Obama-Clinton Case Study

1 Problem

Problems and Subproblems

The main problem of the case study is failure to correctly address the voters in elections as both candidates have made poor choices in approaching the target voters.

Our sub-problem focuses on the Obama campaign's objective to maximize votes. To optimize campaign budgets, campaign efforts should be concentrated among counties that are probable to sway towards voting for Obama. We believe that politically centered counties with an even predicted voting split are easier to sway than deeply Clinton leaning counties.

Target Attribute

As such our target attribute would be the margin of Obama's votes over Clinton. Margins close to zero would indicate these evenly split counties.

2 Understanding the Data

The 2008 Obama-Clinton election dataset is a reliable source showcasing election data and demographic information from the US Census Bureau. It consists of 2868 rows of available county voting data, of which 1131 rows contain unknown data due to unsubmitted votes.

The dataset presents discrete and continuous numerical and categorical attributes. "ObamaPercentMargin" was derived to calculate the marginal percentage difference between Obama and Clinton votes.

Attributes in Focus

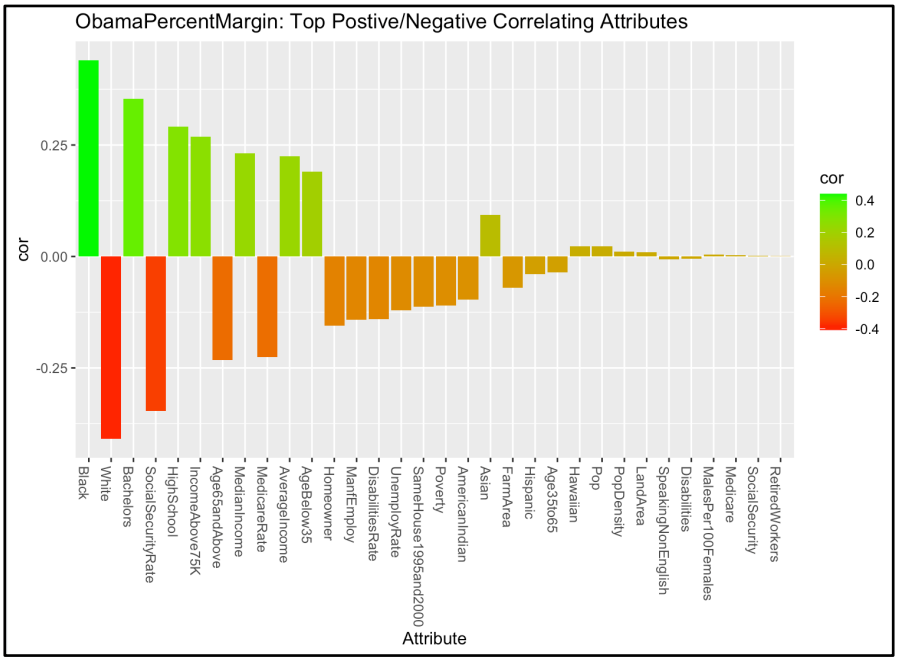

Figure 1: Correlation with ObamaPercentMargin

To predict ObamaPercentMargin, we focus on 10 most correlated attributes with ObamaPercentMargin:

Black
White
Bachelors
SocialSecurityRate
HighSchool
IncomeAbove75K
Age65andAbove
MedianIncome
MedicareRate
AverageIncome

