Table of Contents

Table of Contents

US Presidential Election 2016

Abstract:

The presidential election season is upon us. On November 8, 2016, Americans will head to the polls and choose their president. Through the primary election process, political parties generally hold national conventions at which a group of delegates collectively decide upon which candidate they will run for the presidency. The process of choosing delegates to the national convention is undertaken at the state level, which means that there are significant differences from state to state and sometimes year to year.

1.Problem Description:

(a)Analyze the trends and predictions (Building predictive models using various Machine-Leaning algorithms) of Democratic and Republican Primaries Results.

(b) Social Network (Twitter and Facebook) Analysis of 2016 US Presidential Election Candidates

2.Data Description:

The data is provided in 2 tables- County_facts and Results. Additionally, 2014 cartographic boundary county shapefiles are provided as well which are simplified representations of selected geographic areas from the U.S. Census Bureau's Master Address File / Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) Database (MTDB). These boundary files are specifically designed for small-scale thematic mapping.

2.1 County_facts.csv

This file contains demographic and socioeconomic data for states, counties and cities in USA. It provides statistics for all states and counties, and for cities and towns with a population of 5,000 or more.

PST045214	Population, 2014 estimate
PST040210	Population, 2010 (April 1) estimates base
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	Non-employer 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

2.2 Primary_results.csv

Within this file, you will find fraction of (total party) votes garnered by respective contesting candidates.

We have built predictive models for four candidates namely: (1) Bernie Sanders (2)Hillary Clinton (c)Ted Cruz (d)Donald Trump

3.Preprocessing

3.1 Removing the incomplete/ incorrect data for the states

```
results = results[(results.state != "Maine") & (results.state != "Massachusetts") & (results.state
!= "Vermont") & (results.state != "Illinois") ]
results = results[(results.candidate != ' Uncommitted') & (results.candidate != 'No Preference')]
```

3.2 Giving meaningful names to the variables:

```
demographics = demographics[['fips','area_name','state_abbreviation','PST045214','AGE775214','RHI22
5214','RHI725214','RHI825214','EDU635213','EDU685213','INC110213','PVY020213','POP060210']]
demographics.rename(columns={'PST045214': 'Population', 'AGE775214': 'Age > 65','RHI225214':'Black'
HI725214':'Latino','RHI825214':'White','EDU635213':'HighSchool','EDU685213':'Bachelors','INC110213'
edian Household','PVY020213':'< Powerty level','POP060210':'Population PSM'}, inplace=True)
```

The purpose to rename the variables is for better understanding.

3.3 Calculating state wise total votes and fraction votes:

This is done so that in addition to predicting a candidate's percentage of votes county wise, we have provision to predict state wise as well.

```
#Calculating statewise total votes and fraction votes
votesByState = [[candidate, state, party] for candidate in Dem.candidate.unique() for state in
Dem.state.unique() for party in Dem.party.unique()]
for i in votesByState:
        i.append(Dem[(Dem.candidate == i[0]) & (Dem.state == i[1])].votes.sum())
        i.append(i[3]*1.0/Dem[Dem.state == i[1]].votes.sum())
vbs = pd.DataFrame(votesByState, columns = ['candidate', 'state', 'party', 'votes','partyFrac'])
print(vbs)
```

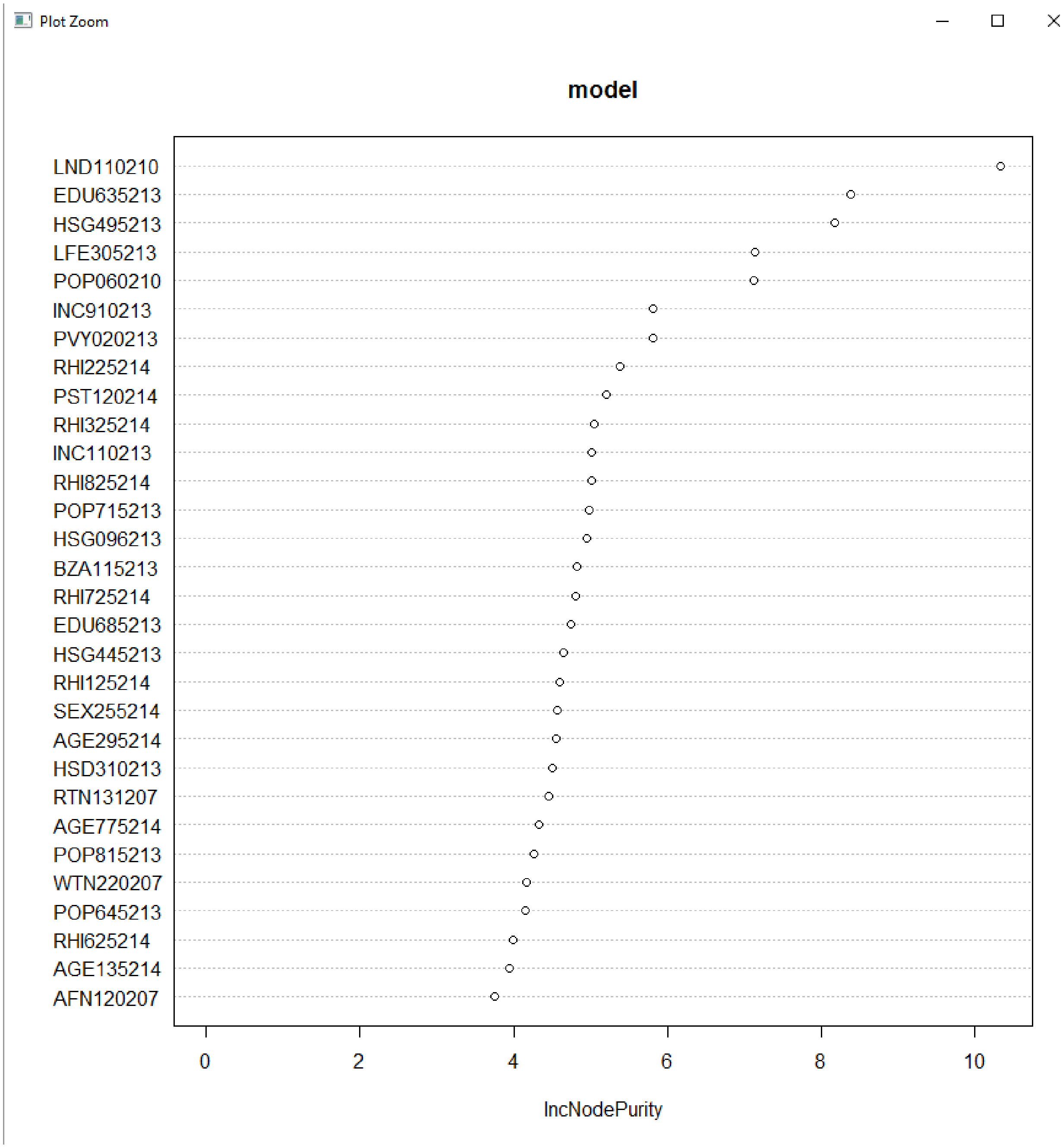
3.4 Merge

The county_facts.csv and results.csv are merged together as a single dataset on which the prediction models were run

state_abbr	fips	state	county	party	candidate	votes	fraction_v	area_name	Population	PST04021	PST12021	POP01021	AGE13521	AGE29521	Age > 65	SEX25521	RHI12521	Black	RHI32521	RHI42521	RHI52521	RHI62521	Latino	White	POP71521	POP64521	P
AL	1001	Alabama	Autauga	Republicar	Ted Cruz	2482	0.205	Autauga Co	55395	54571	1.5	54571	6	25.2	13.8	51.4	77.9	18.7	0.5	1.1	0.1	1.8	2.7	75.6	85	1.6	
AL	1001	Alabama	Autauga	Republicar	John Kasicl	421	0.035	Autauga Co	55395	54571	1.5	54571	6	25.2	13.8	51.4	77.9	18.7	0.5	1.1	0.1	1.8	2.7	75.6	85	1.6	
AL	1001	Alabama	Autauga	Republicar	Donald Tru	5387	0.445	Autauga Co	55395	54571	1.5	54571	6	25.2	13.8	51.4	77.9	18.7	0.5	1.1	0.1	1.8	2.7	75.6	85	1.6	
AL	1001	Alabama	Autauga	Republicar	Marco Rub	1785	0.148	Autauga Co	55395	54571	1.5	54571	6	25.2	13.8	51.4	77.9	18.7	0.5	1.1	0.1	1.8	2.7	75.6	85	1.6	
AL	1001	Alabama	Autauga	Republicar	Ben Carsor	1764	0.146	Autauga Co	55395	54571	1.5	54571	6	25.2	13.8	51.4	77.9	18.7	0.5	1.1	0.1	1.8	2.7	75.6	85	1.6	
AL	1001	Alabama	Autauga	Democrat	Hillary Clin	2387	0.8	Autauga Co	55395	54571	1.5	54571	6	25.2	13.8	51.4	77.9	18.7	0.5	1.1	0.1	1.8	2.7	75.6	85	1.6	
AL	1001	Alabama	Autauga	Democrat	Bernie San	544	0.182	Autauga Co	55395	54571	1.5	54571	6	25.2	13.8	51.4	77.9	18.7	0.5	1.1	0.1	1.8	2.7	75.6	85	1.6	
AL	1003	Alabama	Baldwin	Republicar	Ben Carsor	4221	0.084	Baldwin Co	200111	182265	9.8	182265	5.6	22.2	18.7	51.2	87.1	9.6	0.7	0.9	0.1	1.6	4.6	83	82.1	3.6	

