B565 DATA MINING
FINAL PROJECT REPORT

US PRESIDENTIAL ELECTION PREDICTION SYSTEM

Kushal Sheth (kmsheth@iu.edu)
David Ebenezer (daebenez@indiana.edu)
Vaishali Sainaath (vsainaat@umail.iu.edu)

OBJECTIVES AND SIGNIFICANCE

The goal of this project is to build a prediction system that performs various predictive and classification analysis on the presidential primaries dataset and investigate the accuracy of hypothesis we put forward and certain predictions we make.

The United States presidential election is scheduled to be held on November 9 2016. A series of presidential primary elections and caucasus will occur before the election whose data we have used for training and testing our system. President Obama's re election campaign in 2012 took data science in politics to new heights. Big data was widely credited as a factor in his win. This motivated us to use and apply data mining techniques we learnt in class to build a prediction system that predicts insights in the ongoing presidential election primaries.

Our system performs a prediction, a correlation and a classification task.

1. **Prediction Task** : Based on the counties of the 4 US States that voted and for which results were available, our system predicts who will win the counties in the rest of States for which results were not available. We use k nearest neighbours algorithm with Euclidean distance measure.
2. **Correlation Task** : Based on the counties of the 4 US States that voted and for which results were available, our system finds which candidates in the Republican party are most correlated with the 2 candidates in the democratic party based on the demographic details of counties that each candidate wins. We use Candidate-County matrix with Euclidean distance measure.

3. **Classification Task** : Based on the counties of the 4 US States that voted and for which results were available, our system investigates the impact of specific demographic attributes of counties like population growth, diversity and education on the probability of either Hillary Clinton or Bernie Sanders winning. We use Naive Bayesian Classification algorithm.

BACKGROUND

Political parties for years have been collecting data on their constituents and registered voters, what has changed recently however is how the parties have started deploying and analysing this data to swing elections in their favours. Data analytics apart from being used to predict overall outcomes is being used to make important decisions in every step of the campaign process. Insights gained from data analytics allows candidates to make decisions in real time based on hard data rather than speculations.

Building a predictive system for insights on voter sentiment is not new. During President Obama's rerun for election in 2012, he assembled a 100 person strong analytics team that analysed terabytes of campaign data that integrated data from wide range of sources like opinion polls, surveys and volunteers armed with survey questions who went door to door collecting data. The insights gained from this analysis was used for streamlining political outreach through more targeted advertisements, social media interactions and celebrity endorsements that helped win over swing voters.

A lot of the work in this field has previously focussed on analysis of voter sentiments through their twitter posts. The paper cited here for example focusses on real time analysis of text in twitter posts, calculating their sentiment and correlating them with political events like debates and gauging their effect on voter sentiment.

While papers like these provide real time data on political events, we felt they did not delve deep enough into voter demographics and include their effects on voting patterns. Our project focusses on analysing mainly the demographic data of voters, after the primary elections and attempting to predict their sentiment which will then be used to extrapolate the sentiment of counties in States that have not voted in the primaries yet. Our dataset is fairly recent, It was

updated 3 days before the time of writing the proposal and we do not believe there has been any significant work done on it that is publicly available.

Concepts

Presidential primary elections and Caucuses : Elections that are held in each US state and are part of the nomination process of the presidential elections . For this election cycle this will occur between Feb 1 and Jan 2016. Our dataset has results from the States of Iowa, Nevada, South Carolina and New Hampshire and is recent.

The Republican Candidates represented in the dataset are Ted Cruz, Donald Trump, Marco Rubio, Ben Carson, Cary Fiorina, Rand Paul, Mike Huckabee, Rick Santorum, Jeb Bush, Chris Christie, John Kasich. The Democratic Candidates represented in the dataset are Hillary Clinton, Bernie Sanders and Martin O'Malley.

Previous Work on the Dataset and Our Present Work

The dataset is fairly new and is based on a recent event, hence there have been few analyses performed on them.

Our project investigates in depth the data available to predict which candidate is most correlated with Hillary Clinton and Bernie Sanders. Also, we took a training set of 172 records that had details of counties and candidates who won those counties. And based on the training set, we predicted which candidate has much better chances at winning at each of county for test data of 2970 Records. Also, naive bayes classification is done on the dataset to find which candidate has the highest probability (Posterior Probability) of winning the election based on certain features such as population growth/shrinkage in different counties, gender, percentage of people of age less than 18, race and percentage of people with a bachelor’s degree.

METHODS

DATASET: 2016 US Election

We are using the 2016 US Election dataset from Kaggle. The dataset was uploaded by Ben Hamner who is a cofounder and CTO of Kaggle. The dataset is available in both the csv and SQLite format. For this project we plan on working primarily with the CSV version of of the dataset. The data set that we are working on was downloaded on 02/27/2016. After that the data has been updated on the kaggle. There are 2 important files in the dataset

County_facts:

This file has data on every county in each of the 4 states where the primaries or elections took place. This file has important demographic information which will be our features. Some of the features that we propose to use are age groups for example 'population under 50', race, gender %, percentage of voters who are foreign born, percentage of population that is made up of veterans etc. We will be using the features together to investigate how much of an effect these features have on the election results of each candidate's. The entire dataset had lots of null values and irrelevant data.

The 'county_fact' dataset had data of all the counties of the USA. It had around 51 attributes. We pre-processed the 'county_fact' dataset and removed those attributes which had almost null values and were not relevant with task to be done. So this was our dimensionality reduction task, in which after the discussion with teammates, we removed the irrelevant and repetitive attributes. As a result we were left with 26 attributes. we also aggregated many attributes in one and formed new attributes. This attributes are named "New Feature"in the dataset. This new features were 'Absolute Change' which is a binary feature with 0 for population shrink and 1 for population growth. Next is 'Between 18-65' This attribute shows the population that is in between age 18 and 65. The last feature is 'Homeownership + Median Income' which is a multi-class attribute with values 1,2,3 suggesting the low, medium and high.

primary_results:

This file has county wise information of the number of votes won by each candidate. Each of the 4 state is divided into a number of counties and the number of votes won by each candidate in each county is given along with the fraction of total votes. This dataset had around 8 attributes. There are approximately 13000 such records in this file. The dataset has 172 unique/different counties and for each county it has the details regarding the candidates participating, the votes they get etc. For different type of analysis, this dataset is modified differently.

