Modeling Political Contributions

DS 5559 - Final Project
Group 5

Brad Howlett (bth2g)
Eric Larson (rel4yx)
Hanim Song (hs4cf)

Overview

- The purpose of this project was to explore the relationship between campaign contributions and the outcome of U.S. elections.

- We narrowed this research question to focus on campaign contributions from **individuals** as it relates to predicting **binary** win/loss outcomes for **2016 + 2018 U.S. House of Representative** races.

- **Result:** We were able to build a classification model with **91-92%** accuracy, precision, and recall. Given this approach ignores policy + other factors, we considered this to be a success.

Agenda

- **The Data**
- The Preprocessing + Cleaning
- The Model
- The Evaluation
- The Conclusion

The Data

- The data for this project was sourced from the Federal Election Committee (“*FEC*”) bulk data website and the MIT Election Data and Science Lab.

- FEC Data: the majority of our information
  - Political candidate information by year
  - Political committee information by year
  - Individual contribution information by year (**>$200** for inclusion)

- MIT Data: feature enrichment and to determine race winner
  - U.S. House of Representatives Election Results

The Data

- The FEC Data:
    - Files organized into two year cycles (*2015-16, 2017-18*)
    - Flat text files '|' delimited, separate .csv header file
- Total rows and important features in the files:

42 million, ~7 GB	**13,615**	**14,774**
Individual Contributions	**Committees**	**Candidates**
CMTE_ID	**CMTE_ID**	**CAND_ID**
NAME	Committee Name	Candidate Name
ADDRESS	State	Office *(P, S, H)*
TRANSACTION_AMT	Party Affiliation	Year of Election
STATE	...	State
ZIP_CODE	**CAND_ID**	District
OCCUPATION		Status *(I, C, O)*
		...

