Predictive Analysis on 2016 U.S. Election

Wenxuan Zhang

Introduction

Predicting the winner of the U.S. presidential election has always been diffcult. People's voting tendency depends on various variables and kept changing between the time of data gathering and election day. Statisticians have been investigating patterns and related features and many polls were conducted before the election in order to predict the winner, yet, election 2016 gave us a big surprise whose result was against most predictions. Thus, it is both interesting and challenging to identify significant variables by machine learning models and make predictions as accurate as possible.

In this project, we focus on predicting the county-level winning candidate and identify the demographic variables that distinguish county clusters and associated candidancy preference. We apply logistic regression and random forest model in the prediction task and reach 92.2% acccuracy in test data. We also identify that $White$, $Citizen$, $Employed$, $Professional$, $Minority$ are the most influential variables to the classification models. Then, we use K-Means to cluster counties into 3 groups and analyze their demographic features difference with respect to their general candidate preference, charaterizing the demographic features of the county groups that are in favor of either caniddates.

Data Preparation

We are given three sets of data: 'election_raw', 'census', and 'census_meta'. The 'election_raw' contains the number of votes of each candidate in each area (nation, state, county) which is represented by the unique $fips$ number. Census data contains census information in 2010. census-meta contains the type of each variable in census data.

Since this project is based on the county-level analysis, we first select all county-level observations in election_raw.

county	fips	candidate	state	votes
Los Angeles County	6037	Hillary Clinton	CA	2464364
Los Angeles County	6037	Donald Trump	CA	769743
Los Angeles County	6037	Gary Johnson	CA	88968

Table.1 The county-level raw election data tat includes the votes of each candidates in each county.

Then, we begin to clean census data by removing some variables that are highly correlated and make some adjustments to the data by converting $Employed$, $Citizen$, $Women$ into percentage of total population. Next step is weighting the variables by population and aggregating the census data into county level.

CensusTract	State	County	Women	White	Citizen	IncomePerCap
1.001e+09	Alabama	Autauga	51.57	75.79	73.75	24974
1.003e+09	Alabama	Baldwin	51.15	83.1	75.69	27317
1.006e+09	Alabama	Barbour	46.17	46.23	76.91	16824

Table.2 The aggregated county-level census data

After getting tidied census and election_raw, our final goal is to combine these two into one table. However, before doing that, we need to only obtain the observation of the winner in each county (Trump or Clinton) since our upcoming models are aimed for predicting the winner of each county. Then, we can combine these two tables by the key *county* and *state*, which gives us the final version of data called 'merged_data'.

county	fips	candidate	state	votes	total	pct
autauga	1001	Donald Trump	alabama	18172	24759	0.734
baldwin	1003	Donald Trump	alabama	72883	94261	0.7732
barbour	1005	Donald Trump	alabama	5454	10436	0.5226

Table.3 The aggregated county-level election data combined with census data

In order to fit our models of logistic regression,random forest, and K-means, we need to convert the candidate to factors and remove variables that are not predictive features. Specifically, we do that by letting Trump represent 1 and Clinton represent 0.

candidate	Women	White	Citizen	IncomePerCap	Poverty	ChildPoverty
1	51.57	75.79	73.75	24974	12.91	18.71
1	51.15	83.1	75.69	27317	13.42	19.48
1	46.17	46.23	76.91	16824	26.51	43.56

Table.4 County level feature variables only census data with encoded winning candidates

Methods of Analysis

Our primary goal is to build prediction models for the county-level winning candidate and identify significant predictive variables. We approach this by both supervised and unsupervised learning models.

Task 1: Predict the County Winner: Logistic Regression and Random Forest

For this task we use supervised learning models to predict the winner and identify significant variables by fitting logistic regression and random forest to model the probability of two major candidate won each county. Logistic regression enables us to obtain the probability of each candidate to win the election in each county and find significant variables by $P - value$ and coefficients with high interpretability.

$$\text{logit}\left[P\left(\text{Trump win}\right)_i\right] = \beta_0 + \beta_1 \text{white}_i + \cdots + \beta_{10}\text{income}_i \ , \quad \text{county } i = 1, \ldots, 2148$$

On the other hand,random forest is a bagged decision tree estimator that gives us the class selected by most trees. Although we lose the chance of plotting the tree and some interpretability, random forest makes it up by providing a measure of variable importance in terms of variables interpretation.