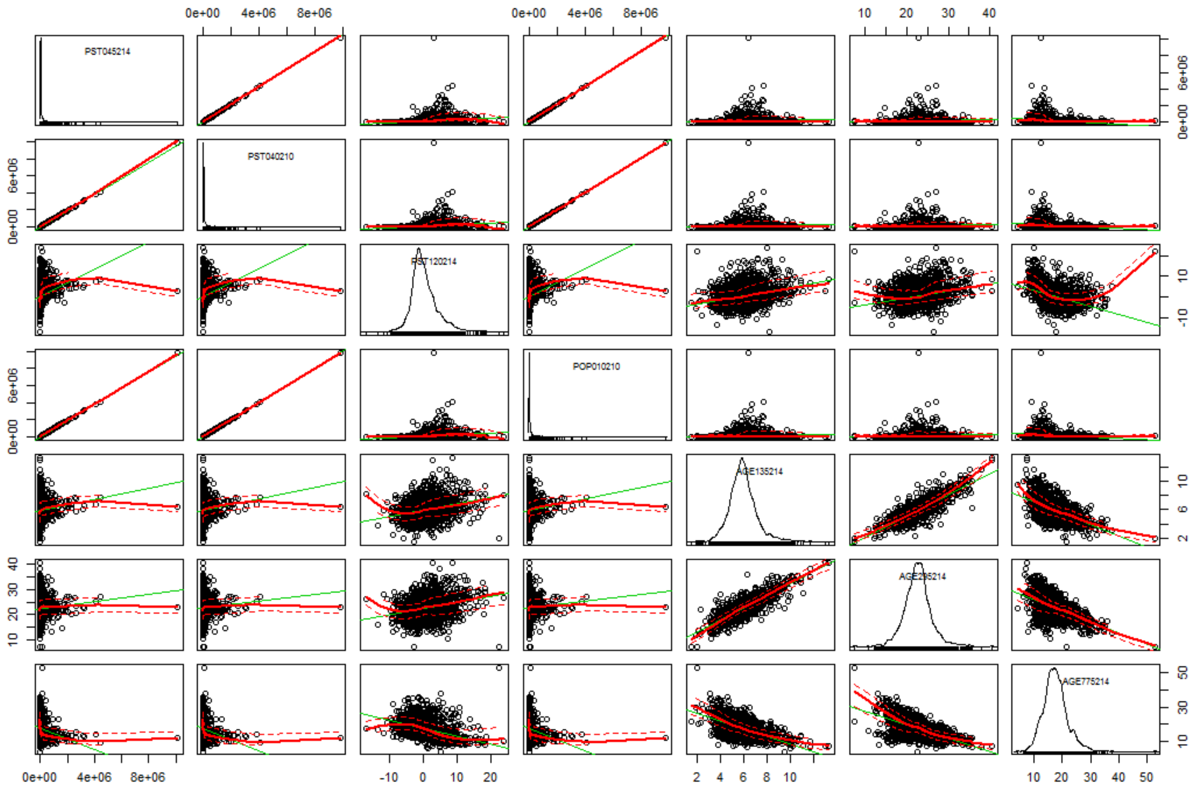
3.5 Exploratory Analysis using Power BI:

Feature Selection:

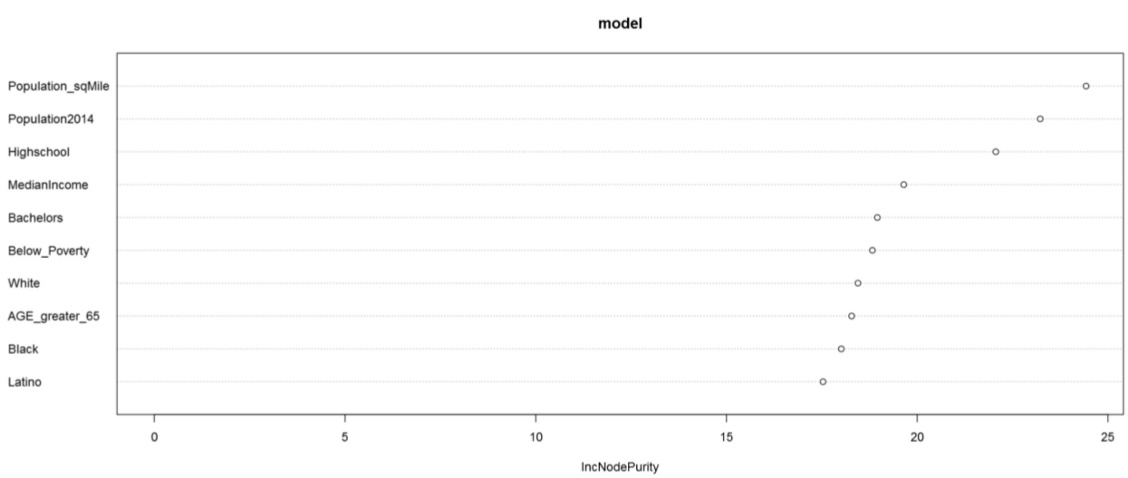


There were 30 demographic features present, to select the features for the predictions we ran the random forest model which gives the features by importance with the Y variable in this case the party fraction.

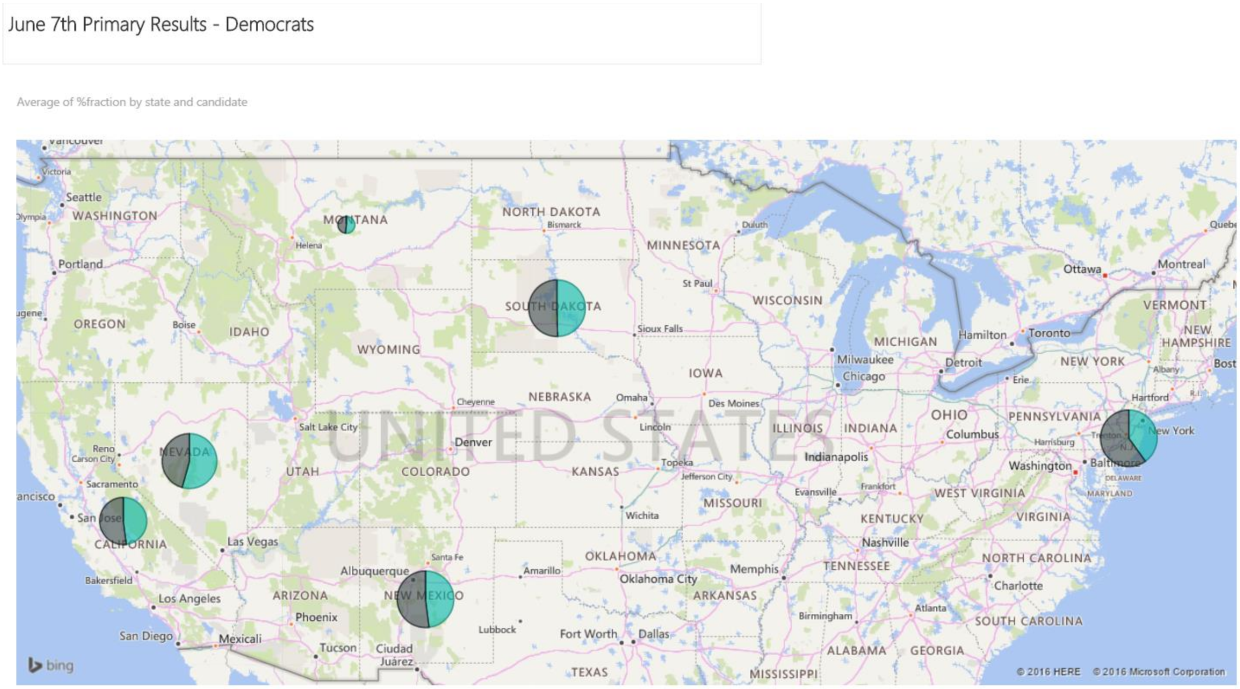
We also generated Scatterplot to establish the correlation between features so that we can remove features with higher correlation.



