ECS171 Project Proposal

Our group plans to develop a system implementation focusing on analyzing historical congressional legislative data to determine what factors correlate with the likelihood of a bill passing in addition to predicting whether a hypothetical bill will pass by looking at these factors. By leveraging machine learning techniques, we aim to identify key factors that impact legislative success and provide insights into the legislative process.

We will experiment with multiple supervised learning models to identify the most effective approach. This project has both a regression, determining the significance of factors, and a classification aspect, predicting the passage of a bill. We can parse the text within a bill to gain a better understanding of the link between a bill's text and the likelihood of its passage. At a surface level, we can see if the length of a bill has a measurable impact on the chances of it passing. We could also observe the differences in the vocabularies of passed vs. failed bills. Most significantly, however, we could use NLP models, such as offerings based on NLTK, BERT, or spaCy, to analyze the sentiment and the topics covered in the bill text. We can also see how the length of a legislator's tenure, the committee to which a bill was first referred to, which House it was first introduced in, what type of legislation it is attempting to pass, the controlling parties in the House, Senate, and/or Presidency, the party membership of its sponsors, etc. relate to the likelihood of a bill's passing. We can find information about the congresspersons using the Congress.gov API.

In our project, we will be primarily developing within Google Colab and Jupyter Notebooks, using libraries such as scikit-learn, panda, matplotlib, and PyTorch. However, depending on the size of the dataset as well as the network we aim to train, these environments may prove to be insufficient, so we may need a multi-core CPU for parallel processing, a GPU for neural networks, and sufficient RAM for large dataset handling. In the event we'd need additional computing power, depending on cost and convenience, we could either run it locally on our devices or rent out a computing instance on AWS or some other cloud computing provider.

We will be using the publicly available dataset "US Congress: Bulk Data on Bills" from PrePublica, which has metadata for every bill introduced between 1973 and 2021. We will supplement this with further data using the Congress.gov API on an as-needed basis. If, for instance, there are bill attributes missing from a certain time period, using the aforementioned API would allow us to retrieve said missing data, instead of having to remove the entire time period from the testing set.

We hope that this project will allow for a better understanding of what qualities make a bill more likely to be made into law. With political polarization higher today than at any other point in recent history, as well as the volatility of protections, programs, and pieces of legislation in today's political climate, an objective analysis of what factors make a bill more or less likely to be passed would aid in ensuring the success of desired legislation. Researchers, policymakers, and political analysts would be able to benefit from being able to identify the most effective drivers of legislative success, which could in turn lead to legislators crafting their bills more effectively. This project could also democratize the legislative process, as it would be able to show the public the measurable impacts that certain choices pertaining to a bill and its sponsorship have on its potential success.

With this project, we foresee three types of risks regarding data quality, model performance, and technical issues. On data quality, as mentioned above, the dataset we plan to use only spans 48 years and does not include the most recent presidency. The data may be considered incomplete as it does not include many significant periods of American history such as World War Two or the Great Depression. However, this latter issue could prove to lend itself to a better model, as the time frame of 1973 and 2021, especially when compared to earlier time periods, are marked by economic and technological conditions that are more similar to today.

Scaffolding (External Resources)

To enhance our project development, we will utilize a range of external resources to deepen our understanding and application of machine learning techniques. We intend to engage with comprehensive tutorials on platforms like Coursera, edX, and fast.ai that offer both basic knowledge and advanced insights. Additionally, educational YouTube channels such as StatQuest, Sentdex, and 3Blue1Brown will provide clear explanations for algorithms, model evaluation processes, and data preprocessing strategies. Moreover, we'll refer to official documentation for libraries including scikit-learn, PyTorch, and NLTK to guide our implementation efforts while adhering to the latest functions and best practices. For troubleshooting assistance or collaborative problem-solving support we'd turn towards community forums like Stack Overflow , GitHub Discussions,and scikit-learn user groups where experienced developers share solutions & advice . These external resources are fundamental components that will help us navigate challenges efficiently ensuring successful completion

Addressing Project Risks

Outliers: National emergencies or unique political events may create anomalies in the data, requiring careful handling and potential exclusion from the dataset.

Presentability

We'll prepare visuals, including images of code and data to accompany the discussion of our project results.

Figure Inclusion