These generally are from three categories: Black/White, Education and Income

Correlation between attributes of interest

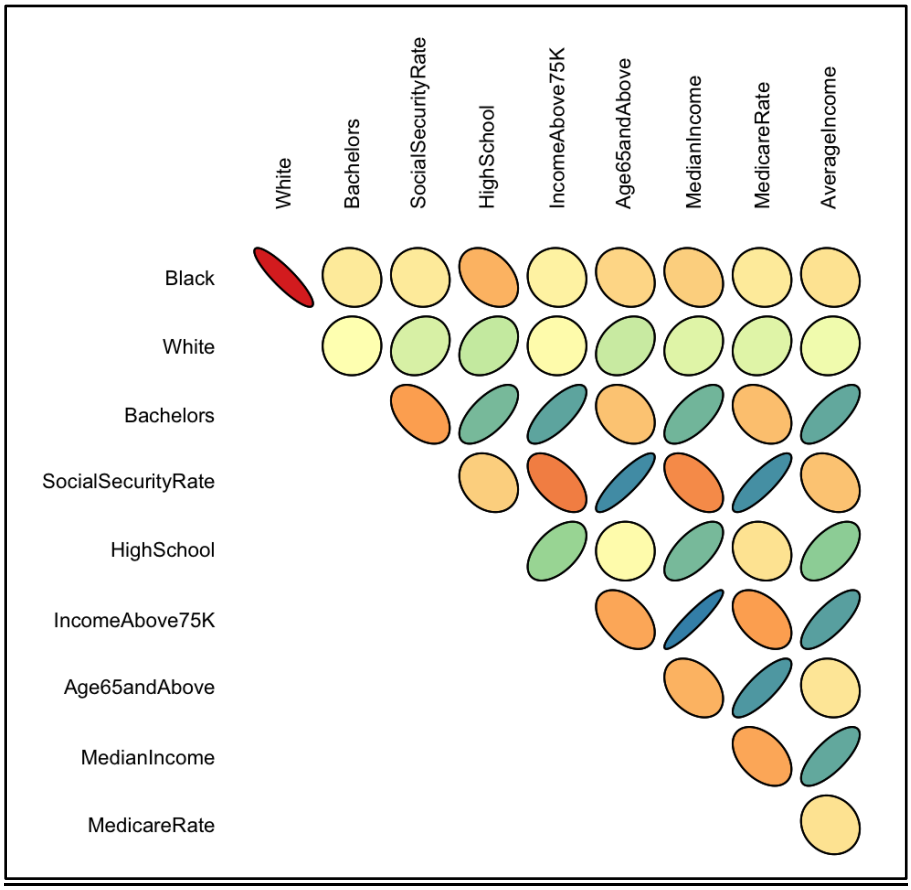

Figure 2: Correlogram

Figure 2 shows strong negative correlation between Black and White and strong positive correlation of SocialSecurityRate with Age65andAbove and MedicareRate and IncomeAbove75K with MedianIncome.

Data Visualisation

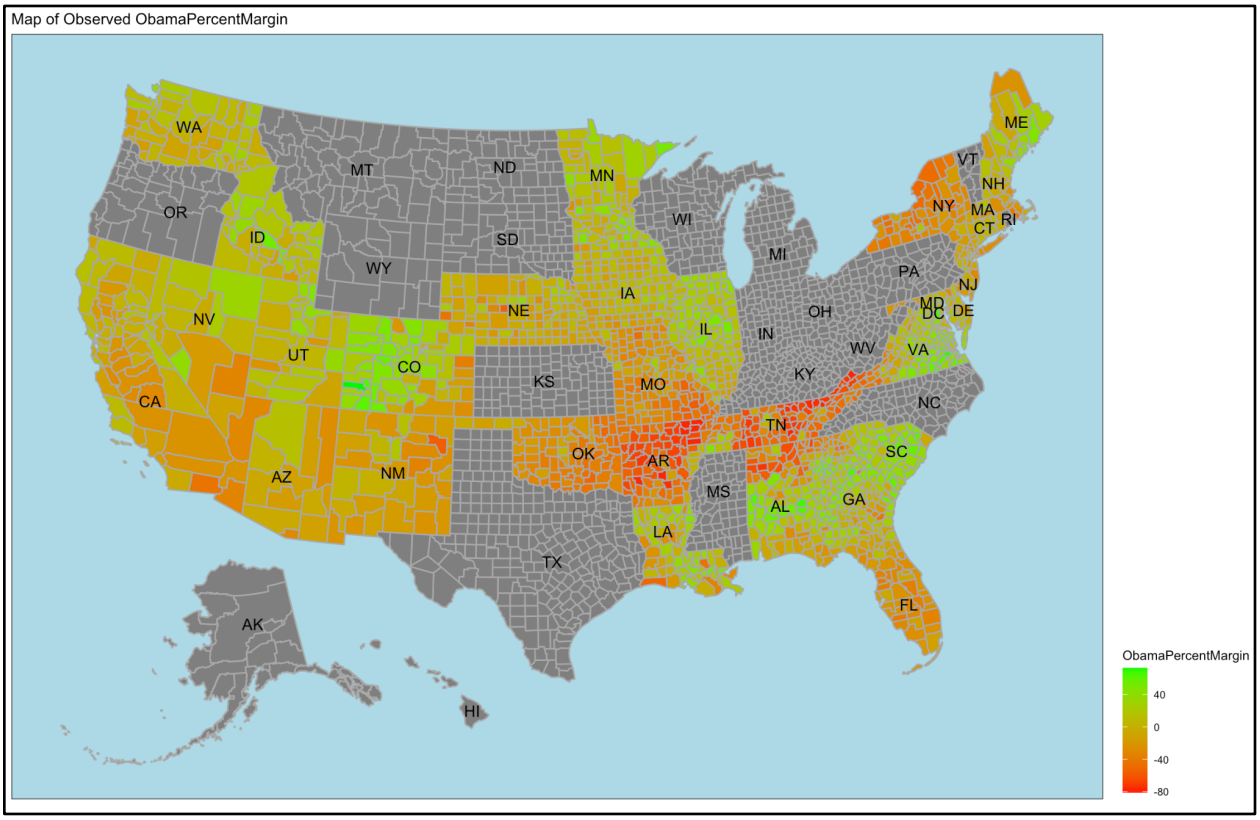

Figure 3: Map of ObamaPercentMargin for voted counties

Figure 3 shows ObamaPercentMargin for the counties that have voted. We see here that some areas (AR, TN) are heavily Clinton leaning while others (CO, AL) lean heavily Obama. There are large variations of ObamaPercentMargin within states and regions with no clear trend.

