# Predicting Propensity to Vote From Demographic and Labor Force Census Data

Tynan Challenor

## What we are predicting

**Predicting** who is likely to vote in an election has been a perennial struggle for campaigns and media organizations. In 2014, the Pew Research Center found that had they known the makeup of the voter pool for the 2014 mid-term elections their predictions would have swung by 7% in favor of republican candidates. Using demographic and labor force data from the U.S. census in 2016 as well as a label indicating whether or not each individual voted I used logistic regression to predict based on the demographic characteristics whether an individual voted or not; the model returned 85.34% precision and 93.59% recall. I clustered the data by features to look for semantically meaningful groups. An open question remains whether demographic data from one year can predict a subsequent vote later.

## Models and techniques

- Logistic Regression:
  - Hypothesis function: $h_\theta(x) = g(\theta^T x) = \dfrac{1}{1+e^{-\theta^T x}}$
  - Log likelihood:

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} log(h(x^{(i)})) + (1 - y^{(i)}) log(1 - h(x^{(i)})$$

  - Rationale: The logistic regression function is well calibrated, so not only can we leverage the predictions, but we can return the probability that an individual will vote; also valuable[3].

- K-Means:
  - Distortion function: $\sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}} \|^2$
  - Rationale: K-means is an unsupervised learning algorithm that clusters the data by features. Analyzing the groups that k-means returns can provide some interesting demographic trends with regard to voting.

- PCA:
  - Goal: to project the elements of x onto the vector u such that variance of the projection is maximized:

$$u^T \left(\frac{1}{m} \sum_{i=1}^{m} x^{(i)} x^{(i)^T}\right) u$$

  - Rationale: After performing k-means, PCA was used to reduce the data to two dimensions in order to view the outcome of the clustering

## Data and features

**The dataset** comes from a U.S. census bureau's survey from November of 2016, in which they added an addendum after the election asking whether respondents had voted[2]. The data came with 380 features, semantically weighted toward labor force information, such as whether you were looking for job, how you were going about looking, and ability to work were you to be offered a job. There were a total of 152,095 responses; for each the last 4 features were removed because they included direct questions about voter registration, leaving a total of 376 features.

## Logistic regression results

**Upon** first running the raw data from the census bureau, logistic regression returned 100% recall and 99.99% precision. Since then four features have been removed corresponding to the following questions:

1. Which of the following was the main reason you were not registered to vote?
2. What was the main reason you did not vote?
3. Did you vote in person or by mail?
4. Did you vote on election day or before?

These questions could only be asked after voting. The model now returns 93.59% recall and 85.34% precision.

