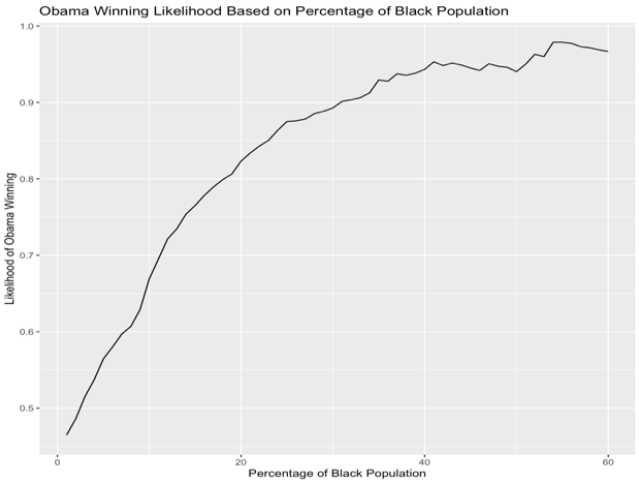
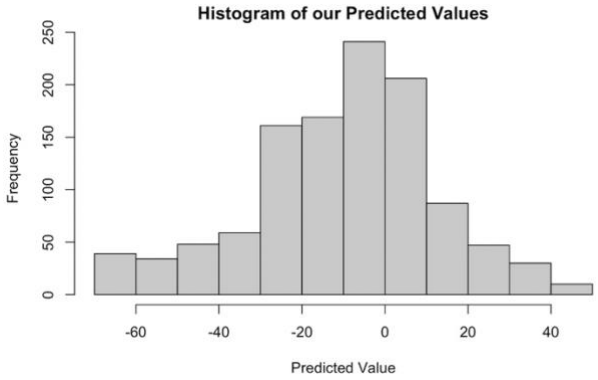
Higher resolution images and full regression equations (if applicable) are available in the R markdown file

Election Outcome Prediction:



After trialing multiple combinations, we have established a relationship between the percentage of black population in a state Vs the likelihood that Obama would win (the "What"). In order to visualize this relationship better, we have plotted the two variables against each other (the "How"). As evident in the plot, there is a logistic relationship between the percentage of black population and the likelihood of Obama winning. This means, the higher the percentage of black population is in a county, the more likely that Obama would win. However, the frequency of this change slows down as the percentage grows higher. We have also considered both the primary and caucuses election vote numbers as an aggregate for this purpose. The question this answers is *"is there a relationship between Obama winning and the percentage of black population in a county?"*



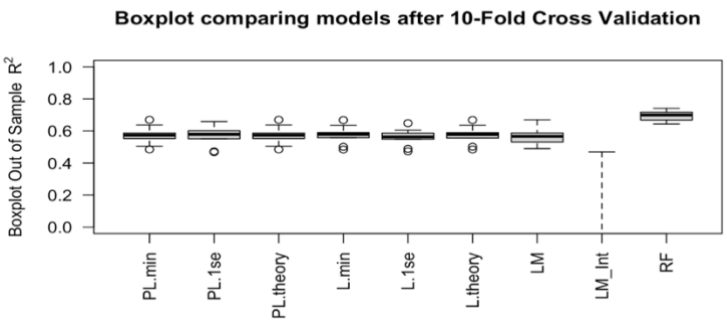
We have the task to predict the winning spread for Obama. We have decided to go with a regression model and selected the target variable of our prediction to be "Obama_margin_percent". The models that we are comparing are the following: Linear Model, Linear with Interaction, Random Forest and Linear Regression with Lasso. Additionally, we have used Cross Validation to find the optimal lambda values for the Lasso model. On using CV to find the optimal lambda values, we noted that the lambda.1se=0.32, lambda.min=0.03 and lambda.theory=0.0304. Thus our value for lambda theory lies between the min and 1se value. The plot can be seen below (figure2). and We used the K-Fold Cross Validation for all our models (10-Fold). Ultimately, we have used the metric out of sample R^2 to compare the models. We plotted the box-plot that compares the models (*Shown below*).

