Capstone Final Report

Leveraging Data Science to promote transparency in elections: The case of Peru

1. **Project problem statement:**
   I will use Data Science to investigate if the controversy around the results of the run-off presidential election in Peru has merits. I will answer if the outcome of the election would have changed had the votes from *alleged controversial* polling stations had been discarded.

2. **Background:**
   Keiko Fujimori and Pedro Castillo participated in the run-off election on June 6th 2021. The results put Pedro Castillo ahead by 44,240 votes. Fujimori has protested these results in electoral court, contesting that hundreds of ballot summary sheets are irregular (i.e. signatures do not correspond to poll workers or are falsified).

   The National Elections Bureau, ONPE, runs a manual process that relies on the honor code to succeed. It assigns three poll workers randomly from the voters from that station. Paper ballots are discarded after they are tallied by the poll workers. These return a handwritten ballot summary sheet to ONPE for tabulation. The sheet is signed by the three poll workers and a representative from each party. Fujimori's surrogates have claimed irregularities in regions where they did not have representatives.

3. **Problem from the users perspective:**
   As an ONPE official, I want to use data science methods to provide the public with transparency, expediency and confidence regarding the results of elections.

4. **Datasets:**
   ONPE published the official results for the first round and run-off elections in csv format as part of the government's digital transparency initiative. The links to the datasets are in the EDA notebook.

5. **EDA:**
   I will use regression analysis for this project. My models need to have voting information from the same polling station for both rounds. EDA was centered on achieving this. Both data sets started with 86488 observations, each corresponding to an individual polling station. Each polling station has a maximum of 300 eligible voters.

   Three levels of EDA were performed on both sets:
   - First level: dropping redundant or empty variables, renaming variables and features of categorical values in English.
   - Second level: consistency checks for null values, duplicates, eligible voters and total voters
   - Feature engineering:
     - Add a regional variable based on state
     - Drop all observations with annulled ballot summary sheets. Some of these observations were easy to track (under the categories "Annulled" or "Did not open" of Ballot_summary_cond). Some were hidden in the "In Process" category.

   After EDA, I merged both clean data sets. I performed some additional feature engineering:
   - Create delta variables for eligible voters and voter turnout subtracting second from first round totals.

   - Drop all observations where the second round eligible voters where not the same as the first round eligible voters. Though the voter registry is supposed to remain close during the election, 400+ polling

places showed a difference between the rounds. Since I'm using the turnout delta as feature, no change in the number of votes should be due to the changing size of the possible pool of voters.

The final result is a merged dataset with 84452 observations, or 97.7% of the initial total.

6. Modelling:

6.1 Linear Regression Analysis:

This analysis seeks to highlight outliers as a proxy for *irregularities.* The proposed methodology is to contrast expected vote count versus actual vote count for each polling place for both candidates. The target variable is votes for each candidate in the second round and the features are the votes for all 18 candidates in the first round and the delta of turnout between the first and second round.

Two step approach to classifying an outlier:

- *A polling place where the residual from the actual value minus the predicted vote belongs to the 2.5th percentile of the residual distribution (i.e. is highly negative) and, simultaneously is in the 2.5th percentile of the distribution of votes for the candidate (i.e. is a low amount of votes).*

Hence, a polling station with a residual of -30 expected votes but 150 votes is less relevant than the same -30 where a candidate obtained only 2 votes. This is to proxy *for potential irregularity from changing the actual results in the ballot summary sheet.*

I constructed the following models:

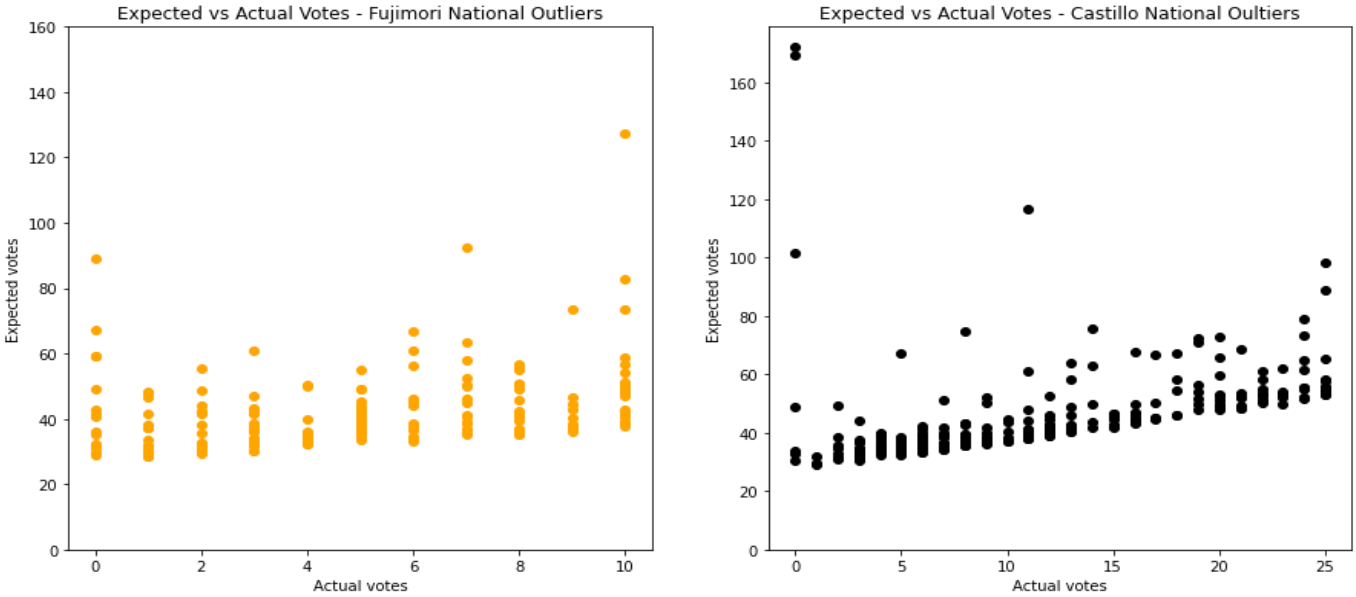
- Fujimori at the national level: one regression using all observations
- Castillo at the national level: one regression using all observations
- Fujimori at the regional level: five regressions Coast, South, Central, East, Abroad. This is an acid test. Regional models will have broader error distributions and will include more outliers than the national model.

