# Detecting Potential Election Fraud with Machine Learning

***Benjamin Juarez***

Mentor: Professor R. Michael Alvarez

Caltech

## INTRODUCTION

- Ensuring the integrity of elections is essential for maintaining the values of democracy in many nations
- With a demand for a better mechanism for identifying election fraud, the use of machine learning models has arisen as a technique for analyzing data from elections
- In this project, the classification of voting locations from a recent election as either clean or at-risk has indicated that there are certain areas appearing to show signs of potential fraud

## HISTORICAL CONTEXT

- Event under analysis: 2019 presidential election in Bolivia
- This election turned controversial very quickly as late-counted votes initially appeared to show evidence of fraud which favored the incumbent Evo Morales (representing the MAS political party) [3]
- Morales was subsequently accused of fraud by his contestant, Mesa (representing the CC political party)
- Resulted in instability and violence along with the resignation and fleeing of Morales to Mexico
- Recent studies of the data from this election are finding new conclusions
- This project provides another perspective on this controversial election

## DATA

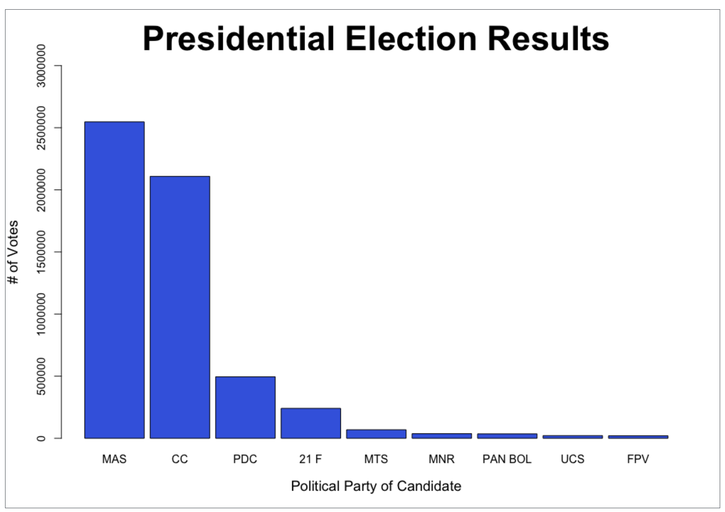


Fig. 1: **This figure shows the official presidential election results from Bolivia in 2019**

- *Election details*: 5,296 precincts containing 34,555 voting booths (*mesas*)
- *Data*: 2019 & 2014 presidential election data, demographic info [2, 4]

## SYNTHETIC DATA GENERATION

***Clean, Synthetic Data***

- Integrated department-level data from Bolivia's 2014 presidential election with components of relevant demographic information through mixed effects regression model
- Predictions of turnout rates and vote shares for Morales (MAS) were used to construct the clean, synthetic dataset

***Manipulated, Synthetic Data***

- Separate parameters defined the probability and extent of either vote stealing or ballot box stuffing occurring in a *mesa*, simulating fraudulant data
- *Mesas* were then labelled accordingly as either clean, at-risk of BBS, or at-risk of VS, producing the necessary labeled, training dataset [1]

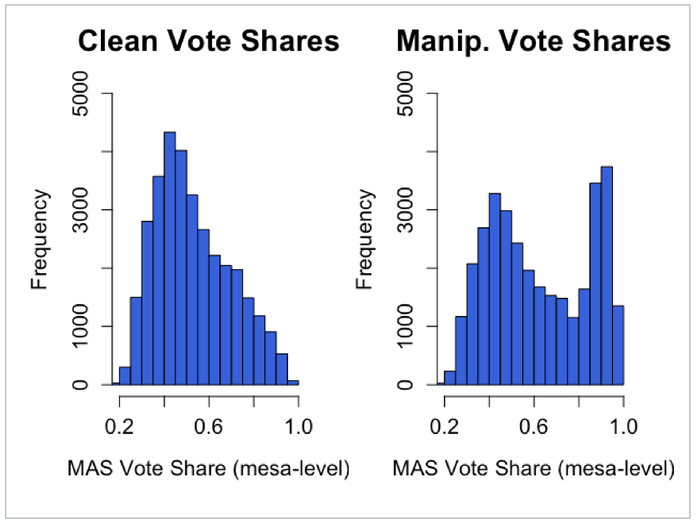


Fig. 2: **Comparison of vote share distributions between clean and manipulated synthetic data**