The Data

- Example: **the Candidates file**
  - Flat text file

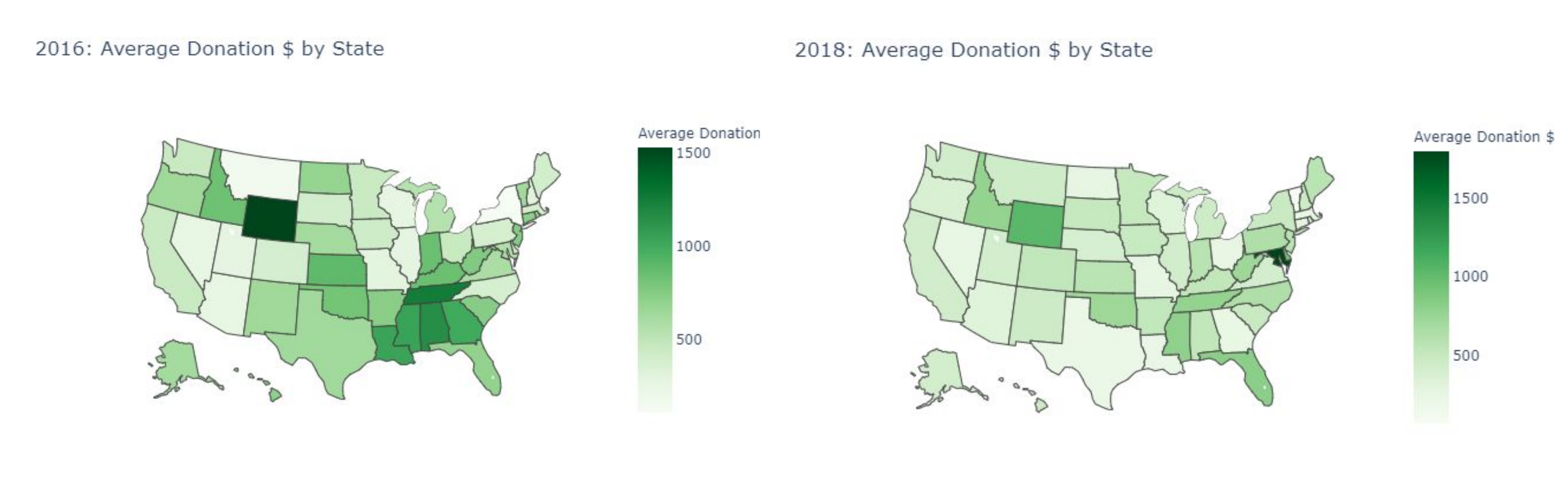
```
H0AL02087|ROBY, MARTHA|REP|2018|AL|H|02|I|C|C00462143|3260 BANKHEAD AVE||MONTGOMERY|AL|361062448
H0AL03192|THOMPSON, HANNAH|DEM|2020|AL|H|03||N|C00681452|2181 N BROADWAY||ALEXANDER CITY|AL|35010
H0AL05049|CRAMER, ROBERT E "BUD" JR|DEM|2008|AL|H|05||P|C00239038|PO BOX 2621||HUNTSVILLE|AL|35804
H0AL05163|BROOKS, MO|REP|2018|AL|H|05|I|C|C00464149|7610 FOXFIRE DR.||HUNTSVILLE|AL|35802
H0AL06088|COOKE, STANLEY KYLE|REP|2010|AL|H|06|C|N|C00464222|723 CHERRY BROOK ROAD||KIMBERLY|AL|35091
H0AL06104|ALLEN, ANDERS POPE|REP|2020|AL|H|06||N|C00681213|123 KATY CIRCLE||BIRMINGHAM|AL|35242
H0AL07086|SEWELL, TERRI A.|DEM|2018|AL|H|07|I|C|C00458976|P.O. BOX 1964||BIRMINGHAM|AL|35201
H0AL07094|HILLIARD, EARL FREDERICK JR|DEM|2010|AL|H|07|O|P|C00460410|PO BOX 12804||BIRMINGHAM|AL|35202
H0AR01083|CRAWFORD, ERIC ALAN RICK|REP|2018|AR|H|01|I|C|C00462374|34 CR 455||JONESBORO|AR|72404
H0AR01091|GREGORY, JAMES CHRISTOPHER|DEM|2010|AR|H|01|O|N|C00472126|510 S LILLY ST||BLYTHEVILLE|AR|72315
H0AR01109|CAUSEY, CHAD|DEM|2010|AR|H|01|O|P|C00475384|205 SOUTH MAIN #203||JONESBORO|AR|72401
H0AR01125|SMITH, PRINCELLA D|REP|2010|AR|H|01|O|P|C00480905|2000 WYNRIDGE COVE||WYNNE|AR|72396
H0AR03022|SKOCH, BERNARD KURT 'BERNIE'|REP|2010|AR|H|03|O|P||21142 KIRKSEY ROAD||ELKINS|AR|72727
H0AR03030|WHITAKER, DAVID JEFFREY|DEM|2010|AR|H|03|O|P|C00468033|PO BOX 957||FAYETTEVILLE|AR|727020957
H0AR03055|WOMACK, STEVE|REP|2018|AR|H|03|I|C|C00477745|91 WOODRIDGE LANE||ROGERS|AR|727563078
H0AZ01184|FLAKE, JEFF MR.|REP|2012|AZ|H|06|C|P|C00347260|4222 E MCLELLAN CIRCLE|UNIT 19|MESA|AZ|852053119
H0AZ01259|GOSAR, PAUL DR.|REP|2018|AZ|H|04|I|C|C00461806|PO BOX 2967||PRESCOTT|AZ|86302
H0AZ01333|GRESSLEY, FORREST DAYL|REP|2010|AZ|H|01|C|N|C00481267|1545 E STIRRUP CT||GILBERT|AZ|85296
H0AZ02166|SCHMIDT II, JAMES A MR.|REP|2020|AZ|H|02||N||P.O. BOX 286|4751 EAST JACKALOPE ROAD|DRAGOON|AZ|85609
H0AZ03248|SCHARER, GENE PAUL|DEM|2018|AZ|H|08|O|N|C00518381|655 WEST 221 DRIVE||BUCKEYE|AZ|85326
```

