Election Forensics using Machine Learning

Ben Juarez

Brief Context

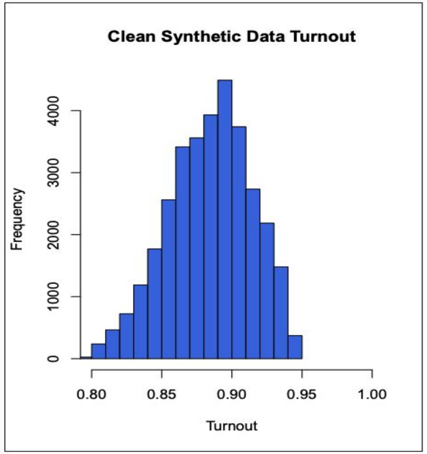
- Revisiting research project conducted in summer of 2020
  - Summer after my freshman year
  - Supervised by Prof. Alvarez through Caltech's SURF program
- Election under investigation: **2019 Bolivian presidential election**
  - Controversial election that had dramatic political consequences
  - Calls for fraud in favor of incumbent, Evo Morales (MAS)
  - Alleged fraud related to late-counted votes
  - Subsequent studies can explain trends without invoking fraud

Overview

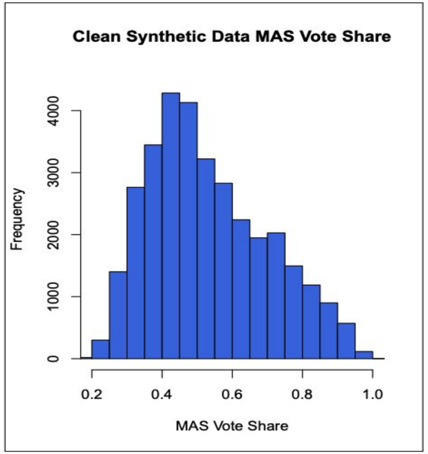
- Research goal = provide alternative perspective to election using machine learning
    - Investigate presence of anomalies that potentially resemble election fraud
    - My first experience with machine learning
- Supervised machine learning model (Random Forest)
    - Used to classify voting booths (mesas) according to potential risk of fraud
        - Clean
        - At-risk of vote stealing (VS) - higher vote shares for MAS
        - At-risk of ballot box stuffing (BBS) - higher turnout rates

Synthetic Training Data

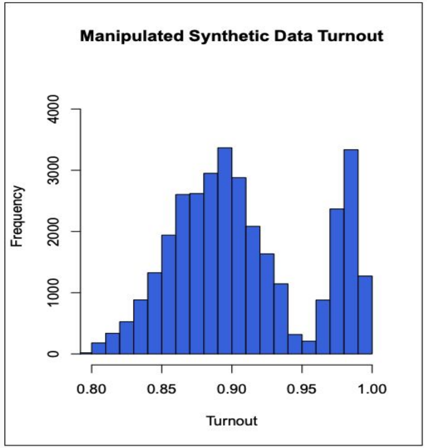
- Mix of 2014 presidential election data & demographic data (Clean data)
  - For each voting booth, generated predictions for turnout rate and MAS vote share
- Generated manipulated, labelled data by simulating VS and BBS
  - Based on techniques used in previous election forensics studies



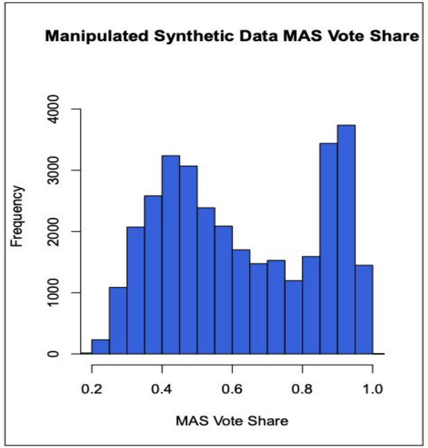
(A) Turnout



(B) MAS Vote Share



(A) Turnout



(B) MAS Vote Share

Past Results

- Validation steps showed that RF model was 98% accurate on training data
  - Trained on 90% of data, tested on remaining 10%
- Trained on full manipulated, labelled data, tested on 2019 data
  - 5,296 precincts containing 34,555 voting booths
  - 84% of voting booths classified as clean (16% at risk)
    - 11% classified as at risk of ballot box stuffing (BBS)
    - 5% classified as at risk of vote stealing (VS)
- Results broken down by department and municipality
- Ran out of time to deeply analyze results, especially within at-risk regions
  - No comparisons with other algorithms

