10-718 Project Report - Team bills1
Predicting Bill Passage in Illinois

Amrit Singhal	**Bo Lei**	**David Robusto**	**Yue Wu**
amritsin@	blei1@	drobusto@	yuewu4@
MSML CMU	MSML CMU	MSPPM-DA CMU	MSPPM-DA CMU

1 Executive Summary

The American Civil Liberties Union (ACLU) continues to provide a unique and essential function with their attorneys and other staff working tirelessly to ensure the protections of civil rights and liberties for all who inhabit these United States. However, as a non-profit organization their resources are spread thin and constantly in demand. Since a primary threat to civil liberties comes from the passage of new legislation, bill monitoring is one of the most important aspects of their work and is where a large part of these resources are spent. As it stands, in order to know if there is a particular bill with civil liberty implications that might galvanize the ACLU to intervene, staff members must manually review all bills introduced in every state and national legislature.

The purpose of this project is to greatly narrow the number of bills introduced in the Illinois state legislature that the ACLU is required to monitor by predicting which bills are most likely to pass without intervention. By only reviewing the bills with a high likelihood of passage the ACLU is not investing time into bills that would never become law. In order to accomplish this prediction task, we employ several machine learning models. These models include RandomForests, logistic regressions, SVMs, and decision tree models which were all trained on 11 years of Illinois bill and legislature information, supplemented by American Community Survey (ACS) demographic data. Our main metric to determine predictive accuracy was precision at the top 30% of bills, for which our best model, a RandomForest, was able to obtain an average of 37.2% across all validation splits.

In addition to absolute measures of performance, we are also interested in the equity implications of our work. To that end, we performed a bias and fairness analysis testing our predictive accuracy on bills introduced by representatives from high proportion African- and Asian-American districts, aiming to ensure our results on these districts were not significantly different from high proportion white districts. We used TPR disparity as our main fairness metric. Due to our high general predictive accuracy, achieving a perfect recall at 30%, we were also able to achieve a TPR disparity of 1, indicating our model performs equally well across all groups. The robustness of our results should give the ACLU confidence that integrating our model into their bill monitoring process would provide them with an accurate picture of the upcoming landscape of the Illinois state legislature. We conclude the report by designing a field trial to validate our results and a discussion of caveats and opportunities for future work.

In accordance with our results and findings presented in this report, we recommend that the ACLU's state affiliate in Illinois does the following:

1. Using a field trial, validate our results on newly introduced bills in the state and measure the resulting process improvements. (see details at 1)

2. Subsequently, integrate our model's predictions into the bill monitoring process for all staff attorneys. (see details at 2)

3. Establish a human-in-the-loop validation process for our model in order to ensure you are incurring an acceptable level of risk. (see details at 3)

Preprint. Under review.

2 Background and Goals

The American Civil Liberties Union (ACLU)[1] is a nonprofit and nonpartisan organization working in courts, legislatures, and communities to defend and protect the civil rights and liberties of all citizens. They work towards this goal by litigating across the nation, lobbying Congress, and educating the public. Before taking actions, they need to examine and identify the bills that are likely to get passed into law so as to target their efforts and allocate their resources in an efficient manner. Unfortunately, this process currently takes significant manual work considering the number of bills introduced at the nation and state levels. In practice, the ACLU tries to select bills that will have the greatest civil liberties impact or that are sponsored by certain legislators/committees based on their domain knowledge. This approach results in the ACLU investing significant time into bills that would not have passed even had they not intervened and still, using expert heuristics means inevitably missing important bills they would want to intervene on. To improve this process, we aim to present the ACLU's attorneys with the bills that are most likely to pass within a given year, saving them time sunk into bills that would end up failing regardless of their intervention.

Our goal is to implement this solution in a way that is efficient, effective, and equitable across different groups. By efficient, we aim to supplement the ACLU's process in a way that demonstrably reduces the amount of time their staff attorneys spend each week on bill monitoring. By effective, we aim to make our solution as generally accurate as possible. By equitable, we aim to ensure our solution is does not perform disproportionately towards any particular subgroup of the population. In this project, we are most aware of the potential trade-offs between effectiveness and equity, as maximizing absolute accuracy might result in a subgroup being asymmetrically excluded (although, see section 7.6 for the main discussion of equity and a note as to why this trade off fortunately does not seem to be a large issue in this case).

Considering time and resource limitations, we chose to narrow down our goal and initially focus on bills in a single state and make predictions of bill passage within a time frame. The reason that we chose state-level legislation instead of federal is that legislation that impacts civil liberties is largely introduced at the state level. In order to maximize the potential effectiveness of the ACLU's interventions, we will only regard bills that passed within a certain time frame as passed and everything else as not passed. The decision of appropriate states and time frames will be based on the number, progress and status of bills that are described in section 5. Overall, our goal is to predict the passage of the bills at a single state or few states level within a time frame.

In terms of potential influence, our project will have an immediate impact on the ACLU's bill monitoring process. The ACLU may need to take efforts to examine the methods and procedures we provide and integrate them into their current working approach. Depending on to what degree they implement our suggestions, our project may have a downstream influence on specific bills that would have an impact on civil liberties.

3 Related Work

A large number of bills are introduced each year, but only a few of them can be passed into law. As introduced bills have a large impact on all US inhabitants' well-being, many people are interested in what factors lead to some bills' success and others' failure. As such, many companies and researches have worked on bill passage prediction and important bill feature identification.

Congressional research has been applied to study legislative process. Some researchers explain why bills are more likely to be passed during the legislative process by studying the legislative effectiveness[1]. Participation in the committee decision making is also studied by some researches[5]. Except for the overall effectiveness, there are also studies on bill passage prediction from individual behavior prediction. A report by Kyle Gulshen, Noah Makow, Pablo Hernandez studies how each congressperson will react to a newly introduced bill. Another paper points out patterns of multiple sponsorship relate to a greater success rate for state legislation[2]. This finding, that increasing the numbers of sponsors can influence the success of legislation, will turn out to be particularly relevant to our results. These studies discuss bill progress from the perspective of legislation and congress members. As for exploration on influencing factors of bill passage from the bills themselves, some previous work focused on bill content[4, 6], building prediction model based on the content and topic of the bill.

