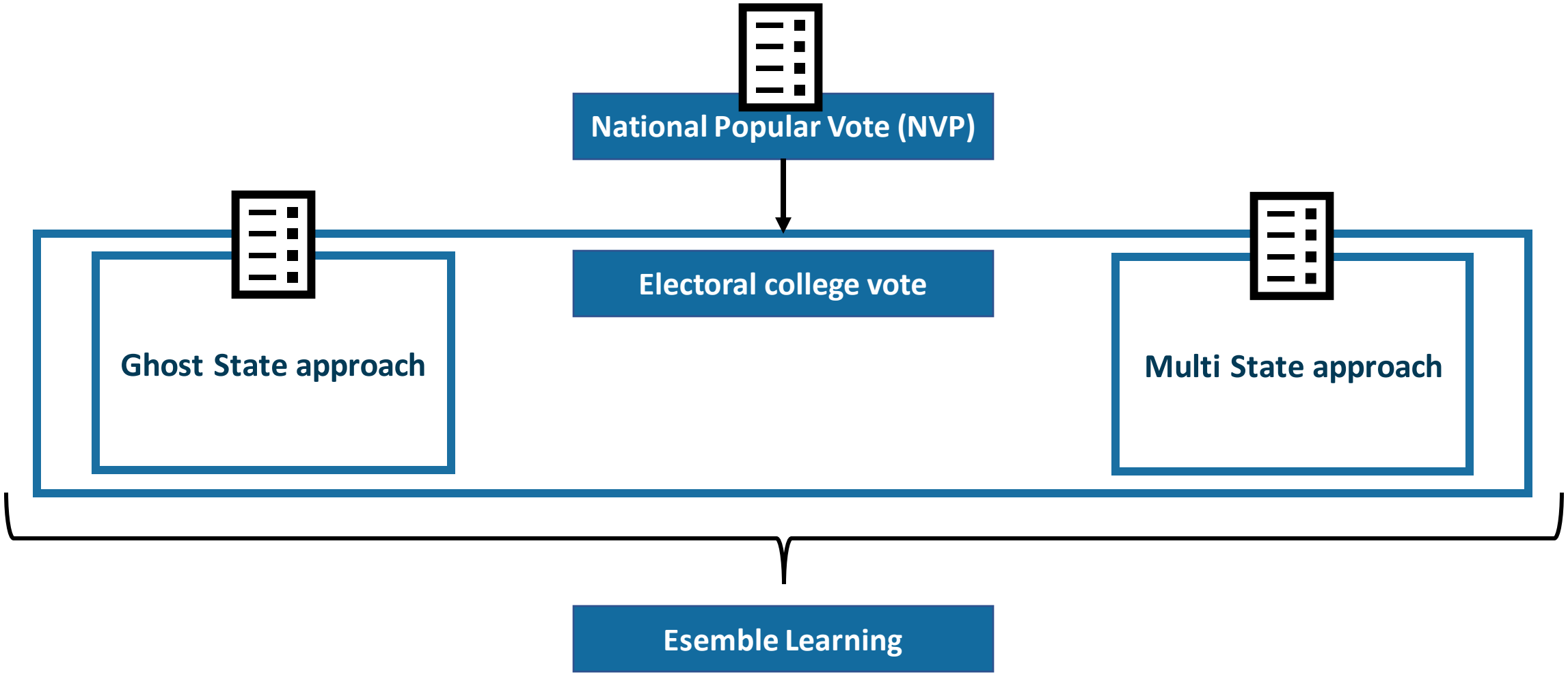
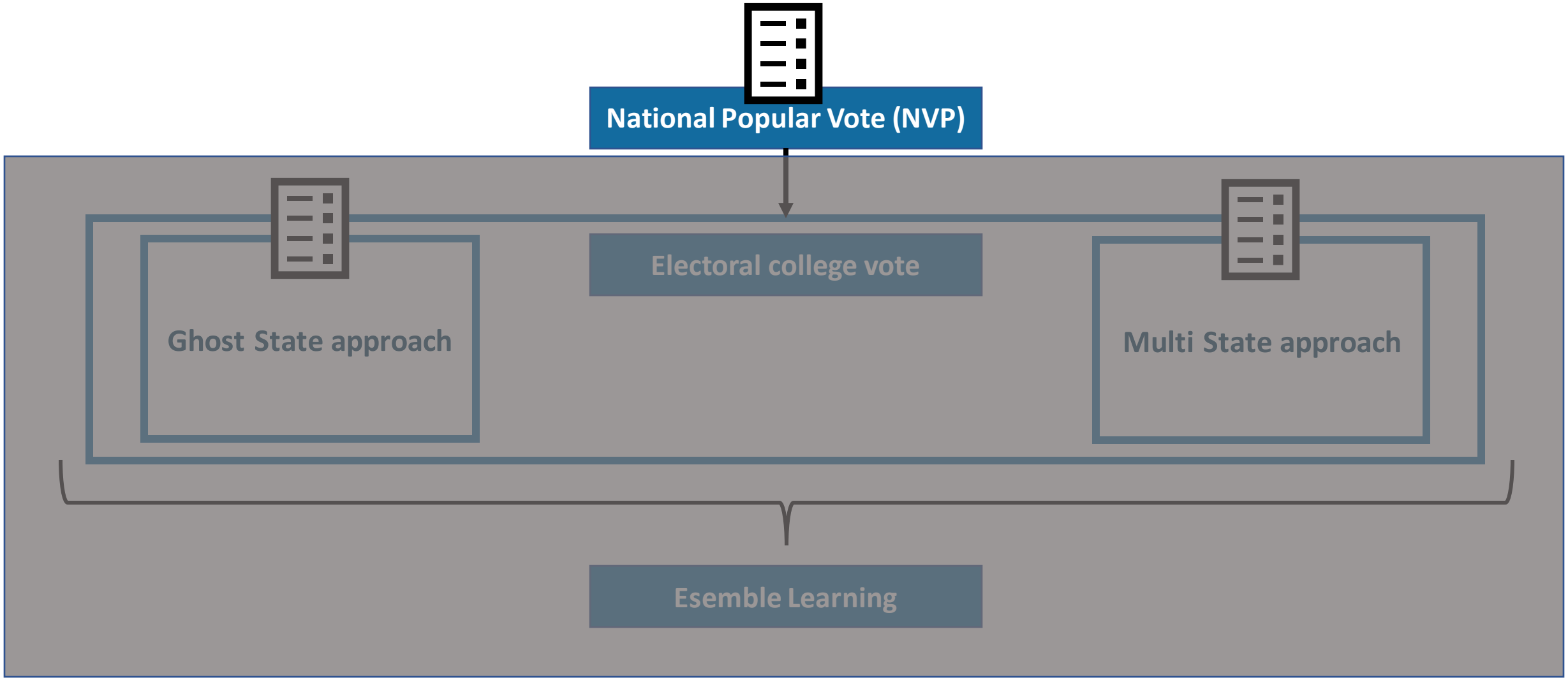
Predicting the US 2020 Presidential Elections

INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

Introduction: Overall Framework



INTRODUCTION: OVERALL FRAMEWORK



National Popular Vote

Database

For every year since 1968:

- MacroEconomic data drawn from FRED
- Polls drawn from 5.38

Dataset: 26 features and 18 observations (one per election year)

How to handle such situation with sparse data ?

Modelling

Favor sparse-like models:

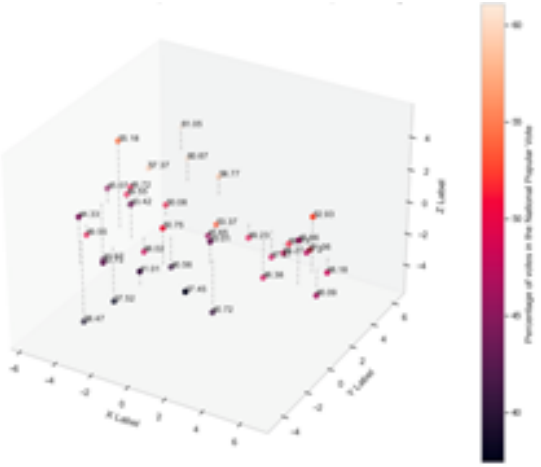
- Feature Engineering combing different predictors
- PCA
- Lasso-Like models
- Cross Validation based on Accuracy