	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

TABLE 3. **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

TABLE 4. **This table displays the predictions from the Random Forest model for each department.**

Municipality	Clean	At.Risk	BBS.Risk	VS.Risk	Avg.Turnout	Avg.MAS.vote	Mesas
Entre Ríos	2.80	97.20	33.30	63.90	0.93	0.91	108
Villa Tunari	8.10	91.90	40.10	51.80	0.94	0.94	197
Puerto Villarroel	8.60	91.40	36.40	55.00	0.93	0.91	151
Tapacarí	11.90	88.10	19.00	69.00	0.91	0.93	42
Palca	14.70	85.30	8.80	76.50	0.92	0.87	34
Totora	18.20	81.80	21.20	60.60	0.92	0.89	33
Cocapata	24.10	75.90	20.70	55.20	0.90	0.86	29
Sacaca	27.30	72.70	45.50	27.30	0.94	0.85	33
Laja	34.00	66.00	60.00	6.00	0.95	0.78	50
Sapahaqui	35.70	64.30	14.30	50.00	0.92	0.84	28
Pucarani	38.40	61.60	56.20	5.50	0.94	0.71	73
Tiraque	38.50	61.50	7.70	53.80	0.91	0.88	52
Colquechaca	40.00	60.00	10.00	50.00	0.89	0.79	40
Pocoata	45.70	54.30	20.00	34.30	0.89	0.81	35
San Pedro	46.50	53.50	4.70	48.80	0.86	0.84	43
Chimoré	47.70	52.30	27.70	24.60	0.92	0.86	65
Mecapaca	52.50	47.50	30.00	17.50	0.93	0.77	40
Calamarca	52.90	47.10	38.20	8.80	0.94	0.73	34
Shinahota	53.10	46.90	31.20	15.60	0.91	0.86	64
Chayanta	53.30	46.70	26.70	20.00	0.92	0.76	30
Arbieto	59.50	40.50	8.10	32.40	0.87	0.82	37
Colquiri	59.50	40.50	32.40	8.10	0.93	0.68	37
Sicasica	60.00	40.00	14.00	26.00	0.93	0.77	50
Ayopaya	61.40	38.60	13.60	25.00	0.88	0.81	44
Colomi	64.30	35.70	7.10	28.60	0.91	0.79	56
Tiahuanacu	64.30	35.70	14.30	21.40	0.93	0.76	28
Aiquile	67.90	32.10	3.60	28.60	0.88	0.75	56
Mizque	68.10	31.90	6.40	25.50	0.89	0.77	47
Achocalla	72.90	27.10	25.70	1.40	0.93	0.71	70
Betanzos	73.60	26.40	1.90	24.50	0.86	0.73	53
Patacamaya	74.50	25.50	12.70	12.70	0.90	0.68	55
Batallas	75.60	24.40	20.00	4.40	0.93	0.68	45
Capinota	77.80	22.20	14.80	7.40	0.91	0.70	54
Sipesipe	77.80	22.20	14.60	7.60	0.91	0.71	144
Coripata	78.00	22.00	22.00	0.00	0.92	0.46	50
El Alto	78.10	21.90	21.90	0.00	0.93	0.55	3022
Copacabana	78.40	21.60	8.10	13.50	0.92	0.74	37
Irupana	78.60	21.40	19.00	2.40	0.91	0.49	42

TABLE 5. **This table displays the predictions from most at-risk municipalities from the departments of Cochabamba, La Paz, and Potosí.**

New Goals

- Compare results across various machine learning algorithms
  - XGBoost, Gradient Boosting, Logistic Regression, …
  - Find “best” method through investigation of results
- More thorough analysis into results
  - Analyze most at-risk locations + explore explanations for trends
  - More detailed graphical results (heat maps, etc.)
  - Examine correlation between vote stealing and ballot box stuffing
- Re-examine process for generating synthetic data
- Emphasis on providing alternative perspective to election
  - Fraud allegations generally considered to be inaccurate

Moving Forward

- Currently working with generating results from other algorithms
- More research into why certain models may perform better than others
- Produce more detailed and complex visualizations of results from best predictors
- Develop deeper understanding of rationale behind potential anomalies (demographic, historical, etc.)
- Investigate relationship between vote stealing & ballot box stuffing