After selecting the variable which the most important and renaming the features the random forest model is as follows:



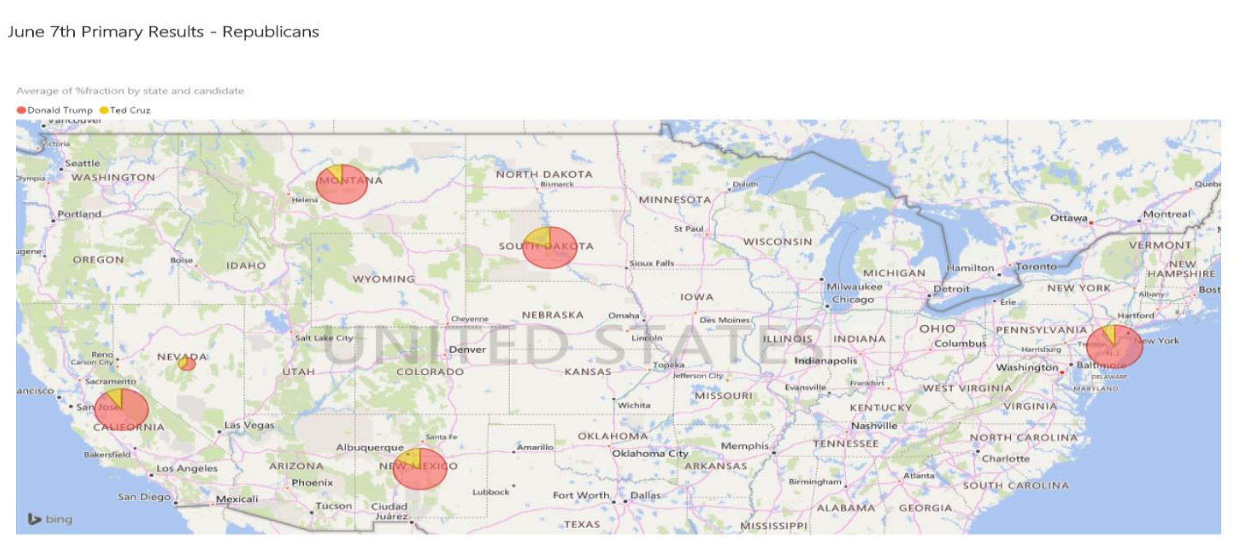
Dashboard 1:



Exploratory Analysis of the percentage of votes for the Democrats Candidates- Hillary Clinton and Bernie Sanders.

This visualization only displays the results of the latest primaries as there is a restriction for Power BI to display the whole dataset

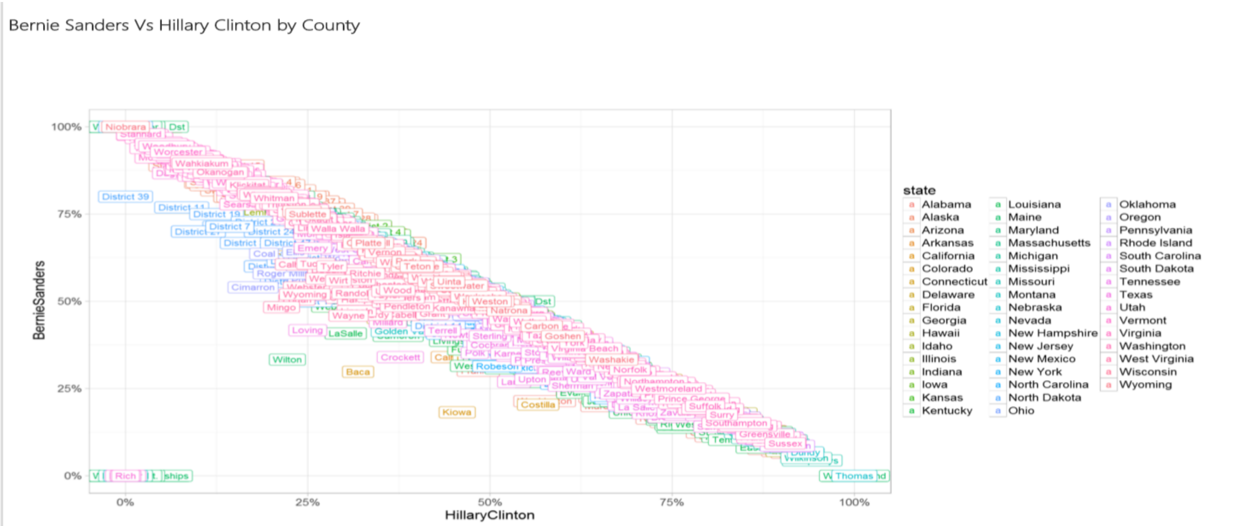
Dashboard 2:



Exploratory Analysis of the percentage of votes for the Republican Candidates- Donald Trump and Ted Cruz.

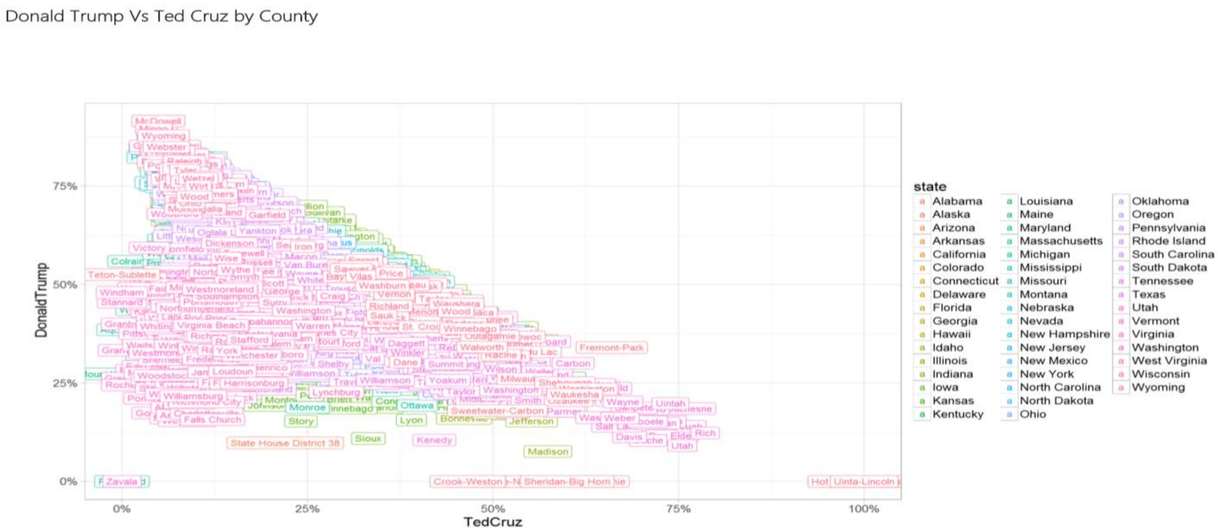
This visualization only displays the results of the latest primaries as there is a restriction for Power BI to display the whole dataset

Dashboard 3:



This visualization shows the county wise percentage of party fraction for Democrat candidates- Hillary Clinton and Donald Trump.

Dashboard 3:

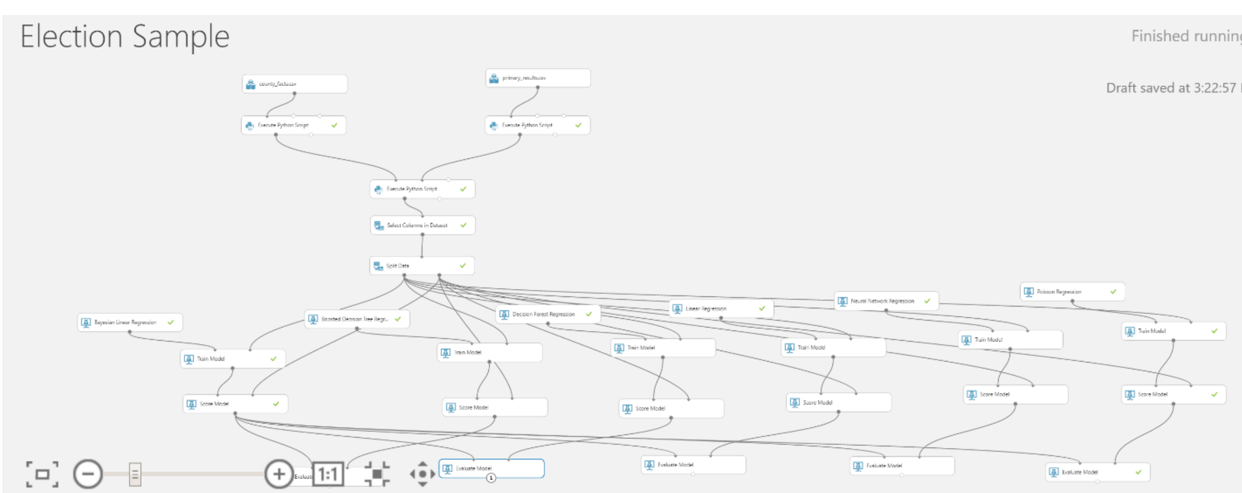


This visualization shows the county wise percentage of party fraction for Democrat candidates- Hillary Clinton and Donald Trump.

