General Assembly, DAT 2
Final Project; Part 2
Project Design Writeup

## Project Problem and Hypothesis:

The primary focus of this project is predicting voter turnout during local elections in the United States. To elaborate and for the purposes of this project, a "local election" is defined as a municipal, county-wide, or state level election in order to distinguish these elections from a federal level election. This project aims to develop a predictive model which will shed light upon the factors which influence individuals to turn out and vote in non-federal level elections.

The variable we are trying to predict in this project is binary as we want to determine an individual's propensity to vote or not vote during a given election cycle based on known variables.

This project could have a strong impact on several fronts. The first of which being that local elections are only beginning to be studied as research into voter turnout in federal/national elections has taken precedence. Thus, this study will make significant inroads into a heretofore understudied segment of the electorate here in the United States. Secondly, this study can be of value for think tanks and nonprofit voter mobilization groups who are attempting to turnout the vote at a local level in any upcoming election. Finally, local campaigns can use the results garnered by this type of study in order to maximize the effectiveness their, at times, limited resources by spending money "where it counts" to turn out their desired constituents on election day.

## Datasets:

The data for this type of project is readily and freely available from the office of the Secretary of State for every state in the United States. Requests for information have been submitted for the states of Georgia, New York, Kansas, Florida, and Hawaii. These data sets are historical voting rolls for every registered voter within the state who participated in the election in question. For this study, I will be focusing on elections in the year 2014. Sample column headers for a dataset from Georgia have been included in Appendix A.

## Domain Knowledge:

Previous political science research at the undergraduate level give me, as a researcher, a bit of insider knowledge. However, I have never dealt with the United States political arena as an area of research. I'll be relying on academic papers in order to gain background knowledge into this area of study. There do exist some studies (Hajnal & Lewis (2003)) but they do not use granular data like individual voter rolls as data sets for their studies. An example of their outcome has been included in Appendix B.

## Project Concerns:

The volume of data is a bit of a concern for this project. With five states worth of voter roll data, I might be significantly backlogged in my efforts to come up with an effective model for this project. Further, I'm planning to employ a linear and/or logistic regression depending on what variables I'll be testing. However, this being said I am concerned that there might be more effective ways to employ machine learning methodologies that I'm not currently aware of.

For this project I wish that I had access to party records so that I can desegregate the data into three different categories: Republican, Democrat, Unregistered/Independent. This way I can measure turnout by party and control for this variable in my model.

There do exist risks for this project. If the data is incorrect, (theoretically) campaigns might misallocate resources in the election cycle going after the wrong voters to increase turnout. This leads to wasted resources that can never be recovered and ultimately may cost a candidate an election.

## **Outcomes:**

My model doesn't necessarily have to be complicated as I currently have limited access to data about individuals found in the voter rolls. More specifically, I have access only to the following data: RACE, GENDER, BIRTHDATE, DATE_LAST_VOTED, PARTY_LAST_VOTED…. As such, the linear model might not need to be more complicated than including these variables.

If the project is ultimately unsuccessful, I anticipate that I will have to find another data set to test and train a model on. In this case perhaps either narrowing or widening the data set and variables will be required.

Appendix A:

*Georgia Local Election Data (2014), Sample Headers*

COUNTY_CODE
REGISTRATION_NUMBER
VOTER_STATUS
LAST_NAME
FIRST_NAME
MIDDLE_MAIDEN_NAME
NAME_SUFFIX
NAME_TITLE
RESIDENCE_HOUSE_NUMBER
RESIDENCE_STREET_NAME
RESIDENCE_STREET_SUFFIX
RESIDENCE_APT_UNIT_NBR
RESIDENCE_CITY
RESIDENCE_ZIPCODE
BIRTHDATE
REGISTRATION_DATE
RACE
GENDER
LAND_DISTRICT
LAND_LOT
STATUS_REASON
COUNTY_PRECINCT_ID
CITY_PRECINCT_ID CONGRESSIONAL_DISTRICT
SENATE_DISTRICT
HOUSE_DISTRICT
JUDICIAL_DISTRICT COMMISSION_DISTRICT
SCHOOL_DISTRICT
COUNTY_DISTRICTA_NAME
COUNTY_DISTRICTA_VALUE
COUNTY_DISTRICTB_NAME
COUNTY_DISTRICTB_VALUE
MUNICIPAL_NAME
MUNICIPAL_CODE
WARD_CITY_COUNCIL_NAME
WARD_CITY_COUNCIL_CODE
CITY_SCHOOL_DISTRICT_NAME
CITY_SCHOOL_DISTRICT_VALUE
CITY_DISTA_NAME
CITY_DISTA_VALUE
CITY_DISTB_NAME
CITY_DISTB_VALUE
CITY_DISTC_NAME
CITY_DISTC_VALUE
CITY_DISTD_NAME
CITY_DISTD_VALUE
DATE_LAST_VOTED
PARTY_LAST_VOTED
DATE_ADDED
DATE_CHANGED
DISTRICT_COMBO
RACE_DESC

LAST_CONTACT_DATE
MAIL_HOUSE_NBR
MAIL_STREET_NAME
MAIL_APT_UNIT_NBR
MAIL_CITY
MAIL_STATE
MAIL_ZIPCODE
MAIL_ADDRESS_2
MAIL_ADDRESS_3
MAIL_COUNTRY

Appendix B:

*Urban Voting Analyses; Hajnal & Lewis (2003)*

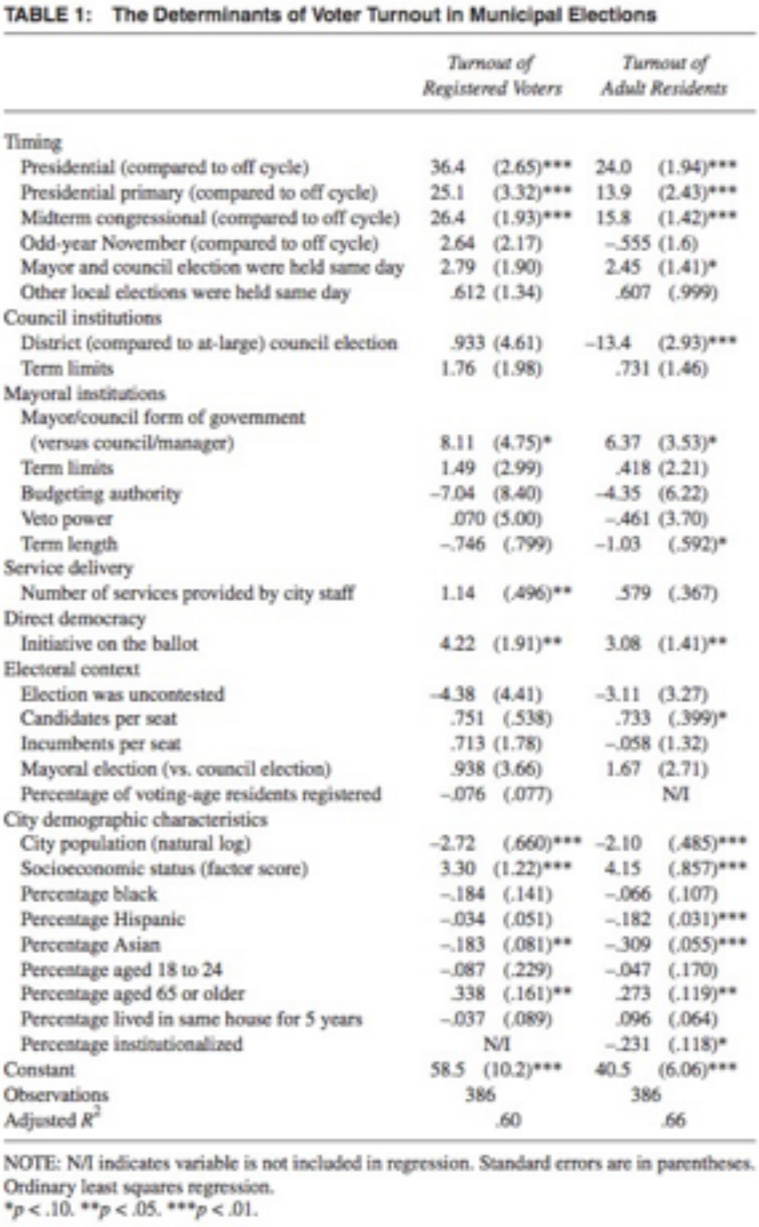

**TABLE 1: The Determinants of Voter Turnout in Municipal Elections**

| | *Turnout of Registered Voters* | | *Turnout of Adult Residents* | |
|---|---|---|---|---|
| Timing | | | | |
| Presidential (compared to off cycle) | 36.4 | (2.65)*** | 24.0 | (1.94)*** |
| Presidential primary (compared to off cycle) | 25.1 | (3.32)*** | 13.9 | (2.43)*** |
| Midterm congressional (compared to off cycle) | 26.4 | (1.93)*** | 15.8 | (1.42)*** |
| Odd-year November (compared to off cycle) | 2.64 | (2.17) | –.555 | (1.6) |
| Mayor and council election were held same day | 2.79 | (1.90) | 2.45 | (1.41)* |
| Other local elections were held same day | .612 | (1.34) | .607 | (.999) |
| Council institutions | | | | |
| District (compared to at-large) council election | .933 | (4.61) | –13.4 | (2.93)*** |
| Term limits | 1.76 | (1.98) | .731 | (1.46) |
| Mayoral institutions | | | | |
| Mayor/council form of government (versus council/manager) | 8.11 | (4.75)* | 6.37 | (3.53)* |
| Term limits | 1.49 | (2.99) | .418 | (2.21) |
| Budgeting authority | –7.04 | (8.40) | –4.35 | (6.22) |
| Veto power | .070 | (5.00) | –.461 | (3.70) |
| Term length | –.746 | (.799) | –1.03 | (.592)* |
| Service delivery | | | | |
| Number of services provided by city staff | 1.14 | (.496)** | .579 | (.367) |
| Direct democracy | | | | |
| Initiative on the ballot | 4.22 | (1.91)** | 3.08 | (1.41)** |
| Electoral context | | | | |
| Election was uncontested | –4.38 | (4.41) | –3.11 | (3.27) |
| Candidates per seat | .751 | (.538) | .733 | (.399)* |
| Incumbents per seat | .713 | (1.78) | –.058 | (1.32) |
| Mayoral election (vs. council election) | .938 | (3.66) | 1.67 | (2.71) |
| Percentage of voting-age residents registered | –.076 | (.077) | N/I | |
| City demographic characteristics | | | | |
| City population (natural log) | –2.72 | (.660)*** | –2.10 | (.485)*** |
| Socioeconomic status (factor score) | 3.30 | (1.22)*** | 4.15 | (.857)*** |
| Percentage black | –.184 | (.141) | –.066 | (.107) |
| Percentage Hispanic | –.034 | (.051) | –.182 | (.031)*** |
| Percentage Asian | –.183 | (.081)** | –.309 | (.055)*** |
| Percentage aged 18 to 24 | –.087 | (.229) | –.047 | (.170) |
| Percentage aged 65 or older | .338 | (.161)** | .273 | (.119)** |
| Percentage lived in same house for 5 years | –.037 | (.089) | .096 | (.064) |
| Percentage institutionalized | N/I | | –.231 | (.118)* |
| Constant | 58.5 | (10.2)*** | 40.5 | (6.06)*** |
| Observations | 386 | | 386 | |
| Adjusted $R^2$ | .60 | | .66 | |

NOTE: N/I indicates variable is not included in regression. Standard errors are in parentheses. Ordinary least squares regression.

*$p$ < .10. **$p$ < .05. ***$p$ < .01.