Age of Congress Project Report

Spring Yan | yan.sp@northeastern.edu | Team 72
DS2500: Intermediate Programming with Data
Dr. Laney Strange
April 2024

Background

Age has been a frequently discussed topic in the news, particularly with the upcoming presidential election in sight. From congressional hearings to calls for term limits, older congress members have been the subject of national headlines. While media coverage is a valuable perspective, it is also essential to use data to understand the full scope of the issue. It is important to discuss and analyze trends in the ages of congress members, especially as the makeup of the US changes and younger generations become more politically active. The current issues that the country and the world are facing depend on the policies our elected officials implement. One does not have to be a political junkie to be concerned about the direction that American politics has been going. My aim for this project is to (1) give greater insight into where exactly Congress is heading in terms of age and (2) how these trends impact congressional productivity.

To address both the age and productivity trends in Congress, I used two datasets with time increments based on congressional class. The first and primary dataset I used was from FiveThirtyEight and it contains the age demographics of every congressmember who has ever served from the 66th to 118th Congress. The second dataset that I used was from GovTrack.us, which tracks the congressional activity of the 93rd to 118th Congress. Below I have outlined the details of the data collection and potential biases of these datasets.

FiveThirtyEight | Congress Age Demographics from the 66th - 118th Congress

FiveThirtyEight compiled their dataset from multiple external sources including the biographical directory of the US Congress, the US House, the US Senate, the US GitHub, and VoteView.com. The dataset has thirteen features, and the ones that I thought were particularly interesting were the cmltv_cong (number of terms served), party_code, generation, and chamber (House or Senate) features. As congress

members are public figures, their personal information (e.g., birthdays and full names) is widely available online. However, to address these privacy concerns, I decided to use specific names and birthdays in a limited manner. Other than using specific congress members' info to evaluate the accuracy of my models, I did not include any personal information in any other area of my project. In terms of data biases, this dataset excludes different subsets of the American population. All the areas where under-representation exists in Congress also exist in the dataset. Additionally, anyone under the minimum age requirements, which are 25 for the House and 30 for the Senate, is not represented in the dataset.

GovTrack.us | Congress Activity from the 93rd - 118th Congress

GovTrack.us directly pulls their numbers from the congressional records stored on the Congress.gov site. The data is frequently updated which is important to note as the current 118th congressional class is still in session until January 3rd, 2025. The data is fairly straightforward with nine features, eight of which pertain to the stage that a bill can be in, such as enacted [passed], voted on, or failed. There are no significant privacy concerns to note, but it is important to differentiate between the types of bills passed. This data from GovTrack.us gives equal weight to every bill, but this is not representative of the quality or the substance of the bill. Naming a monument and passing a national healthcare bill share the same weight which could be misleading if not carefully interpreted. With that being said, I believe it is still valuable to use these numbers to get some insight into the relationship between age and congressional productivity over time. The key distinction is that these numbers are not treated definitively, and that further analysis into the types of bills is necessary for broader application.

Methods

Throughout my project, I used three main methods to analyze both datasets: (1) time series analysis, (2) web scraping, and (3) multiple regression.

Time Series Analysis

The FiveThirtyEight dataset maintains consistent time increments, therefore time series analysis was applicable to visualize the fluctuations in age over time. This foundational approach allowed me to gain

insights into the patterns within the data before going on to introduce more complex methods. Visualizations, particularly line plots, served as a powerful tool to show the trends in mean age, and the differences amongst specific demographics (i.e., House vs. Senate and Democrat vs. Republican).

Web Scraping

Since GovTrack.us was hosted online, I used the BeautifulSoup library to directly scrape the data into a data frame. As straightforward as it seems, cleaning the HTML table was more tedious than I suspected. Once all the data was cleaned, I applied time series analysis to the data and created the visualizations featured in the results section. It is important to highlight that I applied normalization to all of these plots as the mean age and the number of bills are not within the same scale. In order to get an accurate representation of the relationship between these features, normalization was essential.