Before fitting the model, we split the dataset by 80% and 20% to be used as trainning and testing datasets respectively. For both methods, we examine the error rate by computing the confusion matrix of the initial model. Then, in order to minimize the error rate, we use the optimizing threshold by compute the youden statistics that gives the lowest combined false positive/negative rates.

Task 2: County Clustering and Candidate Preference: K-Means

In addition to optimization of prediction accuracy, we also draw the most predictive variables as significant demographic features for this prediction task by interpreting the parameters in logistic model and the measures of variable importance in random forest model. As an initial discovery, we make some exploratory plots based on several most predictive variables.

Moreover, we use K-Means as an unsupervised learning method to cluster the counties with similar demographic features and then analyze the association between the clusters and their winning candidate, looking for patterns of candidate preference across similar groups.

Finally, we compare our analysis for significant predictive variables between the supervised and unsupervised learning approaches. And we then we take one step further by fitting the entire dataset on our random forest model and comparing the election map of prediction and true result.

Summary of Results

To kick off, we take a glance at the class errors of the initial model. As shown in table 5-6, there is a high false positive error rate (around 30%) in both logistic regression model and random forest model, due to the unbalanceness of the number of county won by Clinton and Trump. In the county-level, Trump indeed won far more counties than Clinton did.

y- test y	Clinton	Trump
Clinton	0.6092	0.3908
Trump	0.01894	0.9811

Table.5 Class error rate of initial logistic regression model

	Clinton	Trump
Clinton	0.7586	0.2414
Trump	0.01894	0.9811

Table.6 Class error rate of initial random forest model

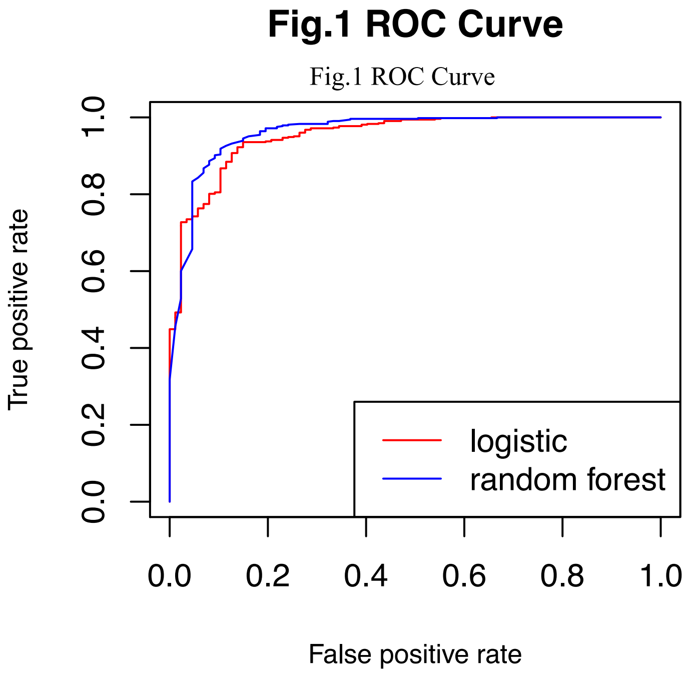
By choosing the threshold that maximizing the corresponding Youden statistics, we significantly lower the false positive rate to 14.9% and 10.3% for logistic and random forest model respectively (table.7-8). We can visualize this process by the ROC plot below (fig.1) where the area under the two curves are approximately the same, which means that both method works well. In addition, combined with the error rate tables, we can see that the random forest predicts better than logistic regression by 1.3% percent.