METHODOLOGIES

The 3 analysis that we are performing are as follows.

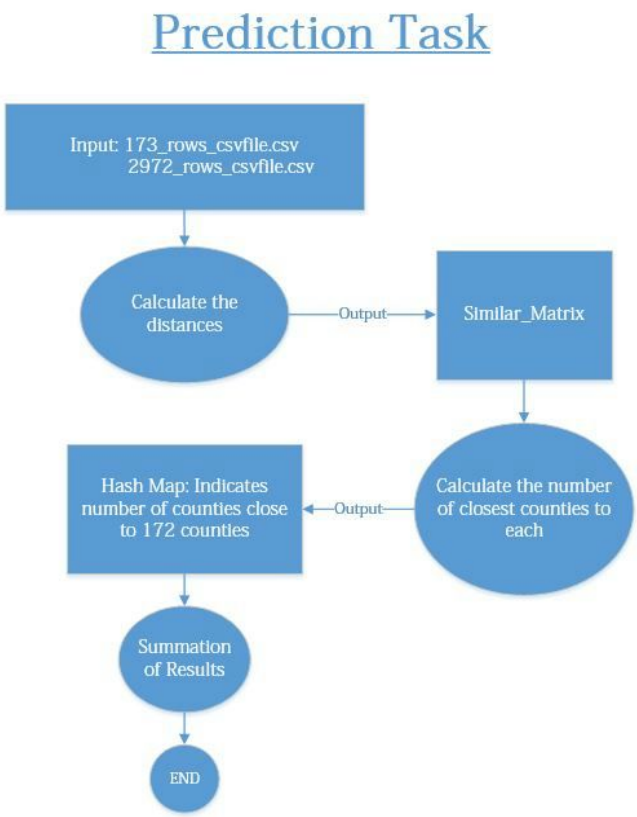
Analysis 1: Predicting the winner

The first important module of our project was making a prediction model that can predict which candidate will win in how many counties. This have to be a multiclass classification because we had the winning candidates as the class. For this prediction, we used Primary Results dataset. This data set has a result of 172 counties. Which candidate won, how many votes they got and which party they belonged to. Firstly we filtered the data of this 172 counties from county_facts dataset and two new datasets were built. One dataset had the county facts (attribute values) for the 172 counties, and the other dataset had the county fact (attribute values) for other remaining counties. This are complete numeric data for the ease of understanding. All the dataset is sorted on basis of the "fips" (the id of the county). This data are in following excel files:

173_rows_csvfile.csv (Data of counties whose results we have)

2972_rows_csvfile.csv (Data of counties whose results we don't have)

The flow of this module is as follows:



Flowchart: Prediction Module

The next part was to find the similarity of each county of one dataset with each county of other dataset. So basically a 2970 X 172 matrix will be formed which will have the values of distances. Here we have used the Euclidean distance to calculate the distances. The code for this is in 'calculate_similarity_counties.py' file. On running this code, We get a 2970 X 172 matrix which contains the similarity measures between all the counties.

	1	2	…	…	172
1					
2					
…					
…					

2970					

Table: Similar Matrix

The next task was to count that how many counties are closer to particular counties, whose results we already have. So we wanted the number of counties which are similar to the 172 counties. For this we made a hash map with 172 keys initialized with 0 and parsed the entire data. Whenever we get the closest county, I would increment the value in hashmap. For example, as shown in above table, we will go through each row from 1 to 2970 and will check that which is the most closest county. So here we are using the k-nearest neighbours approach where k = 1. For example, If the index number 170 is closest to the 1st row, then we'll increment the value of index 170 in the hashmap. This way we will get the values for each of the 172 indexes in hashmap. The code for this is in 'calculate_closed_neighbours.py'. We will store this values in the file Filtered_Set_Result.xlsx.

As all the data is already sorted from the beginning of the process, the values of hashmap will be in the perfect order of the counties in the dataset. Now as we got this results, next task is quite simple. After adding the 'Prediction' column in dataset, we sorted the dataset on 'candidate' attribute which is a multi-class attribute and calculated the sum of the predictions for each candidate. This way we got the prediction of total number of counties, where a particular candidate has chances to win. The results are in candidate_winner_prediction.xlsx. Along with this we also performed the binary classification on basis of the parties. So for this again, we sorted the Filtered_Set_Result dataset on 'party' attribute and again calculated the sum of the prediction for each party. This gives us the result that which party will win in more counties. The results are in party_winner_prediction.xlsx

Analysis 2: Correlation of Presidential Candidates

The second task that we will be performing is to find the candidates in the Republican party who are most closely correlated with the two candidates in the democratic party. We first take the county_facts.csv data file which has all the demographic information on all

counties in all states and use the Euclidean distance formula to find the closest k counties for every county.

