Emerge MA: Surveying Opportunities for Women to Run and Win Elections

Sarah Ferry, Neela Kaushik, Alex O'Connor
ferrys@bu.edu, nkaushik@bu.edu, aoconno8@bu.edu

1. Abstract

Emerge Massachusetts is a nonprofit organization whose mission is to get more female Democrats elected into public office via a strategic and targeted recruitment and training program. We have partnered with Emerge MA to develop a data-backed strategy to increase the likelihood of their candidates winning seats in MA State Senate and House voting districts. The motivating goals of this project are twofold: i) to isolate the factors that make female Democratic political candidates in Massachusetts most likely to win a seat in office; and ii) to integrate these factors into a machine-learning model that can predict the extent to which specific voting districts might be most amenable to electing a female Democrat. We hope that identifying these districts by analyzing their demographic composition and voting behavior might help Emerge MA to strategically prioritize candidate recruitment and campaigning efforts in State Senate and House districts where female candidates might resonate most with the population and have the strongest chance of being elected to office.

2. Data

We use multiple data sources, both public and private, to build our analysis. The public data is available online for download and includes United States census data, PD43+ which provided MA election results per voting district, and OCPF which provides donor contribution amounts per donation. We also use a private data source, VoteBuilder, which was provided to us by our Project Partner after receiving permission from MassDems. Further details about the contents of each dataset and their purpose in our analytical models are discussed below.

2.1. Data Sources

2.1.1. VoteBuilder: This source provides us with state election data and voter-specific data for Republican districts in MA. Due to issues with voter confidentiality, our team was not granted access to the Democratic districts. This data is grouped by district and by party affiliation, and provides summary statistics for the past four elections for the following attributes: i) voter participation in the past four elections; ii) voter race; iii) voter sex. We use VoteBuilder data as an input into a learning model that weighs and reports on the relative importance of these attributes in terms of how strongly they influence whether a district elects a female candidate.

2.1.2. PD43+: This source provides Massachusetts election statistics for all elections at the state/federal level since 1970. Our analysis uses the State Senate and House election results since 2009 in all districts. PD43+ provides data on current and past incumbents name, sex, party, vote count, and vote percent for each individual district. Our model relies on this data as an indicator variable for female Democrats' campaign success.

2.1.3. OCPF: This source provides political contribution data including donation amount and donor district for each year since 2009. We will aggregate the contributions made in districts that are over $500 and use this data to analyze the breakdown of contributions in each of the State Senate and House districts. This data can elucidate in which districts any candidate might most successfully raise money and will be used as feature in our model.

2.1.4. US Census data: This source provides demographic information at the MA State Senate and House district levels for each year since 2010. We use this data to generate a population breakdown with respect to certain key features of interest (political affiliation, gender, race and income) and use them as inputs in our predictive model for a female Democrat's election success in the district. We also run analyses to provide insight on which of these features might be most important in determining what candidate wins the election.

2.2. Collection and Preprocessing

Preprocessing is a critical step to ensure that all of the data collected is of a standardized form such that it can be considered in aggregate as input to our predictive models. Our goal for the data preprocessing stage was to obtain four aggregate data

sets that take into account all of our collected data: ones that represent a historical model (i.e. without VoteBuilder data included) for both House and Senate voting districts, and ones that represent a present model (i.e. with VoteBuilder data included) for State Senate and House districts as well. The structure of these merged data set is as follows: each row (record) represents a House/Senate voting district from a given year between 2010-2016, with the columns representing the features of that given district in that year. The represented features are the concatenation of all of the data sets, merged on the district and year for which they correspond. Therefore, during the data collection process, it was important to identify and standardize the voting district and year that each collected data point represented such that all of the data could be merged on these attributes.

Notable data collection techniques that were used include web scraping from the PD43+ website to obtain the party, gender and margins of victory for each representative elected to the MA State Senate and House from 2009-present, which was used to generate the predicted variable (incumbent female, Democrat representative) for our classification models. Additionally, web scraping was performed on the MA State Secretary's website to obtain mappings of cities and towns to their respective State Senate and House voting districts from 2009-present. This process was needed to map OCPF and VoteBuilder data, which presented results by city, onto voting districts so that the data could be used in our model to predict per-district female Democratic success.