y- test y	Clinton	Trump
Clinton	0.8506	0.1494
Trump	0.06629	0.9337

Table.7 Class error rate of adjusted logistic regression

	Clinton	Trump
Clinton	0.8966	0.1034
Trump	0.08902	0.911

Table.8 Class error rate of adjusted random forest model



Examing the total test error rate table (table.9), we come to an result of 0.078 total test error rate in logistic model with the threshold 0.794 and 0.091 test error rate in random forest model with threshold 0.79. The optimized threhold of both model are very close. In train dataset, the random forest model has lower error rate than that of logistic model, however, the latter performs 1.3% better in predicting on the testing dataset. Therefore, we choose logistic regression model to fit the entire dataset such that we can get a result of predictive election result.

	train.error	test.error	threshold
logistic	0.09406	0.07805	0.7944
random forest	0.01303	0.09106	0.79

Table.9 Total train, test error rate and optimized threshold of both models

By viewing the summary of logistic regression, we can directly identify 10 significant variables by checking small P-value such as $Citizen$, $Services$, $Professional$, $WorkAtHome$, $Service$, $Production$, $Drive$, $Carpool$, $Employed$, and $PrivateWork$. Furthermore, the formula of logistic regression easily guides us to interpret the effect of each variable.

$$\text{logit}\left[P\left(\text{Trump win}\right)_i\right] = -0.1171\, Citizen_i - 0.3295\, Serivice_i - 0.2532\, Professional_i \cdots ;, \quad \text{county } i = 1, \ldots, 2148$$

For example from the formula above, if we have 1 unit increase of citizen, the odds will have a multiplicative change of $e^{-0.1171}$ so that the candidate will be more away to 1 and the county will prefer Clinton more.

From random forest model, we can measure the importance of variable in terms of classification accuracy (0-1 loss) and Gini index. Figure 2 shows the decrease in misclassification rate across trees by order. In other words, the higher the mean decrease of a variable in this plot, the more important and accurate it is in the prediction of this classification task. We observe that

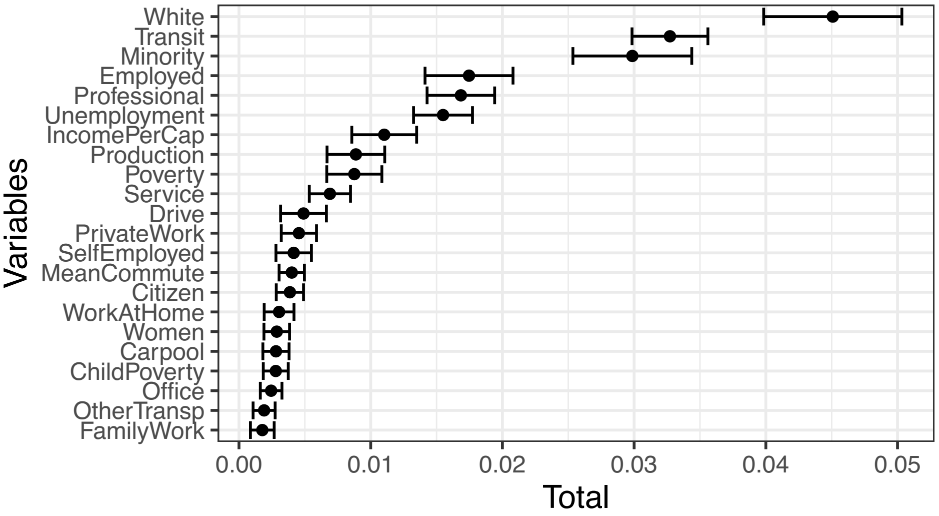


Fig.2 Mean decrease in misclassification rate across trees

Comparing the variables we get from the two model, we find that $Professional$, $Transit$, $Minority$ appear in both approach as important indicators. Interestingly, $White$ appears to be the most significant variable in the decision of random forest model, however, in logistic regression it is not of the same importance. Based on these variables, we conduct several exploratory analysis to visualize their infuence on candidancy preference.

(1) Figure 3-4 visualizes the county distribution with citizen percentages. The distribution patterns for two candidates do not significant differ when we set the total number of counties as the y axis. While we set density as the y axis, the distribution now differs. Among those counties with a low citizen level, there is a higher probabilty that Clinton wins.

