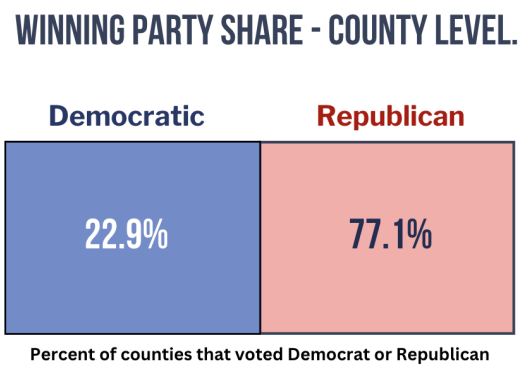
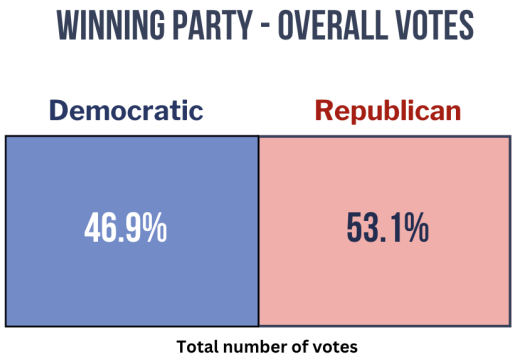
Decoding the 2016 Ballot: Analyzing County Demographics & Voting Partiality

The 2016 presidential election captivated audiences worldwide with its dynamic candidates, political platforms, and goals for the future of American society. As many political scientists and statisticians nationwide attempt to forecast election results by investigating key influencing demographics, we attempt to do the same below. Leveraging 2016 primary election data - we investigated the broader research question: what demographic factors influence voter party tendencies on the county level?

To answer our research question and build a predictive machine learning model, we utilized national primary election result data and county demographic data. Our primary election result dataset consisted of a plethora of interesting information; however, for the sake of simplicity, we leveraged only the following information: political party affiliation, vote count, and fraction votes. For our county demographic data, we looked only to utilize weighted columns that potentially determine county voting tendencies toward Republican or Democratic partiality. Selected weighted columns that we explored via the following general categories include population density, local economic activity, and age.

Overview





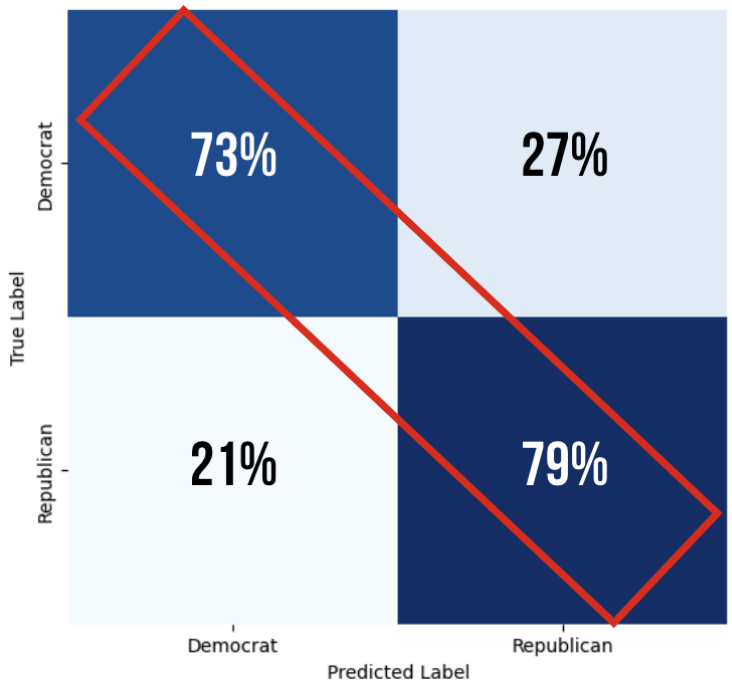
Our analysis demonstrates how cumulative votes were distributed nationally and locally among counties. County-wise, Republicans secured a significant amount of county votes of 77.1% compared to Democrats at 22.9%. However, in terms of the overall vote shares, it was a closer call, with Democrats earning 46.9% of the total share while Republicans earned the remaining 53.1%. This discrepancy highlights a data imbalance of over ¾'s regarding the number of Republic-leaning counties.

