1 and -1 otherwise
    - 20.06% of the total respondents are coded as
    - class-weighting is used when running the machine learning models

Independent Variables (Features)

- Evaluation of the Economic Conditions of the Country and of the Family
- Political Trust in National and Local Governments
- Interest in Politics
- Use of Internet and Social Media
- Individualist and Collectivist Values
- Perception of Treatment by Government
- Perception of Political Rights
- Political Efficacy
- Redistributive Preferences and Family Financial Condition
- Demographics

Machine Learning Models

- **Feature Selection and Regression through Lasso regression**
  - **5-fold cross-validation based on negative root mean squared error (RMSE)**
  - **A function with for loops is written to perform the models on the 14 output variables (listwise deletion and imputation for each of the 7 output variables).**

```
# the following codes will write a function that will produce the lasso cofficients for all the dependent variables
def produce_lasso_table_for_DVs():

    output_table = pd.DataFrame({'Question':predictors})

    for dv in dependent_variables:
        # corresponding observations in X will need to be deleted as well if y is not imputed by median and contains NAs
        X_y = df[~np.isnan(df[dv])]
        X = X_y.loc[:,'q1':'eth_minority']
        y = X_y.loc[:,dv]

        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=1)

        lassoRegr = Lasso(max_iter = 10000)
        alphas = np.logspace(-15, -1, 100)

        #use 5 fold CV to choose the best parameter for alpha
        train_scores, test_scores = validation_curve(
            lassoRegr, X_train, y_train, cv=5, param_name="alpha", param_range=alphas,
            scoring="neg_root_mean_squared_error")

        best_para = parameter_cross_validation(alphas, train_scores, test_scores)

        best_lassoRegr = Lasso(max_iter = 10000, alpha = best_para)
        best_lassoRegr.fit(X_train, y_train)
        y_pred_lasso = best_lassoRegr.predict(X_test)

        coefficients_lasso = pd.DataFrame(best_lassoRegr.coef_, index=X_train.columns)
        coefficients_lasso.reset_index(inplace=True)
        coefficients_lasso.rename(columns={coefficients_lasso.columns[0]: 'Question',
                                    coefficients_lasso.columns[1]: dv}, \
                                       inplace = True)

        output_table = output_table.merge(coefficients_lasso, left_on='Question', right_on='Question')

    output_table.replace(0, np.nan, inplace=True)

    return output_table
```

- **Machine Learning**
  - **Models: SVM, Random Forest Classifier, Logistic Regression, and KNN**
  - **Selecting Parameter and Optimization**
    - **5-fold cross-validation based on accuracy score, precision rate, recall rate, and f-1 score**

```
KNN = KNeighborsClassifier
param_range = [*range(1,50)]
train_scores, test_scores = validation_curve(estimator=KNN(),\
                                             X=X_train,y=y_train,cv=5,\
                                             param_name="n_neighbors",
                                             scoring = 'accuracy',
                                             #scoring="f1",
                                             #scoring='recall',
                                             #scoring = 'precision',
                                             param_range=param_range)
plot_cross_validation_curve(param_range, train_scores, test_scores)
best_para = parameter_cross_validation(param_range, train_scores, test_scores)

KNN = KNeighborsClassifier(n_neighbors=best_para)
KNN.fit(X_train, y_train)
test_accuracy = accuracy_score(y_test, KNN.predict(X_test))
print(test_accuracy)
plot_confusion_matrix(KNN, X_test, y_test)
plt.show()
print(classification_report(y_test, KNN.predict(X_test)))
```

Results: Lasso Regression and Feature Selection

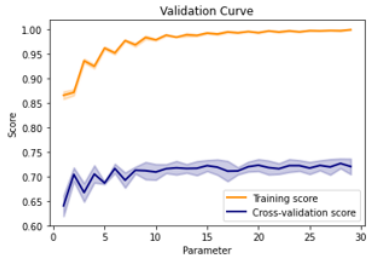
- Some features have consistently stronger effects on the political attitudes in China across all the seven output variables. Consistent Predictors for Political Attitudes in China:
  - better evaluations of the country's current economic conditions
  - positive opinions of the country's income distribution
  - higher trust of the government
  - stronger collectivist values
  - And to a slightly less extent:
    - individual and family economic conditions
    - protection of civil liberties
- These features may be helpful for political scientists and policy makers who are concerned about the more consistent predictors of political attitudes in China at individual level