Results

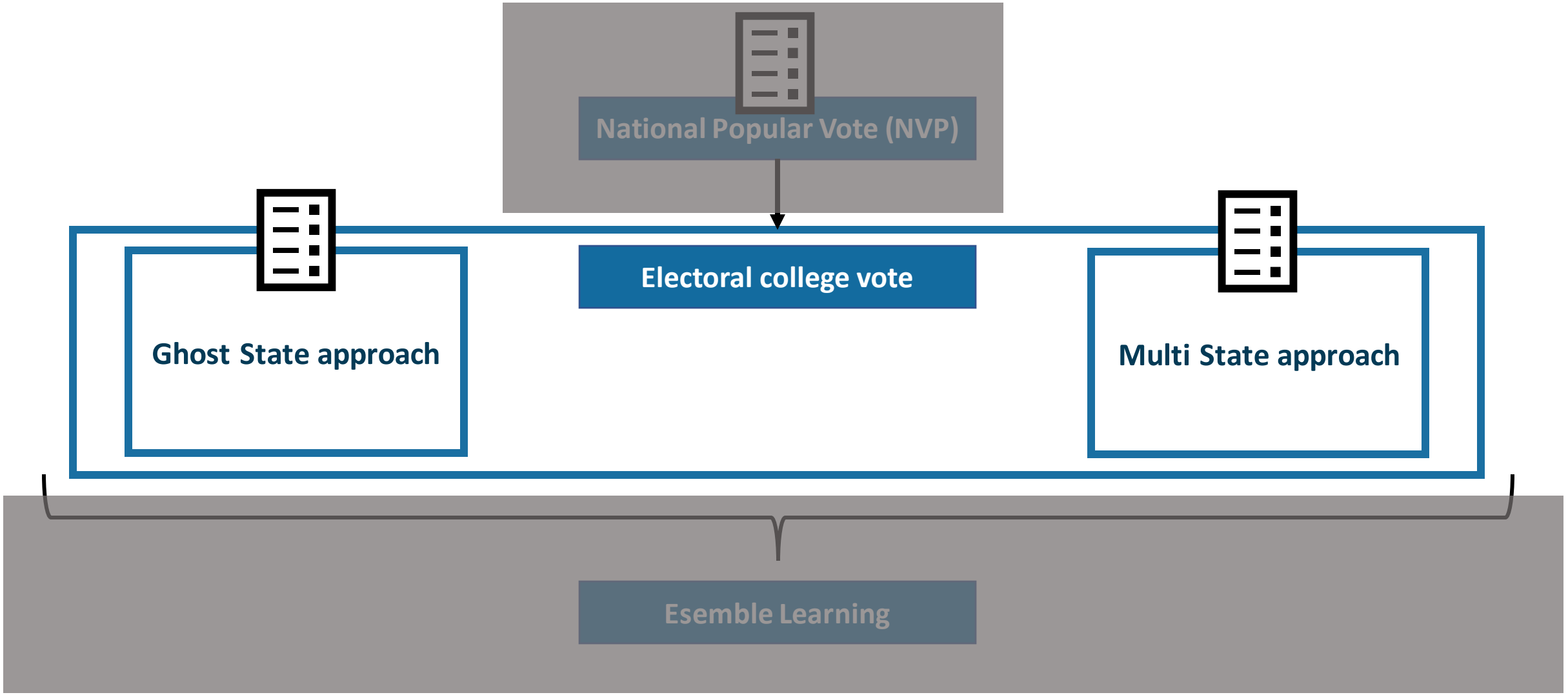
Lasso Regression, with accuracy score 0.73

Predictions:

- For Donald Trump: NPV = 46.525 +/- 1.39
- For Joe Biden: NPV = 52.46 +/- 1.88



INTRODUCTION: OVERALL FRAMEWORK



Electoral College Vote

Database

- 9 Features, 444 Observations (394 for Train/CV, 50 for Test)
- Polling data (5.38), Macroeconomic data (FRED)

Modelling

Tested Models:

- LogisticRegressionCV, DecisionTreeClassifier, RandomForestClassifier, AdaBoostClassifier, NeuralNetwork
- Feature Engineering using Polynomial Features, PCA
- Lasso Regularization, Dropout, EarlyStopping
- Cross-Validation based on classification accuracy for parameters tunning and model selection

	LogisticRegressionCV	DecisionTreeClassifier	RandomForestClassifier	AdaBoostClassifier	SequentialNeuralNetwork
Accuracy	0.88585	0.88075	0.8987	0.87831	0.8992

- Cross-validation accuracies dictated by the data, rather than model's expressive power