  - Header .csv file

CAND_ID	CAND_NAME	CAND_PTY_AFFILIATION	CAND_ELECTION_YR	CAND_OFFICE_ST	CAND_OFFICE	CAND_OFFICE_DISTRICT	CAND_ICI	CAND_STATUS	CAND_PCC

The Data

- Visuals: **Individual Contributions by state by period**



The Data

- The MIT Data:
  - Excel file capturing U.S. House of Representative results from 1976-2018
  - Used the Candidate Votes and Total District Votes to determine race winner
- Total rows and important features in the file:

29,637

U.S. House Races
Year of Election **Candidate Name** State District Candidate Votes Total District Votes ...

The Data

- The combined files used in the project
  - Individuals donate to -> Committees (Inner join on CMT_ID)
  - Committees spend money for -> Candidates (Inner join on CAND_ID)

Individual Contributions	Committees	Candidates	U.S. House Races
CMTE_ID	**CMTE_ID**	**CAND_ID**	Year of Election
Name	Committee Name	**Candidate Name**	**Candidate Name**
Address	State	Office *(P, S, H)*	State
TRANSACTION_AMT	Party Affiliation	Year of Election	District
...	...	State	Candidate Votes
	CAND_ID	District	Total District Votes
		Status *(I, C, O)*	...
		...	

Agenda

- The Data
- **The Preprocessing + Cleaning**
- The Model
- The Evaluation
- The Conclusion

The Preprocessing + Cleaning

- With our population of data defined, we then worked to combine and filter down to the features and target for modeling.

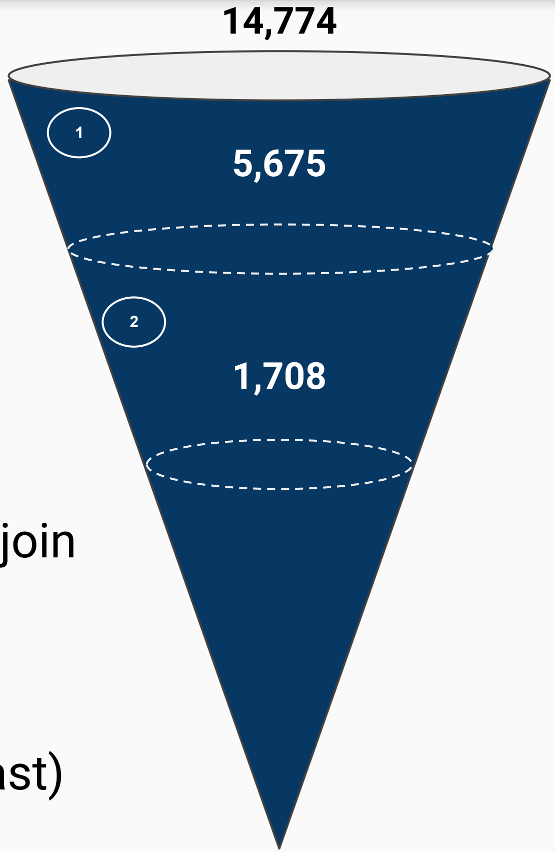
- Target:
  - Candidates for U.S. House of Representatives in 2016 + 2018, **AND**
  - Candidates linked to a Committee for Individual Contributions, **AND**
  - Races we have more than one Candidate, **AND**
  - Races we can define the winner by majority vote

- Features:
  - Categorical: Candidate Status (incumbent (1) and challenger/open(0))
  - Continuous: Individual Contributions
    - Candidate: Total, Count, Average, Large (+ Rel), Out-of-State (+ Rel)
    - Race Relative: Total, Count, Average, Large, Out-of-State