Multiple Linear & Polynomial Regression

Moving on to the machine learning aspect of my project, multiple linear and polynomial regression were the data science algorithms I used for predicting future ages. Throughout my project, I made different iterations of multiple regression models and there are three iterations I particularly wanted to highlight. The first is the multiple regression model I applied to congress members who served or are currently serving in the Senate. This decision came to fruition when I tried creating a regression model on the entire dataset and the mean squared error came back absurdly high. After consulting with multiple TAs and completing additional research, I concluded that narrowing down the data was the next best step. Continuing on that line, the second multiple regression model I created was focused on the congress members representing Massachusetts. Finally, the last regression model I ran was centered on the congress members in the data set who experienced their first term serving at some point within the 66th to 118th congressional classes.

To summarize the high-level steps, I first performed a correlation analysis of all the features against one another. From there I picked out the most strongly-correlated features and normalized the data for regression analysis. Even though I had an inkling that polynomial regression would outperform linear regression, I tested the linear regression model on the training data first before finding the optimal degree

for the polynomial regression model. Once the model was built, I used mean squared error and R-squared to evaluate the model's performance.

In order of most to least accurate, the entry age (EA) polynomial regression model performed the best. Next followed the Massachusetts polynomial regression model and then finally the Senate polynomial regression model (see Figures A & B).

Figure A

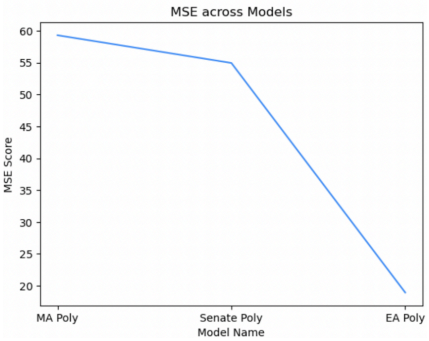
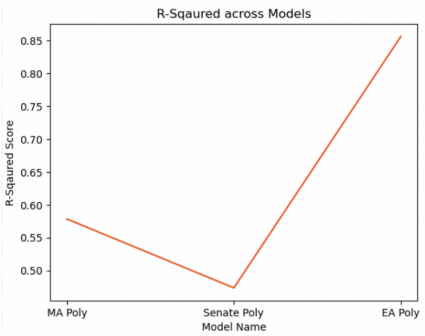
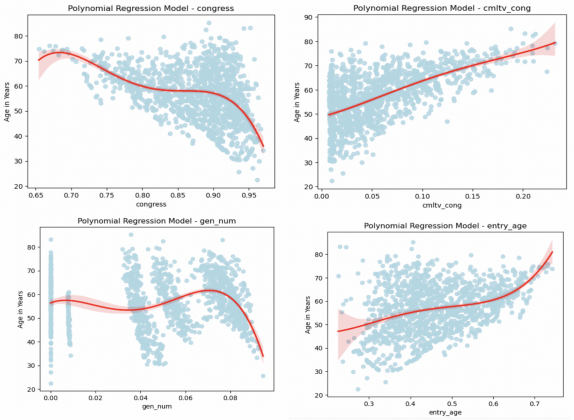


Figure B



The overall trend of the regression modeling demonstrates that polynomial regression was more accurate across the board. This underscores how the relationship between features, such as political party and number of terms served, does not have a linear relationship with age. As for the differences in accuracy between the three regression models, I hypothesize that it is related to the variations in the data. Unsurprisingly, the Senate polynomial regression model performed the worst. The data points were the most widespread, making it hard to find a strong enough correlation between the features. The Massachusetts polynomial regression model performed a bit better because, traditionally, the state backs Democrats and the overall shape of the data was smaller. Furthermore, the entry age polynomial regression performed the best which can be explained by the higher correlated features that were included in the model. Adding the feature of entry age was a major addition to boosting the accuracy of the regression model.