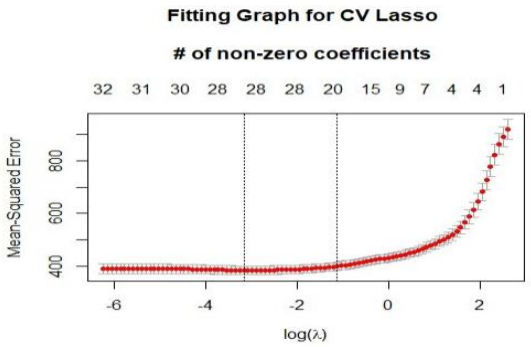
Columns that we have dropped: We drop the columns since they will not help us with our predictive modeling and the columns "County", "State" won't be helpful because the test set has new states that are YET to vote. Hence, they provide no value. "FIPS" is a unique identifier and hence needs to be dropped. The columns " Total Vote", "Clinton", "Obama", "Obama_margin" is what determines what we're trying to predict so they will skew the model, and hence are dropped too.

Core task	**Regression**			
Methodology	**Linear Regression**	**Linear with Interaction**	**Random Forest**	**Linear Regression with Lasso**
Metric used	Out of Sample R^2 is used to evaluate the performance			
Data understanding and limitation of the model	Since there are far too many variables to compare and determine which ones add value, we have used all the variables as dependents (apart from what we have dropped). For better optimization we have used lasso (See column 4)	The linear with interactions also combines all the possible interactions between all the variables after dropping the columns that we don't need.	We understand that the data has many features and that might lead to a larger tree. If there are many leaves then that might lead to overfitting. That is the limitation of the model naturally.	We have used lasso to regularize the model and reduce its dimensions. As you can see in the code, our features were reduced to the following: Lasso Min: 28 features Lasso 1se: 20 features Lasso Theory: 28 features
Overall assumptions	• Both primary and caucuses election data are used together, i.e. The Election Type column would not prove to be valuable in our prediction. • We have also assumed that the OOS R2 value will be 0. Hence, we have not added it in the boxplot for comparison.			

Higher resolution images and full regression equations (if applicable) are available in the R markdown file

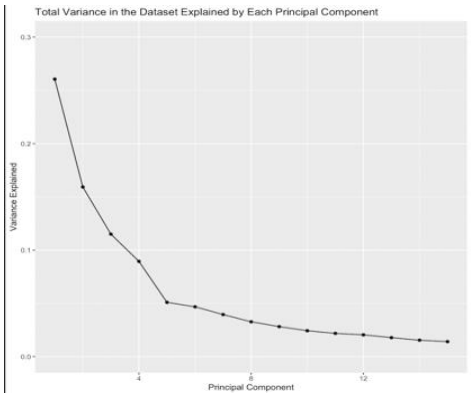
Selection of the model based on OOS R^2 performance: Random Forest

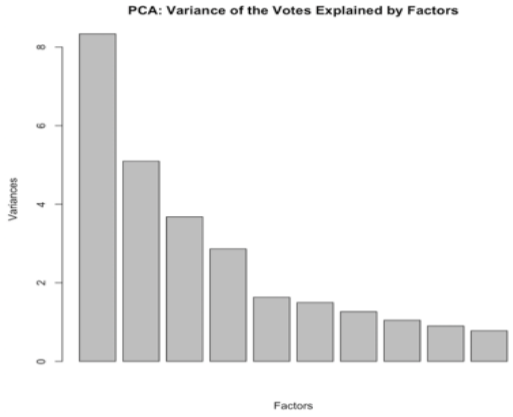




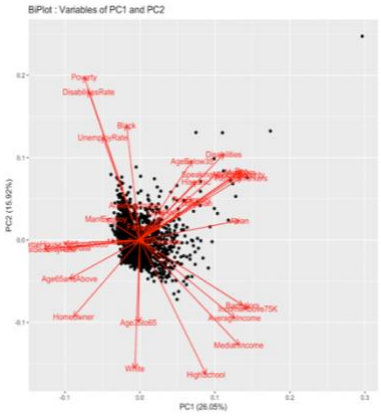
K-fold cross validation: We have applied k-fold cross validation and obtained the above results and selected the random forest model dye to it having the highest out of sample R squared value. We have attached an excel with our predictions of margin spread.

