Predicting United State Presidential Election on County Level

ECON 4130 Economics Analysis for Social Network

Final Term Paper

Yam Tsz Chung

1155122291

Abstract – United State, the 3rd largest populated country in the world has developed a election system that make sure all voice was heard. In this paper, we will use 3 popular classification methods including Logistic Regression, K Nearest Neighbor and Decision Tree to predict which candidate will win in each county. Our result suggested that the Logistic Regression is the most accurate when predicting the election result. The second part of this paper, Decision Tree perform the best when deciding whether the county will change their vote to another party or not.

INTRODUCTION

The 2020 Untied State Presidential Election is still ongoing during the time this paper was typed. The election caught the global attention as this not only affecting the next 4 years of United State, but also the business of the global economic and political environment[1]. While we never have a chance on choosing our own government, it will still be interesting to understand a democracy country voting system and gain some insight from it. Many online articles suggested that there are patterns for voting and thus, this will affect the candidate's strategies. As an economic student, the elections are very fun to analyze because the president will make decisions on trading, fiscal and monetary policies that affecting global business environment and the macroeconomics factors in his/her 4 years. From my own naïve observations, Moderate income and multi-racial tends to vote for Biden. Lower income and white people tend to vote for Trump. This different in preference maybe caused by the economy. If there is really a pattern for voting, then we will be able to do our own prediction on the election result using economic data. More importantly, how accurate will the economic data reflect the voting results. So, this will be a classification process in this paper.

The goal in this paper is to find the most accurate model to predict how will a county vote and try to use the same economic data to predict whether the county will change their vote to another party.

[1] https://www.npr.org/2020/11/03/930722317/how-the-presidential-election-winner-could-effect-the-economy

DATA

We are focusing on the County level of data in this paper. As this year election was not yet finished and still in progress, we will use the 2016 election to train and test our model. There are two major parties in the election, and we will assign “1” as Republican win and “0” as Democrat win on county level in our model as the dependent variable for the first part, predicting which party will win in each county.

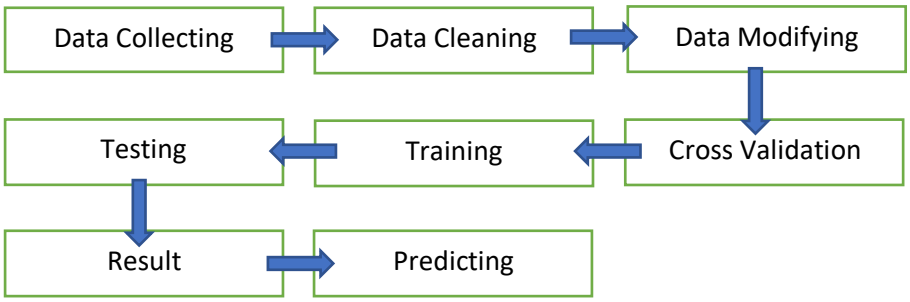
For the second part, to predict which county will flip their vote, we will use another data as dependent variable. We will first compare the voting result between 2016 and 2012. Then assign “1” for those counties who change their vote from Republican to Democratic and “2” for who changed to Republican. The remaining unchanged will be assigned as “0”. There is a total of 238 counties who flipped their vote which 20 of them assigned as “1” and 218 of them as “2”. From these we will apply our model to see if the models can allow us to predict which county will change their vote and changed for which parties.

From the previous data, we can expect that there is a high autocorrelation in election since there was only around 8% of the counties changed their vote from 2012 to 2016. To address this, we will use the result in 2012 as a dummy variable as our independent variable. Same we assign Republican win as “1” and Democrat win as “0”.

The county level of data will also be used in this paper. It is very time consuming to collect the data of county one by one, which Kaggle provided most of the data for us. The data included many economic variables such as population, income, unemployment, racial, age and household. Which the data frame will having 49 columns and 3112 counties. Some counties are missing because data were loss and outdated.

METHDOLOGY

(a) Simple Process Diagram



(b) Modeling Methods

In this paper, we will find the best model that can allow us to achieve our goal. We will compare the performance between three classification models: 1. Logistic Regression, 2. K-Nearest Neighbor and 3. Decision Tree. Since our data is limited and not typically large, we will apply GridSearch Cross Validation function form SciKitLearn with a 10-Fold cross validation. As we want to choose the best model, the parameters for each method to be tuned are as below:

1. Logistic Regression:
   Regularization Constant: [0.001, 0.01, 0.1, 1, 10, 100, 1000]
   Penalty: [L2, L1]

2. K-Nearest Neighbor:
   Number of Neighbors: [1, 3, 7, 9, 11, 13, 15]
   Weights: [Uniform, Distance]
   Metric: [Euclidean, Manhattan]

3. Decision Tree:
   Criterion: [Gini, Entropy]
   Maximum Depth: [2, 4, 6, 8, 10]
   Minimum Samples Split: [2, 4, 6, 8, 10]
   Minimum Samples Leaf: [2, 4, 6, 8, 10]

We split the data with 25% of testing set and 75% of training set for our models. Which a total of 778 counties will be used to test our models and remining 2334 to train.