The data is output in the CountyRank.csv file

County 1	Closest County A	Distance
County 2	Closest County B	Distance
County 3	Closest County C	Distance
County 4	Closest County D	Distance

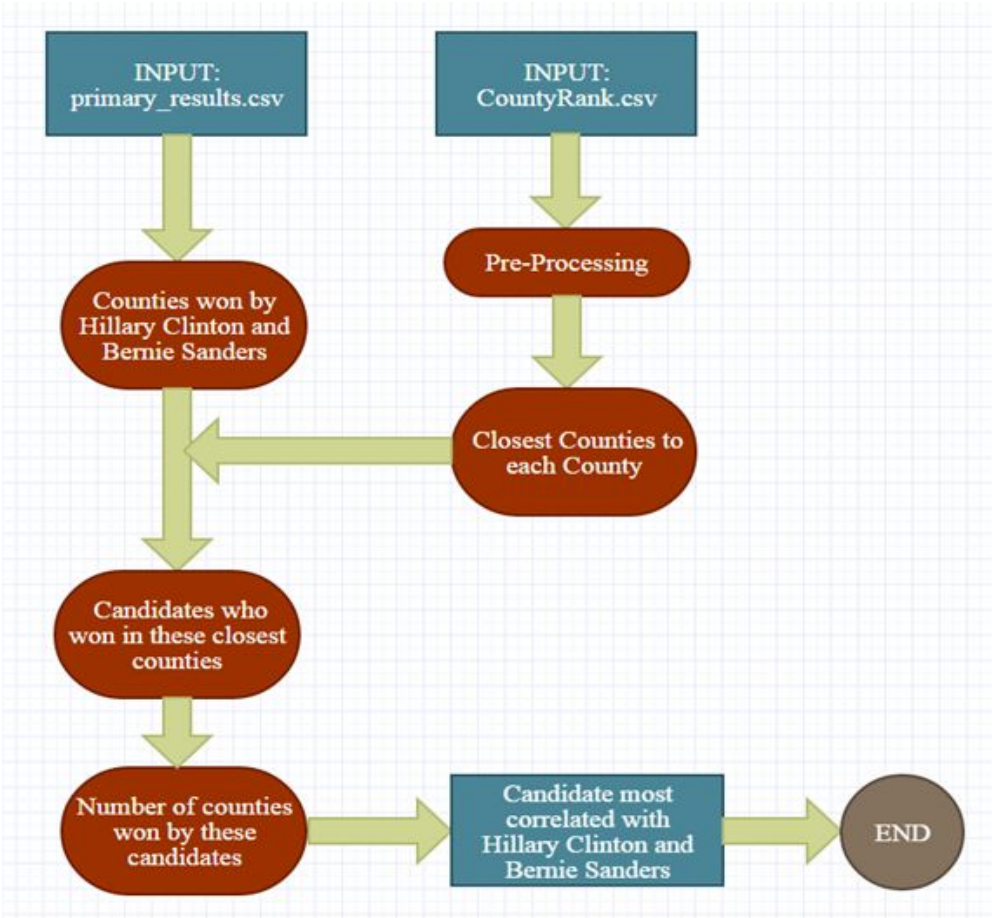
Table: ClosestCounty.csv format

Firstly, we had to preprocess the dataset. Each county is arranged alphabetically in one column and its corresponding four closest counties are recorded adjacent to it in the second column. This makes it easier to traverse through the dataset with just one column and retrieve the closest counties. This considerably reduces the time taken to retrieve data since the dataset consists of about 12000 records, and traversing through each and every column to obtain the desired record would increase the time complexity of the program.

Candidates with whom correlation is being found	Closest County 1	Closest County 2	Closest County 3	Closest County 4
Hillary Clinton	Closest Candidates from this county to Hillary Clinton	Closest Candidates from this county to Hillary Clinton	Closest Candidates from this county to Hillary Clinton	Closest Candidates from this county to Hillary Clinton

Bernie Sanders	Closest Candidates from this county to Bernie Sanders	Closest Candidates from this county to Bernie Sanders	Closest Candidates from this county to Bernie Sanders	Closest Candidates from this county to Bernie Sanders

Table: Candidate Matrix



Flowchart: Correlation Module

- First, for each county, four nearest counties were found based on the Euclidean distance between the features of the counties.
- With the help of "primary_results" dataset that consists of 13213 records, the counties that are won by "Hillary Clinton" and "Bernie Sanders" are found.

- Integrating this result with the previous result (Closest four counties), the closest counties to the counties won by these two candidates is found.
- Now, for all the closest counties to counties won by “Hillary Clinton” and “Bernie Sanders”, the candidates who have won in these counties is found.
- The number of counties won by each of these counties is recorded.
- The candidate who won closest number of counties with “Hillary Clinton” is found to be most correlated with her. And, the candidate who won closest number of counties with “Bernie Sanders” is found to be most correlated with her.
- With this information, the candidate who is next closest to “Hillary Clinton” and “Bernie Sanders” is also found.

Analysis 3 : Investigating the Impact of specific demographic features on the probability of either Hillary Clinton or Bernie Sanders winning the election.

It has generally been observed that democratic presidential candidates perform well in more urban areas where population is high, the percentage of youth population is higher than average, gender and racial diversity is high and there is a greater percentage of population that has a college degree. We investigate this hypothesis to see how well the 2 democratic candidates Hillary Clinton and Bernie Sanders perform relative to each other in counties where these demographic features are more pronounced. We specifically investigate 5 demographic features

1. Did the population increase or shrink between 2010 and 2014.
2. Is there a high percentage of young people living in that area.
3. Is a high percentage of the population female.
4. Is the percentage of racial minorities in that area above average.
5. Does a significant percentage of population have a college degree

Data Preprocessing

In this step we created an intermediary file called NBInput.csv. The Java file DataConversion.java takes as input primary_reults.csv and county_facts.csv and extracts all the 5 features from the 2 files as well as the County name and who won the specific county to create the intermediary file.

The program reads all the items into an ArrayList of ArrayList, extract all entries that are states by using their pincode (states pincode end in 000). Write all of the relevant features and their results into NBInput.csv.

The Intermediary file is in the following format.

