Comparison Analyses of the 2016 US Presidential Elections

Data Application Development

Mina Gaid
School of Computing
National College of Ireland
Dublin, Ireland
x13729979@student.ncirl.ie

For my data applications project, I have chosen the topic of the 2016 US presidential elections. I have selected this topic as I believe big data can be useful in regards to analysing the results.

***Abstract*— On November 8th, 2016, the Republican presidential nominee Donald Trump won the 2016 US presidential elections and became the 45th president of the United States. This paper will examine both Donald Trump and Hillary Clinton's election and primary results to better understand if their performance has improved or declined between the two. The analyses will also provide an insight into whether or not their win in their party's primary gave a forecast of the election results.**

Keywords: Donald Trump, Hillary Clintons, Election, Primary Results, Analyses

1. Introduction

The US elections are certainly interesting and are unique in regards to how they're setup. In order to become a presidential candidate, you most first compete in the primaries. The primaries can be best described as an internal election for the main Parties in which they select a candidate to represent them. Only the winner of the primaries can become the party's nominee.

Though the elections and the primaries are different, they are structured the same way in terms of delegates and states.

As such, I thought it would be interesting to compare the results of the primaries to the results of the election for each of the main presidential candidates (Donald Trump and Hillary Clinton). I wanted to see how they did in comparison, whither they improved on their individual result of the primaries in the election or declined.

Using descriptive statistics, I can achieve this comparison. There were many methods but I have chosen to use the student T test. The particular T-test I will conduct is that of a Paired Student T test in which the past results (primaries) are compared with those of the present (election). Using the parried T test, I can work out the t-stat and use that to retrieve the t-crit. This will help me differentiate my Null and Alterative Hypotheses.

2. Data

This section will outline the source of data that was chosen for this Analyses along with the data dictionary for the datasets used.

Choosing the right data for the analyses was critical for results of the research. It was important that I find datasets that hold the results for both presidential candidates for both the election and primaries. With the election, having just ended, it was difficult tracking down premade datasets in the CSV format that could be used. However, other sources of data were available in which could be used to create datasets. With there being hundreds of different sources of data,

selecting one was easy as most left out certain results for some states. In the end, I have chosen to extract the data from Wikipedia as I had two ways of extracting the data. The first was through web scraping and the second was to manually copy and paste the necessary variables.

To extract the data using R, I intended on using the "rvest" library. The following pseudo code went something like this:

```
library("rvest")
url <-
http://en.wikipedia.org/
```

It took a few attempts to try and implant but the end results were sloppy tables that were unusable. To add to that, too much unnecessary data was extracted such as delegate counts. This would have required an unreasonable amount of pre-processing and cleaning of the datasets. As such, I scrapped the idea of using R to extract the data and went with manually copying the variables into excel.

I First went about this by retrieving the data of both the primaries and the elections. I saved the data into two individual excel spreadsheets titled primaries.xlsx and elections.xlsx. Once saved, I converted the spreadsheets into a CSV format that can be used by R to better manage the data. At this point, I had two separate datasets for the election results and the primary results. Each one contained the results of both presidential candidates in a CSV format.

The following tables show what my dataset looks like:

Column Name	Data Type	Rows	Description
State	Text	56	Name of state
Donald Trump	Double	56	Trump's Primary results
Hillary Clinton	Double	56	Hillary's Primary results

Table 1 Primary's Dataset

Column Name	Data Type	Rows	Description
State	Text	56	Name of state
Donald Trump	Double	56	Trump's Election results
Hillary Clinton	Double	56	Hillary's Election results

Table 2 Election's *Dataset*

Happy with the results, I was ready to begin analysing the datasets in R.

3. Methodology

Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams (machinelearningmastery.com, 2016).

The process of KDD can be described as follows (tutorialspoint.com, 2016):

- Data Cleaning − In this step, the noise and inconsistent data is removed.
- Data Integration − In this step, multiple data sources are combined.
- Data Selection − In this step, data relevant to the analysis task are retrieved from the database.
- Data Transformation − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining − In this step, intelligent methods are applied in order to extract data patterns.

- Pattern Evaluation − In this step, data patterns are evaluated.
- Knowledge Presentation − In this step, knowledge is represented.

In regards Data cleaning, only the necessary data was extracted from the source (Wikipedia) as mentioned in the previous section. There was not much to be done in terms of cleaning. However, the step was undertaken to filter out the % sign from the result fields of both candidates for both datasets. This was done so R can differentiate between the results and identify them as numeric values.

Nothing was done for the Data Integration phase as I was reading out the data from two datasets simultaneously. There was no need to combine the data for any particular reason.