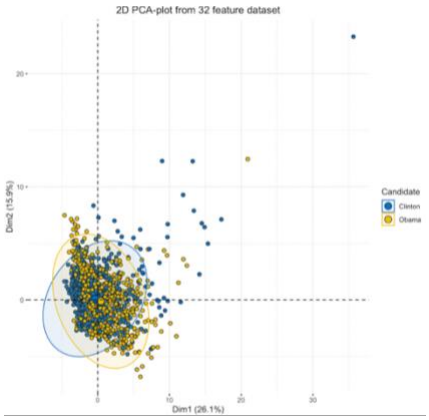
In order to explore the dataset, we have applied principal component analysis in efforts of capturing the variance in just several dimensions. Using the package "plfm" we ran a principal component analysis on the training dataset after dropping the non-numerical and predictor variable columns. The first 15 principal components account for 93.6% of the variability in the dataset. However, we have decided to omit the principal components with eigenvalues less than 1 because those components explain less than a single explanatory variable. The first 8 components have eigenvalues>1 while also explaining 79% of the variability in the dataset. The training dataset had 32 explanatory variables initially (excluding the variables that needed to be predicted and the categorical variables). Through principal component analysis we have successfully reduced the dimensions to 8 principal components





Analyzing the principal components





The BiPlot shows how strongly the variables in the dataset influence the first two principal components PC1 and PC2. The further the distance of the variable is from the center, the stronger its' influence is on the principal component. In order to interpret this more clearly, we have performed a loading analysis to determine which factors exactly have a strong influence on the first four principal components. From the 2D PCA plot we can infer that there is no visible clustering in this dataset with regards to the Obama- Clinton preference (i.e, there is a lot of overlap between votes for Obama and Clinton)

PC1

IncomeAbove75K	Bachelors	Pop
0.2703670	0.2573727	0.2546433
RetiredWorkers	SocialSecurity	Medicare
0.2541063	0.2532532	0.2515439
Asian	MedianIncome	SocialSecurityRate
0.2493214	0.2477901	-0.2390821
AverageIncome	Disabilities	SpeakingNonEnglish
0.2368762	0.2102171	0.1921868
SameHouse1995and2000		
-0.1890212		

PC2

Poverty	DisabilitiesRate	HighSchool	White
0.3718694	0.3383788	-0.3067044	-0.2931444
Black	MedianIncome	UnemployRate	Disabilities
0.2616193	-0.2394588	0.2340859	0.1958560
Age35to65	AgeBelow35	AverageIncome	
-0.1877225	0.1803111	-0.1775049	

PC3

AgeBelow35	Age65andAbove	SocialSecurityRate
0.3658871	-0.3494381	-0.3227037
MedicareRate	Medicare	RetiredWorkers
-0.2993133	-0.2796108	-0.2769322
SocialSecurity	Disabilities	
-0.2735223	-0.2658604	

PC4

FarmArea	LandArea	Black
-0.4278640	-0.3731425	0.3504350
Hispanic	SpeakingNonEnglish	White
-0.3495099	-0.3018136	-0.2763606

The above shows clearly which latent features impact the principal components. For instance, PC1 is heavily influenced by variables such as IncomeAbove75K, Bachelors etc while PC2 is heavily influenced by variables such as Poverty, DisabilitiesRate etc.

Higher resolution images and full regression equations (if applicable) are available in the R markdown file

Providing an estimate for what would have been the average impact on the winning spread for Obama over Clinton (measured in percentage of total voters) had the Hispanic demographic been 5% larger

We have used the below assumptions when approaching this question:

a. If the Hispanic population (or black population) increases by 5%, all the other variables in our data don't change. Meaning Hispanic population and black population are independent of the other variables.
b. The percent increase in both of black and Hispanic populations is uniform across all counties.