3. Approach

As mentioned, the data sources with which we worked are inconsistent with regard to the time frame for which they provide information, reflecting the difficulties of working with real-world data. Namely, while we have historical data from at least 2009-present for the remaining three datasets, the VoteBuilder data only provides information from the 2015-2016 election cycle. Therefore, we build two predictive models: one that takes into consideration the VoteBuilder summary statistics but is limited to the scope of the last two election cycles, and another that provides a more historical overview of the MA districts' voting behavior without the VoteBuilder summary statistics. Both trained models will return the probability of a female Democrat winning a particular voting district based on the district's demographic makeup and features, including contribution information and past voting behavior. Additionally, the present model will return the 30 most statistically significant features in the model. We hope that a combination of these predictors will be a helpful snapshot of both past and present voting outcomes and create a robust model in aggregate.

To develop trained predictive models for the historical model, we split the data into training, testing and validation subsets as follows: all data from 2009-2012 is considered as training data, data from 2013-2014 is considered as validation data and data from 2015-2016 is considered as testing data. All subsets are further split into data for House districts and data for Senate districts. A merged dataset is created for both historical and present models that combines the State Senate and House districts for training, validation and testing.

For the present model, we used a leave-one-out approach for training and testing on the 2015-2016 data since we had a lack of historical data to train from. Table 1 illustrates all of the datasets with which we build, train and test our models.

		Training Data	Validation Data	Testing Data
Historical Model	Senate Districts (SD)	2009-2012 data by SD	2013-2014	2015-2016
	House Districts (HD)	2009-2012 data by HD	2013-2014	2015-2016
	Merged Districts (MD)	2009-2012 data by MD	2013-2014	2015-2016

		Training Data	Testing Data
Present Model	Senate Districts (SD)	2015-2016 by SD	2015-2016 leave-one-out
	House Districts (HD)	2015-2016 by HD	2015-2016 leave-one-out

Table 1. Breakdown of training, validation and testing datasets used for developing Historical and Present predictive models of election results.

4. Experiments

We developed models to analyze both the historical data and the present data containing various different classifiers and approaches. For the historical model, we ran models with logistic regression, K-nearest neighbors, support vector machines, and a decision

tree classifier to determine which would give us the best accuracy. For the present model, we ran logistic regression with recursive feature elimination to extract the most statistically significant features. A breakdown of each model is discussed below.

4.1. Historical Models

The purpose of the trained historical model is to take into account information regarding MA census information, donor information and history of state representatives (i.e. party and gender of elected officials) from 2009-2012, and predict the probability that the incumbent representative of a given district is a female and/or a female Democrat. We create two indicator variables in the data—female and female democrat—that are encoded "1" if the district representative in a particular year is a female and a female democrat, respectively. We use these variables as the binary predictor on our classification model.

We employ four common classifiers in an effort to identify the model that yields the best accuracy for our particular dataset. Namely, we employ logistic regression (LR), support vector machines (SVM), *k*-nearest neighbors (KNN) and a decision tree classifier (DTC). We train models with each of these four classifiers on the Senate, House and Merged training datasets, using both female and female democrat as the indicator variables, separately. Models trained on Senate data are tested on Senate data; models trained on House data are tested on house data; and models trained on the Merged data are tested both on Senate and House data. Predictions for a trained model consist of list of probabilities that correspond to the likelihood of each district in the test being encoded a "1" on the indicator variable in that particular test. These predicted probabilities are then thresholded (i.e. if $p > 0.5$ assign it a 1, else assign it a 0) to yield a list of binary predictions that is then compared to the true values in the test set to compute an accuracy score for each model. Table 2 displays the results of model training for our conducted analyses.

4.2 . Present Models (VoteBuilder)

The purpose of the present model is to incorporate VoteBuilder data with existing data from the historical model from 2015-2016 (census, contribution, past election results) in order to isolate significant features for success, and predict overall success for female Democratic candidates. Similar to the historical model, we create an indicator variable to identify female Democratic candidates and use this as the dependent variable for our model. A logistic regression classifier is used to produce this model.

