*The Dartmouth Undergraduate Journal of Politics, Economics, and World Affairs*
*Volume. I, Issue. IV, 133-143, July 2022*

# Identifying Potential Pork-Barrel Legislation Using Machine Learning: A Preliminary Analysis

SUNIL GREEN

Purdue University

ABSTRACT: Pork-barrel legislation has been criticized by some as an excessive and potentially corrupt use of Congressional appropriations. The task of finding the specific parts of legislation that have been "pork-barreled", however, requires many hours of labor by policy researchers. Using data from government watchdogs and machine learning algorithms, the research explores the idea of creating a model to flag specific line items of appropriations bills for policy researchers to further explore as potential pork-barrel legislation. The model constructed uses data from the *Earmark Database* by Taxpayers for Common Sense, and the *Congressional Pig Book* by Citizens Against Government Waste to attempt to identify line items in appropriations bills as either "potential pork" or "not potential pork". Criteria for the model are based upon Citizens Against Government Waste's seven criteria for pork-barrel legislation as well as their identification of egregious examples of pork-barrel projects. The research focuses on the fiscal years of 2008-2010 when Congressional rules required the disclosure of earmarks. The first prototype of a machine learning model was trained on fiscal year 2008 data and showed limited effectiveness at differentiating pork and non-pork in the 2008 Consolidated Appropriations Act.

## Introduction: Defining "Pork-barrel" and "Earmark"

Pork-barrel legislation is a phenomenon in political science that has been fairly well studied[1,2] . The specific term, "pork-barrel" has a dark history, and arose by comparing the way slaves rushed towards salted pork with how members of Congress rushed to secure federal funding for projects for their constituencies[3]. This hyper-localized appropriations legislation is often viewed as corrupt, has a long history in American discourse, and has drawn the ire of many politicians over time. Similarly, the term "earmark" often has been used synonymously in much political science literature, and American political discourse with "pork-barrel" to refer to potentially corrupt hyper-local appropriations.[4]

Despite a fair amount of research on pork-barrel politics, and despite the existence of multiple watchdog groups who criticize the earmarking process, there is some disagreement as to what each term means. Both terms have been used in the literature and discourse rather loosely. Of the many groups and institutions that research government spending or provide data about it, three bodies largely influenced this research's definition of "earmark" and "pork-barrel". The government watchdog and nonprofit *Citizens Against Government Waste* (CAGW) commonly uses both terms synonymously and provides their definition of what is an earmark or pork-barrel project in many of their reports as well as on their website.[5] *CAGW* identified seven criteria that it believes describe a "pork-barrel project" (or any earmark in their view), of which only 1 is needed for it to be classified as so:

- Requested by only one chamber of Congress;
- Not specifically authorized;
- Not competitively awarded;
- Not requested by the President;
- Greatly exceeds the President's budget request or the previous year's funding;
- Not the subject of congressional hearings;
- Serves only a local or special interest

While CAGW may use both terms interchangeably, CAGW's Director of Research, Sean Kennedy, was quoted as making some small but important distinctions regarding pork-barrel spending. In an article for *FiscalNote*, Kennedy was reported as saying, "Pork barrel spending is in the eye of the beholder, but it can really be any spending that's added in order to win votes."[6] The discrepancies between the CAGW's official definition of both pork-barrel and earmark and Mr. Kennedy's quote are very important. Whereas pork-barrel projects carry with them the connotation that they are used for the purpose of winning votes, every earmark in a piece of legislation may not have been added explicitly for the main purpose of winning a vote. Operationally then, "pork-barrel" and "earmark" should not automatically be treated as synonyms. Furthermore, according to Kennedy pork-barrel is a broad "nebulous" term[7]. As such, it has been used differently by different people and organizations.

*Taxpayers for Common Sense* (TCS), another nonprofit government watchdog, provides different definitions and a clearer distinction between what it believes constitutes pork and an earmark. TCS, like CAGW, advocates for the abolishment of earmarks and criticizes how the process used to fund an earmark is not as competitive as the processes federal or state agencies use to award a project. In an article explaining earmarks and the earmarking process, however, the group distinguishes what it believes constitutes pork from an earmark. TCS writes that:

> An earmark is not necessarily pork. The term 'pork' is often applied to government spending, especially spending done by way of earmarking. But 'pork' is a loaded term, associated with waste, fraud, or abuse. Many representatives bristle at the notion that their earmarks are pork, pointing out how their earmarks fund roads, support schools, or create jobs.

