How important is campaign spending in winning the US Election?

INM430 Coursework - A Tiny Data Science Project

Courtney Irwin, MSc Data Science, City University

Project Domain

US Presidential election campaigns are multi-million-dollar undertakings. Data on election campaign spending can be reimagined as a proxy for election campaign strategy; given the limited resources of an election campaign, how can the funds available be deployed in a manner that maximizes the chance of achieving the electoral college votes required to win the race? This project looks at the choices made by candidates about how and where they choose to spend their campaign funds. The importance of these choices is measured by modelling the relationship between campaign spend and election outcomes; and by demonstrating how effective campaign spend is in predicting election outcomes.

Data Source

The main dataset for the project is the campaign disbursement data publicly accessible from the US Federal Election Commission Website: https://www.fec.gov/data/disbursements/

The campaign disbursements data was extracted for the period of 12 months prior to the election for disbursements related to the eventual Democratic and Republican Nominees for the US elections in 2016, 2012, and 2008.

The outcome of the election by state were copied from tables available from the NY Times website:

- https://www.nytimes.com/elections/2016/results/president
- https://www.nytimes.com/elections/2012/results/president/big-board.html
- https://www.nytimes.com/elections/2008/results/president/votes.html

Finally, the election spending data was supplemented with some data on the race and age population characteristics of each state, from the 2010 census, https://www.census.gov/data/tables/2017/demo/popest/state-detail.html.

Analysis Strategy

The aim is to explore the location (US State) and purpose of election spending and create models that predict the outcome of the election in these States, in order to understand the correlation between election spending and election outcomes.

It's anticipated that election spending will have a high correlation with election outcomes, due to the strategic nature of campaign spending.

For comparison, the election spending data will be supplemented with census data on race and age breakdowns of populations. Its anticipated that this data will offer improved the prediction accuracy of the model.

The main challenge to using campaign spend to predict the outcome of an election is anticipated to be the data from the 2016 election. This election was unique in that the winning candidate spent significantly less on the election campaign than their opponent. However, this data is reflective of the wider challenges in modelling election outcomes.

The plan for analysis is:

- Model the total election campaign spend by State; predicting election outcomes
- Adjust the model to include more detailed information on the use of campaign funds
- Adjust the model to include details of the race and age population characteristics of each state
- Compare the models above based on the average cross validation accuracy score

Findings and Reflection

The ipython computational workbook US Election Spend Analysis - Final.html demonstrates the step by step analysis process undertaken as described above.

Initial Data Wrangling Challenges

The initial plan for the project, as proposed in the progress report, was to use a slightly different dataset, also downloaded from the FEC, but with fewer features. I intended to use some basic natural language processing functions to transform a basic free text field (the disbursement description field) into a small number of disbursement categories. I experimented with this, although I wasn't successful in creating less than 1000 categories for analysis. Further investigation on the FEC website revealed a larger dataset, with a greater number of variables, including a disbursement description field. The initial wrangling undertaken is included in the workbook Election Data - Description Clustering - Initial Workings in the file.

The relevance of the data to the phenomena under observation

The process of undertaking the analysis required several assumptions which need to be considered ahead of interpreting the findings.

Firstly, the model is based on 'spend by state', however, this information was derived from the address of the campaign fund recipient in the disbursement database, creating the assumption that only money received by vendors based in the state was spent on campaigning for that state. It ignores the potential value of money spent centrally (e.g. on campaign staff payroll) may have on the campaign in a state.

Secondly, the model aims to predict a binary outcome; who will win the electoral college votes of that state. However, the states of Nebraska and Maine allow a split of electoral college votes[1]. In these two cases, the model only predicts who will take the majority of the electoral college votes.

Finally, for the information on population breakdown by race, the US census website reports the number of people in each state who identify with a subset of races as at 2010 (the date of the last US census), however, it doesn't offer a projection for how this may have changed since. This meant that the model only reflected the population by race, as at 2010.

[1] https://www.archives.gov/federal-register/electoral-college/faq.html#wtapv

In relation to the data on age, the census website did offer estimates as to how the population may have changes in relation to age in the years since the census, and the model incorporates this. The exception is for 2008, where the data wasn't available, and the age breakdown for 2010 was used as a replacement.

Summary of findings and relevance to the analytical questions asked:

Average Cross Validation Accuracy Score for all Models:

	Naïve Bayes (Gaussian Distribution)	Logistic Regression	Linear Support Vector Classification	Average Score across three algorithms
Model 1: Summary State by State Spend Data	0.5563	0.5796	0.6260	0.5673
Model 2: Detailed Purpose Category and Summary State by State Data	0.5688	0.4900	0.5123	0.5237
Model 3: Census Age and Race Population Characteristics, Detailed Purpose Category and Summary State by State Data	0.5817	0.6682	0.4967	0.5822

The average accuracy score for models using only spend data (model 1 and 2) was approximately 55%. This was lower than anticipated, given the hypothesis that election spending by a candidate could be a proxy for the 'winnability' of a state vote. It may be that the data from the 2016 election skewed this result, and the model could be improved with data from further elections. However, the low score here may be a demonstration of the unpredictable and complex nature of election outcomes.

The improved performance of Model 1 over Model 2 (a model with more detailed information) was also unexpected. This may be due to:

1. It might not be that the type of spend doesn't impact the outcome, it means that candidates already know this, based on their existing analysis, and are spending in line with a strategy aimed at optimizing outcomes. Therefore, the model doesn't reflect an obvious gain or penalty by spending against each category.
2. The categories used are specified by the FEC and are reasonably broad and general in nature. A more detailed analysis (e.g. spend on advertising by type of media) may improve prediction accuracy.
3. The first model may simply predict the outcome based on who had the highest spend. At a national level, the winner was the candidate that spent the most in two thirds of the cases presented in the data. The accuracy rate is reasonably close to this ratio.