The Preprocessing + Cleaning

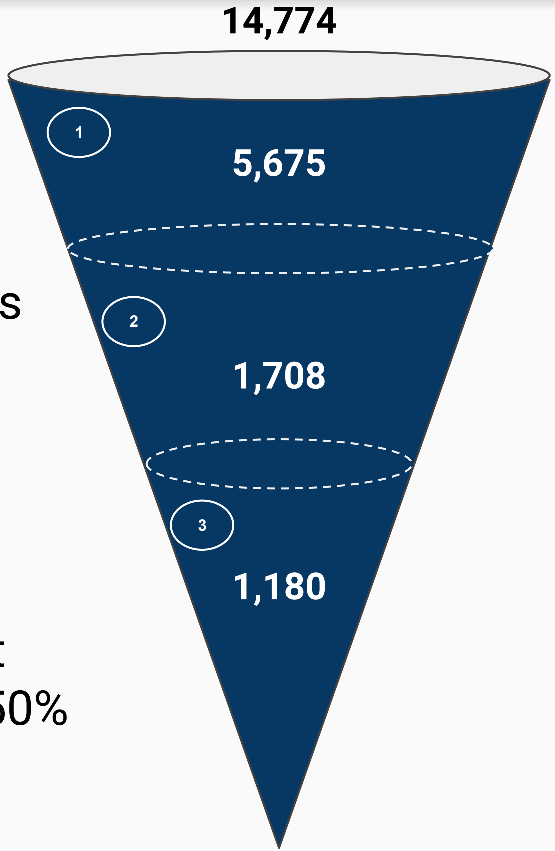
- **Target**:
  - Step 1:
    - Join all candidates from *2015-16* and *2017-18*
    - Filter where race = ‘H’
    - Split + Filter by race year = ‘2016’ or ‘2018’
    - Filter duplicates by CAND_ID

  - Step 2:
    - **fuzzywuzzy** string match Candidate names to join MIT data with FEC data; imperfect approach
      - Scored string comparison out of 100
      - Required manual review
        - Many issues (First, Middle == First, Last)
        - Identical names, different districts