FIPS	County Name	Population %	% Under 18	% Female	% whites	% College grad
10001	Adair County	-3	29	51	70	20

Table : Intermediary File format

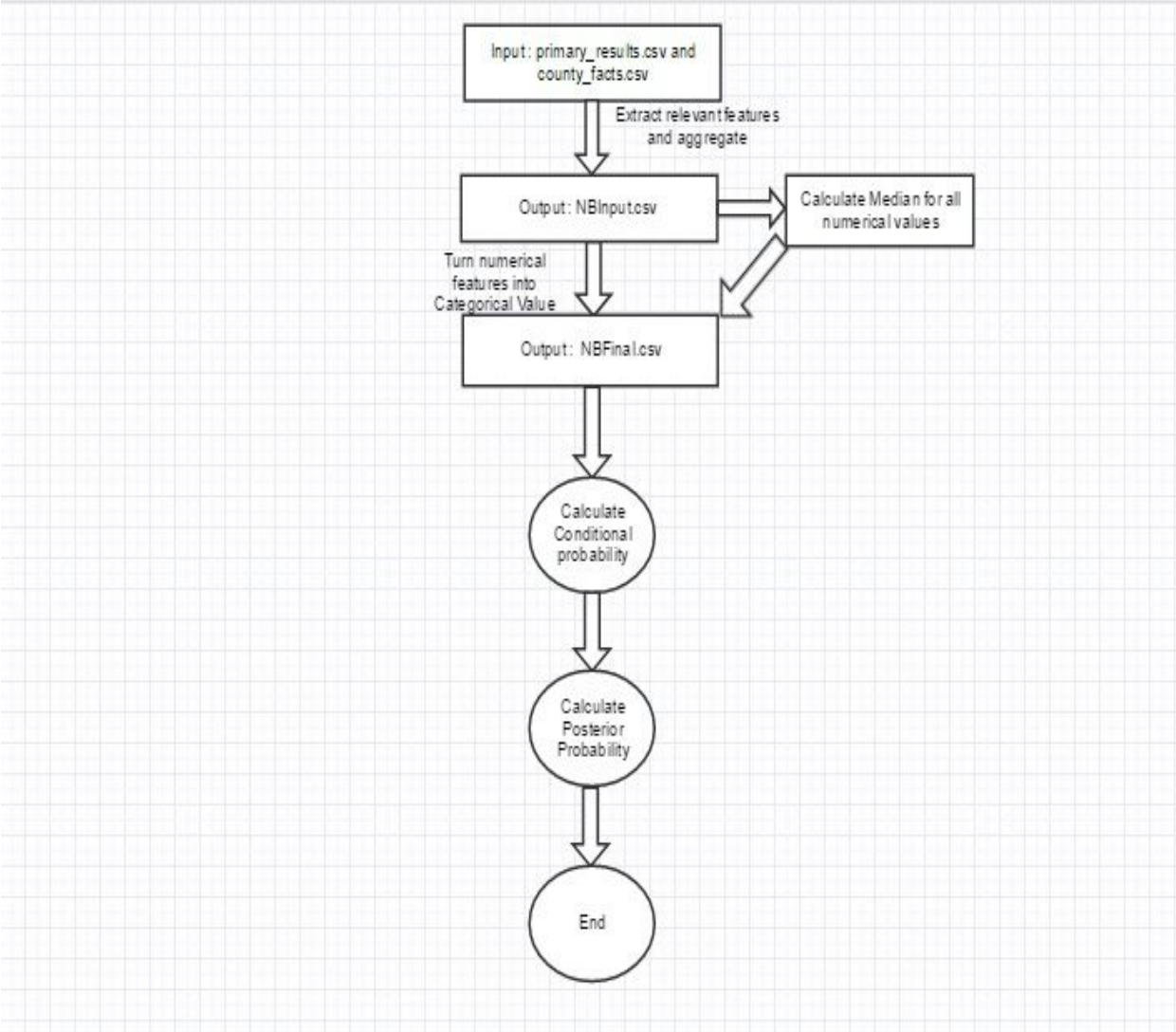
Further we convert the following intermediary file into its final format by

1. Convert population data into a binary class attribute of either increasing or decreasing based on whether they grew or shrunk between 2010 and 2014
2. Calculate the median for other numerical attributes using a TreeSet and classify them into low or high based on whether their value is higher or lower.

The final final NBFinal.csv is in the following format

FIPS	County Name	Population %	% Under 18	% Female	% whites	% College grad
10001	Adair County	low	high	low	high	low

Table : Final table format

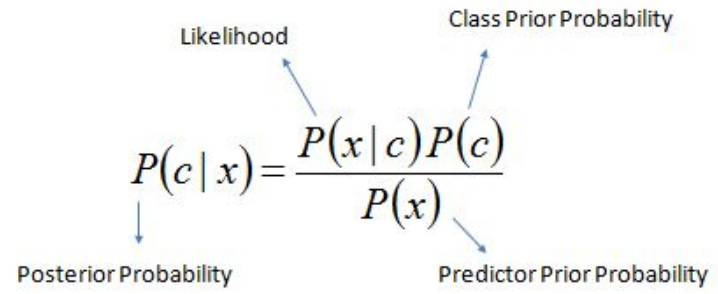


Flowchart: Naive Bayes Classification

Naive Bayesian Classification

Once all of the necessary features are aggregated and converted to class attributes, We perform Naive Bayesian Classification to find the posterior probability of Hillary or Bernie winning given an increasing population, high percentage of young people, high percentage of female, high percentage of racial minorities and high percentage of people with college degree.

We first calculate the probabilities of Hillary Clinton and Bernie Sanders winning in the counties of the 4 states. Next we calculate conditional probability. We use the naive bayesian formula to then calculate the posterior probability.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Naive Bayes formula

Evaluation strategy

For the analysis of the winner, we divided our primary results data set in two part. The first set contains the data of three states and other set contains the data of remaining state. We passed this test and training data to the same codes and got the results. As of which candidate wins. The results were not accurate but they were quite close to the original results.

