Executive Summary

North Carolina maintains detailed data on voter registration by county and makes this data available to the public. This data includes personal data for each registered voter (e.g. age, race, ethnicity, gender) as well as political data (e.g. congressional district, party registration, precinct, state house district).

This report, created by the Reese News Lab Local Elections Data project, examines a subset of demographic factors to explore how these factors influence a registered voter's party affiliation. We use two "classification" modelling techniques to assess the relative weight that race, voter residence location (as represented by state house district), gender, age, and ethnicity have on party affiliation.

Data and Assumptions

The complete statewide voter registration data has been loaded county by county into a table as part of a private relational database accessible only to UNC employees and independent contractors working on the Reese News Lab Local Elections Data project. This table consists of a row for every voter that had recently registered to vote in North Carolina. The table contains 71 columns each representing either personal or political information for each voter.

For our analysis, we selected only "active" voters and ignored inactive, removed or denied voters. This was accomplished by selecting only the rows with column "status_cd" equal to "A"

We then extracted six columns from the database table with the following definitions:

race_code:

B	BLACK or AFRICAN AMERICAN
I	AMERICAN INDIAN or ALASKA NATIVE
O	OTHER
W	WHITE
U	UNDESIGNATED
A	ASIAN
M	TWO or MORE RACES

ethnic_code:

HL	HISPANIC/LATINO
NL	NOT LATINO
UN	ETHNICITY NOT DECLARED

gender_code

M	MALE
F	FEMALE
U	UNSPECIFIED

birth_age

Current age in years calculated by North Carolina based upon birth date

Note: The assumption of accurate age calculation needs to be verified. If this cannot be *verified, approximate age in years can be calculated using the column birth_year*

nc_house_abbrv

The state of North Carolina is divided into 120 House Representative Districts, each containing approximately 83,000 North Carolina residents. These districts are labeled via an integer value between 1 and 120 inclusive

party_cd

DEM	DEMOCRATIC
LIB	LIBERTARIAN
REP	REPUBLICAN
UNA	UNAFFILIATED

Note: After exploring a subset of the data, these were the only parties found in the database table. These could be updated based upon full data exploration

Data Cleaning

Since we can't infer the values of missing data in any columns, we start by dropping any rows where data is missing in any of the five columns listed above. At this time, no rows have been dropped.
Note: This can be updated with percent rows dropped if this changes

After exploring the data, the only item that appeared to have been miscoded is for unaffiliated voters in party_cd. Various codes such as "UN" and "u" were found. We set up a rule that such that if this column contained either an upper or lowercase "U" we substituted "UNA" .
Note: We can always add additional correction rules if these become necessary after further data exploration.

All columns were checked for invalid data values and none were found at this time. If we do find invalid values, we will attempt to create a rule that fixes these values if the true values can be inferred (similar to party_cd above). If values can't be inferred, we will drop the rows.

Data Adjustment

To enable analysis, we created a new column to reflect age groupings. To improve model performance, we used age groupings as follows:18-22, 23-27, 28-32, 33-37, 38-42, 43-47, 48-52, 53-57, 58-62, 63-67, 68 and over. Also, since this analysis covers tendency for specific political affiliation, we removed unaffiliated voters from our data set. Lastly, we also removed voters that affiliated with the libertarian party, since these voters only make up less 0.5% of those that affiliate with a party.

Modelling Methodology

Setup

We chose two classification methods to build models that attempted to predict voter affiliation (either with the Democratic or Republican party) based upon race, state house district, gender, age group and ethnicity.

1. We ran a logistic classifier using the Newton-Raphson method which is equivalent to the iteratively re-weighted least squares algorithm. We iterated until we converged on a solution.
2. We ran a decision tree classifier of essentially infinite depth with a minimum weight of each leaf node of 0.1

To assess the accuracy of each instance of each model, we randomly divided the data into a training set (80%) that was used to build each model and a test set (20%) that was used to evaluate each model.

For each classification method, we built and evaluated a model based upon each of the five demographics (race, state house district, gender, age group or ethnicity) singly. We then built and evaluated another model combining the top two most influential demographics to improve model accuracy, and lastly we built a final model combining all five demographics into a single model for a total of seven models.