One large differentiator of our work relative to previous literature is that we look at the way both bill text features and bill attribute features contribute to likelihood of passage. We compare the performance of text model, no text model and hybrid model. In addition, we include both state level and district level demographic data, and use the district level information to study bias and fairness of the model for different race group, taking into consideration of protecting

[1] `https://www.aclu.org/about-aclu`

racial minorities. For feature analysis, different from Browne's study focusing on the statistics and regression result[2], we compared multiple machine learning techniques to study feature importance.

4 Problem Overview

4.1 Problem Formulation

Our objective is to identify bills that are likely to be passed into law in Illinois. Bill monitoring takes significant efforts in terms of labor and capital and the ACLU has limited capacity across both of these resources. To help ACLU prioritize their advocacy and legislative efforts, the goal of this project is to identify bills that are likely to be passed within one year. Every day, for all bills that are introduced into the Illinois State Legislature in the past year, we provide a list of the top 30% of bills that are most likely to pass within a year from introduction.

4.2 Solution Overview

In our final data table, each row represents one bill with selected features and the label indicating whether the final status of the bill is 'pass' or 'fail'/'mark fail' within one year (details on how we obtain these labels can be found in the appendix in section 12.1). The bill features includes basic bill attributes such as introducing party and introducing body, session-related features such as time since session start and time till session end, sponsor-related features such as total number of sponsors and sponsor pass rate, state demographic features such as median age and gender distribution, and text-based features such as bill description and bill text.

The evaluation metric we applied is precision at top 30% of validation data. We compared the metric across four different machine learning models including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest and the baseline model. The baseline is calculated from top 30% ranking result of total number of sponsors. As our paper uses time series data, traditional k-fold cross validation would not be appropriate. Instead, we use temporal block validation which we attempted with both expanding windows and disjointed blocks. Our bias and fairness analysis was conducted looking at bills introduced in high-proportion African- and Asian-American districts relative to bills in high white districts and the primary fairness metric used was TPR disparity. Further details of our solution are discussed in section 6.

5 Data

5.1 Data Description

We used 2 primary data sources: LegiScan, American Community Survey (ACS), and supplemented with Ballotpedia.

LegiScan is an organization that collects data about legislation. It provides an API to acquire updated legislative sessions and bills data for 52 states and districts of the US. In our model, we focus on 10 years data in one state, Illinois. A cleaned database schema from the original JSON files was created. It includes information on State legislative session information, Bill information, Bill text information, People information, Voting information, Bill event information. We also used Ballotpedia for supplementary information such as the start and end dates of session in Illinois.

We gathered state and district level demographic features from ACS, including age, population, income, race, gender, employment, and education level.

5.2 Data Exploration

Exploratory data analysis is necessary before applying machine learning techniques to predict the final status of the bill. It will help us understand the data distribution, relationships, and to select appropriate features for further analysis.

We performed several types of analysis on the data. In the following section, we list each analysis performed, and how it helps us draw meaningful conclusions from the data.

1. The first important step was to ascertain how much data we have available for each state. This is important to decide what state we should eventually choose to tackle the problem in since we need to have enough relevant data to proceed.

   We found the list of the top 12 significant states which have the highest number of bills on record - NY, IL, TX, HI, MN, MA, OK, MS, VA, PA, CA. The data can be found in figure 19.

2. For each of the significant states, we found the distribution of the number of bills over time. While we expected this distribution to be quite uniform, we found that most states have a clear zig-zag pattern. An example of this is shown in figure 21 for the state of Illinois. This is interesting since we do not see an obvious reason for this although we suspect it might have something to do with different types of legislative sessions.
3. We also did a comparison of the number of entries we have for each `'bill_status'`, which describes the bill's progress through the legislature. The distribution of these labels across all data is shown in figure 20. The database contains a total of 12 statuses for the bills. For this problem, we are mostly interested in the `'Passed'` and `'Failed'` statuses. However, we see that the number of states sufficiently using the `'Failed'` label is quite low, which suggests me might have to do some data aggregation and relabelling for getting our target labels.
4. We did a similar analysis for the labels, as was done in point 3, at the state level as well since different states have different patterns. The distributions we found were similar to what we observed at the aggregate level.

   From the summary, we also found that the number of bills that get introduced is more than the combined number of bills that are either `'Passed'` or `'Failed'`. Additionally, in some states, we also found there are almost no bills that have a `'Failed'` status. Considering this, we are certain we will need to employ some data relabelling to obtain proper data points for each label for our problem.
5. Motivated by the findings in the point above, we also summarized the final status of bills for different states because it is our main target for prediction. This is shown in the figure 23. Ideally, there should only be two main final statuses, but we found that in some states, many bills had a final status as one of the other labels, giving us a total of 12 final statuses as well. To get one prediction label from this data, we need to combine and classify these 12 final statuses into the 2 prediction labels that we will classify on.
6. One method to approach this relabelling was to label all bills that do not pass within a specific time period after their introduction (say, same year, or same session etc) as the negative label. For now, we chose this time period to be one year. This is reasonable as the bills that haven't passed within a year may not be of immediate concern to the ACLU. Figure 24 shows the distribution when we have relabeled all the bills with progress time longer than one year to `'Inconclusive'` in the state of IL. Like in IL, there are many bills that have progress time longer than a year, and those may not considered the most urgent bills to the ACLU.

   An interesting factor to note between the analyses from parts (5) and (6) is that we can compare how many of the total passed bills were actually passed within the specified time period (one year in this case). For example, in the case of IL, we see that the numbers are almost the same, which gives us confidence that most of the bills that pass do so within a year. This would help justify our choice of a time period after which we can consider a bill as failing.