Category	Question (if not specified, higher answers means more positive evaluations or stronger agreement)	System capable of solving country's		Proud of system (1-4)		System deserves the people's		Rather Living under Current System than		Degree of Systemic Changes Need		Support for Open and Competitive		Support of Multi-party (1-4)	
		linlusinr	lmpsitrd	linlusinr	lmpsitrd	linlusinr	lmpsitrd	linlusinr	lmpsitrd	linlusinr	lmpsitrd	linlusinr	lmpsitrd	linlusinr	lmpsitrd
economic conditions of the country and of the family	How would you rate the overall economic condition of our country today?	0.0288	0.0203	0.0255	0.0185	0.0186	0.0171	0.0271	0.0081	-0.0527	-0.0333	-0.0181	-0.0147	-0.0123	-0.0068
	How would you describe the change in the economic condition of	0.0018			0.0069					-0.0115		0.0001			0.0087
	What do you think will be the state of our country's economic	0.0098		0.0187	0.0186			0.0131	0.0130		-0.0135	-0.0267		-0.0005	-0.0112
	As for your own family, how do you rate the economic situation of	0.0030			0.0034		0.0075								
	How would you compare the current economic condition of your	0.0055		-0.0026		-0.0099						0.0004			
	What do you think the economic situation of your family will be a				-0.0055		-0.0112			-0.0017			-0.0072		
political trust	Trust in the national gvt	0.0362	0.0375	0.0634	0.0473	0.0320	0.0341	0.0397	0.0425	-0.0204	-0.0148	0.0263		-0.0720	-0.0448
	Trust in the local gvt	0.0268	0.0180	0.0231	0.0208	0.0172	0.0143	0.0062		-0.0840	-0.0782				
interest in	How interested would you say you are in politics?			0.0018	0.0073	0.0056	0.0003			-0.0004			0.0030		
internet and social media	How often do you use the internet?			-0.0094	-0.0381		0.0116	-0.0189	-0.0145		0.0306			0.0270	0.0354
	Do you currently use any of the following social media	0.0067			-0.0089		0.0237								
	How often do you use the Internet including social media networks to find information about politics and			0.0265	0.0289	0.0177	0.0270							-0.0022	
	How often do you use the Internet including social media networks to express your opinion about politics				-0.0165										
individualism and collectivism	In a group, we should sacrifice our individual interest for the sake of the group's collective interest	0.0434	0.0375	0.0424	0.0467	0.0342	0.0268	0.0283	0.0300	-0.0053	-0.0053	0.0128		-0.0322	-0.0459
	For the sake of national interest, individual interest could be sacrificed	0.0610	0.0282	0.0482	0.0378	0.0293	0.0331	0.0488	0.0606	-0.0162	-0.0086	0.0235	0.0133	-0.0570	-0.0517
Perception of treatment by govt	All citizens from different ethnic communities in China	0.0453	0.0322	0.0268	0.0269			0.0366	0.0201	-0.0214	-0.0151	0.0640	0.0548	-0.0042	-0.0152
	Rich and poor people are treated equally by the government.			0.0038	0.0055	0.0211	0.0072			-0.0109				-0.0013	-0.0011
	People have basic necessities like food, clothes, and shelter.	0.0170	0.0206	0.0097	0.0300	0.0300	0.0239	0.0319	0.0458	-0.0164	-0.0010			-0.0380	-0.0389
perception of civil rights	People are free to speak what they think without	0.0417	0.0310	0.0557	0.0460	0.0121	0.0109	0.0507	0.0436		-0.0015			-0.0323	-0.0265
	People can join any organization they like without	0.0108	0.0001	0.0132	0.0145	0.0472	0.0413	0.0008		0.0133		-0.0001			
political efficacy	I think I have the ability to participate in politics.	0.0000		0.0091	0.0036	0.0021	0.0034	-0.0114		0.0175	0.0298	0.0209	0.0147		0.0041
	Sometimes politics and government seems so complicated that a person like me can't reallu understand what			0.0092		-0.0014	0.0083	0.0230	0.0218		-0.0078	-0.0186	-0.0014	-0.0186	-0.0217
	People like me don't have any influence over what the	0.0241	0.0069	0.0039	0.0038	0.0270	0.0205								-0.0071
redistribution preferences and family financial condition	How fair do you think income distribution is in	0.0420	0.0376	0.0554	0.0307	0.0111	0.0239	0.0457	0.0283	-0.0468	-0.0476	-0.0116		-0.0201	-0.0025
	Do you agree or disagree with the following statement: "It is the responsibility of the government to reduce the differences between people with high income and those with low		0.0097	-0.0082	-0.0170		0.0064			0.0089	0.0176	0.0322		-0.0052	-0.0126
	How concerned are you about the loss of your or your family's major source of income within the next 12 months? (Not at all:4 --Very concerned:1)	-0.0003		-0.0076	-0.0018		0.0141			-0.0017	-0.0244	0.0152			0.0084
	The following is a hypothetical question: If you were unfortunate enough to lose your main source of income, how serious would it be for you and your family?	-0.0037			-0.0044			-0.0002			-0.0080	0.0205			
	Considering all the effort that you and your family members have made in the past, do you think the income that your family currently receives is fair or not fair?	0.0211	0.0206		0.0018	0.0252	0.0141	0.0113		-0.0187	-0.0240	-0.0176	-0.0111		-0.0032

