Sam Coxon

Machine Learning

Final Project Report

23/12/2016

Problem:

The role of the Federal Election Commission (FEC) is to collect campaign finance data and enforce provisions surrounding campaign funding law in the United States. Extremely expensive campaigns in the past decade and the rise of campaign finance reform as a legitimate platform issue by candidates such as Bernie Sanders have thrust the idea of money in politics into the forefront of the national conversation about the state of the democracy in the US and made the role of the FEC all the more important and relevant. But can information collected by the FEC be used to qualitatively classify winners and losers in US elections?

Our challenge is to build build two different classification models for predicting the outcome of congressional elections in the States using solely the information provided to the FEC by campaigns. We will be ignoring presidential elections as the money used in those campaigns vastly outweighs that used for congressional races, and the number of possible examples to base a classification model on is significantly lower.

Approach:

This data set was downloaded directly from Kaggle and simply combines the publicly disclosed finance data from the FEC with election results from CNN. The FEC data comes in the form of a .csv file with columns representing various (mostly financial) details of the campaign such as “sum of contributions of more than $200” or “net operating expenditure”.

Because this data encompasses many different levels of congressional elections there is a large variation in the features that we will base our models on. This data represents tightly contested battleground-county races that will naturally involve more money and resources. But it also includes data on unknown and unsuccessful campaigns that had very little exposure or support from national parties and local forces. Luckily the size of the data (1600 different contests covered) was probably sufficient in balancing out the kinds of races and the outcomes.

The continuous and wide distribution of feature values in this data, combined with the fact that we were attempting to implement a binary classification algorithm made two

techniques seem particularly well suited for classifying this data. Firstly **k-Nearest-Neighbors** seemed to be an appropriate choice given that winning and losing campaigns would logically cluster is a somewhat predictable way. An unknown data point in the midst of many other *winning* campaigns would likely be a winning campaign, and vice versa for a losing campaign.

An obvious second choice for a classification algorithm came in the form of **Logistic Regression**. Not only is LR a typical and robust method for classifications of this type, but with this data there should be a certain inherent threshold; i.e. it is sensible that past a certain point, more money in a given feature category would correspond to a much more likely victory. The decision boundary of LR is built for that kind of threshold and the reliability and commonness of LR speak to its effectiveness.

	can_id	can_nam	can_off	can_off_sta	can_off_dis	can_par_aff	can_inc_cha_ope_sea	can_str1
1	H2GA12121	ALLEN, RICHARD W	H	GA	12	REP	INCUMBENT	2237 PICKENS RD
2	H6PA02171	EVANS, DWIGHT	H	PA	2	DEM	CHALLENGER	PO BOX 6578
3	H6FL04105	RUTHERFORD, JOHN	H	FL	4	REP	OPEN	3817 VICKERS LAKE DRIVE

After deciding on these techniques there was a fair amount of work necessary to clean and prepare the data for the algorithms. Firstly the data initially contained multiple features that were irrelevant to the classification techniques being used, such as the candidate name and address (above).

Secondly certain features had to be converted from string values to floats in order to be able to plot in hyper-dimensional space for training and testing. Features such as win-status had to be changed from "Y" to 1, and information about whether the campaign was a challenger or incumbent had to be similarly mapped to float values (see right)

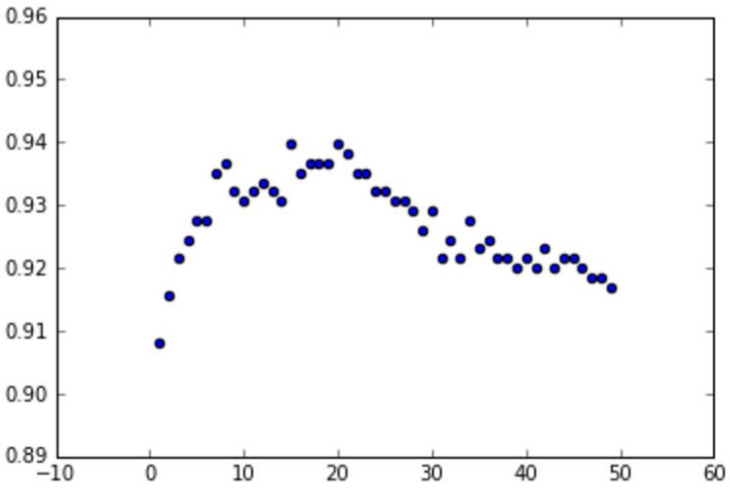
```
def seatStatus(inp):
    if inp == "INCUMBENT":
        return 0
    elif inp == "CHALLENGER":
        return 1
    elif inp == "OPEN":
        return 2
    else:
        return 3
```