All of the above analysis was done for all 12 states. The comparison between all these factors would help us decide on the final state we choose as well for this project.

6 Details of the solution

6.1 Pipeline and Implementation Details

A complete data pipeline was developed to obtain the solution to the problem presented above. The pipeline started with LegiScan data present in a database, along with access to ACS[3] data through a python API. With these, we solve the problem by adequately obtaining and processing the relevant data, making temporal splits for the data, deriving useful features from this raw data, and training various classification models on these features.

The entire pipeline was developed in python. We used the `sklearn` implementation of the classification models. We implemented running the model grid in parallel across multiple cores using python's `joblib` module. The data was stored in `postgreSQL` database, which was loaded into python through the `psycopg2` module. We created a JSON config file, that allows our code to be controlled in almost all decision choices.

The complete codebase for the solution can be found at `https://github.com/dssg/mlpolicylab_fall20_bills1`.

6.2 Temporal splitting

We experimented and evaluated with two types of temporal splits.
For both, we used a validation data obtained from bills obtained over a duration of one year. In both the settings, we had to leave a buffer for one year after the timespan of the training data, and after the timespan of the validation data,

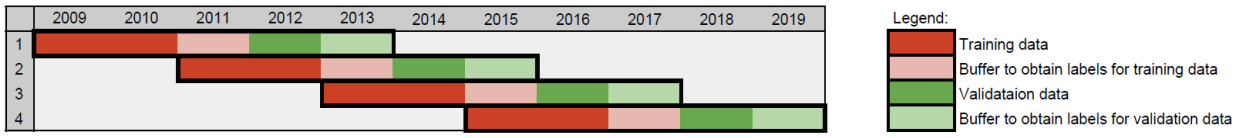


Figure 1: Temporal splits of Type I

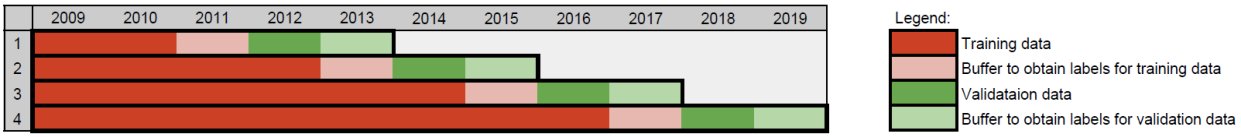


Figure 2: Temporal splits of Type II

in order to allow the labels to be obtained for all the bills in those sets. This duration was obtained from our problem formulation where we assigned each bill, that was inconclusive at the end of one year from its introduction, with the label failed. Thereby, we are guaranteed to obtain labels for every bill with a one year buffer after the ending the respective timespans.

The update frequency for the splits was chosen as 2 years. This was motivated by 2-year periodic pattern we saw in the number of bills for Illinois, as can be seen in figure 21.

The actual temporal splits that we used were as follows.

- **Type I:** Using a fixed window size. These are shown in figure 1.
- **Type II:** Using an expanding window size. These are shown in figure 2.

For both the cases, we end up with 4 temporal splits.

We initially opted to use Type I splits as the fixed window size and complete disconnect between successive training and validation sets minimizes the possibility of leakage. However, the trade-off is that there is less data available in each training set, potentially stifling our models' optimal performance. Eventually though, we decided that the increased potential performance from using Type II splits outweighs what we realized is a very small possibility of leakage. So, our optimal models use Type II splits.

6.3 Features

There is a lot of information present for a bill in the LegiScan dataset available to us. After having extracted only the relevant information through the database operations, we now need to make useful features from that information, such that we can use them to perform the classification task.

Broadly, we have the following sets of features.

1. Bill attribute features - These include the features obtained from the various attributes of a bill, such as the house in which the bill got introduced, the number of sponsors of the bill, the sponsoring party of the bill etc.
2. Bill text features - The included the text information available for the bill. We have a shorter `bill description` or the complete `bill text`. We implemented multiple types of methods to obtained features from these texts, including Bag-of-words, TF-IDF and LDA features.
3. Bill state demographic features - These features were obtained from data extracted from the ACS[3] database. The relevant data columns were downloaded in stored along with the other LegiScan data in the local database. Then, we added the state demographic features such as median age, mean income, gender ratio etc obtained from this data. The idea behind adding these features was to capture any change in the demographic landscape of the state over the years, which could potentially have an impact in controlling what kind of bills pass.

A complete list of all the features obtained is given in the appendix in section 12.2.

6.4 Models

It is a binary classification task. We aim to predict whether the bill will be passed or failed within one year based on features obtained for each bill (as specified in Section 6.3). Four classification models were implemented:

1. Logistic Regression Classifiers
2. Decision Tree Classifiers
3. Random Forest Classifiers
4. Support Vector Machines

These four models are commonly used in modern researches for classification, with different level of accuracy, interpretability and generalizability. We applied model grid for hyper-parameter tuning, and recommended the best model based on comparison result from our evaluation metric.

6.5 Metrics

The evaluation metric used for our models is precision at the top 30% bills.

Although our problem is a classification task, we do not aim to work with the standard scores of labelling all the bills and obtaining the f1-measure for classification performance. This is because that would not help capture the ACLU's need to to be able to get certain bills that it can act on further.

We were informed that ACLU generally has a capacity to manage around 30% of the total number of bills that are presented in a state legislative assembly. So, we aim to provide ACLU with a list of 30% of the bills that are most likely to pass. With this objective in mind, we need to evaluate the quality of the top 30% bills predicted from our model, as those would be the ones that would be acted upon by ACLU.

We need to ensure that we can get as many of those bills that we propose to be bills that do actually end up passing. This would ensure that the effort and resources from ACLU are not wasted on bills that they should not be used on. For this reason, precision at top 30% would be the ideal metric to optimize for our setup.