This visualization clearly shows that Donald Trump won with a greater majority.

4. Building predictive model

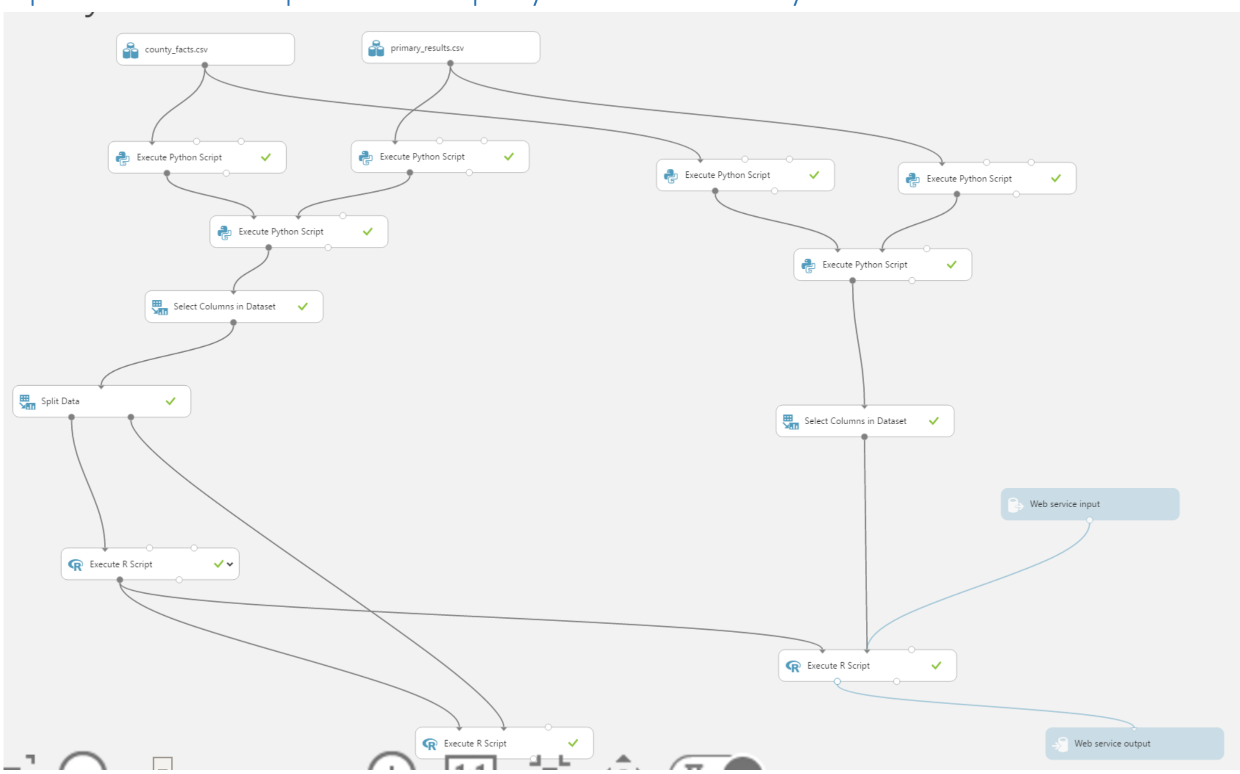
Pipelines built using various models:



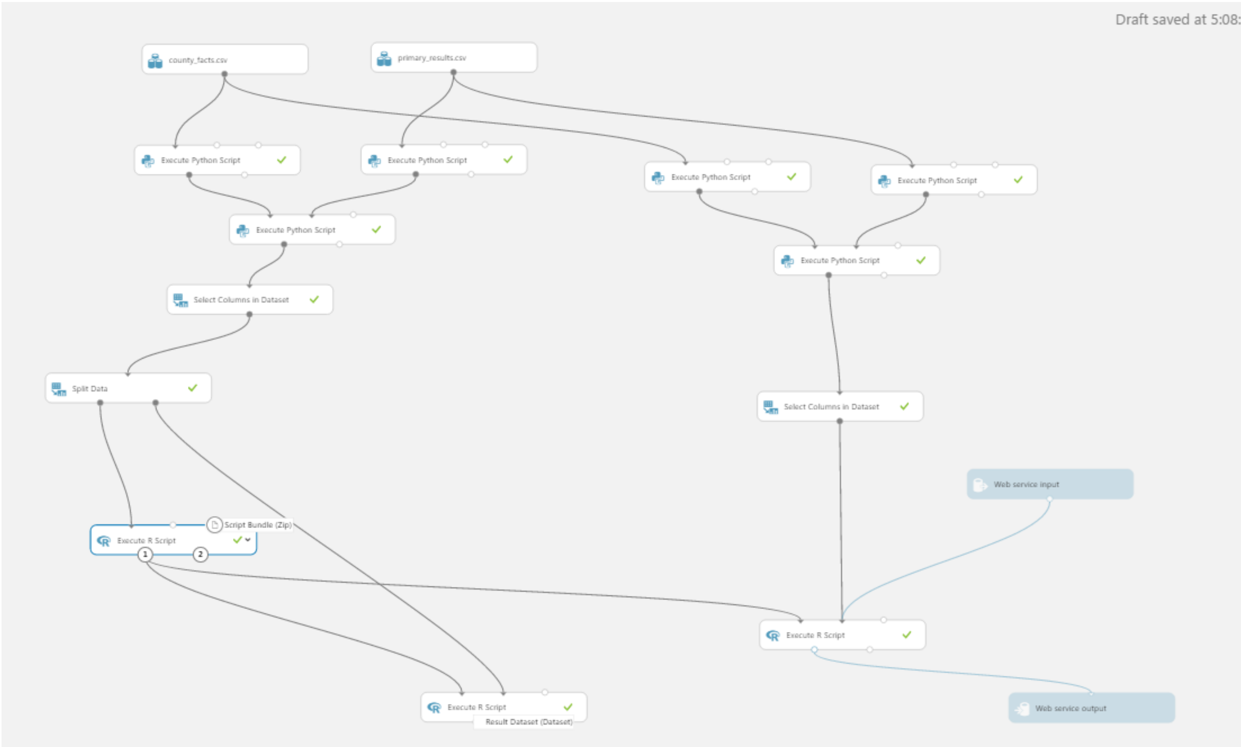
Predictive Model for Hillary Clinton	Bayesian Linear Regression	Random Forest	XGBoost Regression	Logistic Regression
Test RMSE	0.07	0.092	0.03	0.10

We chose XGBoost Regression model in this case as it gives us the least value of RMSE

Pipeline created for prediction of party fraction for Hillary Clinton:



Pipeline created for prediction of party fraction for Bernie Sanders:

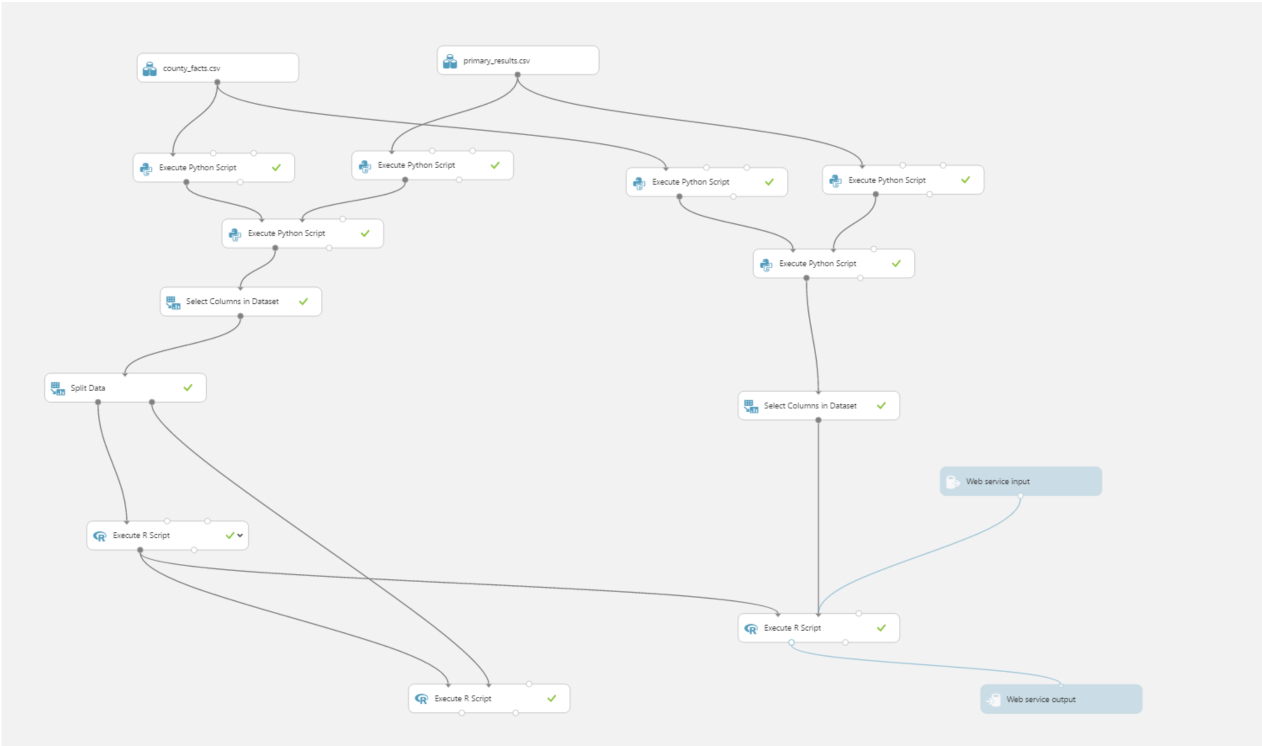


Predictive Model for Bernie Sanders	Bayesian Linear Regression	Random Forest	XGBoost Regression	Logistic Regression
Test RMSE	0.042	0.04	0.037	0.05

We chose XGBoost regression model in this case as it gives us the least value of RMSE.

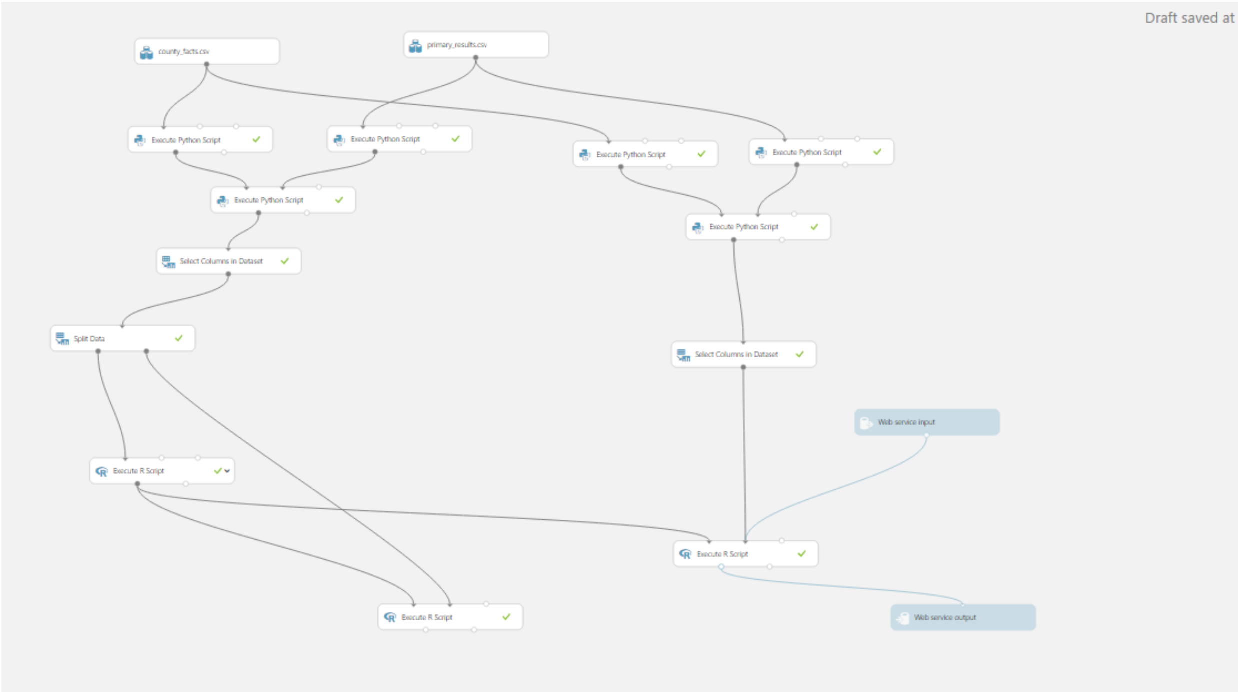
XGBoost parameters tuning was done by trying various permutations and combinations of parameters like number of rounds, max_depth and eta.

Pipeline created for prediction of party fraction for Ted Cruz:



Predictive Model for Ted Cruz	Bayesian Linear Regression	Random Forest	XGBoost Regression	Logistic Regression
Test RMSE	0.157	0.21	0.151	0.156

Pipeline created for prediction of party fraction for Donald Trump:



We chose XGBoost regression model in this case as it gives us the least value of RMSE.

Predictive Model for Donald Trump	Bayesian Linear Regression	Random Forest	XGBoost Regression	Logistic Regression
Test RMSE	0.18	0.20	0.153	0.16

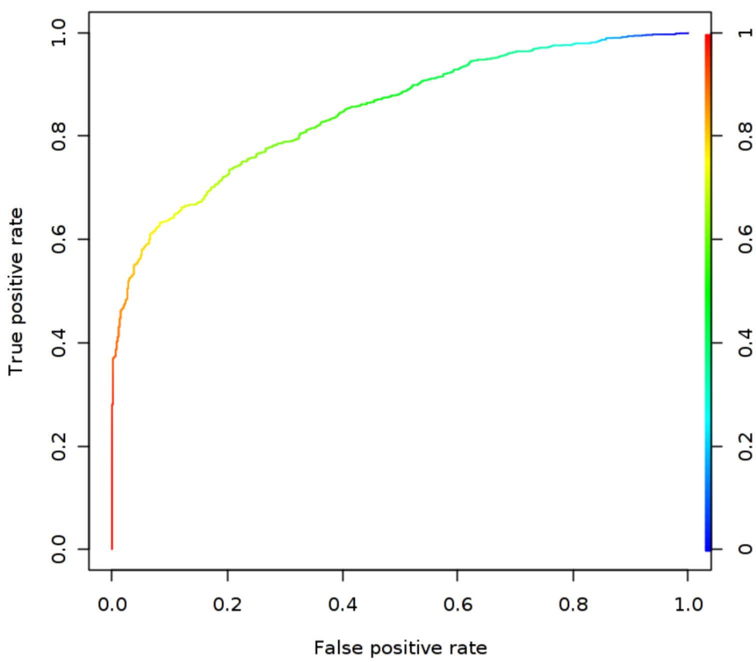
We chose XGBoost regression model in this case as it gives us the least value of RMSE.

Performance Metrics

Data is subset into training and testing datasets. Training data contains the set for which we have the winner available. Test dataset consists the future elections. Logistic model is built with first eight components as predictor variables and the model fit is tested on training dataset.