In the Data Selection phase I needed to select the results associated with each candidate from the datasets so I can compare them individually. For example, I needed Trump's results so I selected his results only from each

4. Implementation

As mentioned before, this analysis was conducted using a paired student t test. To perform this test, I am to follow the formula shown in figure 1.

$$t = \frac{\Sigma d}{\sqrt{\frac{n\left(\Sigma d^2\right) - \left(\Sigma d\right)^2}{n-1}}}$$

Figure 1 Formula for Paired Student T-test

Before starting, I intended to get a visualisation of the data by using several function such as that of the plot() function. The str() function was used to get an overview of the datasets and the summary() function was used to retrieve basic information such as the mean. Plots were then created to better showcase the data as seen in figure 2.

The KDD process was adapted for this report and the following are examples of how it was used.

dataset and put both of them into objects (data frames / subsets) to better manage them. I did the same for Hillary's results.

For the data to be processed and properly analysed, it was important that the Data Transformation stage was applied. This was done so the data can be better understood by R. With R being such a powerful language, it could ascertain that my data was numeric after the % sign was filtered out. It was still none the less important that I declare this in my code for the sack of subduing any future errors that may arise. This set was undertaken in as part of my analyses of the code. Following the mathematical formula for the paired t test, I was required to subtract the election results from the primary. While doing so, I declared and transformed the selected results into numeric values.

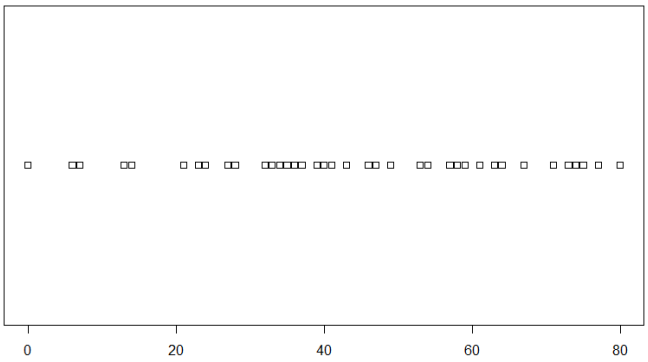

Figure 2Overview plot of Trump_primaries

To begin, I selected the required data for the test, the first being Trump's. As such, the appropriate subsets for Trump were selected. The data was that of the primary and the election results. Using the formula, I was to subtract each state result of the election from each state result of the primary. This was tricky to do as I kept receiving the total combined result. The issue being how the

algorithm was structured. After noticing the issue, I manged to retrieve the results and store them in a subset called td (trump data).

After doing so, I then went about getting the same results to the power of 2. I did this by selecting the td subset and using the carrot sign (^) to retrieve the results. The values were then stored in the subset tdsq (trump data squared).

Using the sum function, I was able to retrieve the sum of both the td and tdsq subsets and store the results in other subsets std (sum trump data) and stdsq (sum trump data squared).

I then went about getting the degree of freedom value by subtracting 1 from the value of tn and storing it in tdf (trump Degree of freedom).

with that of std squared. The value was then stored in tcal1 (trump calculation 1).

The syntax was as follows:

```
tcal1 <- (tn * stdsq) - (std)^2
```

This was then preceded by dividing the result of tcal1 with the degree of freedom (tdf). The result was stored in tcal2 (trump calculation 2).

The next step was to get the square root of tcal2. This was done by using the sqrt() function in R to retrieve the value which I then stored in tcal2_sqrt (trump calculation 2 square root).

The finale step was that of Dividing the sum of td (std) with previous result (tcal2_sqrt) and storing the value in the object trump_tstat (trump t stat). The value calculated was that of the t stat for Trump's results.

The same steps taken for Trump's paired T test were then carried out and applied for Hillary Clinton using the same methods.

To find the value N of the Formula, I had to get the count of each record in the tdsq subset. I attempted to do this using the function nrow() which retrieves the number of rows. This didn't work for me as the result kept coming back as null. Instead, I used the NROW() function. The difference between NROW() and NCOL() is that the lowercase versions will only work for objects that have dimensions (arrays, matrices, data frames). The uppercase versions will work with vectors, which are treated as if they were a 1 column matrix, and are robust if you end up sub setting your data such that R drops an empty dimension (stackexchange.com, 2016). The result was then stored in the value tn (trump n).

After doing so, I continued following the formal by multiplying tn by stdq and subtracting the result

After completing both tests, I checked the results by performing the two tests using excel and they both came back with the same results (when rounded). This concludes that the tests carried out on R are indeed correct.

For the paired T test to have been completed, I required the T-Crit. It was not however possible to get the T-Crit as I required the t-distribution table to do so. As such, this was the end of my analysis.

To finish off, I retrieved the correlation results by using the correlation function in R cor(). I applied "*.test*" at the end to obtain the relevant results of a T test. This was applied for both Hillary Clinton and Donald Trump at the end of both tests.

5. Results

When examining the results, we can see that Hillary's tstat is 3.866348 While Trump's tstat is -0.5758538. This tells us that Hillary Clinton

performed better in the US presidential election then she did in the democratic Primaries. Donald Trump however did better in the primaries then in the US general election.

