Domain Background

The Parliament of Canada provides a website with well-structured data regarding matters of legislation. This data can be used to answer interesting questions, such as which political party has a historically higher percentage of sitting days in parliament, or which party is, on average, quicker at passing bills into law. The data could also be used to build suitable models to generate predictions, such as the likelihood a bill will pass into law, and how long it might take. There is certainly often much economic and political interest in the answer to these questions, such as with the legislation regarding the legalization of marijuana.

Problem Statement

The deliberation process before the passage of a bill can take anywhere from a single day to several hundred days. Here we develop a model to predict the number of calendar days, from the date of the first reading until the date of royal assent, of a bill presented in the Parliament of Canada, assuming that the bill does indeed receive royal assent. There are x number of bills that are proposed before parliament in any given session of parliament.

Datasets and Inputs

The raw data can be obtained in XML format from the website of the Parliament of Canada. There are numerous relevant features that can be extracted from here, such as the sponsoring party of the bill, the party of the current prime minister, the entire text of the bill, the origin of the bill (house or senate), or whether the bill is private or public. It is expected that some of these features should have some relevance to the length of time it takes for a bill to receive royal assent.

Many of the features are categorical variables, such as the political party of the bill sponsor or the session of parliament. One numerical feature that may be relevant is the number of lines of text in the content of the bill. The target variable will be extracted by writing a script to compute the difference in days between the date of first reading and the date of royal assent as a timedelta object. Intuitively, I would expect that the larger bills with lots of text would take longer to pass. I would also expect that bills that are sponsored by a member of the opposition party would take longer to pass. Public bills

would take longer to pass than private bills as more care and thought needs to be put into issues of public concern.

An example of a new input of would be a bill that has recently undergone first reading. Thus, some of the features (such as the statute reference information, and the complete history of events) will not be available to the model to make predictions.

Solution Statement

A linear regression model using appropriate features from the data would be a suitable first step. The target variable, which is a number of days, can be treated as a continuous numerical variable. We shall create a pandas dataframe that extracts most of the metadata from the raw XML data, and after some exploratory data analysis, the subset of features for our model can be further refined and saved in a dataframe of around 6 or 7 features. We wish to keep the number of features relatively low as the size of the available data is small.

Benchmark Model

A simple mean will be used as the benchmark model as we are not aware of any existing solutions to this problem. The evaluation metric can be used to compare the performance of our regression model with the benchmark model. We hope to demonstrate a significant improvement over our rather simple benchmark model.

Evaluation Metrics

The evaluation metric will be R squared, which can be computed as follows:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \bar{y})^2}$$

Project Design

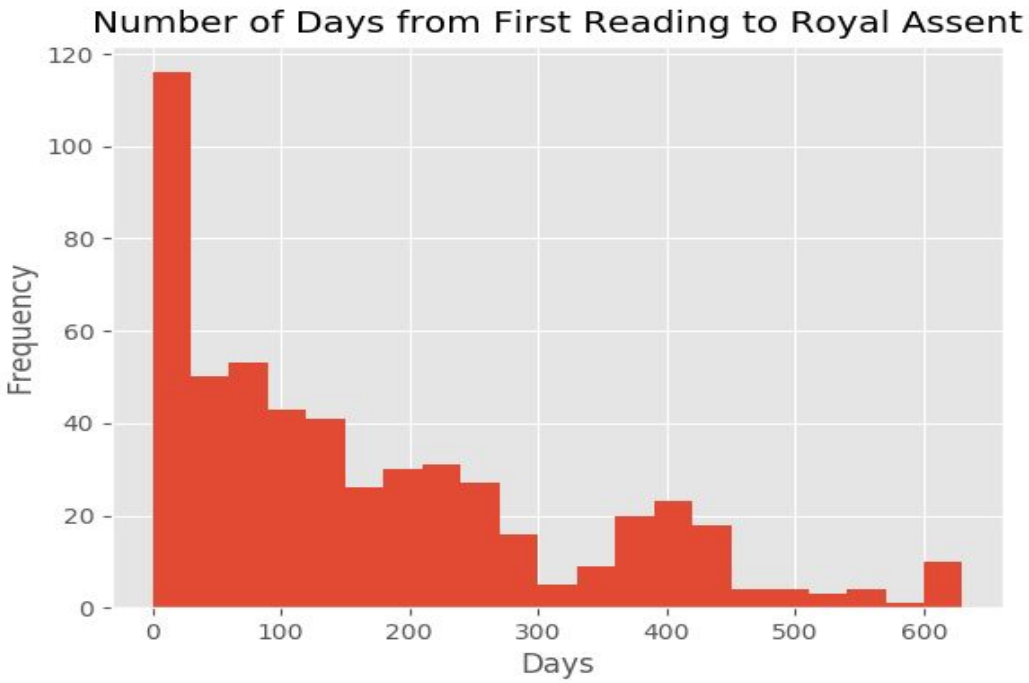
We shall make use of the cookiecutter datascience tool to create a template to organize the project. The raw data files will be downloaded as XML into the appropriate folder and left untouched. Any further processing will be performed on a copy of the data in a separate folder.

The raw data will be comprised of all 537 bills which have received royal assent to date as well as the XML of the individual bills. Initially we shall create a feature set based on all of the metadata available in the XML files. Then we will create a subset of features that should be relevant to solving the problem. To start off, some candidate features are: the data the bill was introduced, the parliament and session number, the bill type, the political party of the prime minister, the political party of the bill sponsor, the size of the bill in pages, and the percentage of sitting days (computed as sitting days divided by calendar days). The target variable is the number of days from first reading to royal assent.

The following script can be used to extract the target variable:

```
bills_xml = '/path/to/xml_file'

tree = ET.parse(bills_xml)

root = tree.getroot()

FIRST_EVENT = 0

LAST_EVENT = -1

start_dates =
pd.to_datetime([parse(bill.find('Events').find('LegislativeEvents')[FIRST_EVENT].attrib['dat
e']) for bill in root])

end_dates =
pd.to_datetime([parse(bill.find('Events').find('LegislativeEvents')[LAST_EVENT].attrib['date
']) for bill in root])

days_until_royal_assent = (end_dates - start_dates).days
```

A future addition that is beyond the scope of this project could include some natural language processing features as there is plenty of textual data available.



The refined feature set will be saved as a CSV file. For input to the model, the data will be loaded into a pandas dataframe and one-hot encoded. At this point we will have sufficiently prepared our data and can continue with exploratory data analysis. We may consider using PCA, t-SNE and k-means clustering algorithms to identify structure in the data. One question worth exploring might be to see if there is a relationship between the time of the year and the target variable.

We now split the data with a 90% training and 10% testing split. We can also create validation set of around 20% for 5-fold cross-validation. Next we shall create a benchmark model by computing the mean of the target variable for all bills. This model can be evaluated by computing the score according to the evaluation metric, which in our case is R-squared.

We shall use a linear regression model using sklearn. The score will be computed using the same evaluation metric as above. At this point we can determine if an improvement can be made upon the benchmark model. We may wish to consider changing some features or experimenting with additional techniques such as regularization, or perhaps a new model entirely. For instance, we may discover that a logistic regression model works better if, instead of using a continuous numerical value, we identify three classes of bills as ‘quick’, ‘medium’ or ‘long’ to describe how long we expect them to pass into law.