The article further elaborates, explicitly stating, "Anything that meets our definition of an earmark is included as an earmark in our databases. Our databases do not separate earmarks for 'good' projects from earmarks for 'bad' projects."[8] What separates a "good" project from a "bad" project is the focus of this research. Taking into consideration the implications of the term "pork-barrel" as being wasteful and corrupt, this research classifies instances of "pork-barrel" as those projects that TCS classifies as "bad." Because of the increased specificity of their definition, this research adopts the TCS standards for separating pork from an earmark.

**Research Question**

The goal of this research project is to use machine learning to build a model that identifies earmarked projects that are blatant examples of pork. According to CAGW, nearly all pork-barrel projects are specified in the appropriations bills (or Omnibus spending bills depending on the year) created during the budgeting process.[9] It is important to note that some studies have looked at earmark line items in other bills.[10] Even with an overbroad definition of pork, however, it is clear

that the focus should be on the earmark data, and not the text of the bill itself. Simply put, the nature and characteristics of an earmark determine whether a given provision is or is not pork.

Additionally, a specific change in congressional rules made a focus on earmarks even more analytically compelling. In fiscal years 2008-2010, Congress passed earmark disclosure rules that allowed for large amounts of data to be collected regarding earmarks. In 2011, earmarks were banned via an "earmark moratorium" imposed by the Republican leadership at the time.[11] Of significant note is that in 2021 the earmark moratorium in Congress was lifted[12]. Given the large amount of potential data available and the unique nature of pork as hyper-local legislation with specific characteristics, the topic lent itself to be explored as a classification problem in machine learning. Accordingly, I ask "to what extent can the identification of pork-barrel legislation be automated, and how?" It is important to note that this research makes no normative claims as to the purported evils of pork-barrel projects compared to other earmarks. It simply aims to build a classifier that can help distinguish between the two.

**Significance of Analysis and Tentative Explanations**

While the pork-barrel legislation phenomenon has been documented by researchers before,[13,14,15] only one other paper could be found that tried to use machine learning to study earmarks and pork[16]. Most research regarding pork has used conventional social science research methods rather than data science approaches. Many studies in the field, including the one that built its own machine learning model, have also viewed pork as largely synonymous with the term earmark. As noted above, the current research adopts the more specific TCS definitions of pork and earmark in the construction of its classifier, thereby reducing the likelihood of false positives resulting from automatically equating all earmarks as instances of pork.

There are at least two reasons why machine learning classifiers and the combination of technology and social science research is limited for this topic. First, the fields of machine learning and data science are themselves new and emerging. Consequently, many social science researchers are not yet fully trained in data science methodologies. The potential for research combining these fields remains high, as there exist many other problems that could use classifiers or other machine learning algorithms to help advance research on the topic. Second, although the research problem is ripe for the application of a machine learning approach, the fact is that data on the subject are not standardized, hard to find, and oftentimes inconsistent. These data limitations will be discussed in greater detail below.

**Literature Review**

Research on pork-barrel projects has been consistently pursued in political science over the past few decades, with studies dating back as early as 1979 and continuing to the present day.[17,18] Much of the research itself focuses on either the history of pork-barrel politics or on the application of pork-barrel politics to other theories in political science. A 2007 paper on pork includes an in-depth study on earmarks found in the 2005 highway bill and its relation to theories about legislative malapportionment.[19] The research asserts that smaller states are overrepresented in the Senate and therefore receive more pork in Senate versions of bills as opposed to House versions. A Cambridge study published in 2018 provides an in-depth overview of the presence of pork-barrel politics in American history and analyzes data regarding early local congressional projects[20].