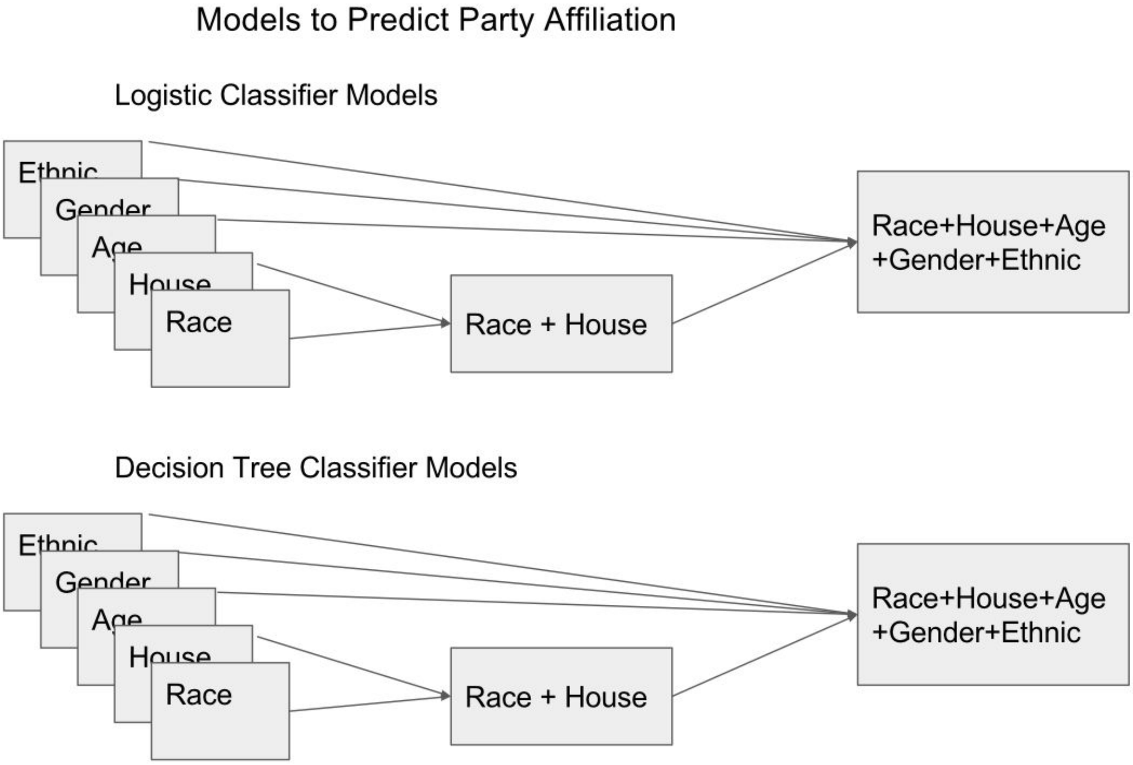

Models to Predict Party Affiliation
Logistic Classifier Models
Ethnic
Gender
Age
House
Race
Race + House
Race+House+Age +Gender+Ethnic
Decision Tree Classifier Models
Ethnic
Gender
Age
House
Race
Race + House
Race+House+Age +Gender+Ethnic

Figure 1

Execution and Validation

Creating a model using a classification algorithm does not yield a closed-form solution. The parameters (weights) are determined using iterative techniques that attempt to minimize a loss function (in this case error).

We chose two different classification methods hoping that each would yield results consistent with other; results that differed significantly might be indicative of data or other issues. We compared the results of each of the seven models created using the logistic classifier with the results of the seven created with the decision tree classifier. Specifically, we looked at prediction accuracy; what percentage of the predictions on the test data are correct. For each of the seven models, the accuracies of the logistic models were within 1% of the accuracies of the decision tree models. We concluded both models were delivering meaningful results.

To select a specific classification method, we looked at the stability of each of the algorithms. We created 40 models using logistic classification and reviewed the variation of the reported accuracy. We did the same with the decision tree classification models. While both methods were stable, the decision tree model appeared to have less variability, so we selected this model for our final analysis.

Our challenge was to estimate each demographic's influence on party affiliation. We decided to use model "accuracy" to measure this. As previously discussed, accuracy is the percentage of correct predictions in the test data set result from the model built on the training data set.

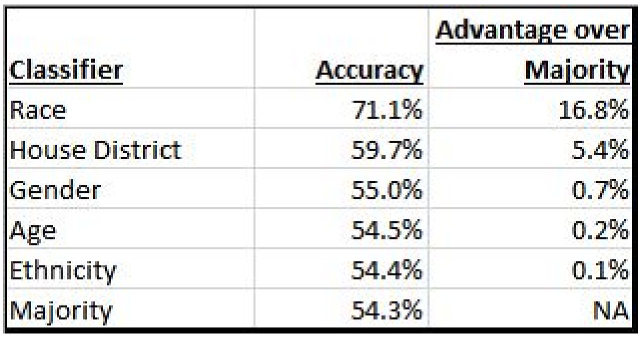
In classification problems, the baseline (or lowest possible) level of accuracy is always the "majority classifier" The majority classifier simply predicts that every observation will belong to the majority class. The result is an accuracy level equal to the percentage of the majority class in the overall population.