Confusion Matrix of the training dataset:

```
Confusion Matrix of the training dataset:

                 Predicted
Truth                0    1
  Bernie Sanders   324  280
  Hillary Clinton  163 1102


Area under the curve is:  0.85

0: Bernie Sanders
1: Hillary Clinton
```



Testing the accuracy of our models

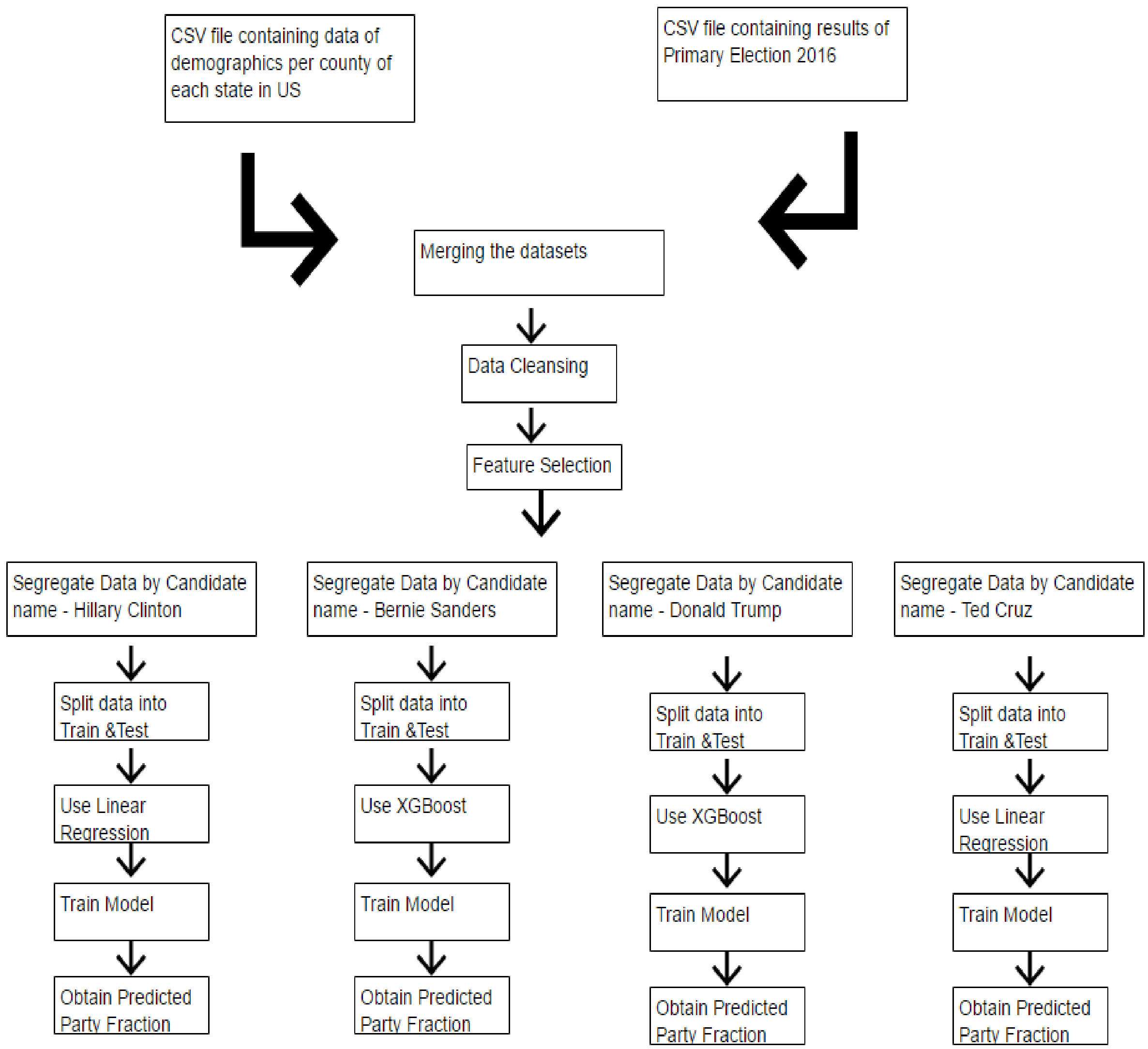
```
#Predicted Results
a=pd.Series(xx, index=['Pennsylvania','Connecticut','Maryland','Delaware','New York'])
#Actual Results
x = np.array([.55,.51,.63,.60,.58])
b=pd.Series(x, index=['Pennsylvania','Connecticut','Maryland','Delaware','New York'])

#Calculate RMS Error
error3 = np.sqrt(mean_squared_error(a,b))

d = {'Predicted' : a,'Real' : b}
final=pd.DataFrame(d)
print (final)
print("Error=",end='')
print(error3)
```

```
              Predicted  Real
Pennsylvania   0.506972  0.55
Connecticut    0.460250  0.51
Maryland       0.576795  0.63
Delaware       0.557297  0.60
New York       0.615027  0.58
Error=0.0451833131275
```

The actual results of the primaries in the respective states are obtained from-

https://en.wikipedia.org/wiki/Results_of_the_Democratic_Party_presidential_primaries,_2016

Workflow of the Models Created



Real-time Twitter sentiment analysis in Azure Stream Analytics

Social media analytics tools help organizations understand trending topics, meaning subjects and attitudes with a high volume of posts in social media. Sentiment analysis - also called "opinion mining" - uses social media analytics tools to determine attitudes toward a product, idea, and so on.

Twitter has become a central site where people express their opinions and views on political parties and candidates. Emerging events or news are often followed almost instantly by a burst in Twitter volume, providing a unique opportunity to gauge the relation between expressed public sentiment and electoral events. In addition, sentiment analysis can help explore how these events affect public opinion.

I. Objective

The objective is to analyze public sentiment for 2 of the presidential candidates in the ongoing 2016 U.S. election namely, **Donald Trump** & **Hillary Clinton**, as expressed on Twitter, a micro-blogging service.

II. Prerequisites

- Twitter account and OAuth access token
- TwitterClient.zip from the Microsoft Download
- Work or school account for Power BI

III. Solution Overview

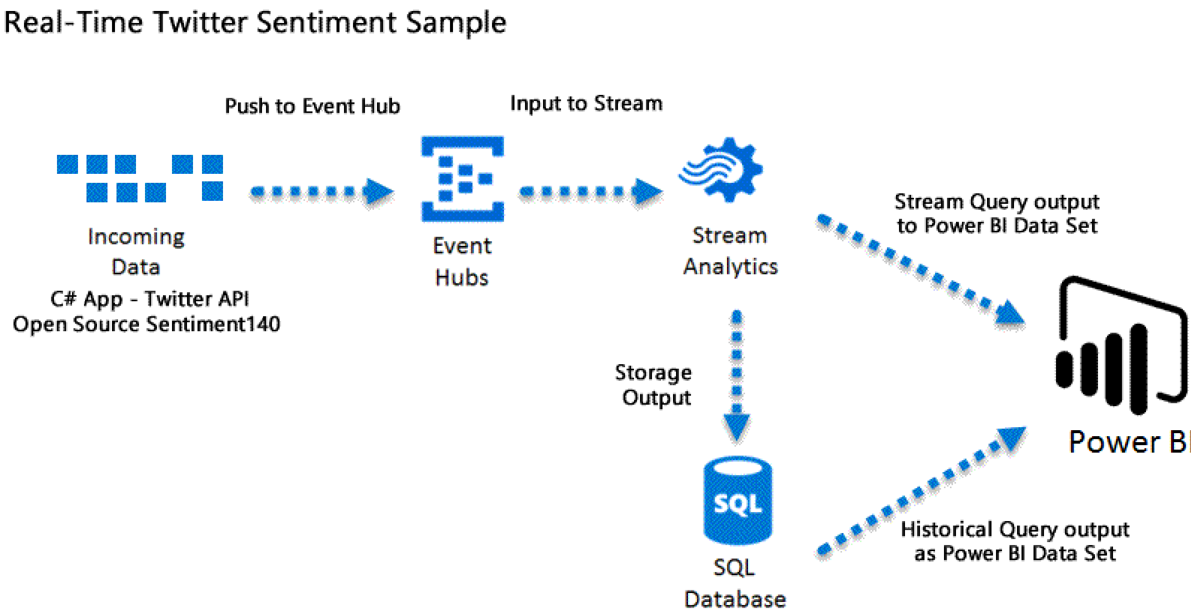
1. Azure Event Hubs is a highly scalable service for ingesting Internet of Things (IoT) event processing data sources. It enables the collection of event streams at high throughput from a diverse set of devices and services. Event Hubs can massively parallel intake millions of events per second via HTTP(S) or AMQP protocols. Once data is brought into Event Hub partitions, you can apply transformations and/or store the event data using any real-time analytics provider or with batching storage adapters.
2. Microsoft has provided a client application that will tap into Twitter data via Twitter's Streaming APIs to collect Tweet events about a parameterized set of topics. The 3rd party open source tool Sentiment140 is used to assign a sentiment value to each tweet (0: negative, 2: neutral, 4: positive) and then Tweet events are pushed to Event Hub.

3. Sentiment140 allows you to discover the sentiments of a brand, product, or topic on Twitter. Sentiment140 uses Maximum Entropy classifier and distant supervision, in which the training data consists of tweets with emoticons where the emoticons are treated as noisy labels.
4. Stream Analytics is great for querying live data streams like Twitter. It is often used to detect anomalies, trigger alerts when errors occur, or feed real-time dashboards like we will be doing with the Twitter Sentiment sample. Stream Analytics provides super simple out-of-the-box integration with Event Hubs and Power BI for developing end-to-end, highly scalable, Internet of Things (IoT) analytics solutions.
5. Power BI is a cloud-based business analytics service from Microsoft that empowers anyone to experience any data – structured or unstructured – via simple drag-and-drop ease. Unlike many other dashboard solutions, Power BI can render live dashboards with moving charts and continuously updated visualizations for monitoring real-time streams from supported data sources.