Data Cleaning:

One systematic challenge posed in our analysis was ensuring consistency with our data quality. Approximately 35.6% of data values were missing. Furthermore, states such as Alaska, Colorado, Connecticut, Illinois, Kansas, Maine, Massachusetts, North Dakota, Rhode Island, Vermont, and Wyoming had missing information. These minor inconveniences were addressed prior to merging our datasets. As for merging our data, we employed the FIPS column in each of our datasets. FIPS stands for Federal Information Processing Standards - essentially it is an identification number for specific geographic areas, in this case that would be our counties.

The Predictive Model & Results

Adopting a **logistic regression model**, we attempt **to classify counties voting tendencies** as either



Democratic-leaning or Republican-leaning **based on demographic** factors. Our confusion matrix and metrics significance are in detail below.

Model Metrics:

- **77.5% accuracy:** Overall **correct** predictions
- **90% precision:** The model is **rarely wrong** when it says "Republican."
- **79% recall:** It captures **most** Republican counties.
- **82% AUROC:** Good **discerning ability** between Republican-voting counties from Democrat-voting counties

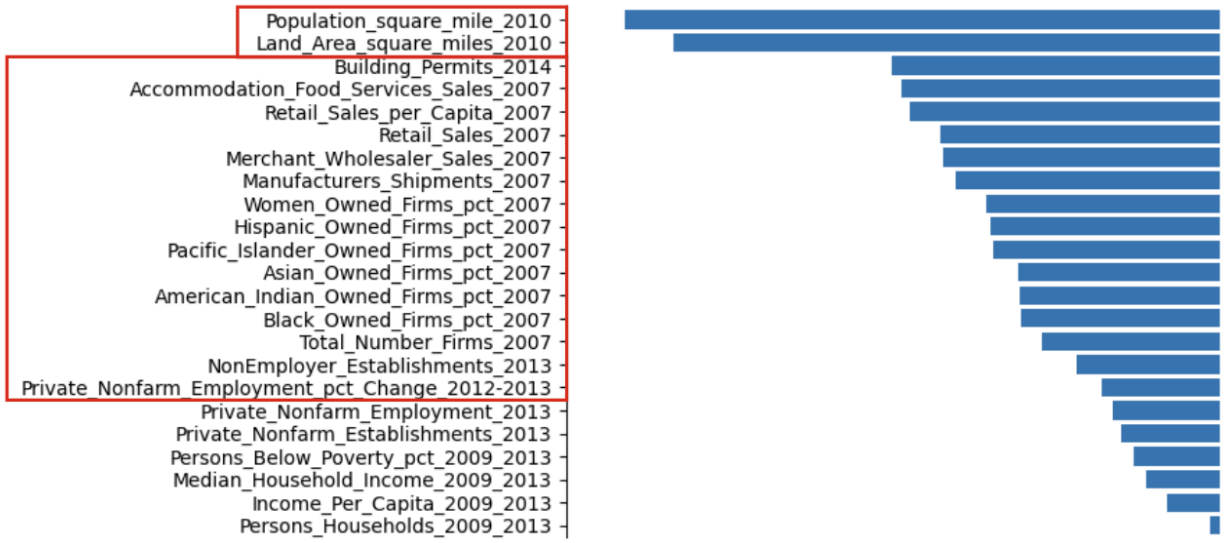
Building The Model

We start by selecting numeric features from our dataset and standardizing them using a **StandardScaler**. Next, we create and train a **Logistic Regression** model with **L1 (Lasso) regularization** and a **balanced class weight in 80%** of our data (the remaining **20% is saved for testing**). The Lasso penalty performs feature selection by shrinking less important coefficients to zero. The balanced class weight compensates for any imbalance in the label distribution (e.g., if there are more Republican-winning counties than Democratic ones, or vice versa).