LogisticRegressionCV chosen as final model for the Ghost State Approach

Electoral College Vote

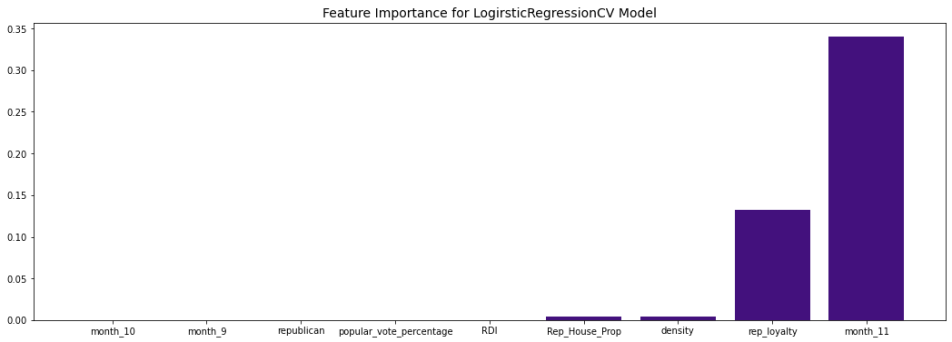
A. Ghost State Approach

B. Multi State Approach

Model Interpretation

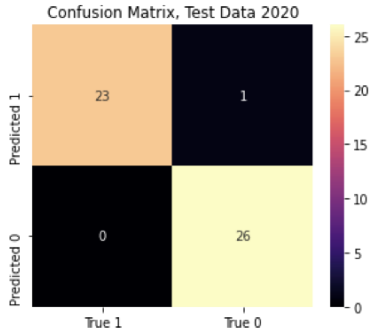
- polls_month_11 – most important feature
- pairwise correlation between polls
- positive relationship: rep_house_proportion, state_loyalty
- negative relationship: population_density
- multicolinearity: popular_vote_percentage

	coef. value
polls_month_10	0.000000
polls_month_11	17.095946
polls_month_9	0.000000
republican/democrat	0.000000
rep_house_proportion	0.693126
state_loyalty	2.972780
popular_vote_percentage	-2.413295
population_density	-1.774068
RDI	-0.868299

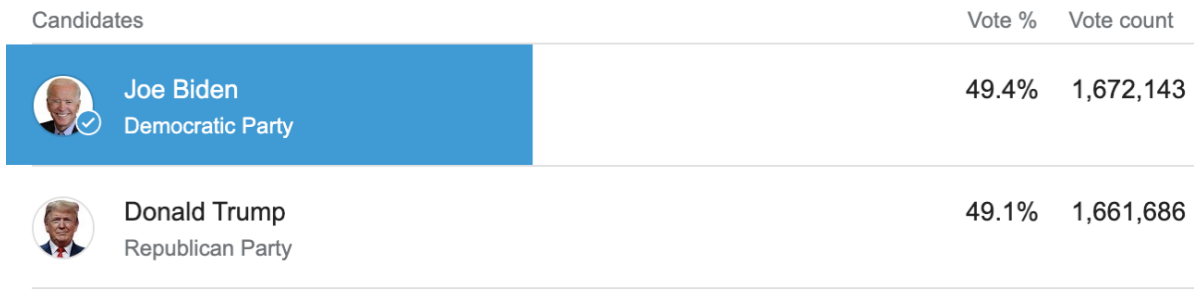


Results & Error Analysis

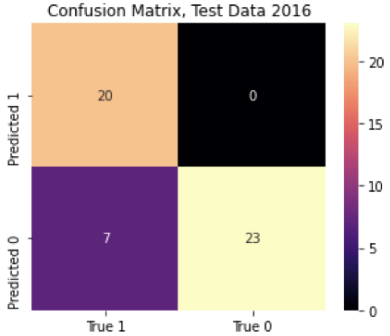
2020 Overall Test Accuracy: 98%



Misclassified state: Arizona



2016 Overall Test Accuracy: 86%



Electoral College Vote

- Predict each state seperately

Not swing states

- Deterministic approach
- Based on loyalty feature of the state

Swing states

Predict each state seperately

Construct the dataframes

- Correlation matrix from The Economist

	WI	OH	NV	NH	NC	MI	IA	FL
WI		0.85	0.28	0.81	0.60	0.86	0.85	0.50
OH	0.85		0.28	0.72	0.71	0.88	0.82	0.52
NV	0.28	0.28		0.25	0.25	0.31	0.27	0.57
NH	0.81	0.72	0.25		0.47	0.75	0.73	0.49
NC	0.60	0.71	0.25	0.47		0.69	0.65	0.39
MI	0.86	0.88	0.31	0.75	0.69		0.81	0.57
IA	0.85	0.82	0.27	0.73	0.65	0.81		0.40
FL	0.50	0.52	0.57	0.49	0.39	0.57	0.40	

The Economist

1=perfect correlation

- More data points → add correlated states
- → Model can learn the correlation between states