Methodology: According to our simple model, a 5% increase in the Hispanic population will lead to a 0.55% decrease in Obama winning margin percent. In order to calculate the change in Obama's marginal percent with our Random Forest model had the Hispanic population been 5% larger, we will compare the mean Obama winning margin percent in our original data with the data in which we have a 5% larger Hispanic population. So, according to our Random Forest model, a 5% increase in Hispanic population will lead to 0.2% decrease in Obama's average winning percent. Comparing our model to the simple model that we were given, we can see that Random Forest is more conservative in its estimations of Obama's marginal percent.

In the case of the 5% Black population increase, the simple model predicts a 4.3% increase in Obama's marginal percent. However, in the case of Random Forest model, a 5% increase of black population will lead to only a 0.6% increase in Obama's marginal percent increase. Again, according to our model the increase in Obama's win percentage would be a lot less than 4.3%.

	5% Hispanic Increase	5% Black Increase
Simple Model (glm)	0.55% decrease in Obama winning margin	4.3% increase in Obama winning margin
Random Forest	0.2% decrease in Obama winning margin	0.6% increase in Obama winning margin

Steps going forward for a chosen candidate:
Our chosen candidate is *Obama*.

Methodology:
We have filtered the testing dataset by state and determined which states are "swing states". This means states where voter preference for each candidate is split evenly and the decision on which candidate is actually preferred could "swing" (E.g.: A state where Obama has won in 5 counties and Clinton has won in 4 counties is a "swing" state). Based on our analysis (detailed code in R markdown file), we have determined the following categories to which the swing states belong to and their respective advertising categories:

Winning margin in state	**Advertising strategy**	**Reasoning**
Obama is predicted to win big	Minimal advertising/campaigning	Obama is predicted to win big in these states and therefore resources should not be allocated heavily towards advertising in these states.
Obama is predicted to win marginally (swing state)	Heavy advertising / campaigning	Obama is predicted to win marginally in these states and therefore we would advise to conduct heavy advertising because there is always a chance that Clinton could win too.
Obama is predicted to lose big	Minimal advertising/ campaigning	From what we know about elections there are some states which are heavily biased. These would be the states where Obama is predicted to lose big and therefore resources should not be allocated heavily towards advertising in these states as the likelihood of winning would not be high even with a lot of convincing.
Obama is predicted to lose marginally (Swing state)	Heavy advertising/ campaigning	Obama is predicted to lose marginally in these states and therefore we would advise to conduct heavy advertising because there is always a chance that Obama could win too.

Assumptions:

a. Already a certain level of campaigning and advertising has been conducted in all states/counties.
b. All counties have equal weightage in votes at the presidential election.

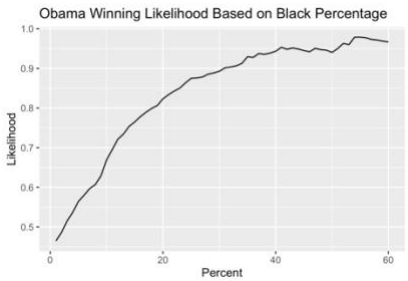
Higher resolution images and full regression equations (if applicable) are available in the R markdown file

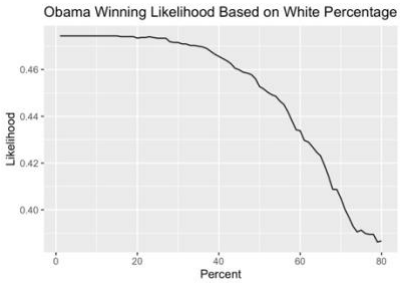
State	Number of counties Clinton wins	Number of counties Obama wins	Total counties	Level of advertising/campaiging
HI	1	3	4	Light
IN	52	40	92	Heavy
KY	115	5	120	Neutral
MS	21	61	82	Light
MT	34	22	56	Heavy
NC	47	53	100	Heavy
OH	61	27	88	Neutral
OR	18	18	36	Heavy (Extreme swing state)
PA	61	6	67	Neutral
RI	2	3	5	Heavy (Extreme swing state)
SD	43	23	66	Neutral
TX	212	39	251	Neutral
VT	3	11	14	Light
WI	23	49	72	Light
WV	53	2	55	Neutral
WY	3	20	23	Light