Figure C - Poly EA Model

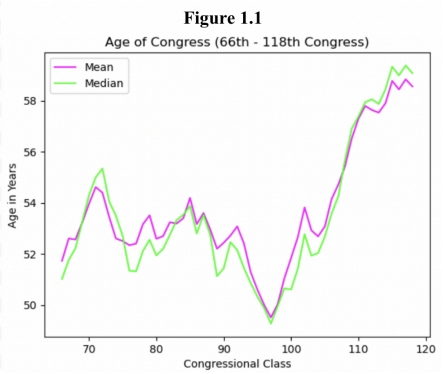


Results and Conclusions

Between the two datasets I selected and the three data science methods I applied in this project, the results can only be properly represented in relation to their original methods.

Results from Time Series Analysis

Figure 1.1



To start with the time series analysis, the visualizations convey that both the mean and median age of congress members are rising (Figure 1.1). Even amongst different demographics, the same trend was reflected. Within the comparison between the mean ages of Democrats and Republicans, one can spot a shift at around the 90th congressional class (Figure 1.2). Before the 90th class, Republicans were consistently older, but after that mark, Democrats have gotten older, especially in the most recent congresses. In regards to the House versus Senate comparison, it is important to note the Senate has been getting older within the past five congressional classes while the House has gotten slightly younger (Figure 1.3). When applying real-world context, this checks out as the minimum age to serve in the House

is five years younger than in the Senate. Younger generations, such as Millennials and Gen Z, have also run in more congressional races as of late (Pew, 2023).

Figure 1.2

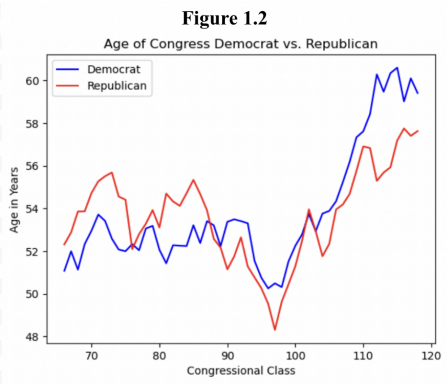
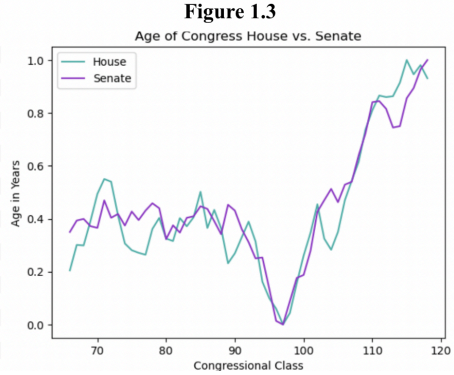


Figure 1.3



Results from Web Scraping Productivity Analysis

From the 66th to 118th Congress, the percentage of enacted legislation has stayed below ten percent. The bar chart below highlights this pattern and offers greater insight into how difficult it is to get a bill passed in both chambers (Figure 2.1).

Figure 2.1

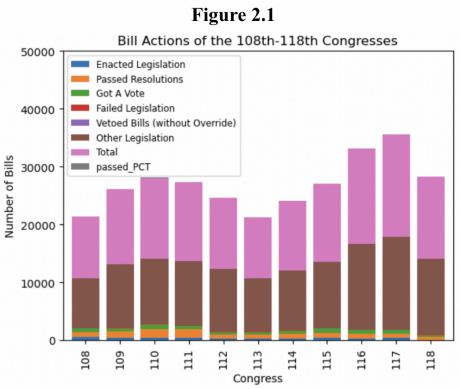
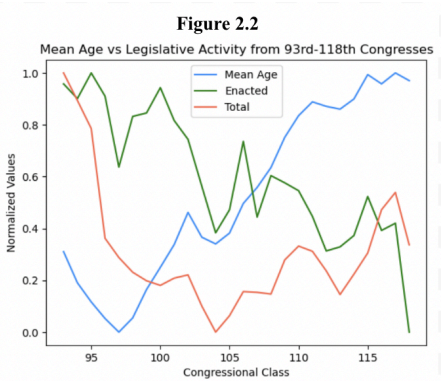


Figure 2.2



From the 115th Congress to now, it is clear that the total and enacted number of bills is going down while the mean age is rising (Figure 2.2). While this is not enough to prove that age is directly correlated with overall congressional productivity, it does suggest that a meaningful relationship exists.