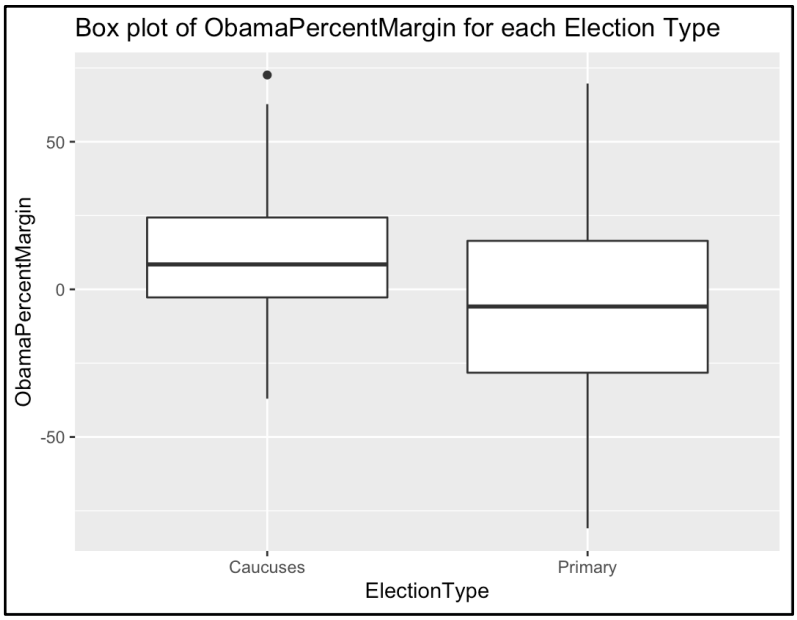

Figure 4: Boxplot of ObamaPercentMargin

Figure 4 shows ObamaPercentMargin positively skewed for caucus elections while negatively skewed for primary elections, therefore, Obama generally performs better in caucuses. The greater variation of ObamaPercentMargin for primary elections could be explained by the larger sample size of 282 caucuses, contrasting with 1454 primaries.

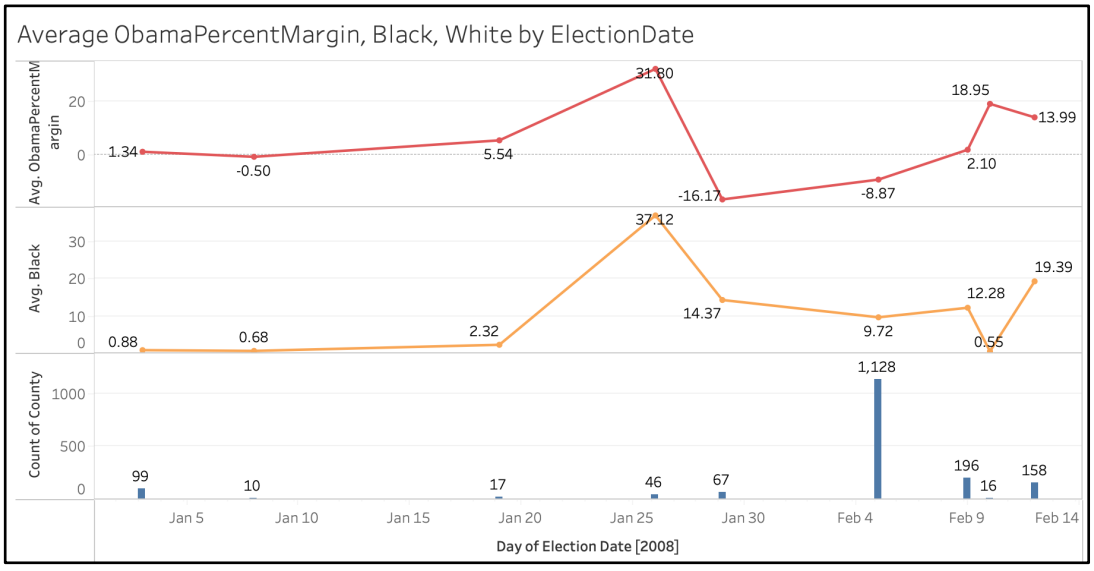

Figure 5: Average ObamaPercentMargin, Black and Election count against ElectionDate

Figure 5 shows similar trends in average ObamaPercentMargin and Black until January 29th, however, after this date, the trend for average Black opposes the average ObamaPercentMargin. The average demographic of counties who voted before January 29th may be unrepresentative of counties who voted after. As such, Black may be a factor that determines ObamaPercentMargin but only in certain demographics.