RESULT

(a) Models performance in predicting vote result in 2016:

Logistic Regression:

C = 0.1

Penalty = L2

With the Training set accuracy: 0.962

F1 score in Testing set = 0.96

K Nearest Neighbor:

Metric = Euclidean
No. of Neighbors = 13
Weights = Distance
With the Training set accuracy: 0.952
F1 score in Testing set = 0.94

Decision Tree:

Criterion = Entropy,
Max_depth = 10,
Min_samples_leaf = 10,
Min_samples_split = 8
With the Training set accuracy: 0.955
F1 score in Testing set = 0.95

(b) Predicting the county vote result in 2016:

After training the models with GridSearchCV, we find out that the logistics regression gave us the best accuracy when predicting which candidates each county will vote for, with 0.962 accuracy in training set and the highest F1 score in the testing set with 0.96. The other two methods gave us the near results in both training and test set with less than 2% loss compare to the best model.

Using the best model and parameters, we then predict all counties vote result with the same independent variables set. Republican (Trump) will win more electoral votes and become the next president. Where Democrat (Clinton) will win in Alabama, Arkansas, California, Colorado, Connecticut, Delaware, Hawaii, Massachusetts, Maryland, New Jersey, New Mexico, New York, Oregon, Rhode Island, Virginia, Vermont, and Washington, which is a total of 201 electoral votes compare to the real history that Clinton got a total of 232. The states which Democrat win in the prediction are those they won in real history but missing several states. Out model and prediction in this case are quite accurate.

(c) Models performance in predicting vote changes from 2012 to 2016

Logistic Regression:

C = 0.1

Penalty = L2

With the Training set accuracy: 0.966

F1 score in Testing set predicting changes:

No change from 2012 to 2016 = 0.97

from Republican to Democratic = 0.00

from Democratic to Republican = 0.63

Overall accuracy: 0.95

K Nearest Neighbor:

Metric = Euclidean

No. of Neighbors = 13
Weights = Uniform

With the Training set accuracy: 0.963

F1 score in Testing set predicting change:

No change from 2012 to 2016 = 0.97

from Republican to Democratic = 0.00

from Democratic to Republican = 0.64

Overall accuracy: 0.95

Decision Tree:

Criterion = Gini

Max_depth = 4

Min_samples_leaf = 10

Min_samples_split = 10

With the Training set accuracy: 0.964

F1 score in Testing set predicting changes:

No change from 2012 to 2016 = 0.98

from Republican to Democratic = 0.00

from Democratic to Republican = 0.70 <<< Highest among other two methods

Overall accuracy: 0.95

(d) Predicting the change in county vote from 2012 to 2016:

The three models are having high accuracy in the training set of around 0.96, and all have a high F1 score in predicting the counties with no changes (“0”) at around 0.97 and an overall F1 score at 0.95. Although all three models seem to have a good prediction accuracy score, all of them failed to predict the changes from Republican to Democrat (“1”). While the best model (Decision Tree) gave us the highest F1 score in predicting the changes from Democrat to Republican (“2”) at 0.7.

Conclusion and Suggestions

The economic variables we used in the models are accurate when predicting which party, the county will vote for. But it is not as accurate when predicting the swinged counties. Even the best model gave us a 70% accuracy only on the swing from Democrat to Republican and all of them failed to predict the change from Republican to Democrat. Under these results, we can rely on the model to predicted who win but not as an election strategy as these models fail to predict the most changes.

To improve the accuracy of the models, controlling for the variables that stable across time such as gender and racial percentage could be an option. We can also input more independent variables when training the models, for example, the election campaigns held by each party and the percentage of people using social media regularly. With enough data and time, we can also account for the past policies that the last president imposed and see if these have any effect on the voting choices after 4 years.

Future Work and Limitation

In this paper we used 2016 result as our dependent variable, if we apply the models to the 2020 result, will the model as accurate as it was in 2016? If the high autocorrelation still appliable in 2020, we could expect the result should be close to what we have predicted in this paper. But there are many changes from this year election with past, a more than 23% increase in voters where 152,507,250 compare with 2016 at 123,716,997. The voting system also changed a lot this year since the outbreak of coronavirus and mailing vote is allowed in many states.

Social media and its influence have grown a lot more compare to 4 years ago, many articles and journals suggested that this will also affect voters and thus our model and the predictions[2].

In 2016 election, Clinton won around 1 million votes more than Trump, but Trump won by 74 more electoral votes and became the next president. So, the election strategy is not

[2] https://journalism.uoregon.edu/news/six-ways-media-influences-elections

winning more people support but win in the states that contain more electoral votes. In the next machine learning paper, by applying game theory, we could try to come out with the best strategy that the candidate may want to do and use neural network to improve the accuracy.

There will be some limitations that we may need to come over, some variables are difficult to estimate, such as the out of expectation events like "Anyone but Trump[3]" this year, the voter fraud also a hot discussion between both parties and finally the media influences also a major part this year election. These events are difficult to estimate but will change how the people will vote and affect our prediction result.

