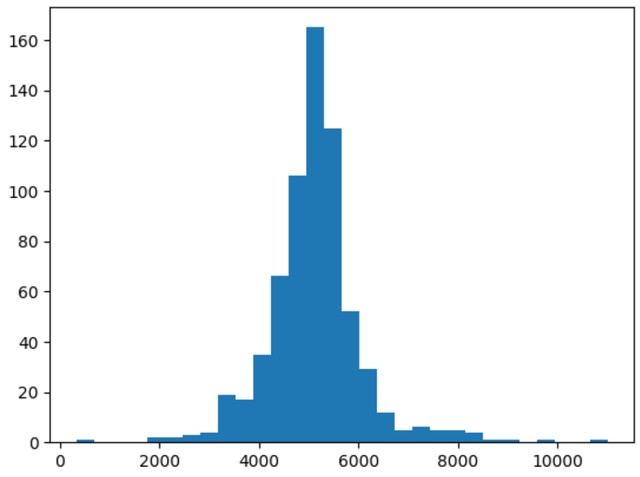
Supreme Court Verdict Prediction via Oral Arguments

Maggie Chen, Carolyn Liu, Eujene Yum

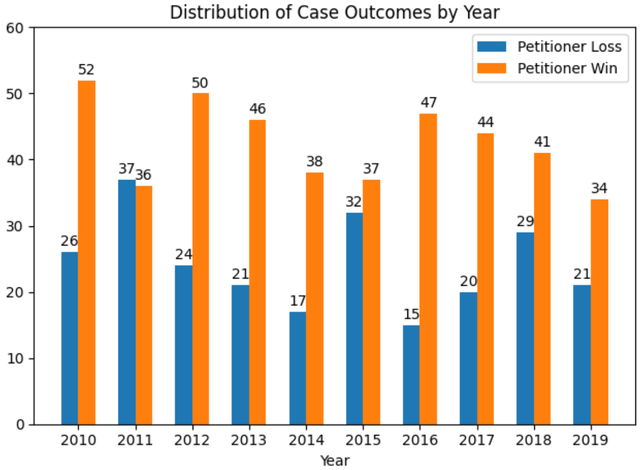
1. Data Description

We use data from 667 cases from 2010 – 2019 from Convokit, including the transcripts of oral argument and the verdict. Our goal is to make verdict predictions based on the utterances of justices and advocates.



The length of oral arguments for each case follows a normal distribution, with a mean of 5000 characters.

This works well for logistic regression, neural network, and tree-based models, but will need to be adjusted for BERT.



From this graph, we can see that our data is **imbalanced**. Therefore we used Random Oversampling to upscale the minority class (petitioner loss) for improved model performance.

2. Model Outcomes

2.1 Baseline model performances

1. We use CountVectorizer to obtain unigrams as input to our logistic regression model and use grid search for parameter tuning.
2. Our Neural Network has 2 hidden layers.
   - Activation Functions: ReLU(), Tanh()
   - Loss Function: binary cross entropy
3. We run two types of tree-based models – Random Forest and Gradient Boosting with parameter tuning. LightGBM gives better performance.

	Precision	Recall	F1
Logistic Regression			
Respondent (0)	0.67	0.76	0.71
Petitioner (1)	0.5	0.39	0.44
Accuracy			***0.62***
Neural Network			
Respondent (0)	0.61	0.35	0.45
Petitioner (1)	0.59	0.81	0.68
Accuracy			***0.60***
LightGBM			
Respondent (0)	0.63	0.75	0.68
Petitioner (1)	0.40	0.27	0.33
Accuracy			***0.57***

2.2 BERT model performances

Model Input

BERT can take a maximum of 512 input tokens, but oral arguments are much longer. To combat this issue, we divide the text into <500-word chunks. Consequently, our data contains multiple rows for each case.

Pretrained model:
bert base model (uncased)

Embeddings

Encoder

Classification Layer

Testing accuracy: ***69%***
Validation loss (for best model): 0.552

3. Conclusions & Next Steps

1. Out of the four models, BERT gives the best testing accuracy. This is most likely because BERT takes care of bidirectional contextual understanding across the text while others do not.
2. Oversampling helps with baseline models but we see a drop in testing accuracy when oversampled data is fed into BERT. This may signal that pretrained BERT models might already be good enough to handle imbalanced datasets.

1. Our current model only uses oral arguments as our input. A promising next step would include bringing in additional features such as divisiveness of topic, measurement of public attention, and political leanings of the justices.
2. We currently use the embedding from BERT's pretrained model. Another possible extension of the project could include building our own embeddings based on the the Supreme Court corpus, which might better represent the text data we have and achieve a higher model accuracy.