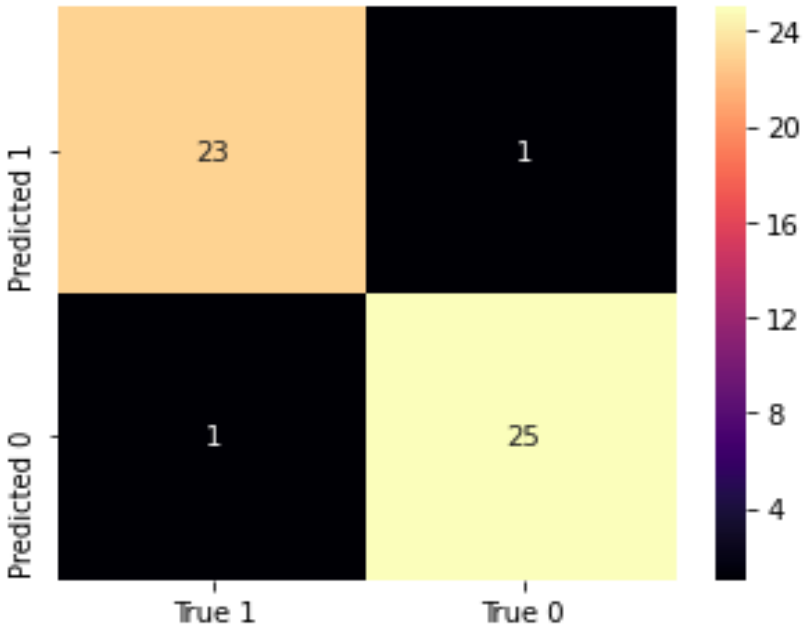
	Texas	Georgia	Arizona	Pennsylvania	Michigan	Virginia	Wisconsin	Nevada	New Hampshire
LogisticRegression(C=0.1, max_iter=10000, penalty='l1', solver='liblinear')	0.888889	0.888889	0.888889	0.888889	0.888889	0.888889	0.777778	0.833333	0.611111

	Florida	Ohio
KNeighborsClassifier()	0.611111	0.888889

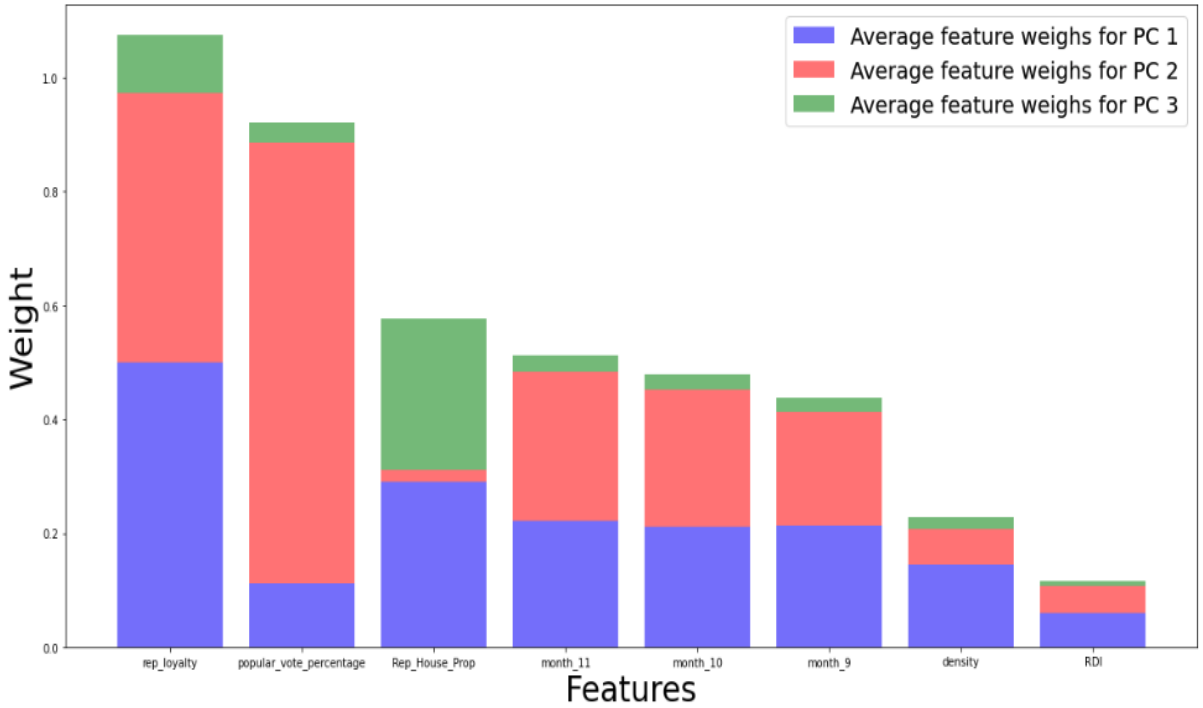
	Colorado	Iowa	North Carolina
LogisticRegression(C=0.1, max_iter=10000)	0.611111	0.666667	0.722222