[3] https://www.nbcconnecticut.com/lx/poll-suggests-young-voters-are-getting-behind-biden-because-he-is-anyone-but-trump/2348342/

APPENDIX I DATASET

fips	Rep2016	Rep2012	PST045214	PST120214	AGE135214	AGE295214	AGE775214	SEX2552
1001	1	1	55395	1.5	6.0	25.2	13.8	51
1003	1	1	200111	9.8	5.6	22.2	18.7	51
1005	1	0	26887	-2.1	5.7	21.2	16.5	46
1007	1	1	22506	-1.8	5.3	21.0	14.8	45
1009	1	1	57719	0.7	6.1	23.6	17.0	50
...	...	...	...	...	...	...	...	
56037	1	1	45010	2.7	7.3	27.0	9.5	48
56039	0	0	22930	7.7	5.7	19.1	12.2	48
56041	1	1	20904	-1.0	7.6	29.8	11.0	49
56043	1	1	8322	-2.5	5.5	23.9	20.1	49
56045	1	1	7201	-0.1	6.5	21.6	18.1	47

3112 rows × 51 columns

APPENDIX II COUNTY FACTS DICTIONARY

PST120214	Population, percent change - April 1, 2010 to July 1, 2014
POP010210	Population, 2010
AGE135214	Persons under 5 years, percent, 2014
AGE295214	Persons under 18 years, percent, 2014
AGE775214	Persons 65 years and over, percent, 2014
SEX255214	Female persons, percent, 2014
RHI125214	White alone, percent, 2014
RHI225214	Black or African American alone, percent, 2014
RHI325214	American Indian and Alaska Native alone, percent, 2014
RHI425214	Asian alone, percent, 2014
RHI525214	Native Hawaiian and Other Pacific Islander alone, percent, 2014
RHI625214	Two or More Races, percent, 2014
RHI725214	Hispanic or Latino, percent, 2014
RHI825214	White alone, not Hispanic or Latino, percent, 2014
POP715213	Living in same house 1 year & over, percent, 2009-2013
POP645213	Foreign born persons, percent, 2009-2013
POP815213	Language other than English spoken at home, pct age 5+, 2009-2013
EDU635213	High school graduate or higher, percent of persons age 25+, 2009-2013
EDU685213	Bachelor's degree or higher, percent of persons age 25+, 2009-2013
VET605213	Veterans, 2009-2013
LFE305213	Mean travel time to work (minutes), workers age 16+, 2009-2013
HSG010214	Housing units, 2014
HSG445213	Homeownership rate, 2009-2013

HSG096213	Housing units in multi-unit structures, percent, 2009-2013
HSG495213	Median value of owner-occupied housing units, 2009-2013
HSD410213	Households, 2009-2013
HSD310213	Persons per household, 2009-2013
INC910213	Per capita money income in past 12 months (2013 dollars), 2009-2013
INC110213	Median household income, 2009-2013
PVY020213	Persons below poverty level, percent, 2009-2013
BZA010213	Private nonfarm establishments, 2013
BZA110213	Private nonfarm employment, 2013
BZA115213	Private nonfarm employment, percent change, 2012-2013
NES010213	Nonemployee establishments, 2013
SBO001207	Total number of firms, 2007
SBO315207	Black-owned firms, percent, 2007
SBO115207	American Indian- and Alaska Native-owned firms, percent, 2007
SBO215207	Asian-owned firms, percent, 2007
SBO515207	Native Hawaiian- and Other Pacific Islander-owned firms, percent, 2007
SBO415207	Hispanic-owned firms, percent, 2007
SBO015207	Women-owned firms, percent, 2007
MAN450207	Manufacturers shipments, 2007 ($1,000)
WTN220207	Merchant wholesaler sales, 2007 ($1,000)
RTN130207	Retail sales, 2007 ($1,000)
RTN131207	Retail sales per capita, 2007
AFN120207	Accommodation and food services sales, 2007 ($1,000)
BPS030214	Building permits, 2014
LND110210	Land area in square miles, 2010
POP060210	Population per square mile, 2010

REFERENCE

1. Hanmer, B. (n.d.). Retrieved from 2016 US Election: https://www.kaggle.com/benhamner/2016-us-election
2. Greene, D. (2020, November). *What Is the Impact of The Presidential Election On The U.S. Economy?* Retrieved from National Public Radio: https://www.npr.org/2020/11/03/930722317/how-the-presidential-election-winner-could-effect-the-economy
3. Brichace, A. *Six ways the media influence elections.* Retrieved from University of Oregon: https://journalism.uoregon.edu/news/six-ways-media-influences-elections
4. *Poll Suggests Young Voters Are Getting Behind Biden — Because He Is Anyone But Trump* (2020, October). Retrieved from National Broadcasting Company: https://www.nbcconnecticut.com/lx/poll-suggests-young-voters-are-getting-behind-biden-because-he-is-anyone-but-trump/2348342/