Lastly unimportant or redundant feature categories had to be removed or ignored. Examples of such features in this data set include columns such as "Fundraising Disbursement" which is

relevant only to presidential campaigns, or much more commonly - features that were simply sums of the previous two or three columns

Choices made and justifications:

The most important choice in setting up the classifications and before training on the data was deciding which features would be used in classifying. Some were ignored for the reason stated above: they were either sums of other features that were used independently, or the unused component terms of a sum that was used. Other features were ignored because they represented innocuous information about the campaign such as the cost of rented equipment, or details of loan repayment.



The choice of classification algorithms and the feature selection were the only choices made without testing. For all parameters regarding the classification techniques such as the number (k) of nearest neighbors in kNN multiple iterations of the classification were run to find the optimal values.

The figure on the left shows the total accuracy of test data (y-axis) vs the choice of k (x-axis). As expected there is an optimal choice of k that maximizes classification accuracy at 94.23% with a 60-40 split of train-test data. K-values smaller than this lead to a kind of overfitting-by-chance, while k values much larger than 20 tend to underfit the data by taking too many points into consideration to determine the classification.

Very few other choices were necessary when training these methods. Logistic Regression needs no extra parameters and k is the only choice needed in defining kNN.

Solution:

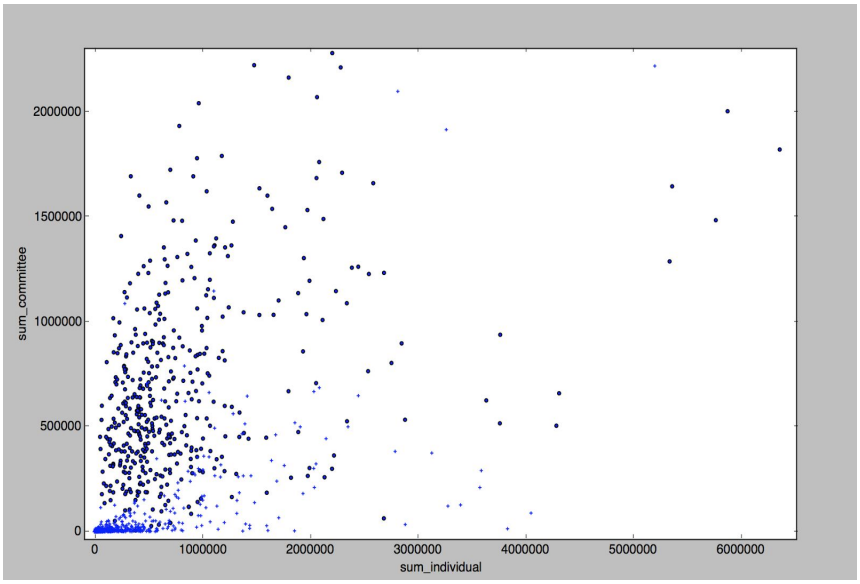
Overall the classification methods were incredibly accurate at predicting outcomes based on the provided data. Both methods had an average accuracy rate of over 90% with Logistic Regression having an average accuracy slightly higher than kNN at 93.8%, with kNN averaging out to 92.6%. This accuracy score is lower than stated above because it reflects the average accuracy score for a k-folds cross validation testing scenario, where instead of splitting the data once into a 60-40 break and training/testing, the data is

segmented and tested multiple times with the accuracy score being the average of the different tests. These results were found using a 5-fold cross validation.