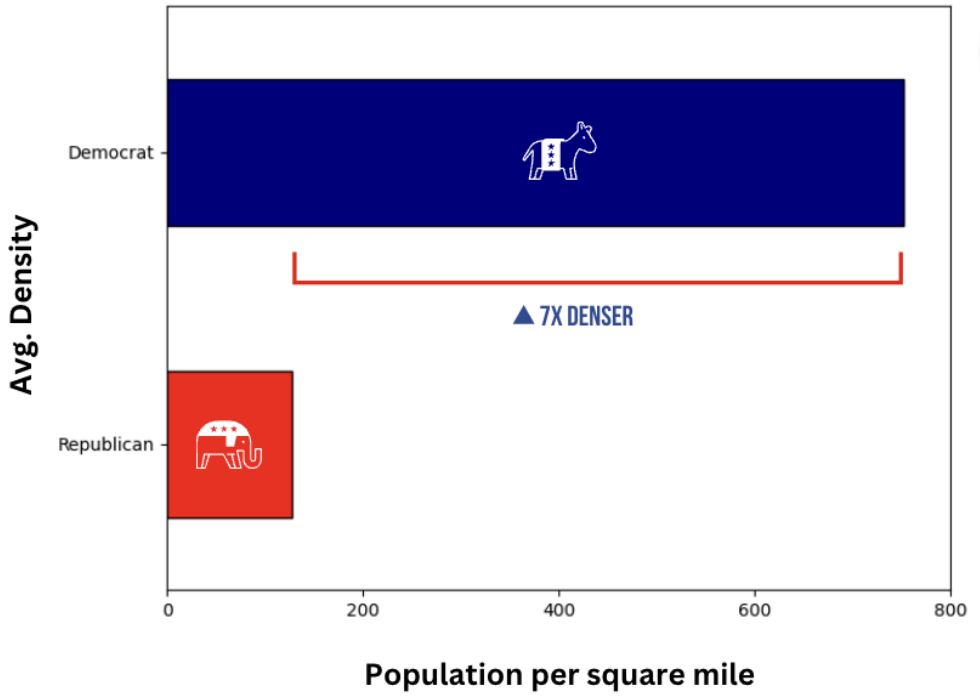
These are the **feature coefficients**, giving us insight into **which factors most strongly push a county toward each party** in the model's eyes.

Democratic Feature Coefficients

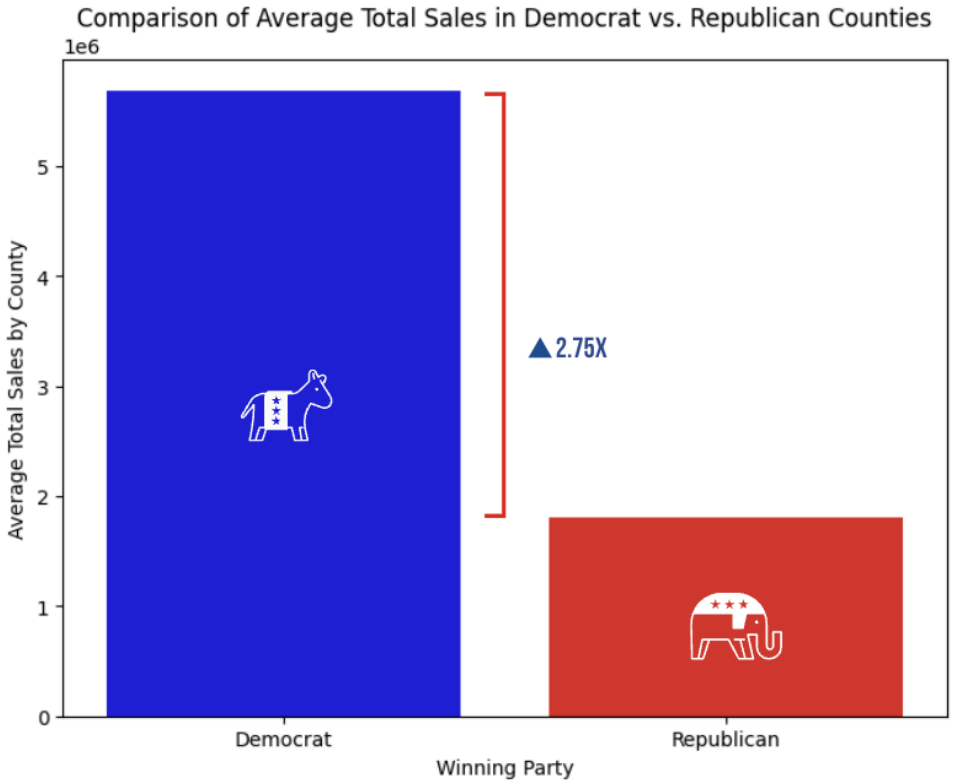
For the democratic winning counties, these are the features with the highest coefficients:



Density (the first red box) **is the strongest group of features for democratic counties**. This analysis revealed that **Democratic-leaning counties** have an **average population density 7 times higher than** their **Republican counterparts.**