RESULTS

Analysis 1: Predicting the winner

As the final output, we get the number of counties in which the candidates have the chances to win. We have results for five candidates, namely Hillary Clinton, Bernie Sanders, Donald Trump, Marco Rubio and Ted Cruz. The following graph shows the numbers of predicted counties for each candidate.

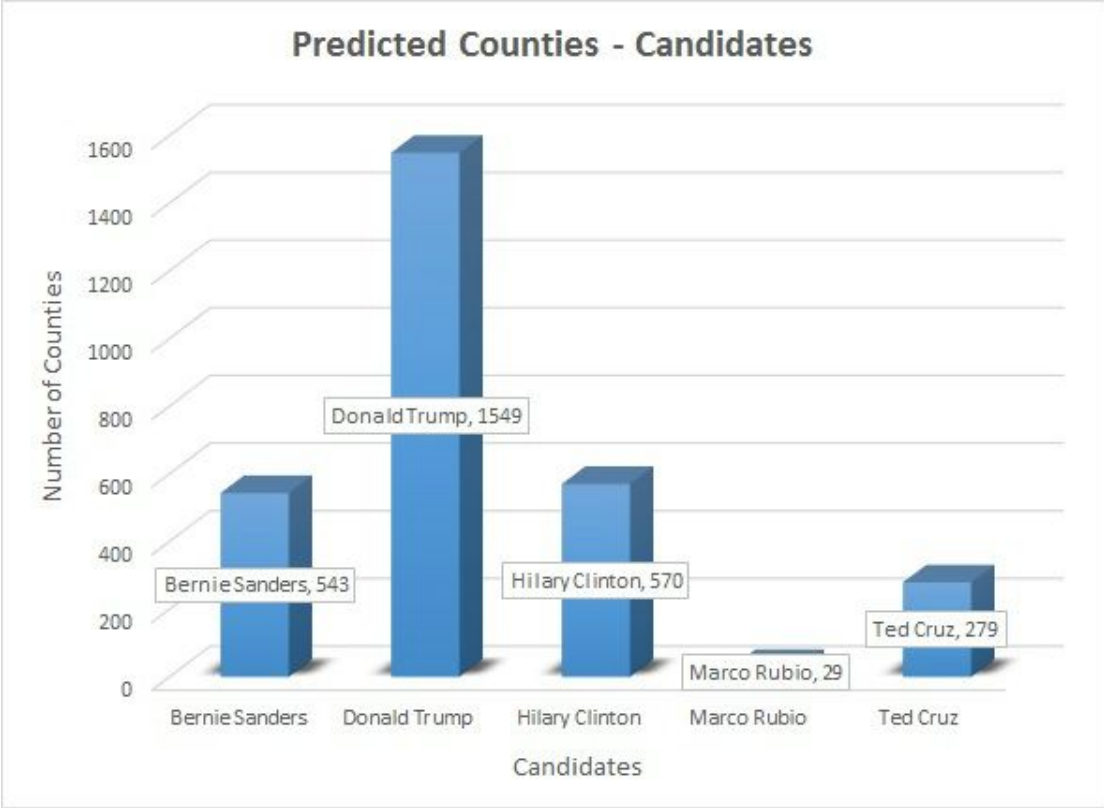


Chart: Candidates vs Number of Counties

The above graph shows that Bernie Sanders may win in more 543 counties and Donald Trump may win in more 1549 counties and so on. This results are part of the multi class classification. From this results we can say that Donald Trump seems to win the Presidential Elections. We have results for two parties, namely Democrats and Republicans. The following graph shows the numbers of predicted counties for each party.

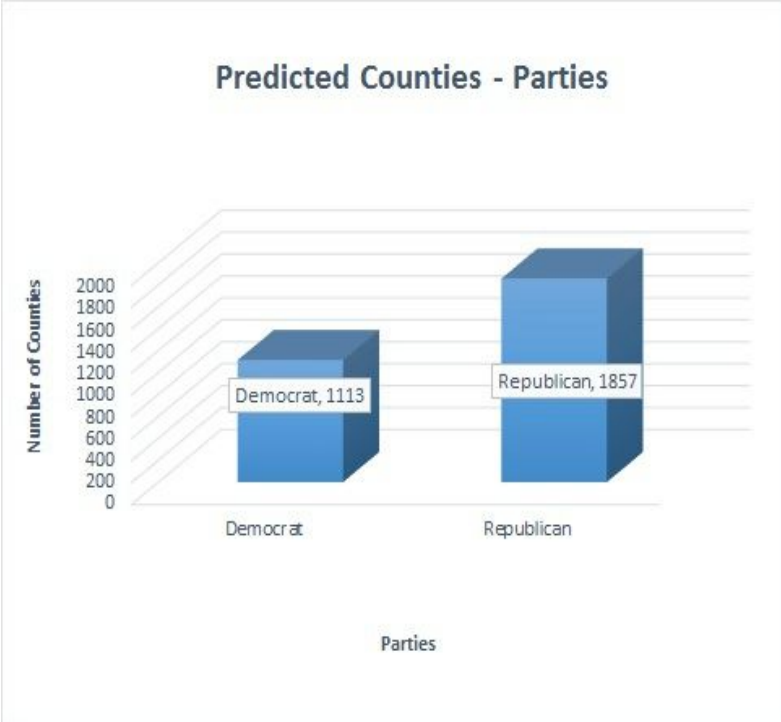


Chart: Parties vs Number of Counties

The above graph shows the Democrats may win in 1113 counties and Republicans may win in 1857 counties that is around 62% of the counties. This result are part of binary classification. This also matches the previous result that Donald Trump may win. This implies that on basis of our result, the Republicans will have more counties under their belt and also the Republican candidate may win the elections.