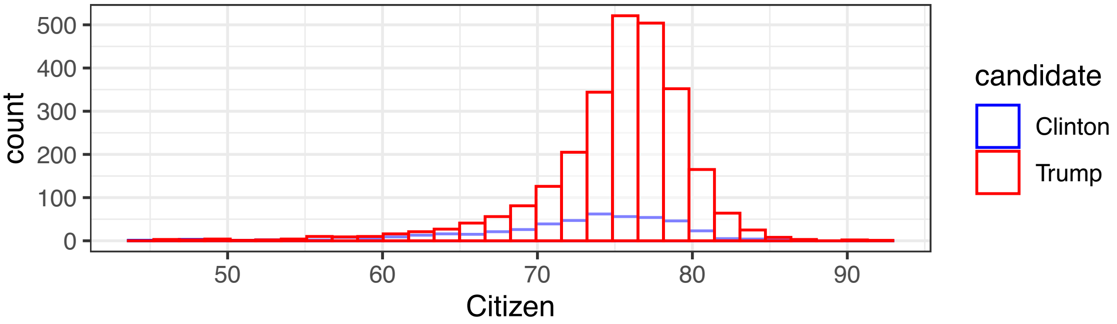


Fig.3 Number of counties in different citizen levels

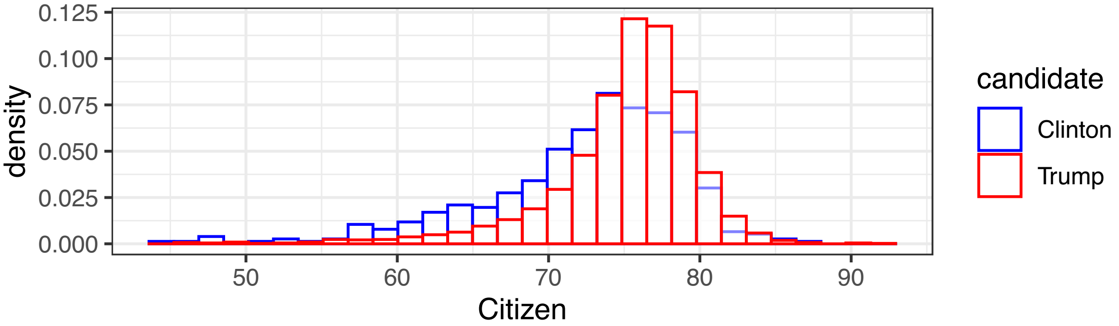


Fig.4 Percentage of counties in different citizen levels

(2) Figure 5-6 shows the distribution of all counties in employment situation. The x axis is the employed rate and y axis is the unemployment rate. Counties that Trump wins are cummulated at the lower part of the plot, which is where the unemployment rate is relatively low. We concluded that counties with a

low unemployment rate are more likely to support Trump.