There is also disagreement in the literature as to whether pork-barrel projects have the effect that they are thought to have. One notable paper broke with the body of research in the field by showing statistically significant results that bringing home pork increased a Congress member's vote share.[21] Many earlier papers lacked evidence for this idea. The effectiveness of "pork" for

members of Congress is therefore thought to be somewhat unclear. Regardless, evidence suggests that members of Congress do actively seek out federal funding for hyper-local projects in the hopes of securing their re-election.[22]

Other researchers have used machine learning classifiers and models to solve problems found in the field of political science. Taiwanese researchers used machine learning models to classify the large amounts of documents from The Parliamentary Library of Taiwan's Legislative Yuan.[23,24] The scope of the problem those researchers faced was far greater than that in this research. Much of the data used in their models had more than two categories. Thankfully, pork-barrel legislation is a simple binary classification problem with categorical independent variables and a discrete binary dependent variable.

The most similar research on this topic emerged from a group called *Data Science for Social Good*. Researchers working in a Carnegie Mellon fellowship were able to identify earmarks using text analysis methods and data from the Office of Management and Budget.[25] They then ran analysis on the characteristics of the earmarks and were able to achieve some success in doing so. Like others who have studied this topic, however, they viewed earmarks and pork as synonymous or nearly synonymous. The current paper uses both different data and a different set of definitions regarding "pork-barrel" and "earmark."

The current state of knowledge regarding the automation of the identification of specifically pork-barrel legislation is virtually nonexistent. While research has been done on the pork-barrel as a political phenomenon, and while some analysis has been conducted on earmarks, no research could be found that uses a machine learning model to separate pork from other earmarks. A number of reasons are likely to blame for this lack of research, including the specificity of the problem, the lack of clear standardized definitions for this issue, and the infancy of the application of data science to the field of political science.

**Quantitative Analysis**

As noted above, the objective of this analysis is to use machine learning to construct a classifier to identify pork from data describing a set of earmarks. More specifically, an earmark's state and specific appropriations bill were used in constructing a model that attempts to separate earmarks from pork. The Classification and Regression Trees algorithm (CART) was chosen to build the classifier model. It uses a measure called the Gini impurity to try to differentiate between positive samples and negative samples.[26] It was hypothesized that using an earmark's state and appropriations bill would be able to improve the CART algorithm's classification accuracy.

**Data**

The earmark disclosure rules that were implemented in Congress from fiscal years 2008-2010 provided watchdogs and researchers with special conditions upon which much data on earmarks could be collected. Members of Congress were required to disclose details regarding any "Congressionally directed spending project" (earmarks) that they requested to be added to the appropriations bills.[27] As such, data from this time are excellent for the study of the history and use of earmarks in American politics.

Two main sources were used in the collection of data to train and test a machine learning algorithm. CAGW's yearly *Pig Book Summary* details the earmarks–which they view as synonymous with pork–that they find in each year's budget and spending bills.[28] The book summarizes what they view as the most blatant examples of pork and compares these data to previous years. The projects explicitly mentioned in the Pig Book do not constitute all projects that

members of Congress were required to disclose. Rather, the projects constitute a subsection of earmarks that CAGW views as "the most egregious and blatant examples of pork." The current project derives its data from the 2*008 Pig Book Summary* and TCS's comprehensive earmark database for 2008.[28, 29] That 2008 was the first year that comprehensive earmark data were made available makes it a logical starting point for the analysis; furthermore, additional research will examine the next few years.

**Data Collection**

Data from the *2008 Congressional Pig Book Summary* were not available for automated electronic entry. Consequently, a manual data coding process was used. Each item mentioned in the *2008 Congressional Pig Book Summary* was subsequently searched for in the 2008 TCS dataset. If it was found, it was labeled as a positive pork sample. The criteria for identifying a positive pork sample were as follows:

1. A project that matches the bill section, contains keywords in the description, and has **one** earmark that matches the dollar amount (or is within 2% of the total cost)

   *or*

2. A project that matches the bill section, contains keywords in the description, and has **multiple** earmarks that when summed together match the dollar amount (or is within 2% of the total cost)
   Note: All earmarks associated with the project are added to the positive pork data, a sample may pick out a specific earmark from that project

It is important to note that in the data collection process, a few notable observations and exceptions were made. First, each project mentioned in the CAGW data did not always correspond to exactly one data point in the TCS data, and some CAGW projects contained many TCS earmarks. Furthermore, some groups of rows of earmarks in the TCS data corresponded to one singular project, regardless of whether CAGW identified that project as positive pork. CAGW claimed that 1,188 projects out of the 11,146 projects they identified as being disclosed were included as blatant examples of pork. By manually counting the number of projects detailed in the *Pig Book's* paragraphs for each appropriation bill, a total of 1229 projects were found to have been mentioned by CAWG as being blatant pork, contradicting the number they reported.