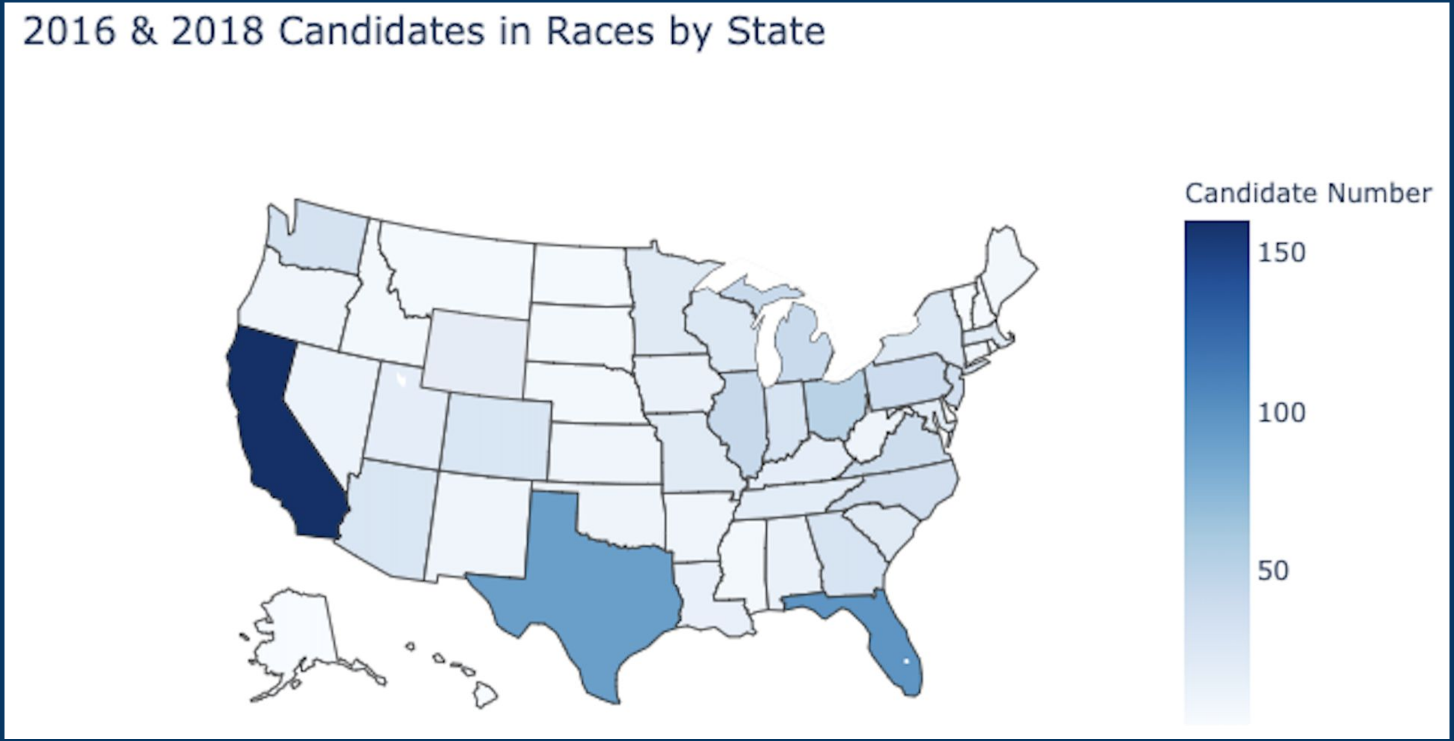
The Preprocessing + Cleaning

- **Target**:
  - Step 3:
    - Clean up resulting Candidates based on each unique race, defined as Year+State+District
      - Filter races if no candidate had +50% votes
      - Filter races where we couldn't connect contributions through a committee
      - Filter races if only a single candidate - this impacted our ability to produce relative features (discussed next)
  - Result was **1,180** Candidates (rows) with the target identified as **WINNER (0,1)** if that Candidate had +50% of the votes



The Preprocessing + Cleaning

- Visuals: distribution of the **1,180 Candidates** in the final data model



The Preprocessing + Cleaning

- **Features:**
  - **Candidate Status** was straightforward to one-hot encode and we decided to consolidate Challenger + Open as we found differentiating from Incumbent
    - Incumbent - *currently holds seat*
    - Challenger - *looking to win seat that is currently held by an incumbent*
    - Open - *looking to win seat that is not held by an incumbent*

	2016 - Total Contributions		2018 - Total Contributions	
Incumbent	$65.6M	57%	$113.2M	35%
Challenger/Open	$50.2M	43%	$210.2M	65%
Total	**$115.8M**	**100%**	**$323.4M**	**100%**