Electoral College Vote

Overall results



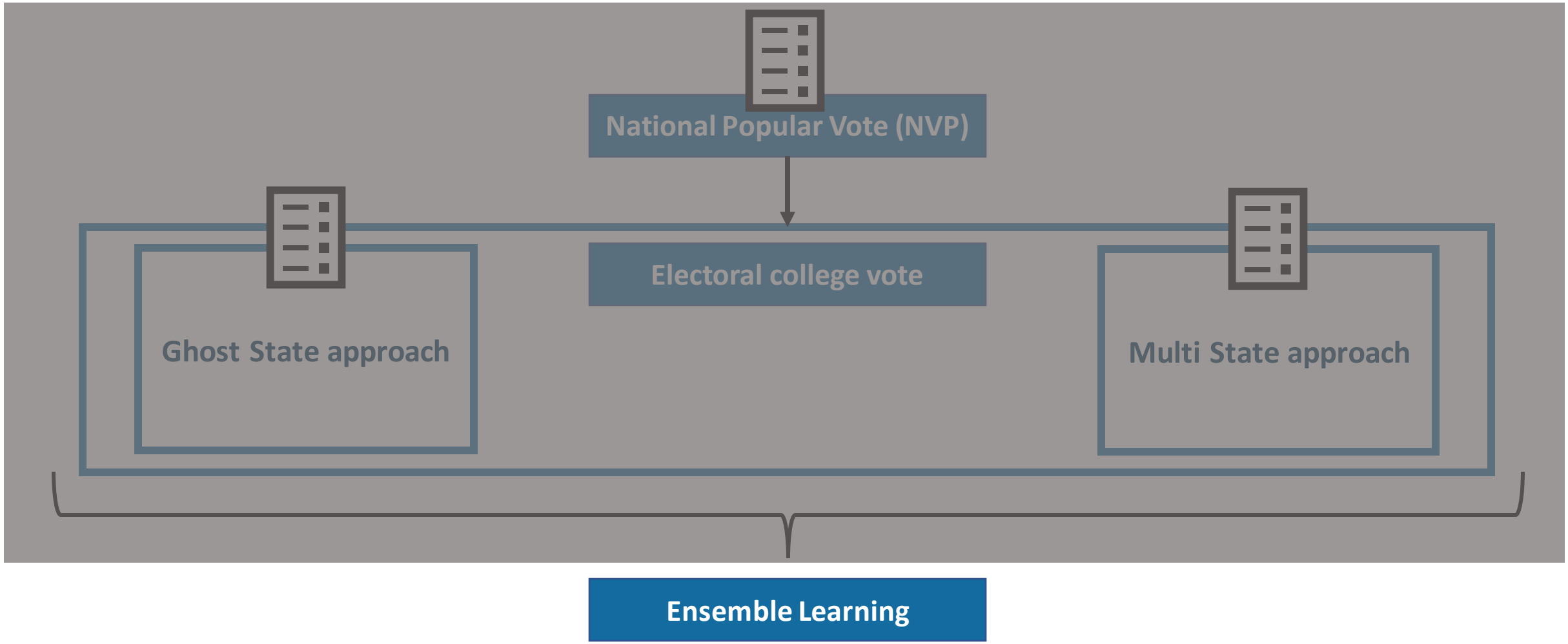
Confusion Matrix
Overall test accuracy of 96%



Average feature importance across all swing states models

➔ **The models doesn't rely on polls**

INTRODUCTION: OVERALL FRAMEWORK



ENSEMBLE LEARNING

Database

Use the output from the Multi State Model and the Ghost State Model

- Probabilities output on every election for every state for the two models

Dataset: 394 examples, 2 features

Modelling

Favor interpretable results:

- Logistic Regression (being a weighted sum of the two predictors)
- Cross Validation based on Accuracy

Results

- Final model: 0.98 classification accuracy
- Misclassified state: Arizona
- Weights to the different models: 4.89 for the single state, 0.92 for the multi state