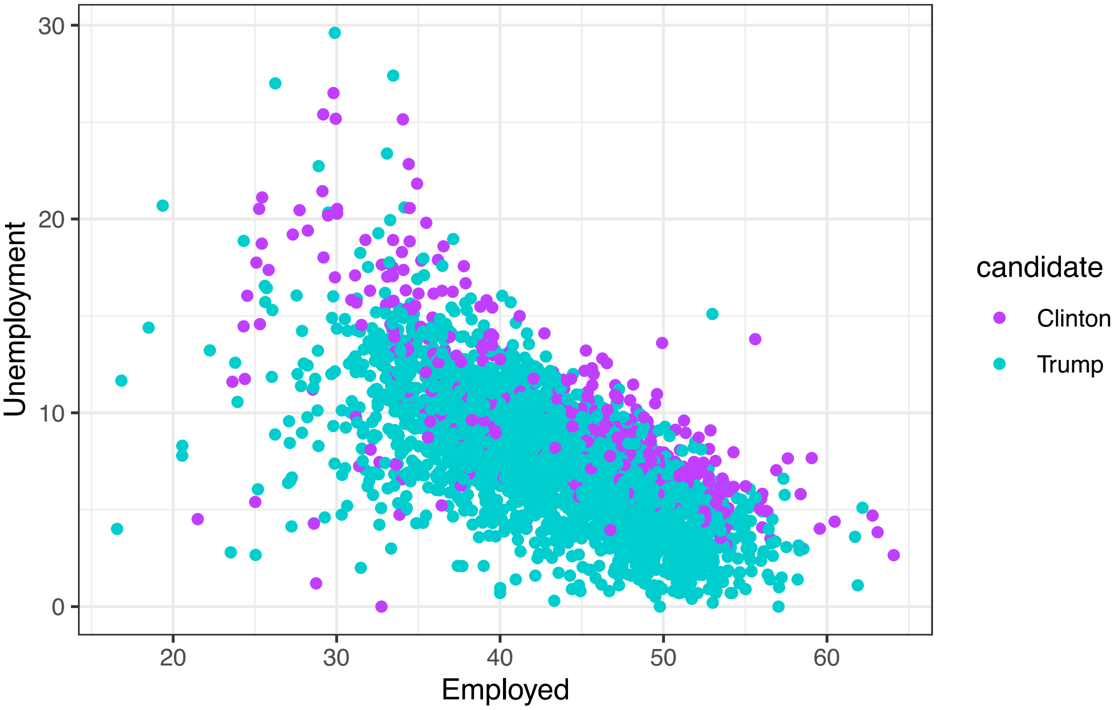


Fig.5 Countiy Unemployment Rate By Winning Candidates

(3) Then we can investigate the professional levels of counties for each candidate. Similarly, figure 7 is the count value of counties, gigure 8 is the density level. The differences are very obvious. Most of the counties that support Trump are in the range between 25% and 35% of professional levels. However, counties with a higher professional level, about 30% to 50%, support Clinton more.