Model	**R2**	**N° of outliers**	**Net total votes for Fujimori if annulled**	**Sum of residuals**
Fujimori National	**0.924**	**184**	**23,826**	**6910**
Fujimori Regional		**212**	**25,955**	**6995**
Fujimori Coast	0.832	71		
Fujimori South	0.921	24		
Fujimori Central	0.872	60		
Fujimori East	0.795	44		
Fujimori Abroad	0.703	13		
Castillo National	**0.902**	**265**	**-25,323**	**9,069**

If the votes from these outlier polling stations were annulled in favor of Fujimori, the results of the election would not change. Likewise, it would be arbitrary to do this and not annul Castillo's outliers. Furthermore, if we take into account the sum of residuals (that is the total votes that the model predicted but did not happen), Fujimori is under the expected vote by less than 7000 votes, while Castillo by 9000. The accompanying Tableau file shows little regional concentration, which counters any theory of a concerted effort to overturn results at the polling stations.

Figure 1 shows the outliers for both candidates versus the actual votes. We see that both candidates have observations where actual votes are very low (below 6), yet expected votes are in the 30s. There are a few outliers for both candidates where the expected vote was much higher than the actual vote, but for the most part outliers are compressed between 20 and 40 expected votes. Furthermore, the 2.5th percentile of actual votes for Fujimori (10) is much lower than Castillo's 25, i.e. there are quite a few polling places with very low vote.

Figure 1

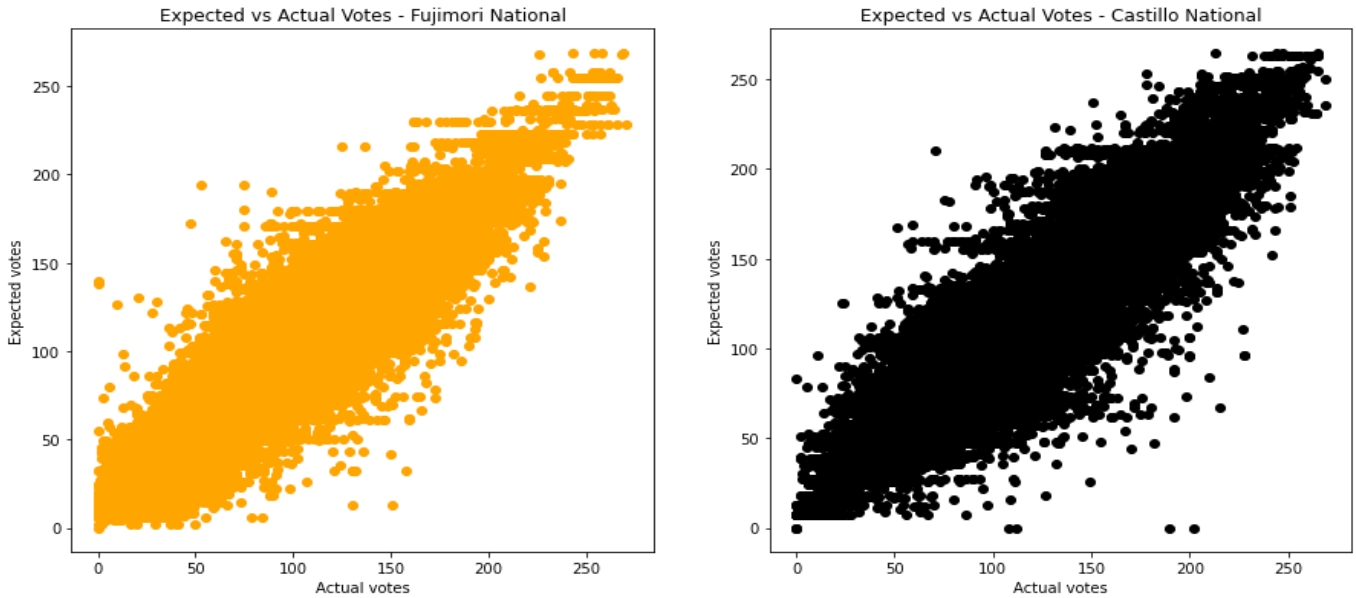


It's important to caveat that the occasional statistically insignificant variable was left to maximize R2. Additionally, errors of all models proved to be non-normal. The tail ends of the distributions proved too broad. Furthermore, errors were not homoscedastic, with higher expected values closer to actual values than lower ones.