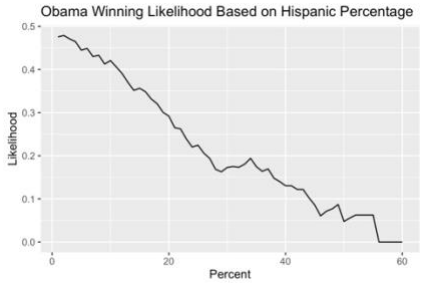
Which states to target: As per the analysis that we have conducted on the testing dataset, the swing states are **IN,MT,NC,OR,RI** and the states where Obama is leading already are **HI,MS,VT**. We recommend that all these states need advertising/campaigning to the level we have recommended.

Which demographic should be targeted in these states:

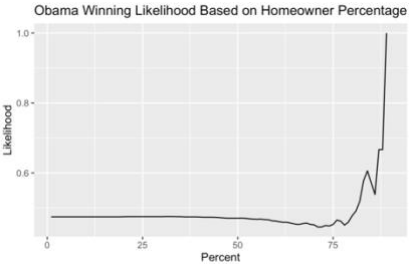
1. Based on ethnicity
a. As per our visualizations below we believe that counties with a larger black community will increase the likelihood of the votes swinging in favor of Obama if advertised as advised.
b. Obama's campaign managers should target counties with lower percentages of Hispanic/white populations as these counties are more likely to swing in favor of Obama.

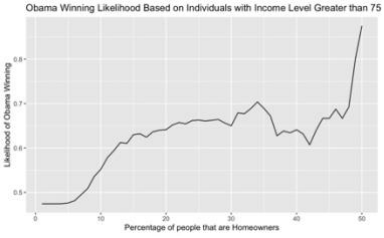






2. Based on income level

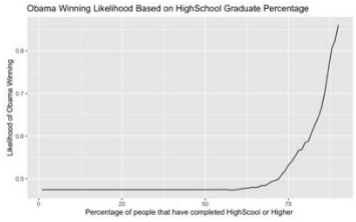


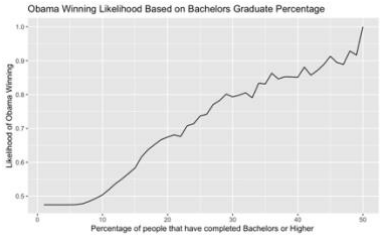


Based on the visualizations, Obama's campaign should target states with a higher proportion of upper middle class and above population. The above two visualizations indicate that the more homeowners/ people with income above 75K there are in the county, the more likely that they'll swing towards Obama.

3. Based on education level

Higher resolution images and full regression equations (if applicable) are available in the R markdown file





Based on our visualizations, we believe that Obama's campaign managers should target states with a higher level of education. I.e, states where a higher proportion of people have graduated from high school/bachelor's degree.

Further steps that could be taken: Diving in deeper

a. After filtering out the swing states in which advertising should take place in, Obama's campaign managers should ONLY focus on the counties where it's reasonable to estimate that the voter mindset could be swung/further convinced in favor of Obama. For instance, if there is a state where 2 out of 5 counties are favorable towards Obama and 3 counties which is heavily in favor of Clinton, it may not make a lot of financial sense to heavily advertise for an Obama campaign because voter mindset may not change regardless although it's filtered as a swing state.

b. There is always an option to dive deeper into the analysis to only target the counties where Obama is winning or losing by a very small margin. Additionally, when we recommend targeting "High-Income Individuals" we suggest a Cost Vs Benefit approach should be conducted to determine the market size and make the decision accordingly. For Example, if we know that a state/county is relatively poorer we would not advertise to High-Income Individuals there, because we incur a cost but there is not a relatively high benefit.