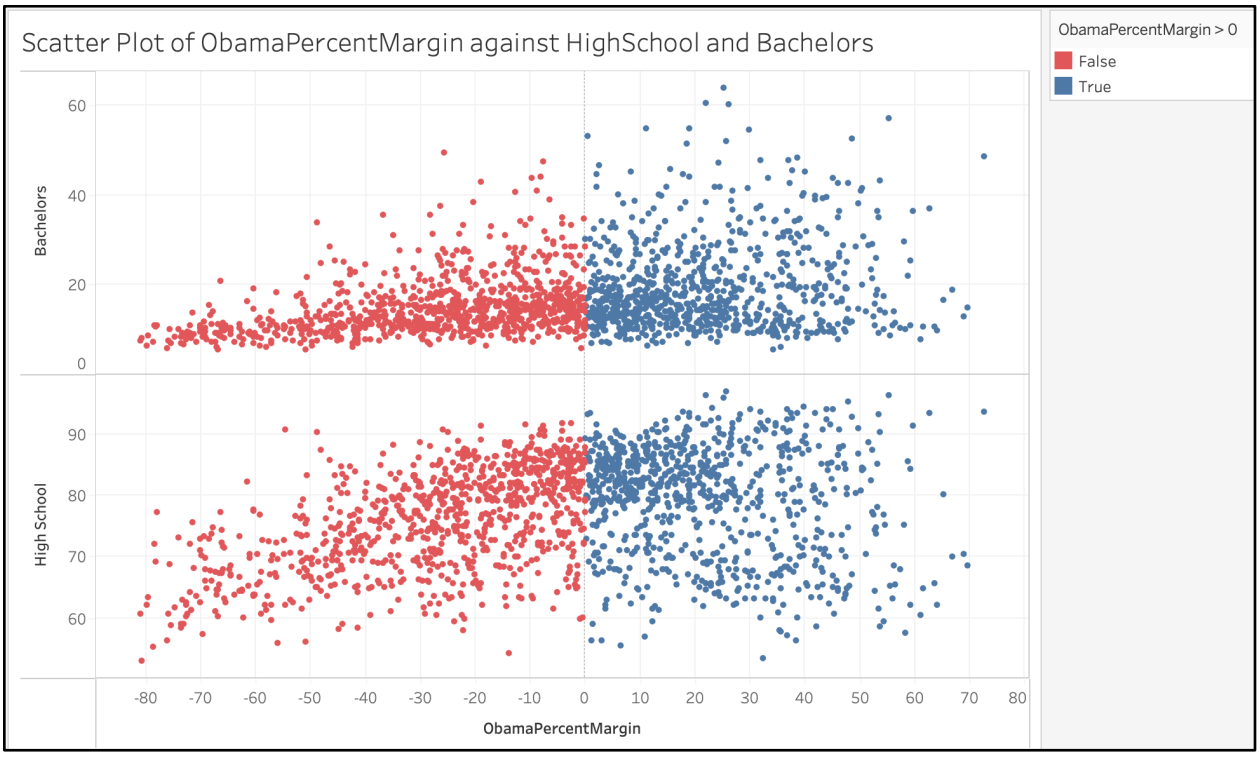

Figure 6: Bachelor and HighSchool against ObamaPercentMargin

The HighSchool plot shows spread out distributions of HighSchool values. Points are concentrated around 80<HighSchool<90 and -30<ObamaPercentMargin<30, with variations at higher values. It is observed that highly negative ObamaPercentMargins occur with lower HighSchool.

The Bachelor plot in Figure 6 shows most points concentrated where Bachelor<20. One-to-many mapping suggests Bachelor is not a good predictor of ObamaPercentMargin due to large variations in points, although there is a weak positive correlation.