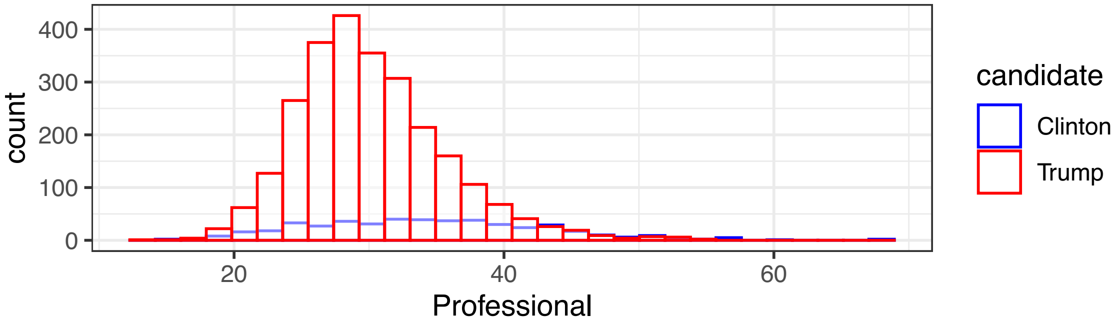


Fig.6 Number of counties in different professional levels

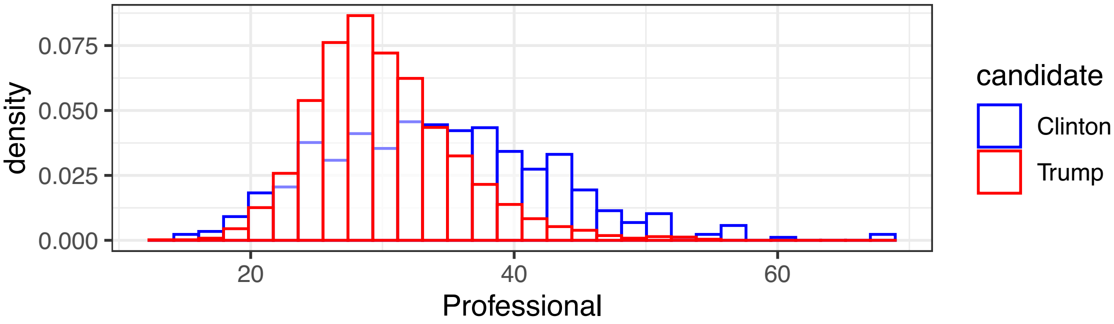


Fig.7 Percentage of counties in different professional levels

After visualizing the variables through exploratoratory plots, we are intereted in clustering the counties with similar demographic characteristics and whether they exhibit patterns in candidate preference. We cluster all the counties into 3 group and calculate the perventage of counties won by Trump. It is obvious that 95% of the counties in cluster 2 were won by Trump! Cluster 1 and 3 have relatively lower proportions of trump won counties.

cluster	trump_win	ncounty	prop
cluster 1	476	691	0.6889
cluster 2	1438	1498	0.9599
cluster 3	693	882	0.7857

Table.10 Trump supporter proportion by clusters.

Ploting out the density distribution of each varibles, we do not see potential clustered peak, however, after coloring by clusters, it's obvious that in varibles such as $WorkAtHome$, $White$, $Unemployment$, $Professional$, $Minority$, $IncomePerCap$, $Employed$, $ChildPoverty$ has quite different distributions across different clusters.