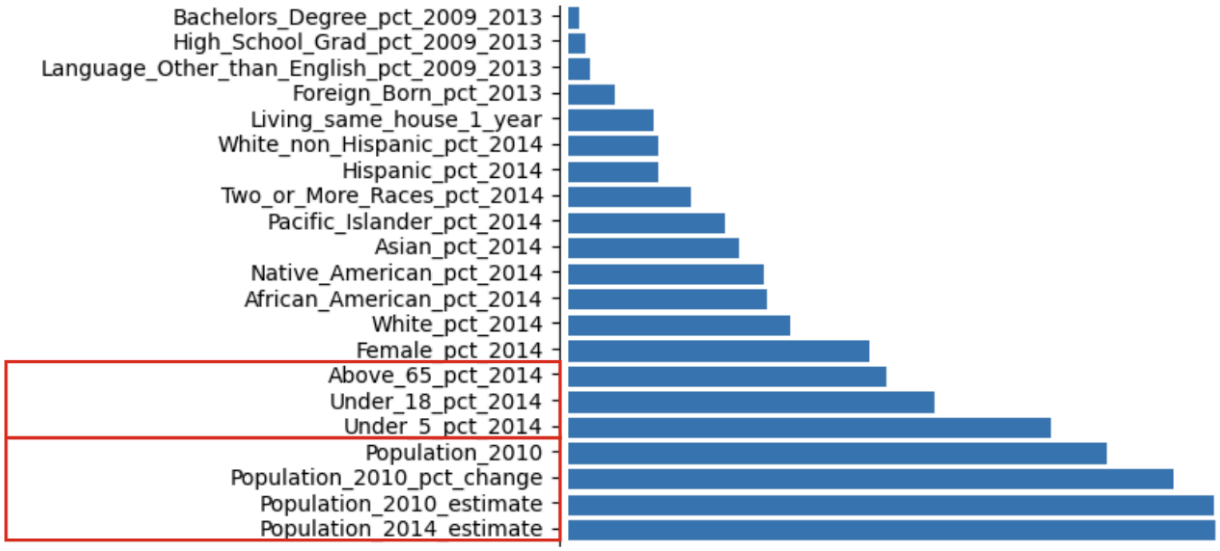
To confirm our analysis, we looked to an external reputable research source. **(Pew Research Center, 2024) confirms** our population density inkling, stating that over the past two decades, there has been **robust support for the Republican party among rural counties.** Nevertheless, in 2016, there was **65% of Democratic party support among urban counties.**



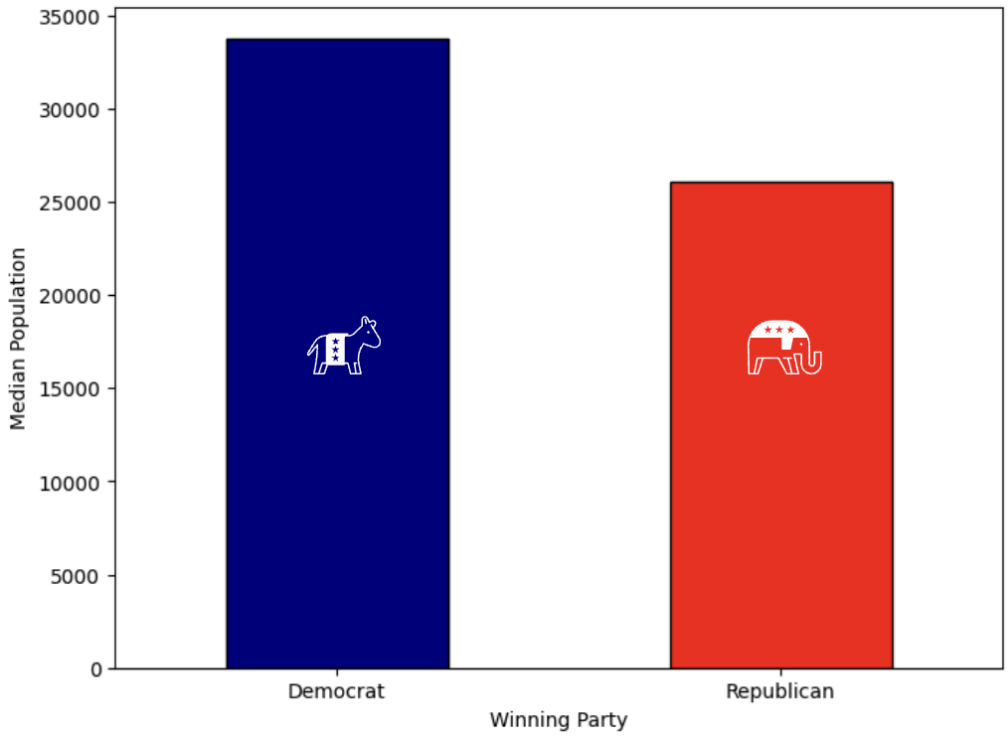
Economic activity (the second red box) features also **highlight** the **concentration of Democratic voters** in **economically active urban counties**, with **2.75 times higher** total business **sales compared to** the **Republican counties.**