## Logistic regression coefficients

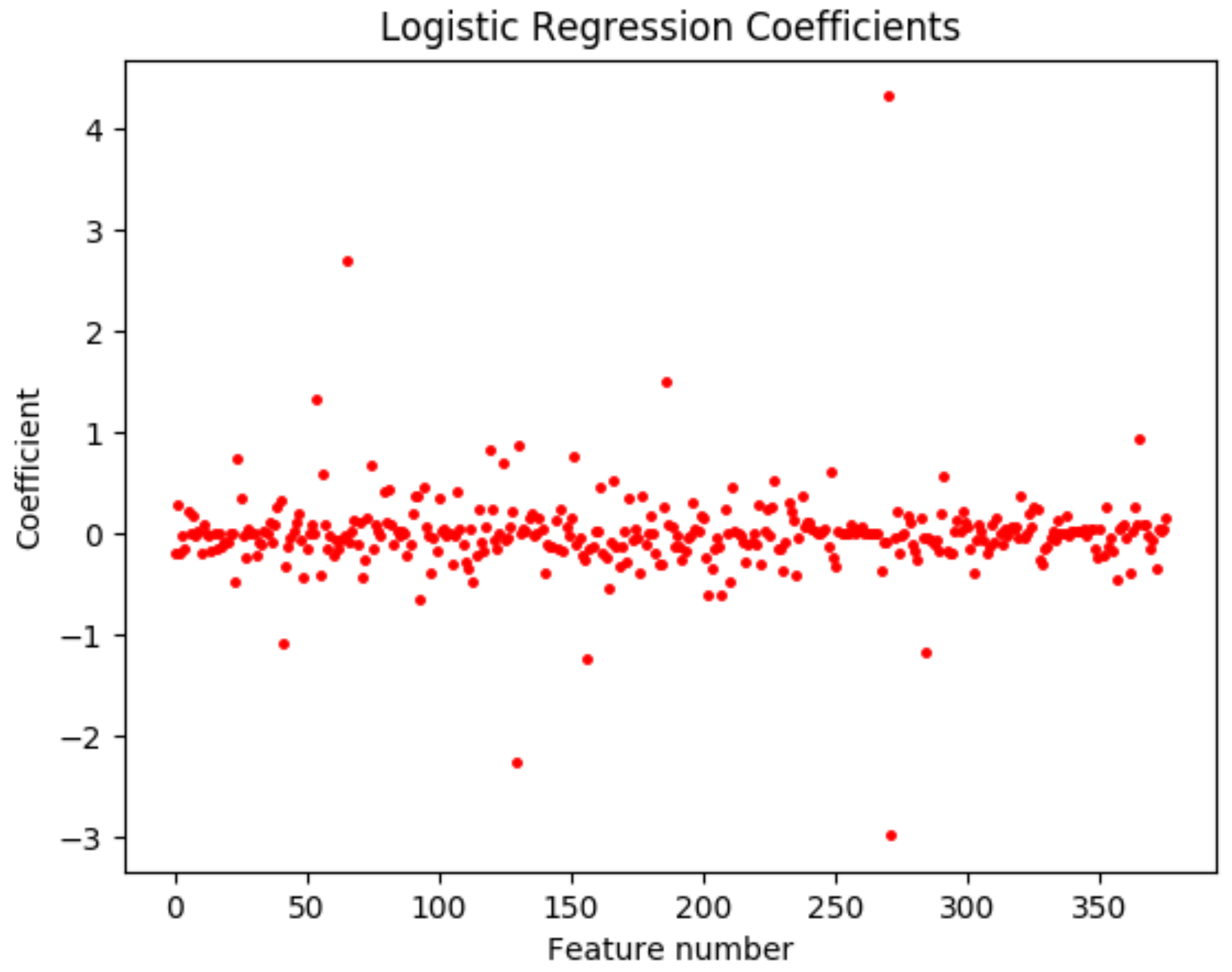

Figure 1. While many of the features have coefficients right around zero, there are a handful with significantly larger values who's semantic meaning from the census survey may help us understand what the model is learning.

| Coefficient (Abs. value) | Semantic Meaning |
|---|---|
| 4.32 | Do you usually work 35 hours or more per week? |
| 2.98 | What is the number of hours you actually worked? |
| 2.70 | Last week you were on layoff from a job – follow up note. |

**Table 1.** The semantic meaning of the features with the highest coefficients could be valuable for ascertaining how to go about collecting useful data in the future for making such a model. The behavior of this model so far has seemingly indicated that economic pressures are the highest indication of voting, although were enriched for labor force enquiries as compared with other demographic indicators such as race or voting district.