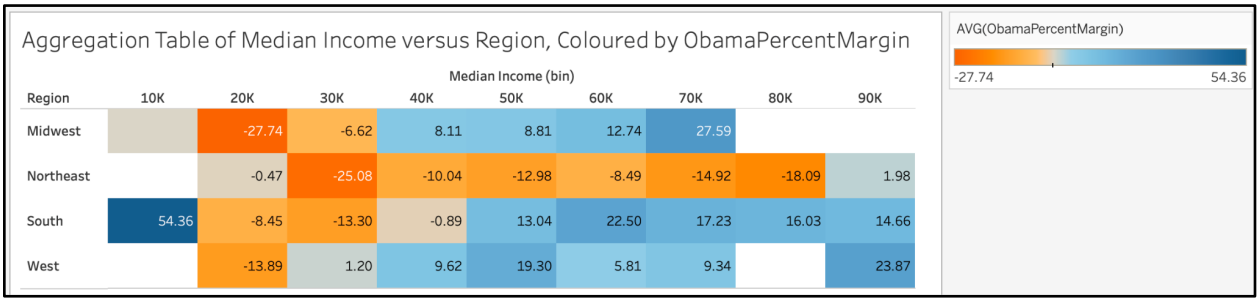

Figure 7: Aggregation of MedianIncome with Region

Figure 7 depicts ObamaPercentMargin generally increasing with MedianIncome. However, the Northeast region presents an opposite trend as average ObamaPercentMargin for most MedianIncomes are consistently negative. This means that for the Northeast, MedianIncome has little effect on voting patterns and Obama performs much worse in the Northeast.

3 Data Preparation

Data Preparation

Our target attribute **ObamaPercentMargin** was calculated like so and added:

ObamaPercentMargin = [100*(Obama - Clinton) / (Total Vote)]

```
# deriving target attribute
elect.df$ObamaPercentMargin <-
  100*(elect.df$Obama - elect.df$Clinton) / elect.df$TotalVote
```

It varies between -100 and 100 with positive and negative values indicating an Obama lead and Clinton lead, respectively. Values close to zero indicate almost even splits of votes.

Plot 1 showed large variation in ObamaPercentMargin within state lines and regions. Plot 4 showed no trend between ElectionDate and ObamaPercentMargin. Thus, State, Region and ElectionDate are not included as predictor variables. TotalVote, Clinton and Obama were already incorporated into ObamaPercentMargin. However, from Plot 2, ObamaPercentMargin seems to vary with ElectionType. With these reasons, we only included demographical data and ElectionType as predictors.

There were many NA values in the dataset. We imputed AverageIncome values with MedianIncome where possible since there was a reasonably strong positive correlation of 0.74. Other NA values were imputed with the nationwide average values.

The data with known votes was randomly split 75/25 to give training and testing datasets.

4 Generate an Evaluate Prediction Models

To predict ObamaPercentMargin, we created three models: a linear regression model, regression tree and random forest.

Advantages and Disadvantages of Models

Linear regressions are easy to implement with fast training while producing measures of significance of each variable. However, it assumes linear relation between the predictor and target attributes, normal distribution of residual errors and non-collinearity between predictor variables.

However, regression trees do not assume linearity and distribution of errors. It efficiently handles collinearity and produces clear visual explanations of predictions, thus more suitable for categorical covariates like ElectionType. Nonetheless, decision trees are prone to overfitting since trees are built to achieve high purity and require pruning.

Random forests are less prone to overfitting since they combine regression/decision trees, producing more generalized and accurate predictions.

Model Improvements

We started with a basic linear model that includes all attributes. From this, we created two more linear models using forward and backward stepwise selection which select attributes until there is no change in AIC. These, however, both resulted in higher MAE and RMSE than the initial linear model.

As linearity is not assumable for ObamaPercentMargin, we created regression trees with a small complexity parameter (cp) to get a large number of different splits. To reduce overfitting, we prune the tree by the cp value that minimizes mean cross-validation error for predictions.

However, as the regression tree resulted in a rather high error, we used random forest which is an ensemble method of regression trees to produce more generalized predictions.

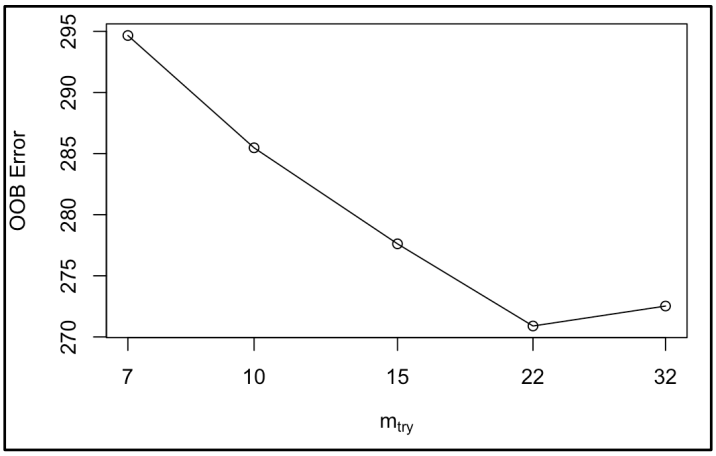

Figure 8: OOB Errors

For this, we used the default number of trees ntree=500 and found the optimal mtry that minimizes Out-of-Bag error to be 22 (see Figure 8).