Projects were separated by each congressional member with approximately one paragraph per member of Congress. CAGW only mentioned a select few of each member's projects, further complicating and reducing the number of positively labeled pork projects. CAGW did not provide the original full data, which would have aided greatly in the collection of training data needed to build the CART decision tree model.

**$15,115,446** for 17 projects by Senate Appropriations Committee Ranking Member Thad Cochran (R-Miss.), including: $3,723,750 for a Natural Products Lab; $2,780,400 for the Jamie Whitten Delta States Research Center; $1,075,419 for the Agricultural Wildlife Conservation Center; $849,015 for genomics for southern crop stress and disease research; $511,395 for biotechnology research; and $229,383 for rural systems research.

**$14,038,041** for 12 projects by Senate Agriculture Appropriations Subcommittee Ranking Member Robert Bennett (R-Utah), including: $5,560,800 for the Agricultural Research Center in Logan; $2,616,555 for a Utah conservation initiative; $1,191,600 for function genomics research; $559,059 for high performance computing; and $186,684 for pasture and forage research.

Example data that had to be manually copied from CAGW's *2008 Pig Book Summary*

In total, only 420 line items in the TCS dataset could be matched with all the projects described in the *2008 Congressional Pig Book Summary*. All other items in the TCS dataset were regarded as not being positively labeled pork. It bears noting that inconsistencies in the data could potentially weaken the model's accuracy and ability to generalize.

**Sampling**

A stratified random sampling approach was used to find training data for the model, with the strata being each type of appropriation bill. The number of projects CAGW claimed to find in each bill was used in a ratio with the percent of total earmarks per bill found in the TCS data. A sample size of n=30 was used, and bills that contained very small ratios of pork to earmarks were rounded up so that at least one data point could be included in the model.

**Table 1: Pork-Barrel Projects Distribution Across 2008 Appropriations Bills**

| Bill | # of Pork-Barrel Projects Found in CAGW Data | % of Total Earmarked Projects Found in TCS Data | Number of Samples for n=30 | Rounded Ip (to include all types of bills) |
|---|---|---|---|---|
| Ag-Rural Development-FDA | 123 | 10.01 | 3.00 | 4 |
| Commerce, Justice & Science | 171 | 13.91 | 4.17 | 5 |
| Defense | 234 | 19.04 | 5.71 | 6 |
| Energy & Water | 147 | 11.96 | 3.59 | 4 |
| Financial Services | 7 | 0.57 | 0.17 | 1 |
| Homeland Security | 98 | 7.97 | 2.39 | 3 |
| Interior | 70 | 5.70 | 1.71 | 2 |
| Labor-HHS-Education | 260 | 21.16 | 6.35 | 7 |
| Legislative Branch | 2 | 0.16 | 0.049 | 1 |
| Military Construction | 10 | 0.81 | 0.24 | 1 |
| State-Foreign Ops | 3 | 0.24 | 0.07 | 1 |
| Transportation and Housing & Urban Development | 104 | 8.46 | 2.54 | 3 |