For North Carolina, of all voters registered as either democrat or republican, the majority (54.3%) are registered as democrats. The "majority classifier" would therefore predict that every voter would register as a democrat, and this classifier would have an accuracy of 54.3%, Therefore, the influence of a specific demographic could be measured by how much larger the accuracy was for the classifier built specifically with that demographic. So, we can define this difference in accuracy as the measure of influence.

Using the decision tree classifier, we pooled the accuracy results of the 40 models created with random splits of the data into training and testing sets and compared this accuracy with the accuracy of the majority classifier.

Results

Figure 2 below describes the difference between the accuracy of the majority classifier and the accuracy of each demographic classifier.

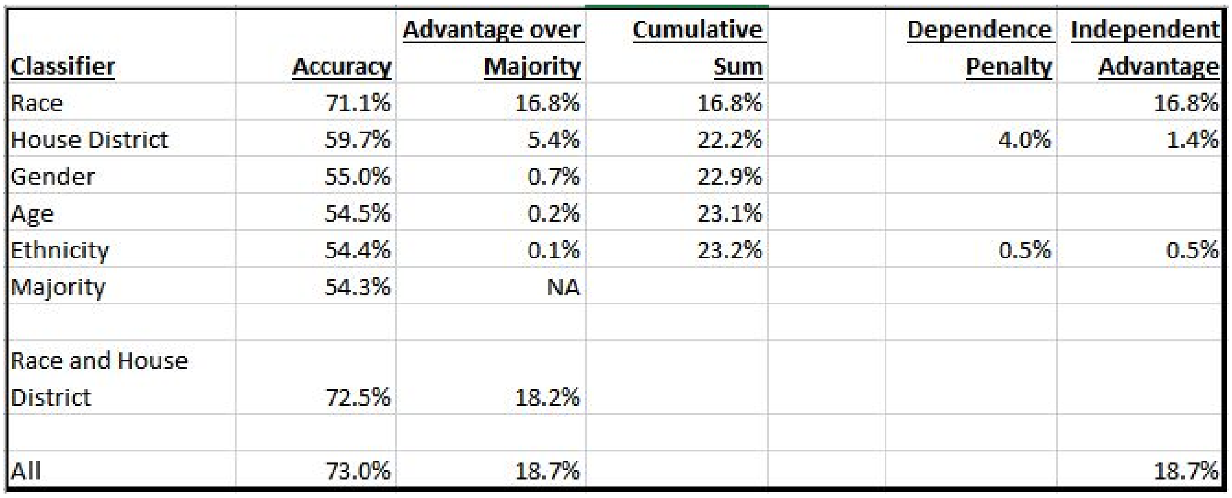


Classifier	Accuracy	Advantage over Majority
Race	71.1%	16.8%
House District	59.7%	5.4%
Gender	55.0%	0.7%
Age	54.5%	0.2%
Ethnicity	54.4%	0.1%
Majority	54.3%	NA

Figure 2

By a large margin, race had the most influence on party affiliation adding 16.8% accuracy by itself. House district was the next most influential demographic, yielding a 5.4% increase in accuracy by itself. Surprisingly, neither gender, age, nor ethnicity provided significant increases in accuracy implying that these three demographics had minimal influence on party affiliation.

Figure 3 below expands the analysis to include a classifier that combines race and house district, and a final classifier that adds gender, age, and ethnicity to the mix.



Classifier	Accuracy	Advantage over Majority	Cumulative Sum		Dependence Penalty	Independent Advantage
Race	71.1%	16.8%	16.8%			16.8%
House District	59.7%	5.4%	22.2%		4.0%	1.4%
Gender	55.0%	0.7%	22.9%			
Age	54.5%	0.2%	23.1%			
Ethnicity	54.4%	0.1%	23.2%		0.5%	0.5%
Majority	54.3%	NA				
Race and House District	72.5%	18.2%				
All	73.0%	18.7%				18.7%

Figure 3

Adding house district to race results in an 18.2% advantage in accuracy, an increase of only 1.4% over race alone. The reason that this is less that the 5.4% advantage house district yields by itself is due to the fact that **race and house district are highly correlated with each other.** In other words, a voter's race has a strong influence on which house district they live in and vice-versa. So this results in a reduction of actual information conveyed to the model by adding house district. I will call the difference between the sum of individual model advantages and the actual advantage of the combined model the "dependence penalty".

Lastly Adding gender, age and ethnicity increases accuracy by 0.5% also implying a correlation between these three demographics and race and house district.

In summary, race clearly influences party affiliation more than any demographic, followed very far behind by house district. Surprisingly, gender, age, and ethnicity have an extremely small (if any) influence on party affiliation.