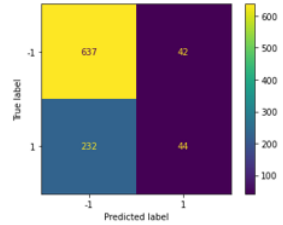
Results: Machine Learning Models

- The models that have been able to produce consistently better results than other models are Random Forest Classifier and SVM.
- The results from the logistic models and KNN are also quite similar.
- The highest accuracy scores are achieved by Random Forest Classifier, which has an accuracy score of 0.71
- The highest F-1 score is achieved by SVM.

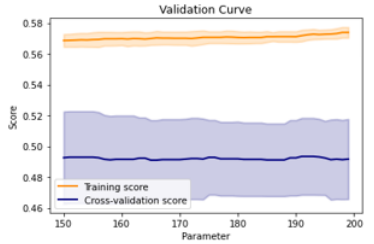
Random Forest Classifier Based on Accuracy Score Using 5-Fold Cross Validation



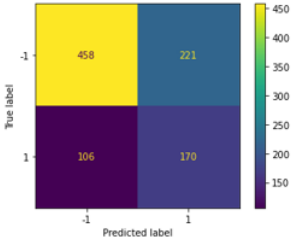
	precision	recall	f1-score	support
-1	0.73	0.94	0.82	679
1	0.51	0.16	0.24	276
accuracy			0.71	955
macro avg	0.62	0.55	0.53	955
weighted avg	0.67	0.71	0.66	955



SVM Based on F-1 Score Using 5-Fold Cross Validation



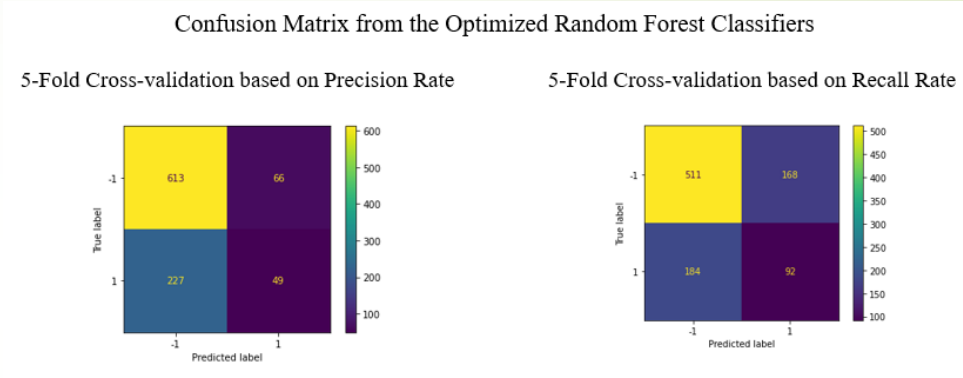
	precision	recall	f1-score	support
-1	0.81	0.67	0.74	679
1	0.43	0.62	0.51	276
accuracy			0.66	955
macro avg	0.62	0.65	0.62	955
weighted avg	0.70	0.66	0.67	955



Discussion and Conclusion

- This study has identified consistently robust predictors of political attitudes in China
  - Better evaluations of the country's current economic conditions
  - Positive opinions of the country's income distribution
  - Higher trust of the government
  - Stronger collectivist values
- The accuracy rate of 0.71 is high but not sufficiently high even for the best model
  - Political attitudes may be hard to predict very accurately
- Generally low precision score across models albeit relatively high recall and accuracy score
  - High False Positives
  - May be explained by preference falsification of the respondents

- Trade-off between precision and recall



  - Similar to theories of individual deterrence and general deterrence in political science:
    - Preference for high precision will make the government miss chances to identify political dissidents, this may be dangerous for the government
    - Preference for high recall will make the government unnecessarily target people who are non-dissidents, this may backfire
  - The importance of F-1 score: similar to the importance of making the trade-off between individual deterrence and general deterrence