The Preprocessing + Cleaning

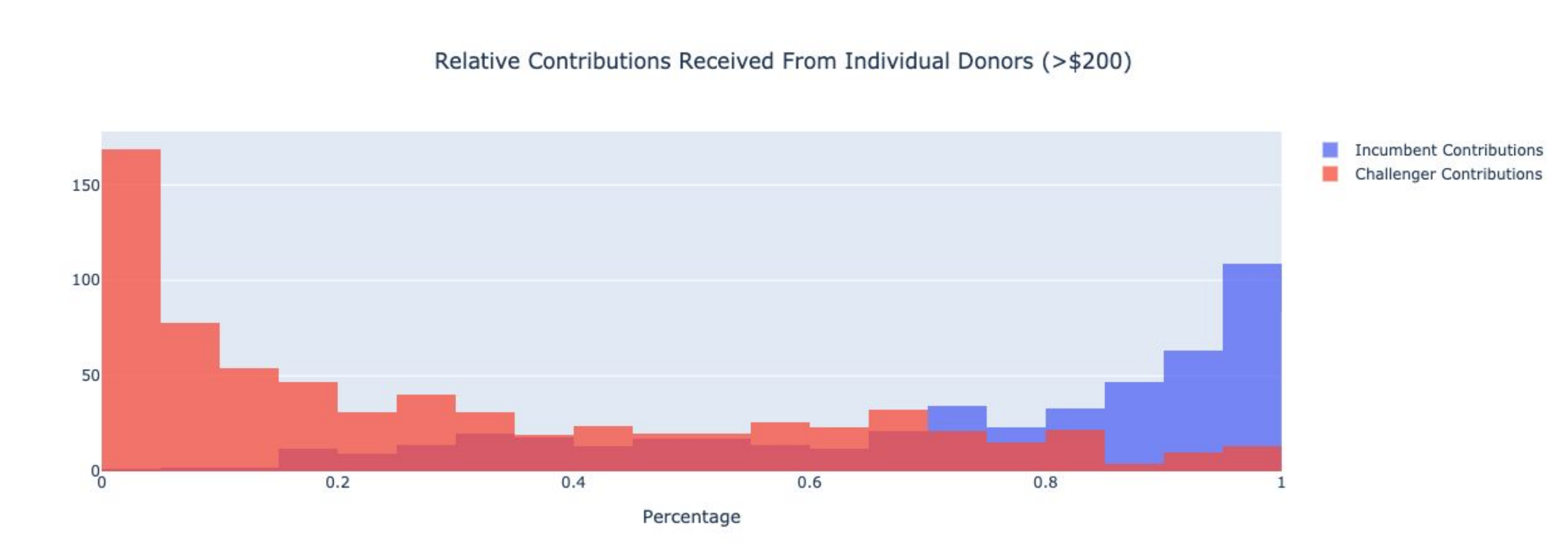
- **Features:**
  - **Individual Contributions** was more complicated due to the volume and the need to normalize the variability across states, districts, and years
    - To deal with the **volume**, we used summary statistics by candidate we thought could be informative:
      - Total $ donations
      - Total # donations
      - Average $ donation
      - Total # large donations (defined as $400+)
      - Total # out-of-state donations

The Preprocessing + Cleaning

- **Features:**
  - **Individual Contributions** was more complicated due to the volume and the need to normalize the variability across states, districts, and years
    - To **normalize** the variability in the absolute numbers, we:
      - Decided upfront to evaluate U.S. House of Representative races as there would be greater quantity and smaller districts for smoothing
      - Decided to look at the **relative** values comparing candidates in a defined race (% out of total contributions for 2016_VA_01)
      - Also included some **relative** values for a given candidate for certain features:
        - % of Total # Donations that were Large
        - % of Total # Donations that were Out-of-State

The Preprocessing + Cleaning

- Visuals: histogram of the **relative** contributions received by candidate status



Agenda

- The Data
- The Preprocessing + Cleaning
- **The Model**
- The Evaluation
- The Conclusion

The Model

- Selected the **binary classification** problem of whether a candidate won or loss the race as our target. This is designated as **WINNER** in the model.