An important point to note here is that though the theoretical upper limit on a precision@30% value is 1, in our case the actual optimal maximum value is quite a bit lower. This is because the base rate of bills passing is only around ~ 0.14 (varying across temporal splits), and so no matter how good of a prediction model we have, a list of 30% of total bills will always have bills that will not pass. So, we need to keep in mind the actual maximum possible metric value while evaluating our models, which we find to be 38.8%.

7 Evaluation and Results

7.1 Base rate

A very small number of total bills pass in Illinois. Calculating the mean base rate across all validation splits, we find the value to be at 11.6%.

7.2 Smart Baseline

After having run through all of our models, and having looked at the features that our model found most useful, we were able to obtain the most important feature begin used by our model. As we will see in the later section 7.5, the most important feature that our models used was the number of sponsors of the bill. So, the smart baseline to use was just ranking the bills in decreasing order of number of sponsors, and picking the top 30% of the the bills.

Using this baseline, we can get pretty decent results. The mean precision@30% that we obtain from the baseline is 35.4%.

The PR-k curve obtained from our baseline for the most recent validation split is shown in figure 3.

7.3 Model Grid

A large number of model configurations were trained for each type of model type described in Section 6.4. The complete list of all the model configurations is given in the appendix in table 2 in section 12.3.

Furthermore, different types of feature configurations were tried for the bills with each model configuration as well. More specifically, we had three sets of feature types: (1) `no_text`, (2) `bill_text` and (3) `bill_description`. The details for these features are described in the appendix in section 12.2.

The all models were tried for both types of temporal splits. As explained earlier, we are more concerned about the performance obtained from Type II splits, so those are the results that we present and analyze here. The best performance

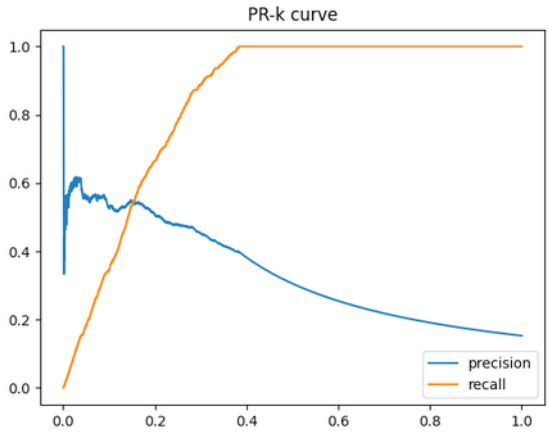


Figure 3: PR-k curve for the smart baseline

was obtained using the `no_text` feature types. The results for other feature types and for the type I temporal splits have been given in the appendix in section 12.5.

The results from the model grid described above is shown in Figure 4. Figure 5 shows the model grid with only the top 3 models configurations for each model type. From the plots, it is clear to see that the performance of our best models is the optimal performance at 30% for almost all of the temporal splits.

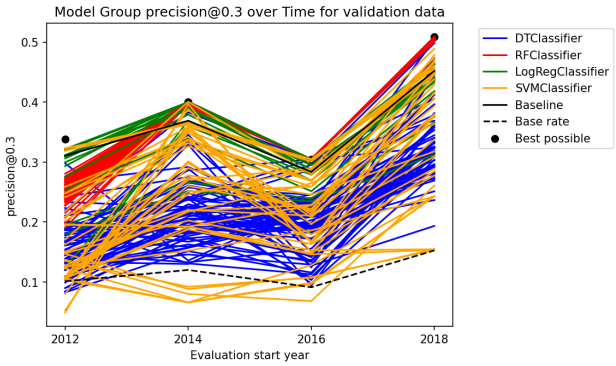


Figure 4: Model Grid

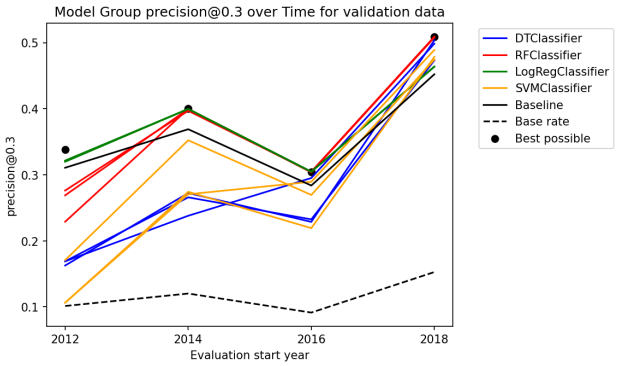


Figure 5: Model Grid top 3 models of each type

However, we see a general decrease in the performance of the models for the first temporal split. As we will see further, our best performing models are using the time based features about when during a sessions duration a bill is introduced, as one of the most important features. This information is missing for all of the data points from 2009, which do not a have a date associated with them. These data points constitute the majority of the data points in the first temporal split, which leads to the imputation performed on the data to be quite ineffective as well. Subsequently, our models are unable to perform well on the first split.

As we go towards the later temporal split, we get more and more data, with full information. This allows us to have more meaningful imputation values, as well as lowering the impact the missing data can have on the overall statistics. This two fold reason causes the impact of the missing data to subside in the later splits. The Random Forest based classifiers clearly outperform all else in all these splits. In fact, if we look at figure 16 in the appendix, we can find that all the top 10 models in our model grid are Random Forest based classifiers.

7.4 Best Performing Model

Having all these model configurations, we need to ascertain which of these is the best model among all that should be used for future predictions. To ascertain this, we quantity we look at is the mean precision@30% across all temporal splits. The model with the highest mean precision@30% would be the model that performs the best on average across the time splits. This is reasonable as we want a model that performs good across all times.

Maximizing this metric, we get that the best performing configuration is a Random Forest Classifier using the `no_text` features. The complete details about the best model setup is given in table 1. This model achieves an average precision@30% value of 37.2%.

This is extremely close to the best possible mean precision value of 38.8%. This best value is not the standard value of 1 as explained earlier in section 6.5.