Analysis 2: Correlation between Presidential Candidates

As mentioned, the counties closest to counties won by Hillary Clinton and Bernie Sanders is found. The main goal is to find which candidates are most correlated with "Hillary Clinton", and which ones are most correlated with "Bernie Sanders". The different candidates close to Hillary clinton and Bernie Sanders and the number of counties each have participated is obtained as follows:

Candidate Close to Hillary Clinton and Bernie Sanders	Number of Counties Each Candidates participated In

Rand Paul	85
Rick Santorum	85
Martin O'Malley	85
Chris Christie	92
Jeb Bush	126
Hillary Clinton	949
John Kasich	927
Donald Trump	926
Mike Huckabee	85
Marco Rubio	908
Bernie Sanders	903
Carly Fiorina	92
Ben Carson	517
Ted Cruz	871

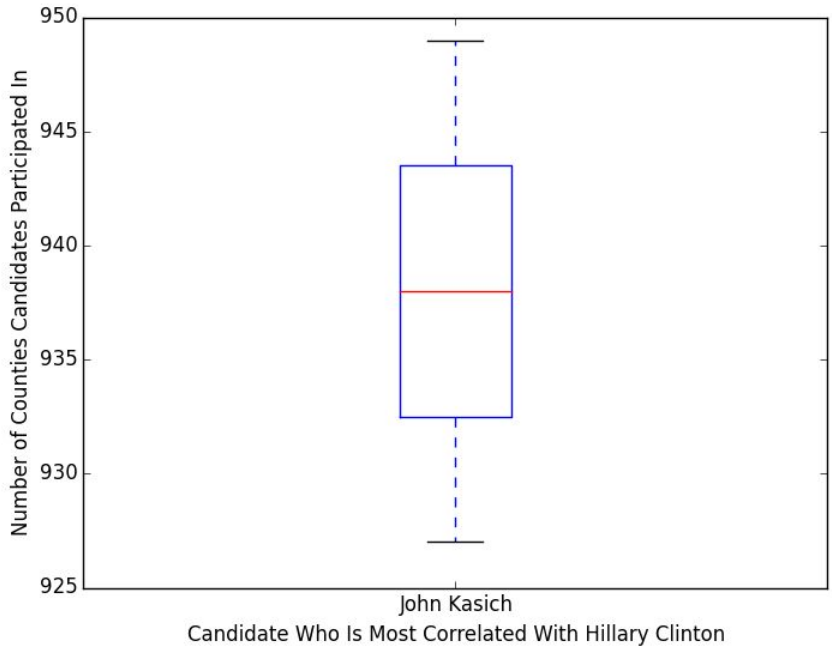
Table: Correlation

From the above results, the final result obtained is ass follows:

Candidates Being Compared With	**Most Correlated Candidate**	**Second Most Correlated Candidate**
Hillary Clinton	John Kasich	Donald Trump
Bernie Sanders	Marco Rubio	Donald Trump

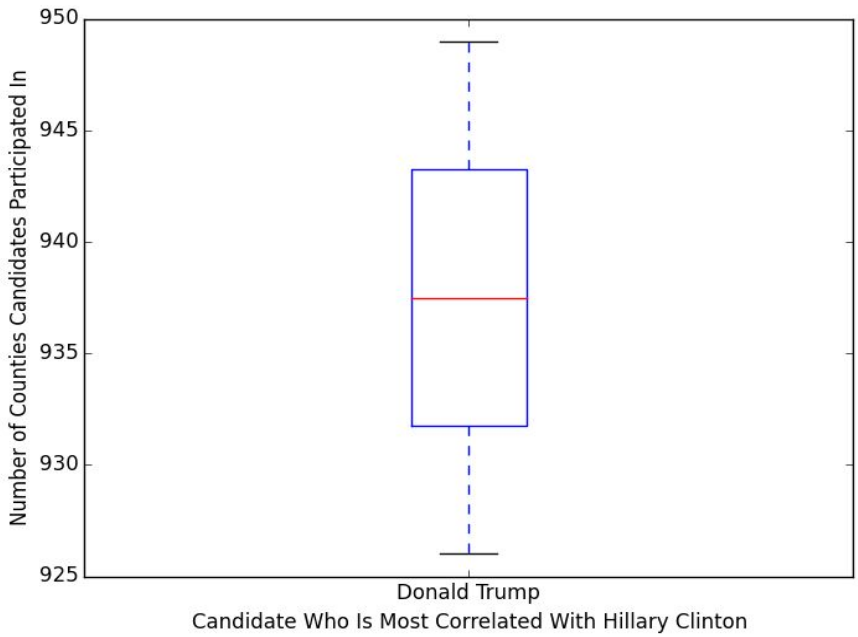
Table: Most Correlated Candidates to Hillary Clinton and Bernie Sanders

The below figure shows the BarPlot showing the candidate most correlated with Hillary Clinton and number of Counties participated in both of them.



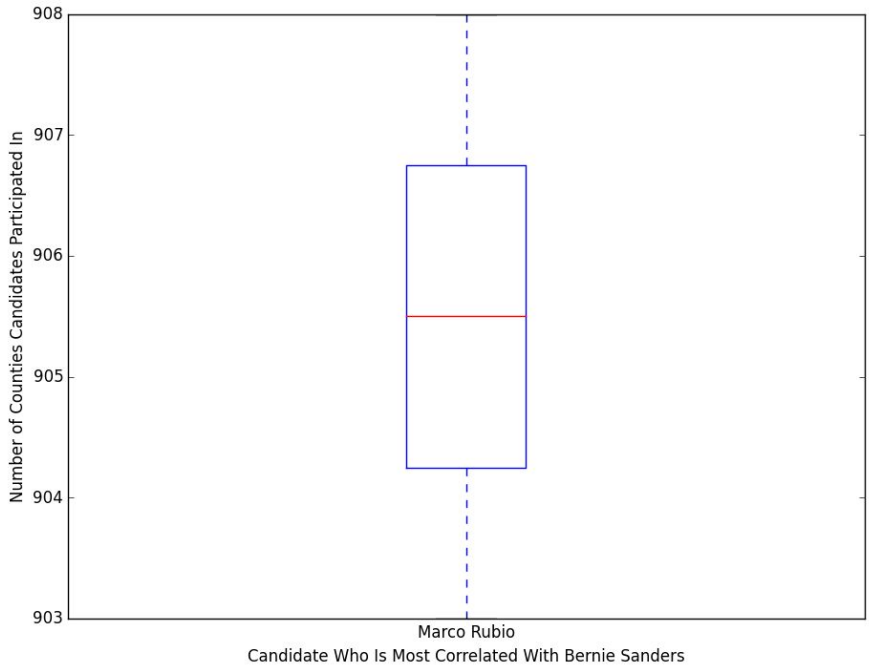
Barplot: Most Correlated Candidate for Hillary Clinton

