Will the Bill Make It Through Capitol Hill Project Summary

The Problem

Every year, the United States Congress introduces thousands of bills. Of these myriad proposed bills, less than 0.5% end up becoming law. The aim of the project is to analyze which features affect the likelihood of a bill's passage and use this to build a predictive model for bill passage. In particular, we investigate whether modern techniques using natural language processing can extract useful information about whether a bill will pass directly from the bill's text.

Data Sources

1. congress.gov Much of the information about any given bill is recorded on congress.gov such as the bill's full text, sponsors, whether it passed the house, senate, and other intermediate bill actions. This data is available for free using an API, though the amount API requests is limited to 5000 per hour. This request limit creates a bottleneck due to the sheer number of bills multiplied by the number of requests needed to collect each bill's text, sponsors, actions, etc.
2. opensecrets.org Each year, organizations spend billions of dollars lobbying congress to influence which bills pass. Detailed information about each lobbying agency's expendature are collected by opensecrets.org.
3. Social Media How is a bill's passage correlated to mentions about it on popular social media platforms such as Reddit and Twitter? Unfortunately, acquiring this information can be costly. Unlike congress.gov, using social media platforms' APIs is not free.
4. News Coverage In addition to social media mentions, we wondered about the connection between a bill's passage and how much it was covered by news outlets. Many news outlets don't provide direct access to this information, but we were able to obtain data for free from The Guardian.

Results

Data sources other than congress.gov proved to be unhelpful for modelling likelihood of bill passage.

Lobbying Data

Our goal was to use the lobbying data from opensecrets.org to collect the amount of money lobbied for each bill. After the initial EDA contradicted our expectations, we researched the way that lobbying data is officially reported and how opensecrets.org catalogs this information. This investigation revealed that it is simply not possible to assign a one-to-one correspondence of lobbying amount with each bill from the information reported by the lobby agencies. This is because agencies report lobbying for certain issues, not for any particular bill. The way that opensecrets.org connects this to a bills is by cateloging which issues are relevant to that bill. However, any given bill may be reported as related to many separate issues while, conversely, one issue can be addressed by several distinct bills and lobbying amount is overcounted and then averaged over many related bills.

Media Mentions

The vast majority of bills aren't mentioned on social media. This is perhaps unsurprising since 99% of bills never make it past being introduced. However, even among the small fraction of bills that *do* get mentioned, there was very little meaningful correlation with bill passage. Bills which got mentioned more on social media had the same rate of passage compared with those that didn't. This result was somewhat surprising since we expected bills that were expected to pass to recieve more attention.

Bill Text

Since lobbying data was not possible to connect directly to individual bills and media mentions proved to be unhelpful, we ultimately used the bill's raw text as the main feature.

Final Model

We passed the bill's text through MPNet to encode the text as vectors in a high dimensional vector space. The resulting vectors were passed into several models including a loggrithmic regression, SVM and XGBoost. We selected XGBoost since it outperformed the other models.

Limitation and Next Steps

The extreme class imbalance of the data severely limits the predictive power of all the models we tried. The baseline model of picking by random chance succeeds less than 1% of the time. Even though our model was ssignificantly more successful than random chance, the predictive poower is still not nearly aaccurate enough to be useful in practice. For future work, it would likely be more useful to make a descriptive model rather than a predictive one.