Confirming our suspicion for understanding how economic activity influences voter tendencies above, we looked to (Muro, Whiton & Maxim, 2020) for the **2016 political-economic context** regarding the overall final presidential results. (Muro, Whiton & Maxim, 2020) gives us reason to believe that counties with higher economic activity tend to vote Democrat, while less economically involved counties vote Republican. As noted in 2016, despite fewer **Democrat-voting counties overall**, they **made up the majority of the national aggregate economy at 64%, with only 472 counties voting blue**. **Republican-leaning counties comprised 36%** of the national aggregate economy, **with 2584 counties voting red.**

Republican Feature Coefficients



Population (the first red box) **is the strongest group of features for republican counties**. **However**, after looking into our data, we noticed that **higher population counties also vote democratic.**

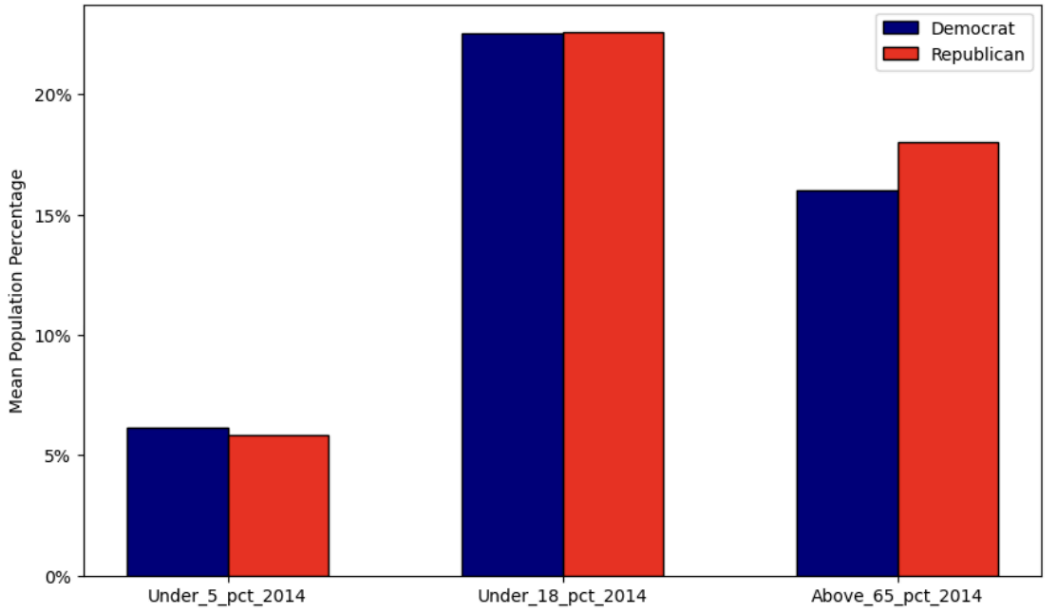


This may make it seem like our model is contradicting itself, but there could be multiple reasons why this occurs.

Why is this the Case?

- **Data Imbalance:** Because there are many small counties that vote Republican, the model can end up associating higher population with a Republican outcome.
- **Other Variables:** Once we account for factors like race, poverty, and age groups, the partial effect of population on the model's predictions can reverse direction.
- **Population vs. Density:** Counties with a large total population aren't necessarily urban or densely populated, so raw population alone might not capture the urban–rural divide.