- Split the data set **80:20** between training and hold-out

Training Set

WINNER	count
1	446
0	489

Hold-out Set

WINNER	count
1	114
0	131

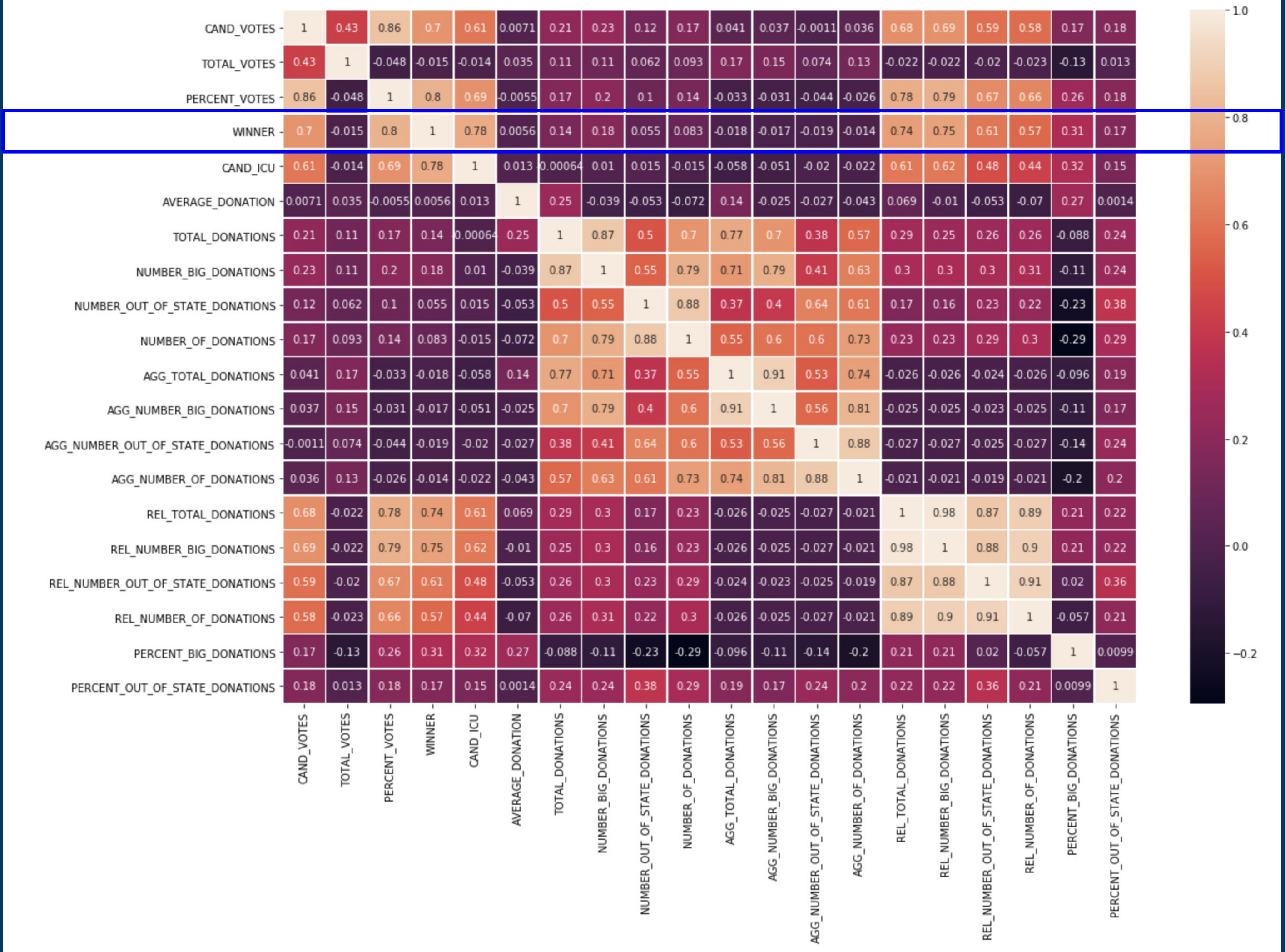
Total

WINNER	count
1	560
0	620

- Next we selected a subset of features to use in building the model by looking at the correlation matrix

The Model

- Visuals: Correlation Matrix



The Model

- Identifying our **benchmark** model:
    - Based on the high correlation features, we built univariate logistic regression models with default parameters for each of **7** features
        - Incumbent status
        - Total donations (relative sum + relative count)
        - Big donations (relative count + % candidate total count)
        - Out of state donations (relative count + % candidate total count)
    - We evaluated across 3 metrics: (1) accuracy, (2) precision, and (3) recall and selected the univariate logistic regression model using the **CAND_ICU** as our benchmark model.