Moreover, the percentage passed versus mean age plot shows that an increase in mean age does not lead to a decrease in the percentage of bills passed (Figure 2.3). It is no secret that Congress is becoming increasingly polarized which is evident in the percentage passed versus the two major political parties plot. The farther the distance is between the number of Democrats and Republicans, the more the

percentage passed seems to drop (Figure 2.4). These three plots demonstrate that congressional age and productivity cannot be directly associated, and other external factors such as political polarization influence the number of bills passed (Solender, 2023).

Figure 2.3

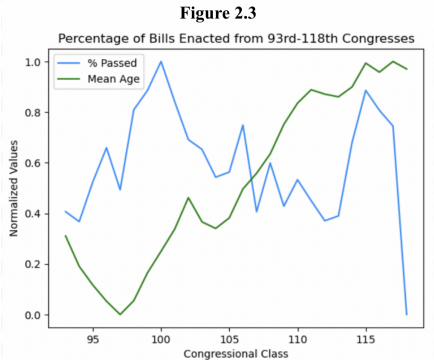
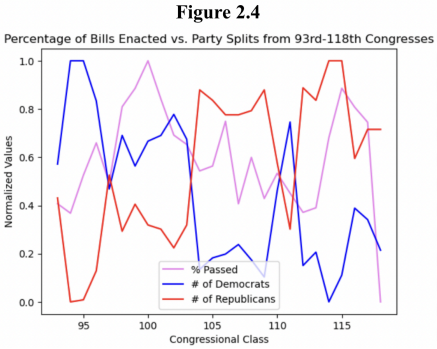


Figure 2.4



Results from Multiple Regression Modeling

As I touched upon in the data science approaches section, the accuracy and results from the multiple regression varied from model to model. These variations can be explained by the composition of the subsets of data themselves, and the lack of correlation between some of the features utilized in the models. As a whole, age is challenging to predict without features that are not strongly correlated with age. This was evident in the drastic differences in R-squared and MSE scores between the Senate, Massachusetts, and entry-age regression models (Figures 3.1 & 3.2).

Figure 3.1

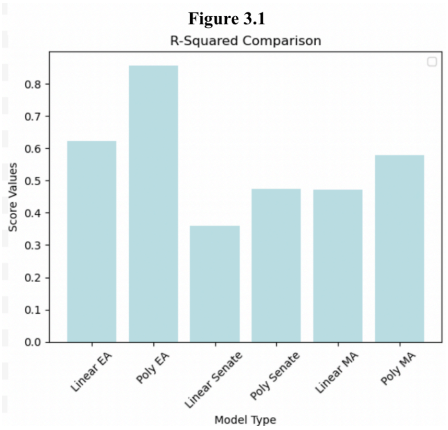
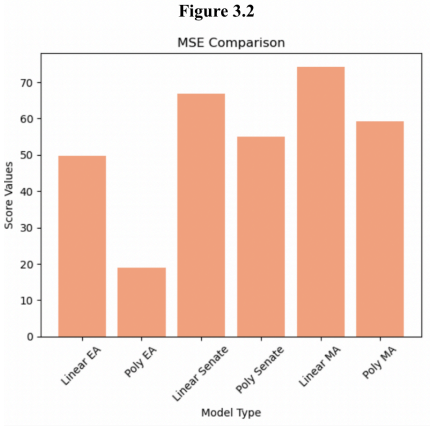