Best Model & Insights

```
> errors[order(errors$RMSE),]
    MAE  RMSE              Model
6 13.06 16.82                 rf
3 14.11 18.30  lm.step.forward
1 14.20 18.37           lm.all
2 14.22 18.39 lm.step.backward
5 16.17 21.27           rt.all
4 16.68 21.53       rt.all.min
```

Figure 9: MAE and RMSE errors

The random forest model resulted in the least MAE and RMSE out-of-sample errors (see Figure 9).

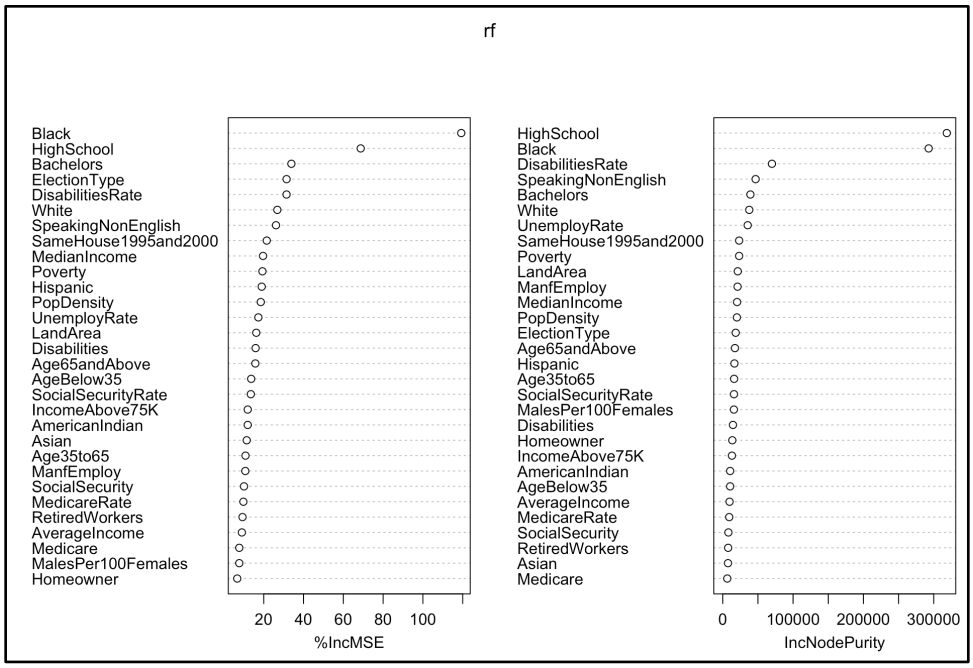

Figure 10: Feature Importance by MSE and Node Purity

The two crucial factors included Black and HighSchool with the most significant effect on both Gini purity and Mean-Square-Error (see Figure 10). These variables have more than 70% impact on MSE if removed from the model, showing their importance in predicting ObamaPercentMargin.

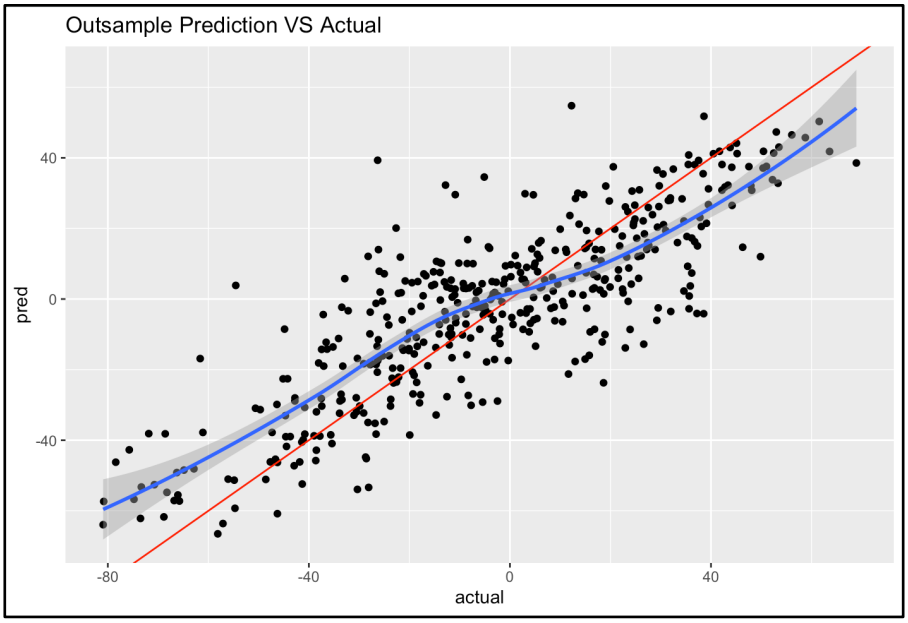

Figure 11: Out-of-sample Prediction against Actual with trendline of points

From the plot of out-of-sample predictions against actuall, we see less positive and less negative predictions compared to actual values (see Figure 11). As such, we can expect the distribution for actual ObamaPercentMargin to be more spread out than the predictions.