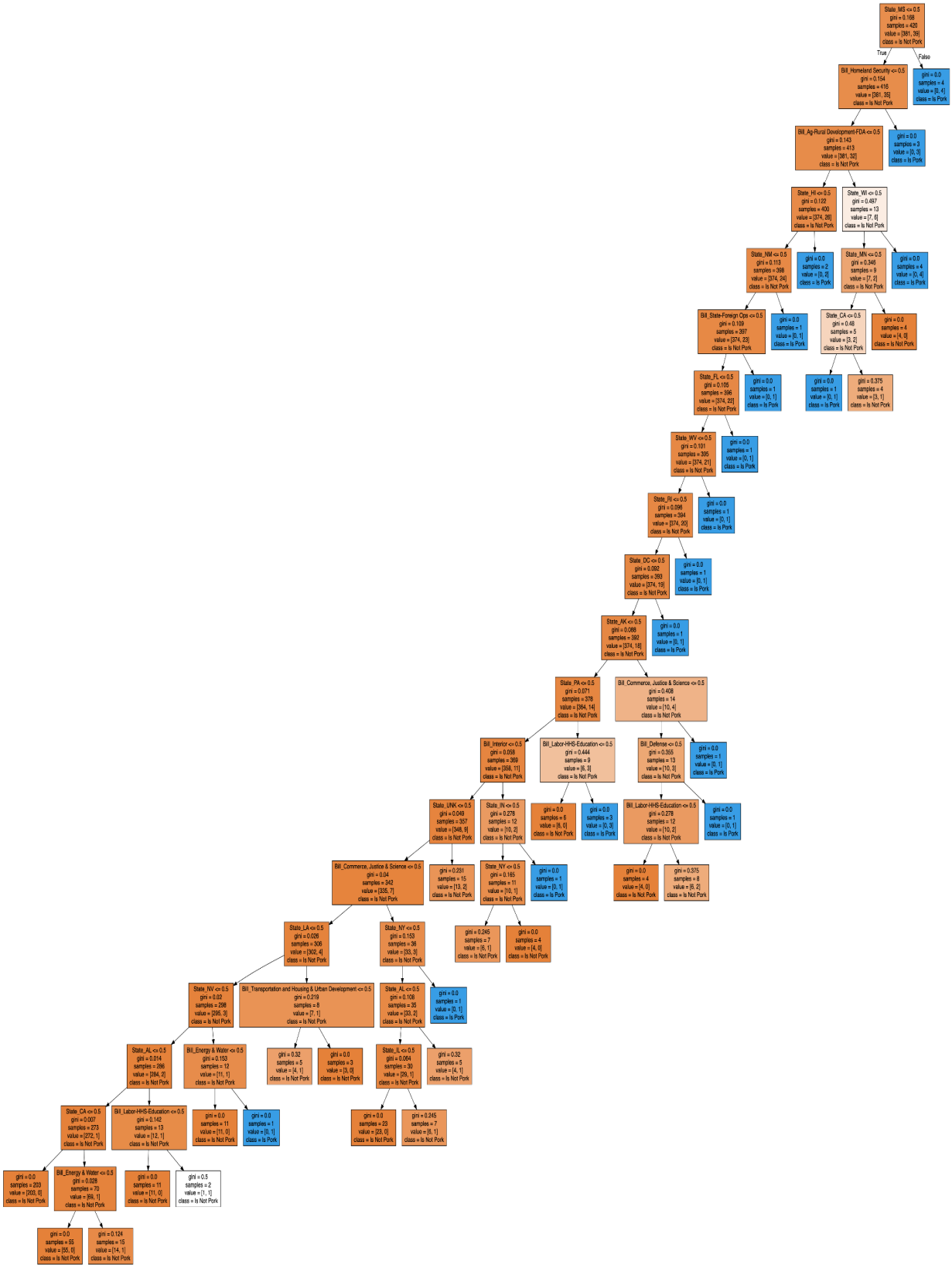
Note: Numbers rounded to 2 decimals places.

## Data Cleaning

Many challenges were also faced in cleaning the data to prepare it for the training of a machine learning model. Due to a breakdown in the pipeline that cleaned the data and then sampled it randomly for the machine learning model to be trained, the positive pork data and TCS earmark data had to be manually merged. After the complete dataset was merged, the categorical data had to be encoded such that CART could run on it. To do so, each possible value for each column was broken out into a separate column and coded as 0 or 1 depending on whether a row contained that value. This also had to be done somewhat manually, as the data pipeline for cleaning and sampling data was not fully functional.