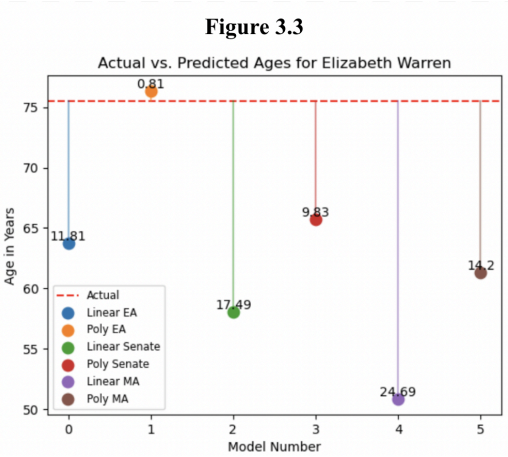
Figure 3.2



Even for the entry-age model, it was still tricky to predict age using demographics, such as a chamber or political affiliation. With these challenges at hand, I learned how to stay persistent and continuously fine-tune regression models. The central pattern I observed was models performed better with features

that were more strongly correlated with age. It sounds intuitive, but without the several regression models I tested, I would not have picked up on this pattern. In addition, I tested the same prediction across the three models, which was forecasting Senator Elizabeth Warren's age at the start of the 119th Congress. The predicted result from the entry-age polynomial regression model was 0.81 years off the actual age Senator Warren would be (Figure 3.3). The other regression models were off by over nine years, which is quite substantial considering the average length of a single generation. Although the entry-age polynomial model performed the best, it is important to mention how the nature of the data needed to predict future ages is faulty if there is missing data or a lack of data.

Figure 3.3



Upon reflection, the results of my project have led me to grow a broader curiosity for the demographic makeup of Congress, beyond simply age or political affiliation. Looking into the future, I hope to expand my project to encompass more text-based data and machine-learning methods. It is equally important to gather insight into how the American public feels about the growing age in Congress and what should be done to address this issue. According to a 2023 Pew Research Center study, seventy-nine percent of US adults favor putting a maximum age limit in place and eighty-seven percent favor imposing term limits (Pew, 2023). With these statistics in mind, it would be worthwhile to conduct a sentiment analysis on the longest-serving congress members and compare the sentiment score of articles and tweets written about them over time. Along a similar vein, it would also be compelling to get the TF-IDF scores of current issues, such as tech regulation and climate change, featured in enacted bills over time. These two new

methods of analysis could offer a broader and more nuanced perspective into why the age of Congress matters to the American people.

References

Congress Demographics (2024). *data/congress-demographics at master · fivethirtyeight/data*. GitHub. [Data Set]. FiveThirtyEight. https://github.com/fivethirtyeight/data/tree/master/congress-demographics

Historical Statistics about Legislation in the U.S. Congress (2023). [Data Set]. GovTrack.us. https://www.govtrack.us/congress/bills/statistics

House gets younger, Senate gets older: A look at the age and generation of lawmakers in the 118th Congress. (2023, January 30). Pew Research Center. https://www.pewresearch.org/short-reads/2023/01/30/house-gets-younger-senate-gets-older-a-look-at-the-age-and-generation-of-lawmakers-in-the-118th-congress/

Pew Research Center. (2023, September 19). *10. How Americans view proposals to change the political system*. Pew Research Center - U.S. Politics & Policy. https://www.pewresearch.org/politics/2023/09/19/how-americans-view-proposals-to-change-the-political-system/#:~:text=Term%20limits%20for%20members%20of,%3B%20just%2012%25%20are%20opposed

Skelley, G. (2023, April 3). *Congress Today Is Older Than It's Ever Been*. FiveThirtyEight; FiveThirtyEight. https://fivethirtyeight.com/features/aging-congress-boomers/

Solender, A. (2023, December 19). *Capitol Hill stunner: 2023 led to fewest laws in decades*. Axios; Axios. https://www.axios.com/2023/12/19/118-congress-bills-least-unproductive-chart