The Model

- Identifying our **benchmark** model:

	accuracy	precision	recall
CAND_ICU	0.861224	0.931373	0.778689
REL_NUMBER_BIG_DONATIONS	0.836735	0.815385	0.868852
REL_TOTAL_DONATIONS	0.816327	0.793893	0.852459
REL_NUMBER_OUT_OF_STATE_DONATIONS	0.77551	0.768	0.786885
REL_NUMBER_OF_DONATIONS	0.746939	0.734375	0.770492
PERCENT_BIG_DONATIONS	0.628571	0.656566	0.532787
PERCENT_OUT_OF_STATE_DONATIONS	0.538776	0.569231	0.303279

The Model

- Identifying our **champion** model:
  - The models used were:
    - Logistic Regression
    - RandomForest
    - Gradient-Boosted Tree
  - Each model used all 7 features explored in the univariate analysis
  - Built each using **mllib** and **ml** libraries to explore the impact of cross validation and parameter grid
    - mllib - basic 80:20 split on default parameters
    - ml - used 5 fold cross-validation on training, parameter grid

The Model

- Identifying our **champion** model:

	accuracy	precision	recall
Gradient-Boosted Tree Model CV	0.926531	0.913793	0.929825
Logistic Regression Model CV	0.918367	0.912281	0.912281
Random Forest Model	0.914286	0.890756	0.929825
Gradient-Boosted Trees Model	0.910204	0.910714	0.894737
Random Forest Model CV	0.910204	0.903509	0.903509
Logistic Regression Model	0.902041	0.909091	0.877193

Agenda

- The Data
- The Preprocessing + Cleaning
- The Model
- **The Evaluation**
- The Conclusion

The Evaluation

- All classification models were evaluated using 3 criteria:
    - Accuracy: (tp + tn) / total
    - Precision: tp / (tp + fp); *of predicted wins, how many were correct*
    - Recall: tp / (tp + fn); *of actual wins, how many were correct*

- Confusion matrices were used to visualize the above evaluation criteria

- Selected models based on even performance across the criteria

The Evaluation

- All together:

	accuracy	precision	recall
Gradient-Boosted Tree Model CV	0.926531	0.913793	0.929825
Logistic Regression Model CV	0.918367	0.912281	0.912281
Random Forest Model	0.914286	0.890756	0.929825
Gradient-Boosted Trees Model	0.910204	0.910714	0.894737
Random Forest Model CV	0.910204	0.903509	0.903509
Logistic Regression Model	0.902041	0.909091	0.877193
CAND_ICU	0.861224	0.931373	0.778689
REL_NUMBER_BIG_DONATIONS	0.836735	0.815385	0.868852
REL_TOTAL_DONATIONS	0.816327	0.793893	0.852459
REL_NUMBER_OUT_OF_STATE_DONATIONS	0.77551	0.768	0.786885
REL_NUMBER_OF_DONATIONS	0.746939	0.734375	0.770492
PERCENT_BIG_DONATIONS	0.628571	0.656566	0.532787
PERCENT_OUT_OF_STATE_DONATIONS	0.538776	0.569231	0.303279

Agenda

- The Data
- The Preprocessing + Cleaning
- The Model
- The Evaluation
- **The Conclusion**

The Conclusion

- **Result:** We were able to build a classification model with **91-92%** accuracy, precision, and recall. This shows clear improvement over the benchmark. Given our approach only uses two base features (status, individual contributions), while ignoring policy + other factors, we considered this to be a success.

- Areas for further research or to extend the model:
    - **Refine** the candidate funnel to increase candidate count - e.g., *use 2018 candidate status to inform winners from 2016 races*
    - Test transferability of the model to more concentrated races - e.g., *Senate, Governor, President*
    - Test transferability of the model to **regression** for predicting vote %

Thank you