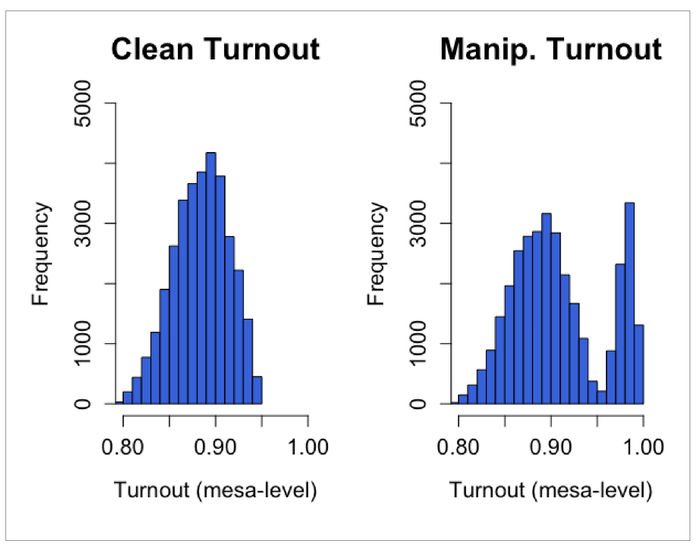


Fig. 3: **Comparison of turnout rate distributions between clean and manipulated synthetic data**

## MACHINE LEARNING IMPLEMENTATION

- Accuracy of the Random Forest model was determined by training the model on a portion of the labeled, training data and subsequently testing the model on the rest of the training data
- Performance test showed that the model correctly categorized about **98%** of the remaining portion of the training data
- With turnout and MAS vote shares as accurate predictors, the model was then trained on the full training dataset, followed by the implementation of the model with the real data to discover any possibly anomalies

## RESULTS

- **84%** of the *mesas* classified as clean with **16%** classified as at-risk

| | Clean | At.Risk | BBS.Risk | VS.Risk | Avg.Turnout | Avg.MAS.vote |
|---|---|---|---|---|---|---|
| All Mesas | 84.30 | 15.70 | 11.10 | 4.60 | 0.90 | 0.46 |

Fig. 4: **This table displays the predictions from the Random Forest model.**

| Department | Clean | At.Risk | BBS.Risk | VS.Risk | Avg.Turnout | Avg.MAS.vote |
|---|---|---|---|---|---|---|
| Cochabamba | 78.20 | 21.80 | 10.10 | 11.70 | 0.90 | 0.57 |
| La Paz | 82.50 | 17.50 | 14.40 | 3.00 | 0.92 | 0.53 |
| Potosí | 83.00 | 17.00 | 7.90 | 9.10 | 0.89 | 0.46 |
| Santa Cruz | 86.20 | 13.80 | 12.20 | 1.60 | 0.89 | 0.34 |
| Oruro | 88.50 | 11.50 | 9.40 | 2.10 | 0.91 | 0.47 |
| Tarija | 90.30 | 9.70 | 6.80 | 2.90 | 0.87 | 0.39 |
| Chuquisaca | 92.00 | 8.00 | 2.30 | 5.80 | 0.88 | 0.40 |
| Pando | 95.80 | 4.20 | 2.70 | 1.50 | 0.84 | 0.43 |
| Beni | 96.80 | 3.20 | 2.00 | 1.20 | 0.86 | 0.34 |

Fig. 5: **This table displays the predictions from the Random Forest model for each department.**

## CONCLUSIONS

- Ability to synthetically generate datasets which simulate some event under analysis will continue to be an essential component in research, particularly in the field of election forensics
- Project results bring about some cause for concern since a fair portion of *mesas* were classified as at-risk
- Further work would allow for the implementation of other methods along with a deeper analysis of at-risk locations

## REFERENCES

[1] Alvarez, Levin, and Zhang. "Election forensics: Using machine learning and synthetic data for possible election anomaly detection". In: (Oct. 2019).

[2] *City Population: Bolivia*. 2020 (accessed June 2020 - Sept 2020). URL: https://www.citypopulation.de/en/bolivia/.

[3] Idrobo, Kronick, and Rodriquez. "Do shifts in late-counted votes signal fraud? Evidence from Bolivia." In: (June 2020).

[4] *Órgano Electoral Plurinational (OEP)*. 2020 (accessed June 2020 - Sept 2020). URL: https://www.oep.org.bo.