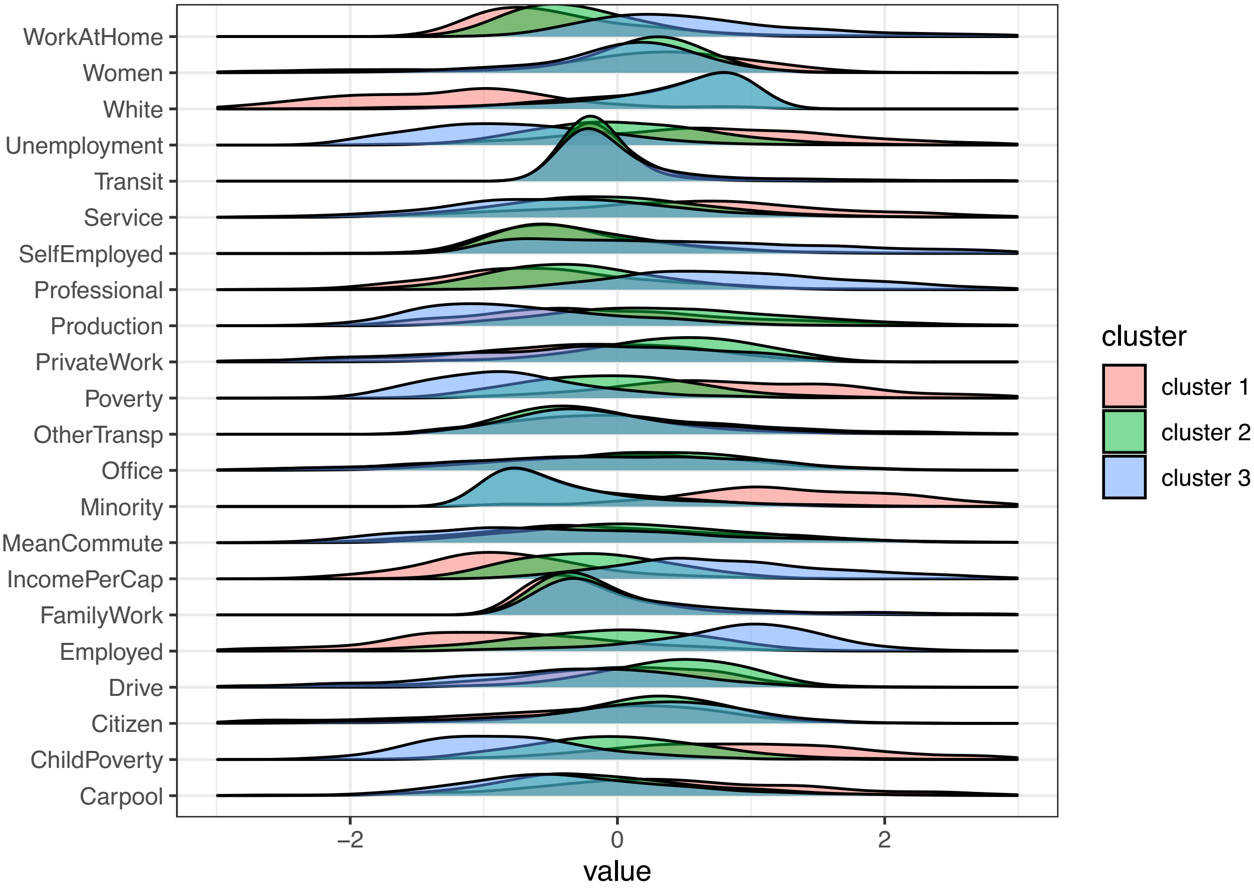


Fig.8 The Density Distributions of Variables by Clusters

Specifically, the cluster 1, the group with lowest Trump won counties percentage, has significant lower $White$, $IncomePerCapita$, $Employed$ and higher $Poverty$, $Minority$, $ChildPoverty$. On the other hand, the cluster 3 where the second lowest Trump won counties percentage, has higher $White$, $Professional$, $IncomePerCapita$ and $Employed$. The cluster 2, in which 96% of the counties hace trump as winner, seems to have most variable density distribution between cluster 1 and 3, except for the $White$, $PrivateWork$, *Production* are the highest.

It seems that cluster 1 depict those counties with poor social benefit conditions that has high poverty and unemployment rate where minority takes up large part of local population. Cluster 3 depicts counties with good economic, employments conditions, high percentage of professional workers where the major population are whites. However, the group that most support Trump is cluster 2 whose population are dominant by whites and have median social benefit and economic development. These counties are feartured by high driving rate, private work and production sections.

Comparing the several significant variables we draw from the previous classification models, we find that $Minority$, $Professional$, *Employed* both appeared in logitic regression, random forest, and cluster analysis, which imply that they are distinguishable variables that best characterize the candidate preference of a county.

For a holistic review on the prediction task, we fit the logistic regression model to predict the whole election result. Fig.9-10 exhibit the map of true election and prediction result based on the logistic model we have build. We can observe that the two map conform in the most part of the map, though there is slightly different result in some counties.

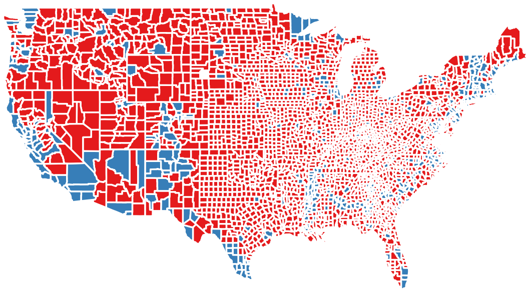


Fig.9 True map of county-level winning candidate

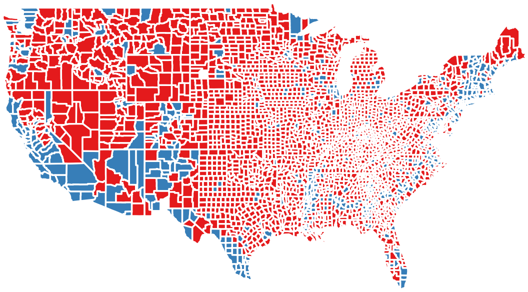


Fig.10 Predicted map of county-level winning candidate

Discussion

Through the analysis from prediction and clustering model, we find accurate prediction models for predicting the winning candidate in each county and identify the relationship between essential predictive census variables and associated candidate preference. The above analysis provides a solid basis for potential further investigation on presidential election. Prospectively, we are able to aggregate our county-level result to state and nation-level, conbining with extra electorial information to make the prediction for state or final election wining candidate. Furthermore, we are also going to dive deeper into our analysis on the association between demographic features and candidate preference with more procedure, such as adding a step of original data cleaning process or introducing other methods such as PCA to help our further analysis.