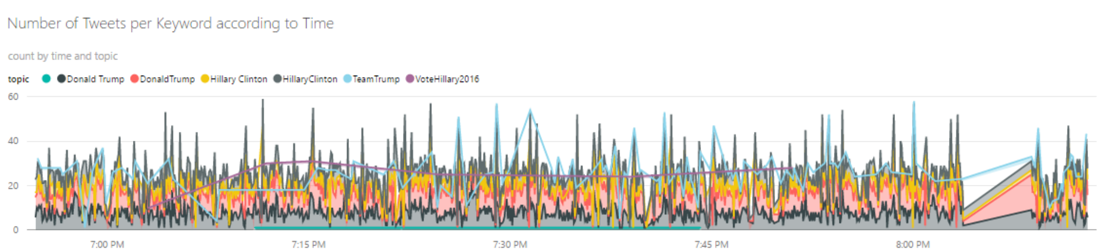
IV. Steps Involved

1. Create an Event Hub input and a Consumer Group
2. Configure and start the Twitter client application
3. Create Stream Analytics job
    - Provision a Stream Analytics Job
    - Specify job input
    - Specify job Query
    - Create Output Sink
    - Specify job Output
    - Start Job
4. View Output for Sentiment Analysis in Power BI

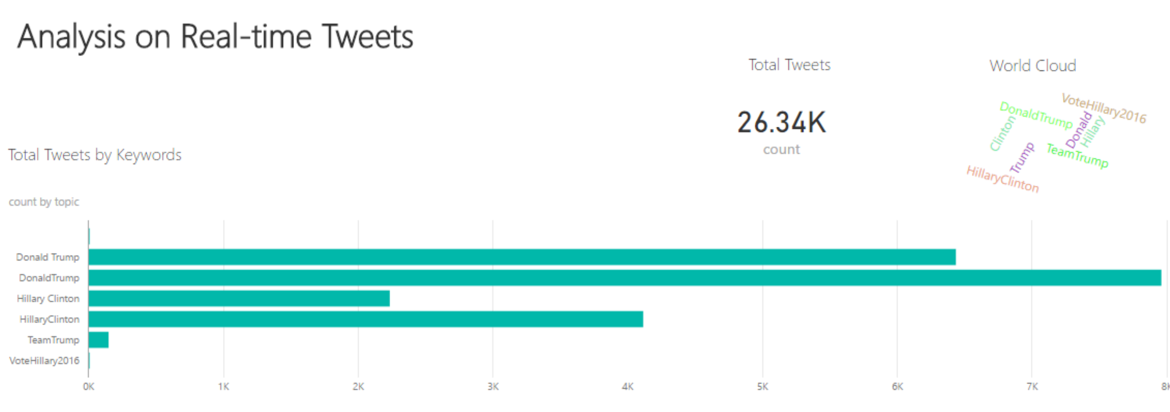
V. Putting it Together



VI. Analytics Dashboard for Streaming data – Power BI

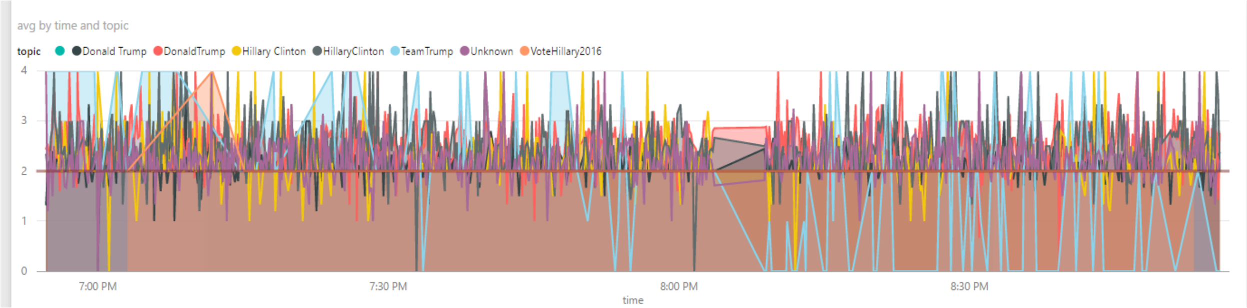


The above visualization shows the number of Tweets by Time for the keywords “Donald Trump”, “Hillary Clinton”, “DonaldTrump”, “HillaryClinton”,” VoteHillary2016”,” TeamTrump”.

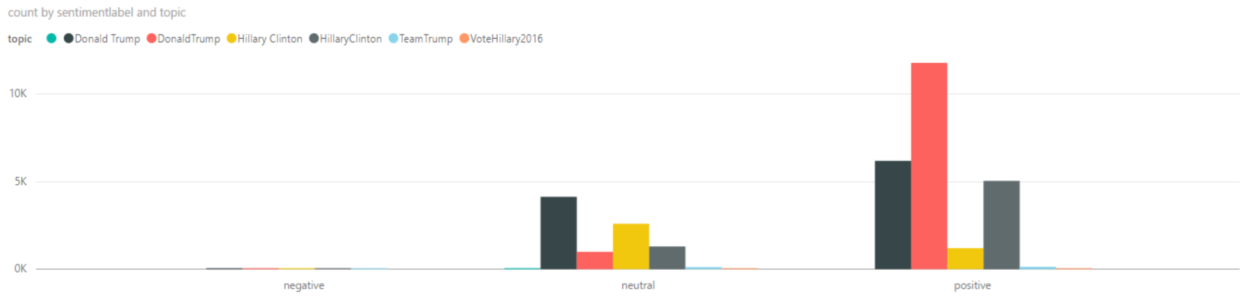


The above dashboard, the first visualization shows Total Number of Tweets that have been recorded since the Azure Stream Analytics Job started and a Word Cloud.

The second visualization also shows the count of Tweet for both the keywords “Donald Trump”, “Hillary Clinton”, “DonaldTrump”, “HillaryClinton”,” VoteHillary2016”,” TeamTrump”.



The above dashboard shows the Sentiment Score of each Tweet 0 being negative, 2 being neutral and 4 being positive.



The above dashboard shows the number of tweets that were under the categories of neutral, positive and negative for the candidates “Donald Trump”, “Hillary Clinton”, “DonaldTrump”, “HillaryClinton”,” VoteHillary2016”,” TeamTrump”.

Analyzing the US Elections with Facebook and R

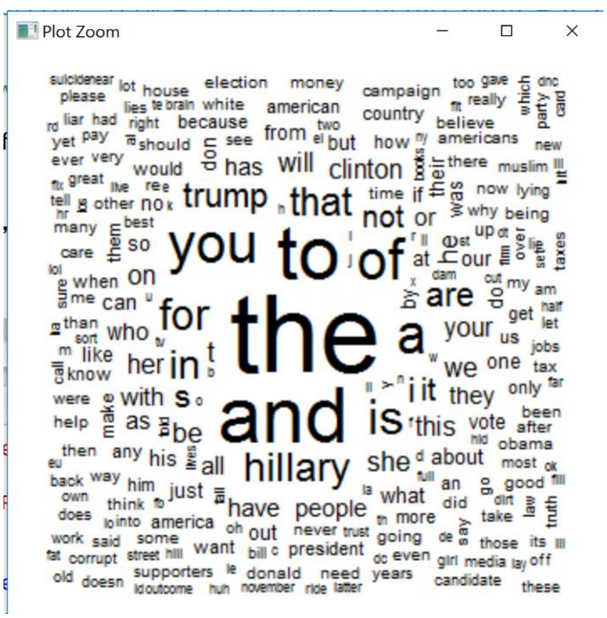
We crawled the nominees’ public page Facebook data, starting May 01, 2015 until May 31, 2016 via R ‘Rfacebook’. Specifically, we request all posts and corresponding comments for the entire time period (Clinton: approx. 1.2m comments / Trump: approx. 1.4m comments). Following this, each comment was analyzed separately with respect to emotional and psychological constructs (the categories are based on the LIWC dictionary) with R ‘tm’ and ‘quanteda’.
Here is a stylized example of the basic code (the code is limited to one candidate (Hillary Clinton), one day (2016-07-07), and refers to a public available dictionary (positive/negative word). The original analysis is based on the LIWC dictionary.