Due to the lack of historical data for VoteBuilder, the accuracy of this model will vary from the historical model. Consequently, we decided to shift our analytical focus to identify the most significant features in the model (race, party, income, etc.). The first step for identifying these features involved recursive feature elimination. Trained with logistic regression, this eliminator selected 30 features from the data by recursively pruning features that were either not significant or not independent from other features. These 30 features were then used to train and test the model.

On top of its limitations with respect to historical data availability, VoteBuilder also only provides information for State Senate and House districts that are currently held by Republicans. Due to the lack of an extensive dataset on which to perform standard training, validation and testing, we implement a leave-one-out method of cross-testing on both Senate and House district data. This process is as follows: the data is split into a training set and testing set before creating the logistic regression model. The training set consists of all but one district from the data, and the testing set consists of the final district. Once split, the model is fit with the training data, and an outcome is predicted from the testing data. This process is repeated for every district in the data until each district has served as the testing set. Each predicted outcome is then used to return an overall accuracy for the model (Table 3).

The χ^2 goodness-of-fit test is used to verify significant variables in the model. In each iteration of leave-one-out testing, the model outputs the χ^2 statistics and corresponding probability values. After

normalizing all p-values from all testing, we use these results to identify all significant variables in the model (Table 4). At an $\alpha = .05$ level, all 30 variables which were originally selected by recursive feature elimination are confirmed to have $p < 0.05$, indicating significance in the model.

5. Results & Visualizations

5.1. Historical Models

The historical model outputs the accuracy on the test data from 2015-2016. We output accuracies for each classifier as well as on the Senate data, the House data, and the Merged data and determine that the SVM classifier performs best when taking into account model accuracy as well as diversity of predicted probability, which we use to generate visualizations for Emerge MA to interpret these results. Additionally, we determine that the models performed best individually, as opposed to merged, so we train only on senate data to predict in the senate districts and only on house data to predict in the house districts. The final models whose predictions are visualized are shown in bold in Table 2.

		Test on Senate		Test on House	
Training Data	Method	Fem.	Fem. Dem.	Fem.	Fem. Dem.
Senate	LR	0.6512	0.6977	N/A	N/A
	SVM	0.7442	**0.7674**	N/A	N/A
	KNN	0.8837	0.8604	N/A	N/A
	DTC	0.7209	0.6047	N/A	N/A
House	LR	N/A	N/A	0.6325	0.7771
	SVM	N/A	N/A	0.7048	**0.8253**
	KNN	N/A	N/A	0.8133	0.8554
	DTC	N/A	N/A	0.6867	0.7430
Merged	LR	0.5814	0.5814	0.5481	0.6325
	SVM	0.6977	0.7442	0.7530	0.8253
	KNN	0.5814	0.6047	0.7711	0.8433
	DTC	0.5581	0.3953	0.6506	0.6687

Table 2. Accuracy of historical model trained with various classifiers on predicting indicator variables (female, female democrat) for Senate and House test data. Bolded data indicates results used for visualization purposes.

For visualization of the results of the historical model, we created heatmaps using the Python package Folium and shapefiles produced by the The Official Website of the Executive Office for Administration and Finance of Massachusetts. Our heatmaps for both the House and the Senate districts indicate districts where female democrats are likely to succeed with a very dark green, and districts where they are unlikely to succeed with a very light green (Figure 1).