6.2 Decision Tree Regressor

I applied an alternative approach to linear regression by using a Decision Tree Regressor with the same targets and features. I used the hyperparameter optimization method with a 5-fold validation, maximizing a depth of 10 nodes for the Fujimori national model and 11 nodes for the Castillo national model. The R2 of both improved when compared to linear regression, with Fujimori reaching 94.2 and Castillo 93.2. Furthermore, when projecting the expected vs actual votes for both distributions, the same result arises. There are outliers on both sides and, in some cases, Castillo's are more notorious.

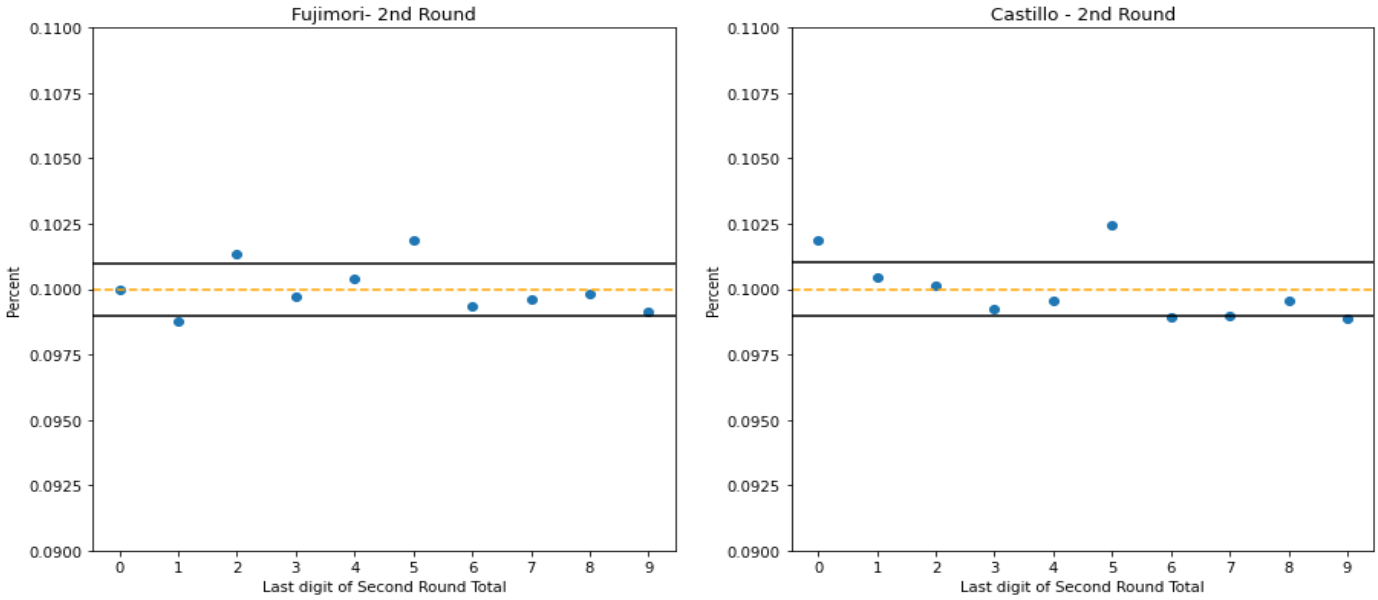
Figure 2



6.3 Applying Benford's Law to the second round

To account for the loss of nearly 2000 datapoints from the second round in the regression analysis, I used the framework of Beber & Scacco (2012). They posit that a fair election should produce returns where last digits occur with equal frequency, since laboratory experiments indicate that individuals tend to favor some numerals over others, even when subjects have incentives to properly randomize. Since in the Peruvian election, ballot summary sheets are, ultimately, handwritten, any systematic bias from poll workers "inventing" the numbers could potentially show up under this lens. However, this was not the case. Applying Benford's law to the last digit of total votes for both candidates in the run-off we see very similar behavior. Numbers hover around the confidence interval, though 5 is slightly above in both cases.

Graph 3



7. Summary of business applications:

As elections are face higher scrutiny from social media, it's becoming easier to disseminate biased or non-accurate information. A gov-tech company can specialize in data science methods to improve data collection, processing, tabulating, and validation of the results in real time and provide this service worldwide. Such effort can provide higher transparency, expediency, and confidence in the results to alleviate mistrust.

For the particular case of Peru, ONPE could install statistical checks such as the ones performed here to validate results in real time. They could do this with historical information to serve as a benchmark. Most importantly, they could have the tools to preempt the social media onslaught with facts.

Total words: 1402