Age groups (the second red box) also **highlight** how **Democratic voters had a higher percentage of children under** the **age of 5,** while **Republican voters** tended to **have** a **higher percentage of teens under** the age of **18 and seniors over** the age of **65.**



Feature Engineering

Variance Inflation Factor (VIF)

We wanted to **fine tune the model by reducing multicollinearity in** our **data**.This is because **a high VIF could mean that a feature is redundant** because it shares too much information with other features. This, **in turn, can make** the **model unstable.** We **condensed multiple highly correlated features into three general ones**. The **correlation threshold** we used was **= 0.8.**

These are the three condensed columns along with what they contain:

- **County_Size_Index**

Population 2010	Households 2009_2013	Population 2014 estimate	Housing_Units 2014	...

- **Youth_pct_2014**

Under_5_pct_2014	Under_18_pct_2014

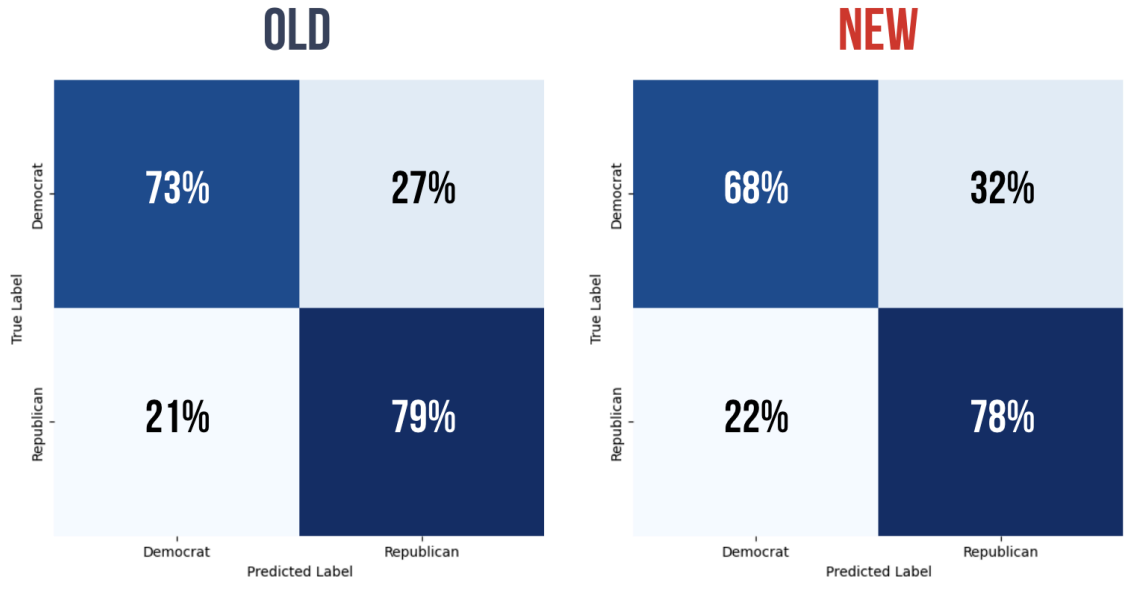
- **Income_Index**

Median_Household_Income 2009_2013	Income_Per_Capita_2009_2013

With this cleaned data we built the model again.

Logistic Regression Model v.2

However, after comparing the **new model's performance** with the old one, we see that it **actually is worse.**



Although the **features** are highly correlated, they still have a **significant impact individually.** So **dropping or merging the features hurts the model predictive power.**

Concluding Thoughts:

Overall, our analysis found that Democratic-leaning counties tend to have 7 times more substantial population density than Republican-leaning counties. We also found that, on average, Democratic-leaning counties have 2.75 times more commercial activity compared to Republican-leaning counties. This evidence leads us to believe that more urbanized and wealthier countries tend to have Democratic voting tendencies and vice-versa for our Republican county counterparts. Nevertheless, we did notice that for our Republican-leaning counties, households with residents under 18 in their households and above 65 leaned Republican. At the same time, "newer" families with family members under five years old tended to vote slightly more Democratic-leaning. Lastly, we found that our predictive machine learning model tended to predict with an accuracy level of 77.5% between political party county winners. However, we firmly believe that due to the data imbalance of having more Republican-winning counties, our model tends to predict Republicans with differentiating population counts inaccurately. Last but not least, we wanted to clarify our motives for this project were solely for our academic development as we combined our data analysis skills with more advanced statistical machine-learning techniques.

Github Repository: 2016-US-elections-prediction-ML