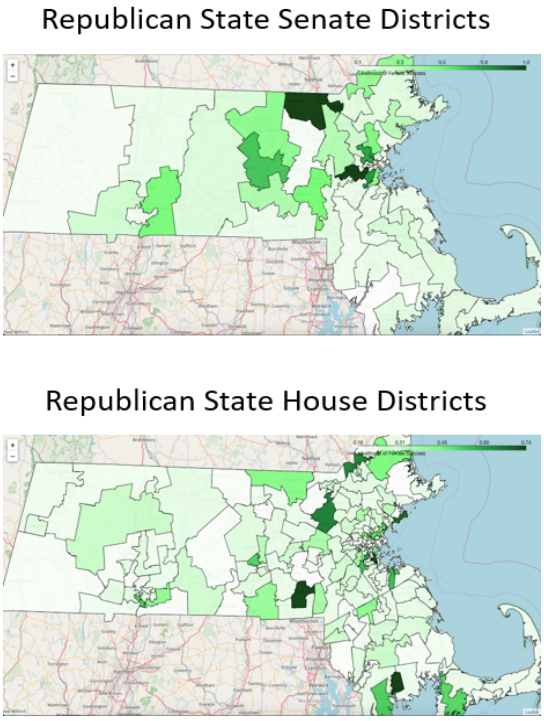


Figure 1. Predicted outcomes for female Democrats in both MA State Senate and House districts. Darker green indicates higher probability of success.

5.2. Present Models (VoteBuilder)

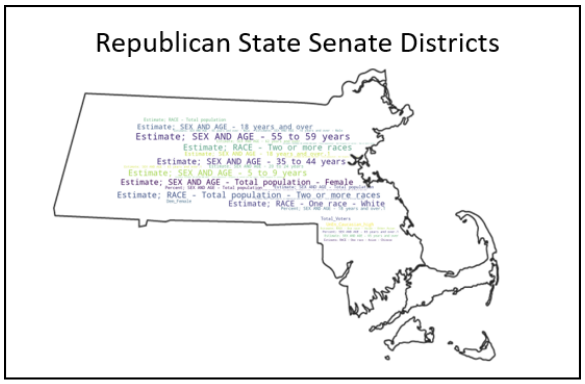
Each iteration of the leave-one-out procedure creates a prediction for each district's outcome, and each prediction is stored to report an overall accuracy of all tests. These accuracies indicate the percent correctness that the model can identify a female democratic winner for a given Republican district, which can be seen below in Table 3.

		Female Democrat
Present Model	Senate Districts	0.7778
	House Districts	0.9333

Table 3. This table includes the accuracies of the Logistic Regression models for Republican State Senate and State House Districts, respectively.

All features that are statistically significant must have probabilities $< \alpha$ where $\alpha = .05$. The χ^2

goodness-of-fit test determined the significance of all features in the models and output results for both the State Senate and House districts, as shown in Table 4.



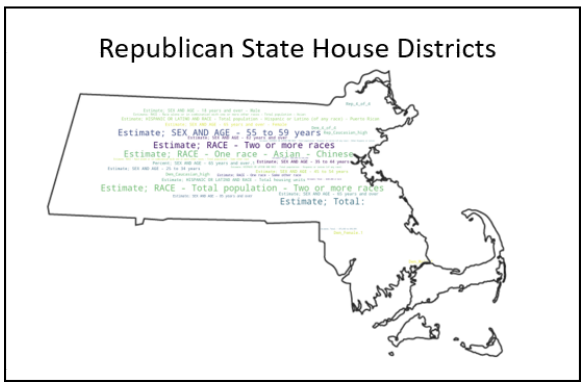


Figure 2. Word cloud visualizing statistically significant features in both the Republican State Senate districts and House districts. The larger a word is, the more statistically significant it is relative to other features in the model.

It is interesting to note that while there are many features that overlap, our model predicts that some features are more significant in the House districts and others are more significant in the Senate districts. We generate word clouds to visualize and highlight all significant features (Figure 2).

6. Conclusion & Future Directions

Based on our results, we believe that these generated models predict probable success for female Democratic candidates in State Senate and State House districts throughout Massachusetts to a reasonable level of accuracy. The historical model is the better of the two for predicting success according to trends from previous years as well as current data, while the present model is useful to identify present day features that are important for a candidate's success.