Analysis done on Hillary Clinton’s public Facebook Page

```
            from_id      from_name
1:   889307941125736 Hillary Clinton
2:   889307941125736 Hillary Clinton
3:   889307941125736 Hillary Clinton
4:   518810038260540   Linda Ziegler
5:   578972082228360   Ansab Tanveer
6: 10152532981077412     Bob Burwell
```

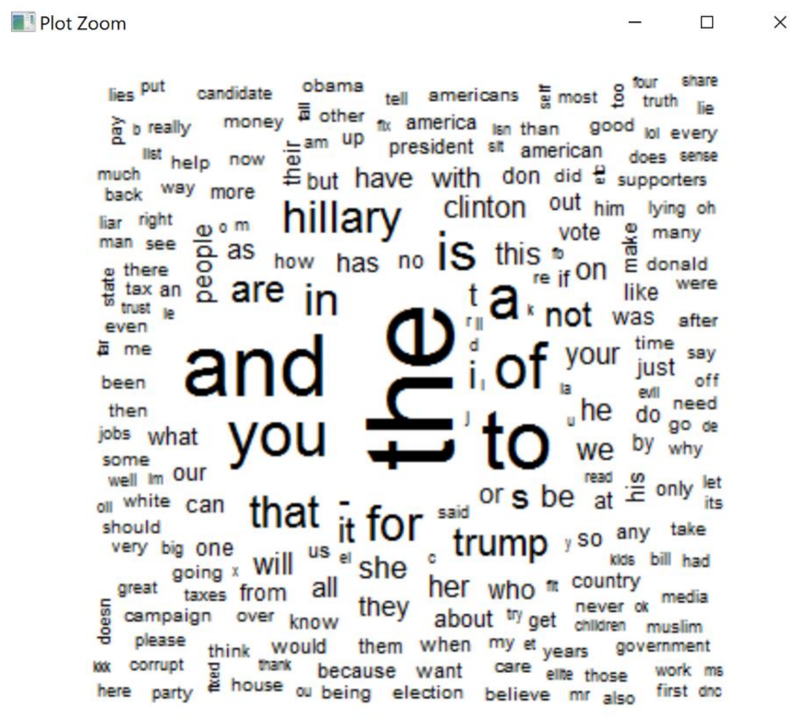
```
                                      link
1: http://www.teamusa.org/News/2016/August/17/100-Meter-Hurdlers-Claim-Team-USAs-First-Ever-
Womens-Track-And-Field-Olympic-Sweep
2: http://www.teamusa.org/News/2016/August/17/100-Meter-Hurdlers-Claim-Team-USAs-First-Ever-
Womens-Track-And-Field-Olympic-Sweep
3: http://www.teamusa.org/News/2016/August/17/100-Meter-Hurdlers-Claim-Team-USAs-First-Ever-
Womens-Track-And-Field-Olympic-Sweep
4:
                                        NA
5:
                                        NA
6:
                                        NA
                                        id       variable L1 value          POST_ID ch
1:   889307941125736_1218943038162223    likes_count  1 33559  889307941125736 15
2:   889307941125736_1218943038162223 comments_count  1  1629  889307941125736 15
3:   889307941125736_1218943038162223   shares_count  1  2171  889307941125736 15
4: 1218943038162223_1218974814825712    likes_count  1   423 1218943038162223 16
5: 1218943038162223_1218958604827333    likes_count  1    85 1218943038162223 16
6: 1218943038162223_1218966508159876    likes_count  1    84 1218943038162223 16
```

Word-cloud

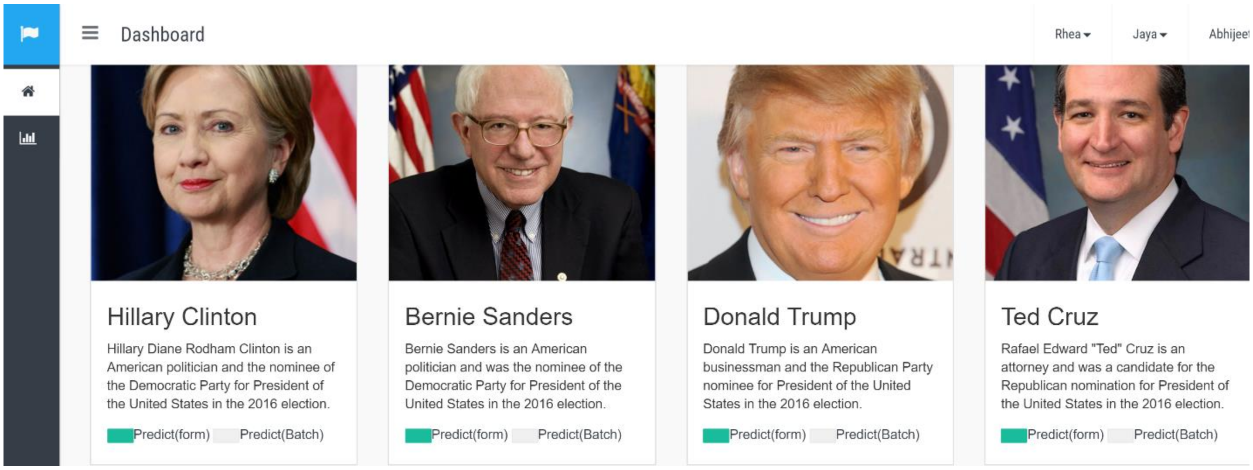


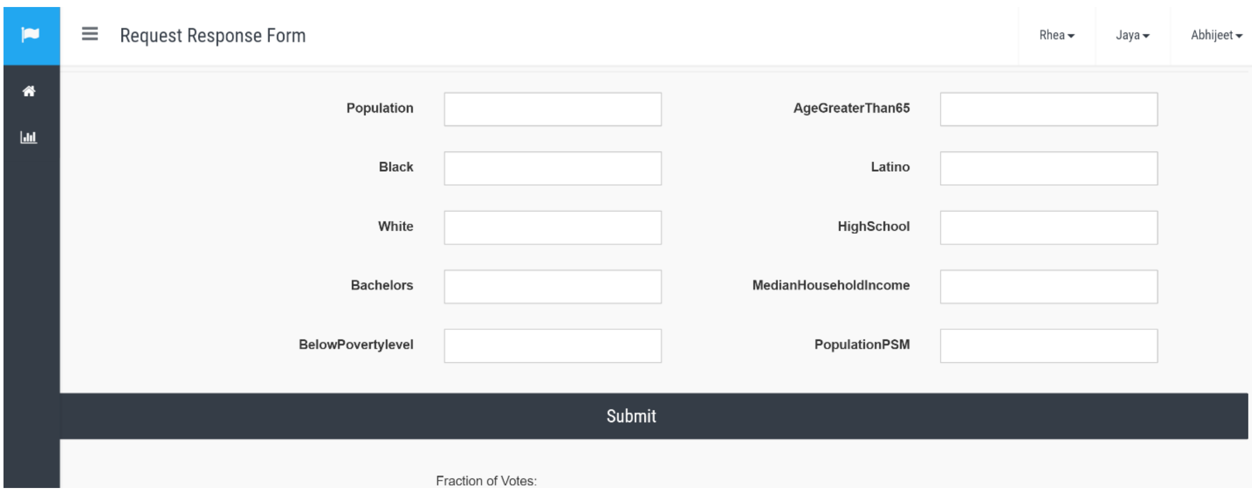
Analysis done on Donald Trump's public Facebook Page

Word-Cloud




```
          Half the people that vote for Hillary haven't been alive long enough to see how c
orrupt she really is...Horrible Crooked Lying Hillary
               created_time type link                                 id    variable L1 va
lue
504 2016-08-18T22:01:19+0000 <NA> <NA> 10157521443780725_10157521458570725 likes_count  1  2
121
505 2016-08-18T22:01:34+0000 <NA> <NA> 10157521443780725_10157521461140725 likes_count  1  1
307
506 2016-08-18T22:01:29+0000 <NA> <NA> 10157521443780725_10157521460815725 likes_count  1
400
507 2016-08-18T22:04:53+0000 <NA> <NA> 10157521443780725_10157521474865725 likes_count  1
332
508 2016-08-18T22:06:18+0000 <NA> <NA> 10157521443780725_10157521484185725 likes_count  1
235
509 2016-08-18T22:01:29+0000 <NA> <NA> 10157521443780725_10157521460860725 likes_count  1
461
              POST_ID ch       date      date1 tempID positive negative
504 10157521443780725 17 2016-08-18 2016-08-18      1        1        3
505 10157521443780725 17 2016-08-18 2016-08-18      1        1        0
506 10157521443780725 17 2016-08-18 2016-08-18      1        3        0
507 10157521443780725 17 2016-08-18 2016-08-18      1        4        9
508 10157521443780725 17 2016-08-18 2016-08-18      1        3        0
509 10157521443780725 17 2016-08-18 2016-08-18      1        1        3
>
```





Outlook
Username : adssummer2016team1@outlook.com
Password : team1team1

Website link : http://adsteam1.azurewebsites.net/webform1.aspx

Github link : https://github.com/abhijeet-s/ADS_Assignment/tree/master/Assignment3