The results were surprising because Donald Trump ran against 17 (5 after the first few caucus) candidates in the primary and Hillary ran against only 2 (one after the first caucus). Donald Trump's results for the election should have overlapped his primary wins. This was not the case as shown by the analyses.

It showed be noted however that Hillary's Primary results included super delegates which helped win her the Democratic nomination. These super delegates were not included in our analyses as we only examined the results of voting per state. Hillary's win of the primaries was due to the delegates and as such her election results came back and showed an improvement because of this. None the less, she did in fact improve on her primary votes.

Donald Trump's results could be explained by the candidates he ran against in the primary's. He was clearly the more favourable candidate. It does not mean Hillary outpaced his results of either the election or primary's but it showed his performance as being better when competing against fellow Republicans.

6. MapReduce

After I was finished with my main analyses, I planned to implement MapReduce. This was so that I can get a better understanding of the presidential candidates results regarding their top performing states. To do this, I applied a mapper and a reducer to run and analyses my data set to retrieve the top ten results.

The mapper and reducer were written in Python 2.7. So, my first step was to download and configure python on my windows 10 machine. I download the required version of python online (Python 2.7). I then went about the installation process. While installing, I made sure to select the option to add python to my Windows path. When the installation was complete, I ran CMD (terminal / command prompt) and simply typed "Python" to test if it was properly configured. It was and I was ready to start running python on my machine.

I started off by downloading the mapper we were given in class. I then modified the mapper we were given. I changed the row in which the mapper analysed, added in the necessary variable to the mapper and replaced \t to a comma so it can run with CSVs. I then copied the mapper and renamed it to reducer.py. In the end, I had both a mapper.py and a reducer.py.

With the data sets being identical in terms of layout, length and headings, I applied the mapper and reducer on both my election dataset and my primary dataset.

The following is a step by step guide to running both the mapper and reducer with the election dataset (for Trump's top ten states).

```
➢ cd C:\MapReduce
➢ type C:\MapReduce\election.csv
➢ type C:\MapReduce\election.csv |
  C:\MapReduce\topTenStatesMapper.py
➢ type C:\MapReduce\election.csv |
  C:\MapReduce\topTenStatesMapper.py
  | sort
➢ type C:\MapReduce\election.csv |
  C:\MapReduce\topTenStatesMapper.py
  | sort |
  C:\MapReduce\topTenStatesReducer.py
➢ type C:\MapReduce\election.csv |
  C:\MapReduce\topTenStatesMapper.py
  | sort |
  C:\MapReduce\topTenStatesReducer.py
  >> C:\MapReduce\TopTenElection.txt
```

As shown in the commands above, I ran the mapper, sorted the results, ran the reducer and outputted the finale results to a text document. The results should then show

Trumps top ten performing states in the election.

(Note, on some machines, the results might come back as empty because of the % symbol in the dataset. I was not able to solve this issue)

7. Conclusions

In conclusion, we have learned and identified that although the results of the primaries don't mean a certain win, in this case, Trump's astonishing results of the primary gave a forecast of the election results. Of course, it's important to note that politics is uncertain and a range of different factors between the primary and election could alter, change and determine the race for better or for worse for both candidates. Trump's lead in the Primary's was so significant when compared to his rivals results that we can determine he will be the election winner.

If the analyses were carried out again, I would factor in delegate counts for each state. I would also include data on other candidates both from the US election e.g. Jill Stein, and the primaries – e.g. Jeb Bush and Bernie Sanders. This would provide a more detailed analysis in which we can better identify the winner beforehand and recognise the reason for their win.

If I had more time in this analyses, I would use it to compare the results of both candidates with one another and factor in other candidates as mentioned above. This would again give a more detailed understanding of the presidential race and the results. The more data available, the more I can do in my analyses. It is possible that I would include more datasets, even mine tweets during both the primary and the election and identity whether they're positive or negative regarding each candidate. This would determine the mood and popularity towards each candidate. With all this factored in with my analyses, it would provide better results.

Bibliography

(Note: This Bibliography was made using citethisforme.com – Harvard Referencing)

[closed], H. (2016). *How do I get the number of rows of a data.frame in R?*. [online] Stats.stackexchange.com. Available at: http://stats.stackexchange.com/questions/5253/how-do-i-get-the-number-of-rows-of-a-data-frame-in-r [Accessed 21 Dec. 2016].

Brownlee, J. (2016). *What is Data Mining and KDD - Machine Learning Mastery*. [online] Machine Learning Mastery. Available at: http://machinelearningmastery.com/what-is-data-mining-and-kdd/ [Accessed 21 Dec. 2016].

www.tutorialspoint.com. (2016). *Data Mining Knowledge Discovery*. [online] Available at: https://www.tutorialspoint.com/data_mining/dm_knowledge_discovery.htm [Accessed 21 Dec. 2016].

YouTube. (2016). *How To... Calculate Student's t Statistic (Paired) by Hand*. [online] Available at: https://www.youtube.com/watch?v=BPbHujvA9UU [Accessed 21 Dec. 2016].