Furthermore, we find that the model achieves a perfect recall@30% of 1. This means that our model is able to successfully retrieve all the passed bills within the top 30% bills predicted b it.

Model Type	Random Forest Classifier
Number of tree estimators	100
Max depth	10000
Splitting criterion	entropy
Max features[2]	log2
Class weighting used	None
Feature Type	`no_text` [3]

Table 1: Details of the best performing model configuration

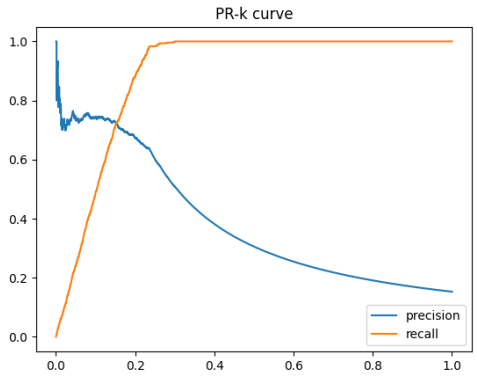


Figure 6: PR-k Curve for the best performing model

7.5 Model Interpretation

As stated earlier, our model is able to achieve the best possible performance possible at 30%. To understand this high effectiveness of our model, and to explain the predictions that our model will give, we perform futher analysis on the model to understand how our model is able to arrive at its predictions.

To this end, we firstly look the feature importances of all the features in our best model. Figure 7 shows the top 20 most important features for our best model.

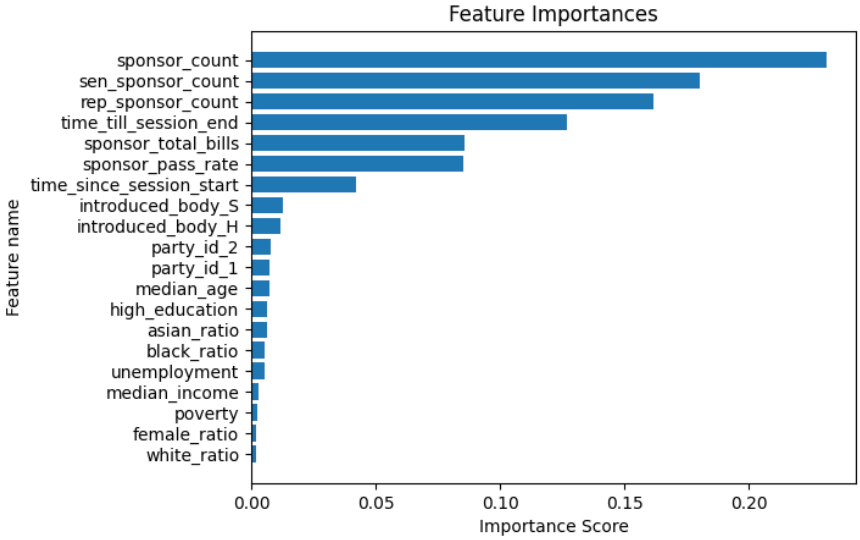


Figure 7: Feature Importances for the best model

[3] how many maximum nodes to consider while splitting at any node

[3] the exact features in this configuration can be found in the appendix in section 12.2

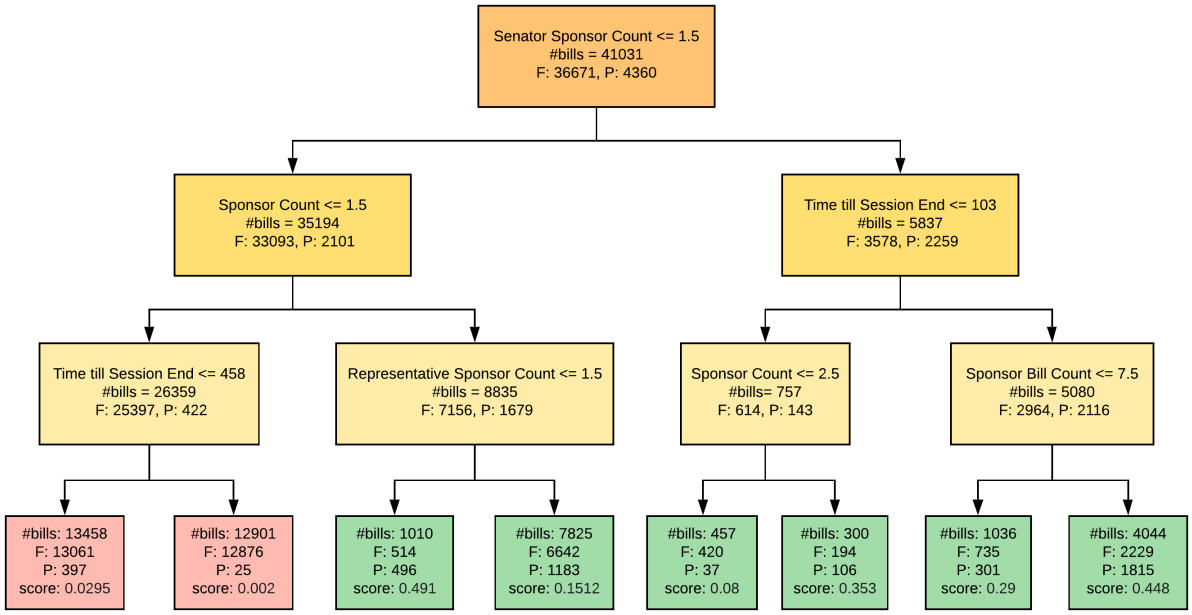


Figure 8: Decision Tree on the top features

Looking at the feature importance, it is obvious that the top few features account for almost all of the importance. This means that these top few features are almost exclusively enough to achieve near perfect predictions for a task load of 30%. To verify this result, we train another small Decision tree (having a max depth of 3) on only the top 5 features from the above ranking, i.e. total sponsor count, number of senator sponsors, number of representative sponsors, time till session end from the date of introduction of the bill and the total number of bills from the sponsor.