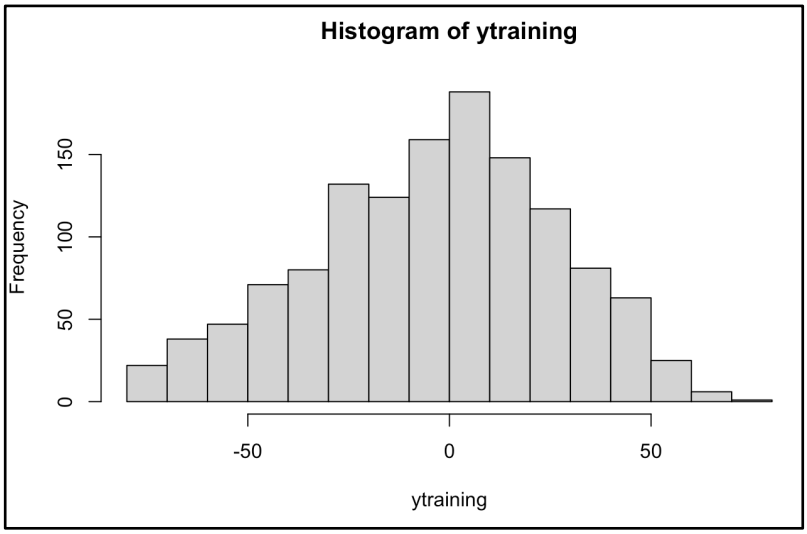

Figure 12: Distribution of ObamaPercentMargin in training dataset

The higher number of inaccurate predictions at extreme ObamaPercentMargin values could be due to a lack of similar data points in the training dataset. This is evident in the distribution of ObamaPercentMargin in the training dataset few at extreme values (see Figure 12).

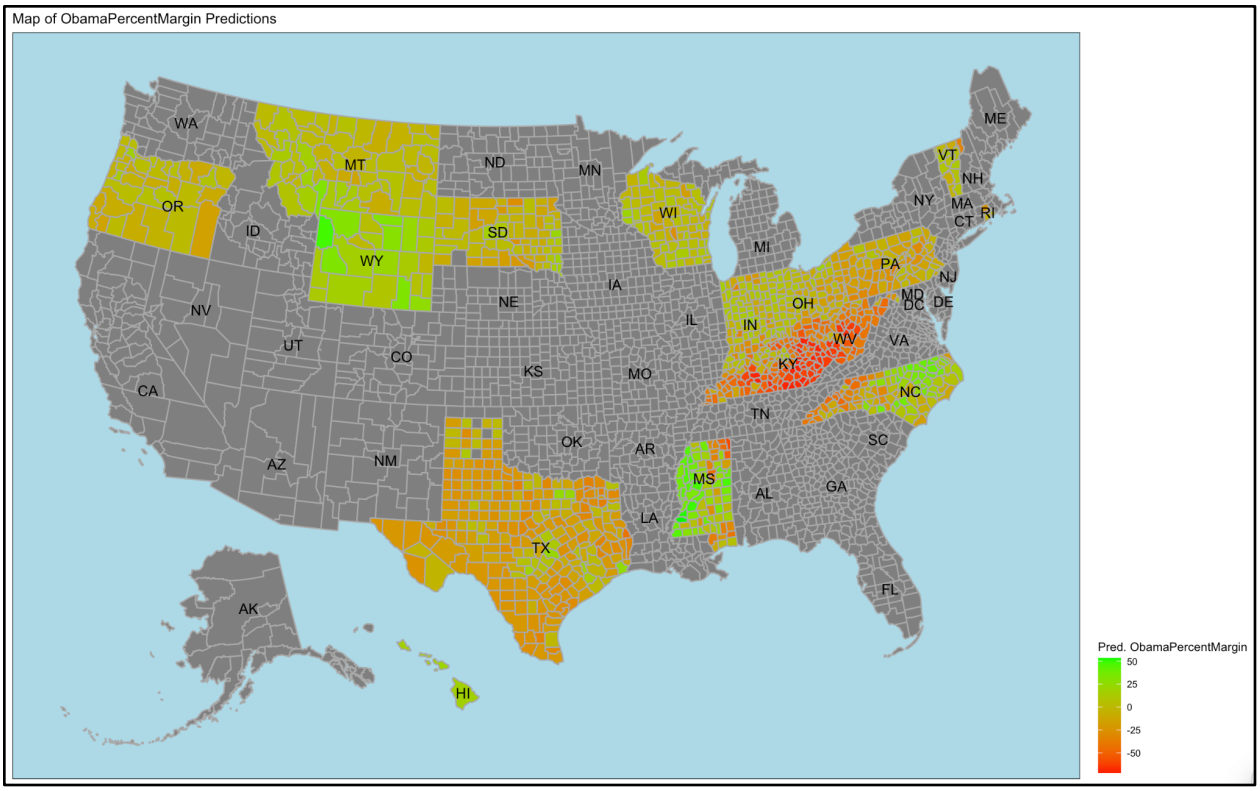

Figure 13: ObamaPercentMargin Predictions

From the predictions, we see that states such as WV and KY are strongly leaning towards Clinton while others such as WY and MS are strongly towards Obama.