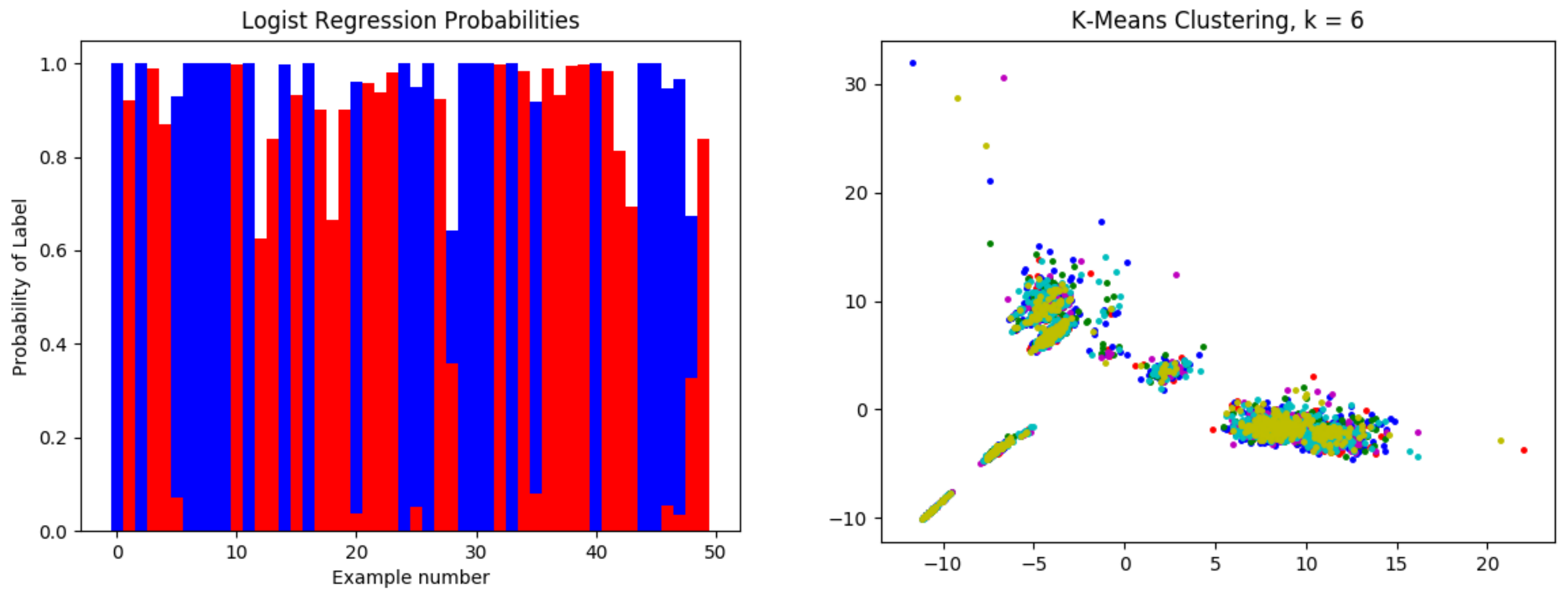

Figure 2, 3: On the left, the probabilities for the first 50 examples, useful for further work in likely voter model predicting. On the right a k-means clustering of the data with k=6. First the data were clustered using k-means and then plotted after shrinking the dimensionality to two dimensions with PCA. There do not appear to be any useful patterns from k-means at this stage of the process.

## Discussion and future directions

**Although** the logistic regression model is by no means perfect, with an untargeted set of demographic indicators that were skewed heavily toward questions of employment, the model was still able to predict with moderately high precision and recall whether an individual would vote. Tuning this model using the 2016 census bureau survey and an indication of voting in the 2016 election would be a nice step to predicting likely voters. Census bureau data is easy to access (if difficult to parse), and it is publically available data such as this that might help campaigns more effectively target the voters they need to.

For the near future, this model can be improved by reading through the semantics of the census data and whittling down the features. It would also be useful to keep working on a clustering algorithm that might shed light to how features are grouping. In the long term, I would want to extend the model to use demographic data from 2016 to predict voting in other elections – like a 2017 senatorial special election for example. How general can we make the model?

## Contact

Tynan Challenor
Email: tynan@stanford.edu
Phone: 510 861 6264

## References

1. 1615 L. St NW, S. 800 Washington, and D. 20036 U.-419-4300 | M.-419-4349 | F.-419-4372 | M. Inquiries, "Can Likely Voter Models Be Improved?," *Pew Research Center*, 07-Jan-2016. .
2. "Current Population Survey FTP Page." [Online]. Available: https://thedataweb.rm.census.gov/ftp/cps_ftp.html. [Accessed: 11-Dec-2017].
3. N. Silver, "How The FiveThirtyEight Senate Forecast Model Works," *FiveThirtyEight*, 17-Sep-2014.
4. N. Silver, "The Real Story Of 2016," *FiveThirtyEight*, 19-Jan-2017. .

.