Senate

Feature	Significance (p-value)
UnEn_Caucasian_high	<.0001
Dem_Female	<.0001
Total_Voters	<.0001
Estimate; SEX AND,AGE - Total population	<.0001
Percent; SEX AND,AGE - Total population	<.0001
Estimate; SEX AND,AGE - Total population - Female	.000285
Estimate; SEX AND,AGE - 5 to 9 years	.002275
Estimate; SEX AND,AGE - 20 to 24 years	<.0001
Estimate; SEX AND,AGE - 35 to 44 years	.002228
Estimate; SEX AND,AGE - 55 to 59 years	.006683
Estimate; SEX AND,AGE - 18 years and over	<.0001
Estimate; SEX AND,AGE - 62 years and over	<.0001
Estimate; SEX AND,AGE - 65 years and over	<.0001
Estimate; SEX AND,AGE - 18 years and over.1	<.0001
Percent; SEX AND,AGE - 18 years and over.1	<.0001
Estimate; SEX AND,AGE - 18 years and over - Male	<.0001
Estimate; SEX AND,AGE - 65 years and over.1	<.0001
Percent; SEX AND,AGE - 65 years and over.1	<.0001
Estimate; SEX AND,AGE - 65 years and over - Male	<.0001
Estimate; RACE -,Total population	<.0001
Estimate; RACE -,Total population - Two or more races	.006567
Estimate; RACE -,One race - White	0.00002
Estimate; RACE -,One race - Asian - Chinese	<.0001
Estimate; RACE -,One race - Asian - Other Asian	<.0001
Estimate; RACE -,Two or more races	.006567
Estimate; RACE -,Race alone or in combination with one or more other races - Black or African American	<.0001
Estimate; HISPANIC,OR LATINO AND RACE - Hispanic or Latino (of any race) -,Other Hispanic or Latino	<.0001
Estimate; HISPANIC,OR LATINO AND RACE - Not Hispanic or Latino	<.0001
Estimate; HISPANIC,OR LATINO AND RACE - Not Hispanic or Latino - Two or more,races	<.0001
Estimate; HISPANIC,OR LATINO AND RACE - Not Hispanic or Latino - Two races excluding Some other race	<.0001

House

Feature	Significance (p-value)
Dem_4_of_4	<.0001
Rep_4_of_4	<.0001
Dem_Caucasian_high	<.0001
Rep_Caucasian_high	<.0001
Dem_Male.1	<.0001
Dem_Female.1	<.0001
Estimate; SEX AND,AGE - 25 to 34 years	<.0001
Estimate; SEX AND,AGE - 35 to 44 years	0.000004
Estimate; SEX AND,AGE - 45 to 54 years	0.000002
Estimate; SEX AND,AGE - 55 to 59 years	0.000056
Estimate; SEX AND,AGE - 85 years and over	<.0001
Estimate; SEX AND,AGE - 62 years and over	<.0001
Estimate; SEX AND,AGE - 65 years and over	<.0001
Estimate; SEX AND,AGE - 18 years and over - Male	<.0001
Percent; SEX AND,AGE - 65 years and over.1	<.0001
Estimate; SEX AND,AGE - 65 years and over - Female	<.0001
Estimate; RACE -,Total population - Two or more races	.0231354
Estimate; RACE -,One race - Asian - Chinese	0.00003
Estimate; RACE -,One race - Some other race	<.0001
Estimate; RACE -,Two or more races	.0231354
Estimate; RACE -,Race alone or in combination with one or more other races Asian	<.0001
Estimate; RACE -,Race alone or in combination with one or more other races - Some other race	<.0001
Estimate; HISPANIC,OR LATINO AND RACE - Hispanic or Latino (of any race)	<.0001
Estimate; HISPANIC,OR LATINO AND RACE - Hispanic or Latino - Puerto Rican	0.000014
Estimate; HISPANIC,OR LATINO AND RACE - Hispanic or Latino -Other Hispanic or Latino	.006567
Estimate; HISPANIC,OR LATINO AND RACE - Total housing units	0.000001
Estimate; Total:	.009597
Estimate; Total: -,60,000*to*74,999	<.0001
Estimate; Total: -,75,000*to*99,999	<.0001
Estimate; Total: -,$200,000 or more	<.0001

Table 4. Significant features and probability values for Republican State Senate districts (top) and House districts (bottom).

While political races can rely on a fair amount of subjectivity, we hope that our models provide a more concrete and objective approach to identifying districts for Democratic women to succeed.

Steps for future work and improvement on these models involve more data collection for years ranging further back than 2009. Specifically for the present model, it can be refined and made more accurate in the future by including more historical VoteBuilder data from previous election cycles.

7. Link to GitHub repo

https://github.com/ferrys/cs-506-emerge-MA