The below figure shows the BarPlot showing the candidate who is second most correlated with Hillary Clinton and number of Counties participated in both of them.



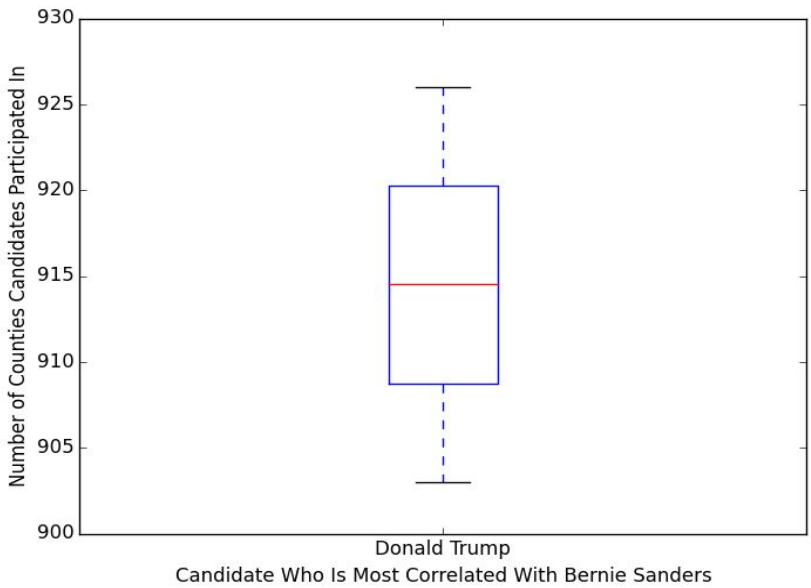
Barplot: Second Most Correlated Candidate for Hillary Clinton

The below figure shows the BarPlot showing the candidate most correlated with Bernie Sanders on and number of Counties participated in both of them.



Barplot: Most Correlated Candidate for Bernie Sanders

The below figure shows the BarPlot showing the candidate who is second most correlated with Bernie Sanders on and number of Counties participated in both of them.



Barplot: Second Most Correlated Candidate for Bernie Sanders

Analysis 3 : Investigating the Impact of specific demographic features on the probability of either Hillary Clinton or Bernie Sanders winning the election.

Calculating the posterior probability for Hillary Clinton given increasing population, high percentage of people under 18, high percentage of young people, high percentage of people of female, high percentage of racial minorities and high percentage of people with college degree we get 0.0011032569274773557

Calculating the posterior probability for Bernie Sanders given increasing population, high percentage of people under 18, high percentage of young people, high percentage of people of female, high percentage of racial minorities and high percentage of people with college degree we get 0.01181114719609321

We see that Bernie Sanders perform slightly better in areas where the 5 features highlighted are higher.

A visualization of the numerical values of various features and candidate results are shown. This data is available in NumericalInput.csv. We see from the visualization that Bernie Sanders performs well in areas with a better percentage of educated college degree holders as well as areas where there has been an increase in population. Meanwhile Hillary Clinton performs well in areas where there are higher % of females, racial minorities and young children. This is consistent with what the media has been claiming about these two candidates.