The Decision tree learnt from these features is shown in figure 8.

In this decision tree, we find that the green nodes account for the top 35% of the bills. Using these bills, if we calculate the precision@35% from this value, we find that we are able to achieve a value of 0.268, which is quite close to the optimal possible value of 0.299. Furthermore, if we calculate the recall at 35%, we can see that we have a recall of >0.9, which means even this simple model is able to include almost all of the passed bills within its top 35% predictions.

This model reaffirms our finding that these top few features are indeed super informative in calculating whether a bill will pass or not in the state of Illinois.

We can see that these top features primarily contain two broad categories of features - the sponsor count features and the session time features. We want to verify whether there exists patterns on just these two features in the data. For this reason, we obtain a heatmap with total sponsor count on the x-axis and the time between the start of the session and data of introduction of the bill on the y-axis. This is shown in figure 9.

In the heatmap, the first number in each cell corresponds to the number of passed bills in that setup, and the second number is the number of total bills. The color ranges from red indicating the least percentage of bills passing, to green indicating the highest percentage of bills passing. As we can see, there are quite clearly defined patches to red and green, which indicates that there are very clearly defined patterns in the data for which setups of these two features lead to bills passing vs failing.

Most interestingly, we see that a bill with a single sponsor has never passed in the Illinois state legislation, which is a very clear indicator to disregard so many of the bills from consideration for the ACLU.

7.6 Fairness analysis

Our models are going to directly impact the choices that ACLU makes regarding the bills that are going to pass or not. This in turn impacts which bills the ACLU chooses to intervene against. Because of this reason, we need to make sure that our prediction model is free from any bias against any group of people, and the predictions are fair for everyone.

Time Since Session Started (in Months) \ Total Number of Sponsors (Senate and House)	1	2-5	6-10	11-20	21-50	>50
1	0/14353	787/3020	373/935	274/637	180/411	50/99
2	0/6760	599/3152	286/701	237/488	152/287	43/63
3	0/47	0/31	1/4	0/3	1/8	0/2
4	0/50	0/28	0/6	0/3	0/7	0/4
5	0/61	1/26	0/7	0/9	0/7	0/2
6	0/30	0/25	0/4	0/3	0/5	0/1
7	0/25	0/16	0/5	0/7	1/9	0/2
8	0/33	1/26	2/12	1/11	2/4	0/0
9	0/50	3/28	1/11	0/6	2/10	4/6
10	0/150	10/94	4/33	4/23	4/23	2/4
11	0/102	7/70	6/15	3/15	5/13	1/7
12	0/482	32/166	28/65	13/34	14/28	1/3
13	0/5174	471/1414	183/382	147/263	87/154	9/24
14	0/2863	280/1338	150/304	105/189	69/119	16/24
15	0/59	3/407	3/6	1/8	2/9	0/3
>15	0/394	2/101	1/35	1/21	1/17	0/1

Figure 9: Heatmap across the top two feature categories

In our specific context, we ensure that our models are from bias against any racial group. We consider two racial groups, African-American and Asian, along with the the reference group of white population.

However, a bill by itself does not hold a racial identity. Thus we use a proxy for this. We identify each state legislative district with a racial group. For the African American population, we choose those districts that have the African American population in a majority. This gives us 8 out of 59 upper chamber and 16 out of 118 lower chamber districts associated with the African American racial group. For the Asian population, we find that there are no upper or lower chamber districts that have Asian population as the majority population. Therefore, for this racial group, we pick the top 5 out of 59 upper chamber and the top 5 out of 118 lower chamber districts based on the number of Asian people. All the remaining districts were attributed to our reference group of white people.

Each legislative district associated with one legislative member in the respective house of the state legislation. Having racial groups for the legislative districts, we associate those racial groups with that legislative representative from that district. Now, for any introduced bill, we say that is belong to a racial group if it has some sponsor associated with a non-reference racial group. In case a bill has multiple sponsors coming from different non-reference racial groups, the higher number is chosen as the bills identity. All the remaining bills are associated with the reference group of white population.

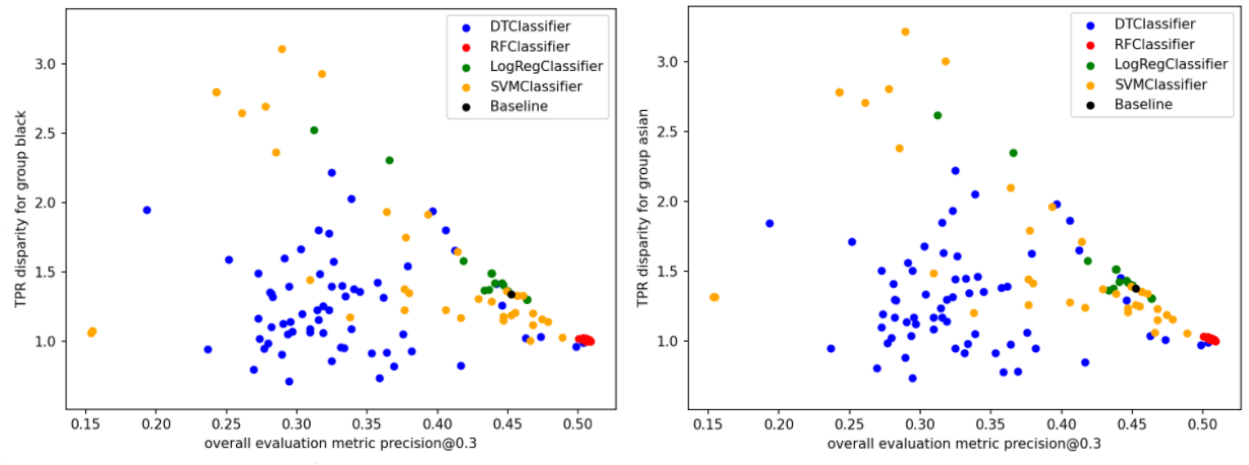
Since this is a proxy for the actual factor of race that we are grouping on, we needed to verify whether the bills across groups actually differ from the reference in any significant regard. To this end, we counted the (non-stop-)words in the high African- and Asian-American districts' bills' descriptions that were not present in the descriptions of bills from our reference group. We found there to be a high number of occurrences of telling words that exist in both our protected groups' bill descriptions such as 'eviction' (n=46) or 'asian' (n=32), but are missing from the reference group. More unique telling words can be found in section 12.9. This suggests that there is some underlying difference in the types of bills originating from these groups.