5 Conclusion and Recommendations

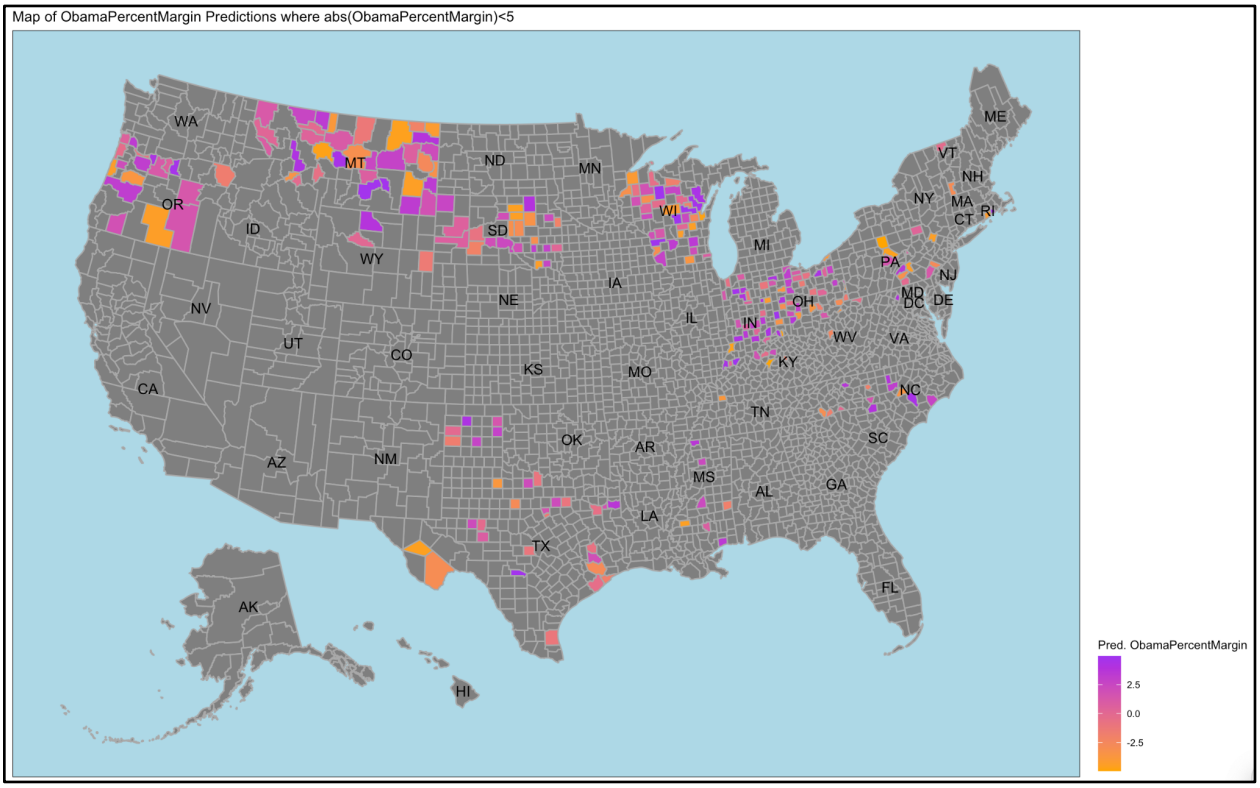

Figure 14: PredictedObamaPercentMargin where abs(ObamaPercentMargin)<5

The data exploration results supported the findings of the final model. The lower feature importance of Bachelors compared to HighSchool confirms that Bachelors may not be a significant predictor (see Figure 6). Our predictions for Northeastern counties are consistently leaning towards voting Clinton, matching findings in Figure 7. Feature importance of the model reflected the high positive correlation of HighSchool and Black with ObamaPercentMargin, combined, we can infer that high HighSchool and high Black leads to a higher ObamaPercentMargin.

To solve the main problem, we identified the counties for Obama's campaign to focus on by looking at counties with -5<PredictedObamaPercentMargin<5. As such, the recommendation would be to focus campaign efforts in OR, MT, WI, SD and OH (Figure 14) as voters could be more easily swayed.