## Operationalization of Variables and Creation of the First Model

Had there not been a breakdown of the original data pipeline, more than two independent variables (state and appropriations bill) would have been used in the creation of a preliminary model. They were the state that the earmark was from and the specific appropriations bill to which it belonged. The dependent variable used in the model was whether or not CAGW classified the earmark and its data as pork (1 = yes).



The first decision tree produced by the CART algorithm

While the selected independent variables were hypothesized to improve classification accuracy, as data was collected it was also hypothesized that the specific appropriation bill would be the largest indicator used by the model to classify an earmark as pork. This is because the ratios of pork to earmark projects for bills such as Financial Services, Legislative Branch, Military Construction and State-Foreign-ops were much smaller compared to the other types of bills.

**Results and Discussion**

The model itself surprisingly used Mississippi as its primary classifier of pork. This is likely due to the large amount of pork from Senator Thad Cochran identified by CAGW.[30] Testing of the model on unclassified earmark data proved to be a difficult task. Like the creation of training data, testing data had to also be manually sampled and created. Using a stratified random sample of 30 rows that were not used for training, the model was able to predict 26 of the 30 samples given. If the model were to predict every sample given as not being pork, it would have predicted 24 out of 30 samples correctly. Because the sample was stratified, only three of the data points were positive pork, and the model correctly identified two of them. A much larger testing size is needed in order to get an accurate estimate of the model's accuracy. A baseline can also show that high accuracy may not always result in a model that is effective. The results of the model's ability to classify pork suggest that states and appropriations bills may be useful in classifying earmarks, but the model is otherwise inconclusive. Further work on the model is needed before any conclusions can be drawn. It is of the utmost importance to obtain far larger amounts of training data before the hypothesis can be definitively accepted or rejected.

**Conclusions**

With an inconclusive model being generated, the original hypothesis was not proven correct. That being said, the process of undertaking this research produced other notable findings. Notably, data regarding pork-barrel legislation remains scant and hard to analyze. More efforts in analysis of the original legislation would aid greatly, as the data from CAGW was not easy to analyze or use. Also of note was a lack of data available from multiple previous administrations' Office of Management and Budget. The archived sites of the former Presidents did not have downloadable data, which could've aided greatly in this research. A cause for concern would be the lack of preservation of that important public policy data in a centralized source. The limited research done in the past regarding automating the process of earmark analysis uses said data. Future work on the model will focus on restoring a reliable and automated data pipeline to clean and create large amounts of training and testing data. Without the full automation of this component, the model is limited to analysis on only a few variables, and testing on only a small dataset. This provided a model that was not a truly effective identifier of pork-barrel appropriations.

Future research on the issue should also think carefully about the difference between the terms pork-barrel and earmark and the different uses of those terms. Standardization of the terminology in the field of political science would greatly benefit future scholars attempting to study the budgeting processes, especially with the return of earmarks in the most recent Congress. What makes pork-barrel legislation objectively bad is not clearly defined. Critiques levied on the entire earmarking process itself by groups like CAGW are broad, and specific research on the varying levels of how bad pork-barrel legislation is and its relation to the degree of effectiveness for politicians should be considered. CAGW may also not be the best source to identify pork-barrel

projects. They are one of only a few groups who study the issue in-depth, and they take a stance that labels all government projects created through the earmarking process as pork. As such they are inherently biased against all earmarks, which may not necessarily be "bad" pork. It should also be noted that they are incredibly biased against government spending in general. Much of the work in their "Pig Book" focuses on name-calling politicians and critiquing all spending.

Future planned technical work on the project includes (upon restoration of the data pipeline) the inclusion of more data sets. Should the OMB earmarking data be found, its inclusion into either testing or training data will be considered. The OMB, as a component of the executive branch, took a cynical stance on earmarks and their data may be valuable in identifying other pork-barrel projects (especially ones which they especially took issue with). Other years of TCS data will likely be taken into consideration as well, especially if a classifying model for earmarks in the current Congress is to be built. There are many ways in which data from other years may be used, including but not limited to the use of a random forest algorithm, which would effectively be a combination of decision trees.