With this setup, we calculate the TPR disparity for all our models. This metric is more important for our case than the FDR disparity. This is because we think that a model that is systematically missing more bills in one racial group from being predicted as passing, that are actually gonna pass, is more dangerous to our model as compared to having the model obtain equal quantity of false positives across the different racial groups. The TPR disparity plots are shown in figures 10a and 10b respectively.

As we can see, for many of the models that we trained, the results are actually quite biased, ranging from a maximum disparity of greater than 3, to a minimum value of around 0.8. This further reaffirms our decision that there is indeed some inherent difference in the type of bills that belong to our different groups.

However, looking at the random forest classifiers that perform as our best models, we can see that those look very fair in the predictions that they make. They obtain balanced TPR and FDR disparities very close to 1. This is expected behaviour since we find that the best model we have performs optimally, being able to recognize all the passed bills. Thus, irrespective of the group, we will obtain a TPR disparity of 1.

Furthermore, we can see that the performance of our models in terms of TPR disparity is quite similar in the African American and Asian cases. This is also rooted in the data, as we can see, while there were words present in the



(a) African American vs White group

(b) Asian vs White group

Figure 10: TPR Disparity

African-American group's bills that existed in neither other group (such as paternity (n=25), incarceration (n=18), and torture (n=17)), the words in the Asian-American group's bills did not exist in the reference group, but were largely a subset of the African-American group's. This can be seen in the exact numbers presented in the appendix in section 12.9.

Seeing as our models perform quite fairly on the TPR metric, we also calculated the FDR disparity for our models, which is given in the appendix in section 12.8.

7.7 Discussion of the results

Overall, the main takeaway from our specific results is that this particular formulation of the problem is quite simple. As we saw in our shallow decision tree 7.5, the problem can be solved at a quite high degree of accuracy with just a handful of rules. With that in mind, if the ACLU decides at any point that the maintenance of our models is no longer feasible, they could still achieve quite good performance in Illinois by just looking at the number of sponsors a bill has and how far into a legislative session it is introduced.

8 Design of the field trial

In order to validate the results, we recommend the ACLU state affiliate in Illinois allows us to conduct a field trial, which would take the form of an A/B test. This trial would occur throughout 2021 in the state and is designed to validate process and outcome improvements made by our model. In order to conduct such a trial, we would need to be able to randomly divide the ACLU's staff attorneys into two subgroups. A treatment group (that is given access to our best model's predictions during the bill monitoring process) and a control group (that completes the process as they historically have). This randomization, assuming we could properly keep the groups in isolation, would allow us to effectively test the impact that our model has on the process.

If we assume that each staff attorney works in isolation, then our unit of analysis is the performance of each attorney. The general question this trial is interested in answering can be phrased as, "is there a difference in the performance of ACLU staff attorneys who have access to our model's predictions relative to those who do not?" As mentioned, our treatment and control are whether or not the attorney received our list of likely-to-pass bills. There are several possible outcomes of interest that could be tested in this trial. Since the original goal of the project was to help the ACLU save resources by expediting bill monitoring, it would make sense to test the difference in the self-reported number of hours each group spends on the bill monitoring process within a time period (say, one week) and compare these averages. An extension of this would be to test see if there is a difference in the average number of bills lawyers were able to actually intervene on throughout the year in the treatment vs control group. Another quantitative metric would be to capture what percentage of bills that each group intervenes on that actually pass. During this period, we would also want to validate that our precision @30% remains high when tested on new data. If our assumptions hold and randomization is successful, the change in these metrics across groups should be the causal impact of integrating our model into the

bill monitoring process. In order to determine the most feasible and effective specifics of the test and confirm our assumptions, we would need to engage in significant dialogue with the ACLU state affiliate in question.

9 Policy Recommendations

The main goal of our project is to help the ACLU better allocate their resources during the bill monitoring process by narrowing their focus to bills that are likely to pass, absent outside intervention. As our best models were found to be highly accurate, we have the potential to deliver significant efficiency gains via our solution. In pursuit of this goal, we recommend the following actions:

1. **The Illinois state affiliate of the ACLU should verify our results via a field trail.**
   While our models perform well on historical and held-out data, the only way to confirm that they are effective and practically useful to staff attorneys is to conduct a field trial. The details of a possible trial that could be conducted in 2021 in the state of Illinois are laid out in the above section. 8
2. **Provided the field trial results in an improvement of identified metrics, integrate our best model's predictions into the bill monitoring process for the remaining staff attorneys.**
   Our product for the ACLU is a list of the 30% of bills that, at the day the models are run, have the highest predicted chance of passage within a year from introduction. As of now, staff attorneys at the ACLU need to manually review all introduced bills and use their expert heuristics to gauge 1. whether the bill has relevance to the ACLU's mission, and 2. whether they expect a bill to pass or not. Assuming our model is able to retain similar levels of accuracy on new bills, it should ameliorate the majority of consideration 2. We therefore recommend that the ACLU presents their lawyers with our list of high likelihood to pass bills from our best model 7.4 and telling them to only scan for mission relevance, cutting out a significant component of the monitoring work. It is our hope that this saves the ACLU precious time and resources that can instead be spent on completing more interventions more quickly.
3. **Establish a validation or fail-safe process for our model.**
   While our model was able to achieve high precision on our training and validation data, no machine learning model will ever be perfect. There exists the possibility, although hopefully a small one, that a bill that would cause egregious civil rights violations might be predicted incorrectly with our model. With that in mind, part of the new process should acknowledge that our model is meant to be a supplement to human judgment, not a replacement. As such, there should be some level of human-in-the-loop decision regarding the bills that we predict do not have the highest likelihood of passage. What exactly this decision is depends on the level of the level of risk the ACLU is willing to take on and how many resources they have to spare. As discussed above, our model is particularly wary of bills with only one sponsor. So, one example of this process could be to have an expert at the ACLU use heuristics, bill subjects, and other factors to quickly scan the single sponsor bills our model does not put forth and flag ones that would have a civil rights impact that is too large to ignore. Again though, there are trade-offs here that only decision makers at the ACLU are able to answer about resilience vs resource usage.