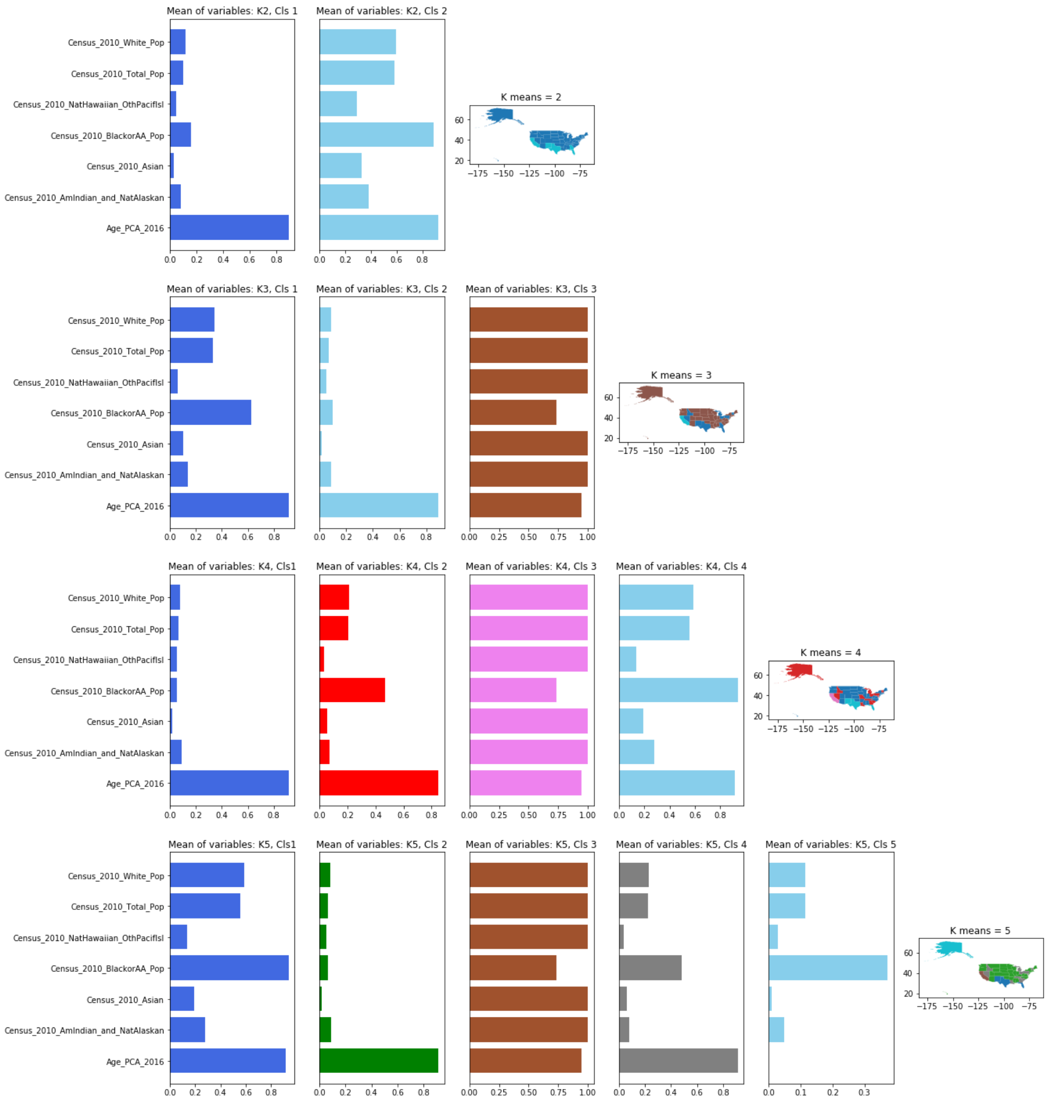
The addition of the census data in Model 3 did not provide a dramatic improvement in prediction. It may be that much more information on state population characteristics (e.g. Religion, Economic data such as employment rate, etc.) is required to see a significant improvement in prediction. Furthermore, it may be that candidates undertake a wide range of analysis ahead of investing campaign budgets in states, and that this basic information has already been accounted for when distributing campaign spending.

Clustering and Multidimensionality Challenges

As a final additional step in the investigations, I attempted to cluster states based on the data from model three (above). The intention of this work was to identify groupings of states with similar attributes, where candidates may utilize a similar campaign approach.

The outcomes of this work are included at the end of the workbook. The first set of clustering models developed offered little variety. After increasing the number of clusters to 7, many states were still clustered under a single category. This effect was traced to the inclusion of unnormalized columns. Once removed, the second set of clusters demonstrated greater variety. However, due to the high dimensionality of the data, it's difficult to determine the source of variety across the clusters.

As a final approach, I clustered only the census data (reducing the dimensionality). This enabled a study of the variation between the clusters. The results showed a high variation between the population characteristics between states. Utilizing a clustering method like this could be a basis to understand variations and similarities in campaign strategy between states.



Further work:

The findings presented here are an initial investigation into the nature of election campaign spending. The small sample size (three elections) may mean that using these models to infer wider realities regarding the outcome of elections is challenging. However, given the infrequency of elections (every four years), and the rate of change of society, data from earlier elections may not enhance models seeking to predict the future. Further analysis work on optimizing campaign spending to achieve electoral outcomes could incorporate 'deeper' data; such as more detailed information on the nature of the spend, and more information on state population characteristics. Additionally, work could be done to review the usefulness of regression models, rather than binary classification models, that seek to predict a percentage of how much of the vote would be won by each candidate.