For any researcher interested in this project, for which work is ongoing, the code base for the project can be found is online.[31] It is in the interest of the field of political science for technical and non-technical scholars alike to collaborate on open source projects. Assistance could be used in the identification of pork-barrel projects from FY 2009 and 2010, as well as in programming the actual code itself. Collaboration is encouraged and the code will remain open source.

## Endnotes

[1] Gordon, S. C., & Simpson, H. K. (2018). The birth of pork: Local appropriations in America's first century. *American Political Science Review*, *112*(3), 564–579. https://doi.org/10.1017/s000305541800014x
[2] Hauk, W. (2007). Small states, big pork. *Quarterly Journal of Political Science*, *2*(1), 95–106. https://doi.org/10.1561/100.00005048
[3] Maxey, C. C. (1919). A little history of pork. *National Municipal Review*, *8*(10), 691–705. https://doi.org/10.1002/ncr.4110081006
[4]Stowe, L. (2021, November 11). *Earmark and pork barrel spending & why you should care*. FiscalNote. Retrieved February 16, 2022, from https://fiscalnote.com/blog/earmark-vs-pork-barrel-spending
[5] Kennedy, S. (2021). (rep.). (T. A. Schatz, Ed.) *2021 Congressional Pig Book Summary*. Citizens Against Government Waste. Retrieved August 23, 2021, from https://www.cagw.org/sites/default/files/pdf/2021PigBook.pdf
[6] See Note 4
[7] See Note 4
[8] Taxpayers for Common Sense. (2010, March 10). *Earmarks and earmarking: Frequently asked questions*. Taxpayers for Common Sense. Retrieved January 22, 2022, from https://www.taxpayer.net/budget-appropriations-tax/earmarks-and-earmarking-frequently-asked-questions/
[9] Kennedy, S. (2015). (rep.). *All About Pork: The History, Abuse, and Future of Earmarks*. Citizens Against Government Waste. Retrieved February 19, 2019, from https://www.cagw.org/sites/default/files/pdf/2021PigBook.pdf
[10] See note 2
[11] See note 9
[12] See note 4
[13] See note 1
[14] See note 9
[15] Stratmann, T. (2013). The effects of earmarks on the likelihood of reelection. *European Journal of Political Economy*, *32*, 341–355. https://doi.org/10.1016/j.ejpoleco.2013.08.001
[16] Heston, M., Khabsa, M., Vora, V., Wulczyn, E., & Walsh, J. (2014). *Using Data For A More Transparent Government*. Data Science For Social Good. Retrieved February 14, 2022, from https://www.dssgfellowship.org/2014/12/04/using-data-for-a-more-transparent-government/
[17] Weingast, B. R. (1979). A Rational Choice Perspective on Congressional Norms. *American Journal of Political Science*, *23*(2), 245–262. https://doi.org/10.2307/2111001
[18] See note 1
[19] See note 2
[20] See note 1
[21] See note 15
[22] See note 15
[23] Lin, F, Hao, D., & Liao, D. (2016). Automatic content analysis of media framing by text mining techniques. *2016 49th Hawaii International Conference on System Sciences (HICSS)*. https://doi.org/10.1109/hicss.2016.348
[24] Lin, F.., Chou, S.-Y., Liao, D., & Hao, D. (2015). Automatic content analysis of legislative documents by text mining techniques. *48th Hawaii International Conference on System* https://doi.org/10.1109/hicss.2015.263
[25] See note 16
[26] Singh, S., & Gupta, P. (2014). COMPARATIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM: A SURVEY. *International Journal of Advanced Information Science and Technology*, *27*(27). Retrieved February 22, 2022, from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.685.4929&rep=rep1&type=pdf.
[27] See note 9
[28] Williams, D. E., & Kennedy, S. (2008). (rep.). (T. A. Schatz, Ed.) *2008 Congressional Pig Book Summary*. Citizens Against Government Waste. Retrieved August 23, 2021, from https://www.cagw.org/sites/default/files/pdf/2008_Pig_Book.pdf
[29] Taxpayers for Common Sense. (2012). *Earmark database*. Taxpayers for Common Sense. Retrieved January 22, 2022, from https://www.taxpayer.net/budget-appropriations-tax/earmark-data/
[30] See note 28
[31] https://github.com/sunilgreen/pork