Evaluation of Approach and Solution:

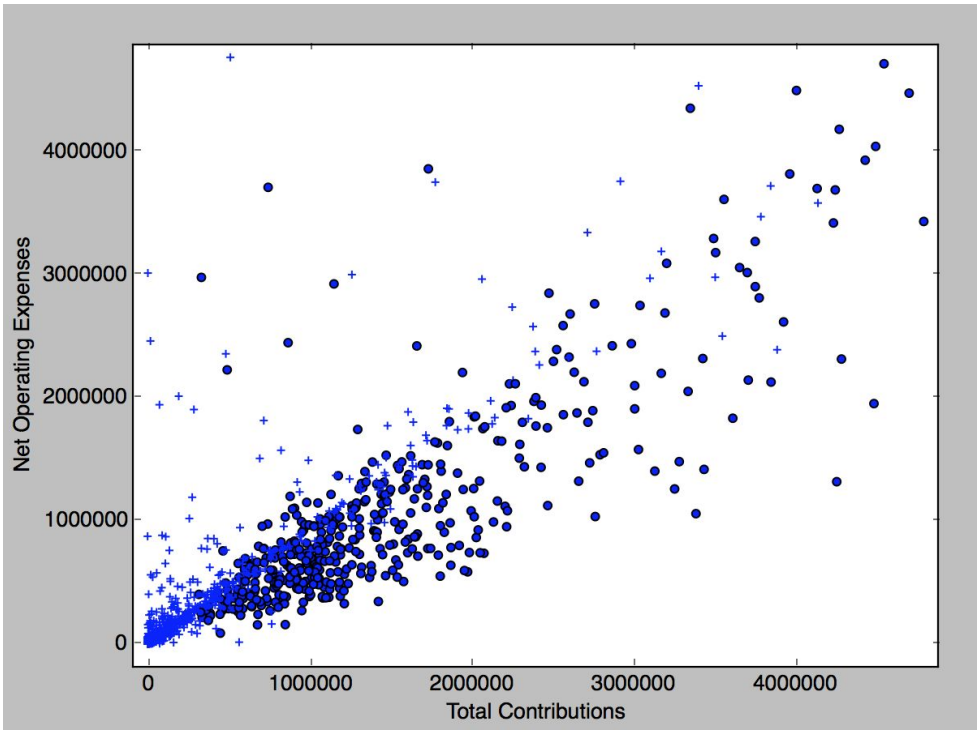
The success of predicting the outcome based on the financial details is both surprising and obvious. The obvious aspect of these predictions is twofold. Firstly it is logical that having more money would correspond to winning an election. More visibility, press, staff, merchandise, etc. all lead to an understandable benefit for the campaign that can finance that. Secondly it is important to note the difference between a correlation and causation in this context. Money does not come from a vacuum; popular and well known candidates are extremely more likely to have more money than small, unknown, or highly unpopular candidates. The amount of money garnered by a candidate can be as much of an indicator of popularity and fame as it is a predictor of success, and the success of said candidate has to be viewed in the frame of those other factors as much as it is in regard to the money.

Still the consistency and success of predicting based solely on financial data is somewhat surprising (and slightly depressing). This outcome states that given just financial information on a candidate there is roughly a **94%** chance of being able to predict whether that candidate will win their congressional election. Don't ever let people tell you that money doesn't matter in US elections.



There are two interesting plots that I think highlight this predictability. To the right is displayed a plot showing committee (such as the national party of a candidate) contribution to a campaign vs individual (such as you and me) contributions. The light blue corresponds to a loss, while the darker blue implies a victory in the race. As can be seen a feature like "committee contribution" has an outsize role in determining a victory compared to individual contributions. In contrast the plot below shows the net total contributions (x) vs the total expenditures (y) to give an idea of the just how much

the average winning candidate tends to consistently have higher numbers in both those categories.



Overall this was an interesting and enlightening look into this side of the electoral process and the predictability of elections. I am pleased with the success of prediction of the classification methods that I chose. However that very same success implies something somewhat unfortunate about the state of campaign finance in the US.