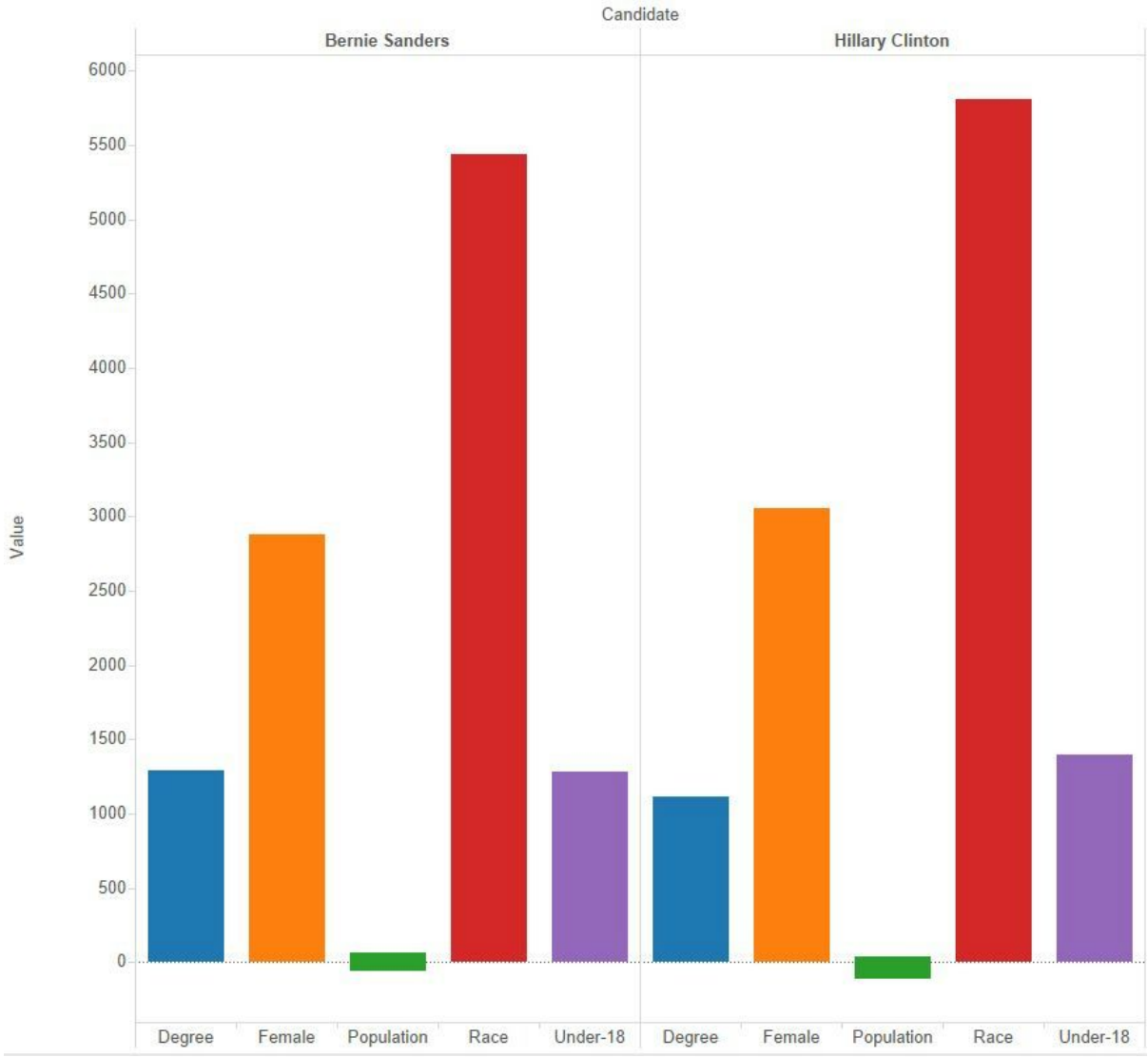


Chart: Hillary Clinton vs Bernie Sanders (Comparison)

CONCLUSIONS

The entire project revolves around the three main analysis modules named Predicting the Winner, Correlation Analysis and Using Naive bayes techniques to analyse how different features contribute towards the chances of the candidate winning or losing. For the prediction module, on seeing the results we conclude that the binary classification and the multi-class classification correlates with each other. We find that the Republican candidate will get the majority and the Republicans will come out as the winner. The approach not being the most accurate one, we can use a better approach and use more data such as the location of counties (Latitudes and Longitudes) for better classification on basis of the location. Also the other

factors apart from demographic data such as campaigning of every candidate plays major role in the election. So the data of this kind could also be considered for accurate prediction. Though not done yet, we plan to implement it and continue with the project as far as possible.

The Naive Bayesian Classification results agrees with what is being reported in the media, that Bernie performs well with highly educated electorate and Hillary performs well with women and racial minorities. A more effective method for converting numerical attributes into class attributes can be followed, probably by using multi class attributes in place of binary class attributes being used currently to more accurately capture the numerical range of the data values.

INDIVIDUAL TASKS

Task 1: Predicting the Winner

Kushal Sheth

- Preprocessed the data and building two separate datasets that will work as an input file to calculate the similarity between counties.
- Produced an 2970 X 172 matrix that contains the similarity values between each counties.
- Produced an output that shows the number number of counties that are similar to the counties of which we have the results.
- Calculated the number of counties in which each candidate have chances to win. This is part of multi-class classification.
- For binary classification, found the results on basis of the parties, that which parties have chances to win in how many counties
- Created visualizations and analysis for candidates and parties.
- All the codes and generated files are in folder ‘KUSHAL’

Task 2: Correlation between the Presidential Candidates

David Ebenezer - Closest County

- Pre-Processing : Using Euclidean distance to find the k closest counties for each county using 29 demographic features.[ClosestCounty.java]

Vaishali Sainaath

- Pre-processing of CountyRanks.csv dataset to arrange the counties alphabetically in the first column, and its corresponding closest counties in the adjacent column.
- Found the counties won by Hillary Clinton and Bernie Sanders.
- Compared this with the CountyRank dataset, and found counties closest to the counties won by Hillary Clinton and Bernie Sanders.
- Found the candidates who won these closest counties.
- Found the candidates who are closest and correlated with Hillary Clinton and Bernie Sanders.
- All codes can be found in the folder 'VAISHALI'

Analysis 3 : Investigating the Impact of specific demographic features on the probability of either Hillary Clinton or Bernie Sanders winning the election.

David Ebenezer

- Data Preprocessing to extract relevant features, aggregate them and create intermediary file.
- Create visualization based of the intermediary file to show which demographic data favors which candidate.
- Data Preprocessing to convert numerical attributes into class attributes as a preparation step for Naive Bayesian Classification.
- Built A Naive Bayesian Classifier to calculate posterior probability for Hillary Clinton and Bernie Sanders. Output is a numerical probability value for each candidate which shows all of the demographic features favouring one candidate over the other.
- All code can be found in the folder 'DAVID'

REFERENCES

1] https://www.kaggle.com/benhamner/2016-us-election

2] http://www.saedsayad.com/naive_bayesian.htm

3] "The Data behind Democracy: Analytics and the 2015 ..." 2015. 29 Feb. 2016 <//www.matillion.com/insight/thedatabehinddemocracyanalyticsandthe2015generalelection/>

4] Wang, Hao et al. "A system for real time twitter sentiment analysis of 2012 us presidential election cycle."

Proceedings of the ACL 2012 System Demonstrations10 Jul. 2012: 115120.

5] "United States presidential primary Wikipedia, the free ..." 2011. 29 Feb. 2016 <https://en.wikipedia.org/wiki/United_States_presidential_primary>