10 Caveats

1. **Models built for the Illinois legislature in particular might not generalize well to other states.**
   As discussed in the subsection on model interpretation 7.5, our models rely heavily on the use of variables related to sponsor number seeing as nearly zero bill sponsored alone pass within a year. However, this might not be the case outside of Illinois. In other states that are less dominated by a single party, building coalitions and getting co-sponsors might be significantly more difficult, meaning this would be a more rare and less predictive feature. Similarly, for the entirety of our training data, democrats have held a sizable majority in both the Illinois House and Senate. This is certainly not the case in all states so when applied to a different state, the fact that a bill was introduced by a House democrat for example might have a significantly different impact on its likelihood of passage.
2. **Our model output does not offer a probability of any particular bill's passage.**
   It is important to note that while our model offers which bills it expects to have the highest likelihood of passage, this is not a probabilistic prediction. Meaning, you will never be able to interpret the predictions of our model as "bill x has a 63% chance of passing."
3. **Our current process is agnostic to the bill's relevance to the ACLU's mission.**

As of now, our models only take into account the likelihood that a bill will pass, it does not at all consider how relevant these high-passage bills are to the ACLU's mission of protecting civil liberties. As such, experts at the ACLU will still need to triage bills on our list with this metric in mind. Despite this caveat, the main purpose of our project is to provide time and resource savings for the ACLU. It is our hope that having 70% fewer bills to parse for mission relevance will still provide a significant improvement to the process. It is also possible that this caveat could be overcome entirely through integration with another piece of in-development work. See point 5 in section 11 for details.

4. **In Illinois, harmful bills might take more than one year to pass.**

   Our current formulation in section 4.1 has us predicting bills that are likely to pass within one year to reflect the official duration of legislative sessions in the state. All bills that are still pending after a year from introduction are considered to have failed. While Illinois' legislature does technically operate on one-year sessions, there are a fair number of examples of bills passing across sessions but within the same general assembly (in between elections). While this analysis is a potential subject of future work, it seems plausible that these bills that have met significant delay before passage could be controversial. As it also seems plausible that bills with a high civil liberty impact could be controversial, there exists a situation in which these slow-to-pass bills might be of high interest to the ACLU and also are missed by our model. To combat this, it might be worth keeping a record of bills introduced that have not officially failed after one year.

11 Future work

1. **Consider how a bill's likelihood to pass changes as it advances through the legislature.**

   Currently, our formulation (as described in section 4.1) only evaluates a bill's likelihood of passage at the time of its introduction. This is not reflective of reality as bills often undergo many stages beyond just introduction then passage/failure. For example, a bill might be introduced, referred to a committee, referred to a different committee, vetoed, then had that veto overwritten for a final result of a pass. It is likely that these subsequent stages of progress a bill might make through the legislature could provide information useful to predicting its eventual outcome. So, in the future we would hope to incorporate these updates to bill status, which would require prediction to occur every time this happens.

2. **Work to increase generalizability beyond Illinois.**

   For reasons described in the caveats section 1, our models are placing extremely high predictive value on several factors that we expect to be relatively unique to Illinois. This would make it difficult to achieve the same level of predictive accuracy if our models were tasked with explaining bills from another state. While it would be easy enough to re-train and tune the models on another state's data, depending on the preferences of the ACLU, this might be undesirable. They might instead want a single model that performs as well as possible on bills, agnostic of the state. While this would likely result in an accuracy trade-off and would require significantly more time to train much larger models, it is a potential avenue for future work.

3. **Implement more advanced model types.**

   While we are happy with the results of our current models, it can only help provide more clarity into our problem and open avenues for better results by introducing more model types. Examples include XGBoost, as we have seen great success on this problem using tree-based models, or neural networks.

4. **Create hybrid models that utilize both bill text-based features and bill attribute features.**

   As described in 7.3, we attempted models with features generated from the text of bills themselves, but ultimately went with the better performing models that just used features on bills' attributes. However, had we more time, we would like to integrate the two model types into a hybrid model. These hybrid models would theoretically provide the best of both worlds in terms of features available. The major barrier to implementing these models at this time is the dimensionality reduction that would be required in order to accurately compare bill text and bill attribute features appropriately.

5. **Integrate models with the other DSSG project in order to consider bill relevance.**

   Happening in parallel with our project is another Data Science for Social Good initiative. This project explicitly aims to automate the process of identifying a bill's relevance to the ACLU's mission. If such a system were to become functional, it would be greatly in our interest to integrate our models with theirs. That way, we could address the two main vectors of the bill monitoring process simultaneously and save the ACLU even more time and resources.

References

[1] William D. Anderson, Janet M. Box-Steffensmeier, and Valeria Sinclair-Chapman. "The Keys to Legislative Success in the U.S. House of Representatives". In: *Legislative Studies Quarterly* 28.3 (2003), pp. 357–386. DOI: 10.3162/036298003x200926.

[2] William P. Browne. "Multiple Sponsorship and Bill Success in U.S. State Legislatures". In: *Legislative Studies Quarterly* 10.4 (1985), p. 483. DOI: 10.2307/440070.

[3] U.S. Census Bureau. *2009-2019 American Community Survey 5-year Public Use*. 2020.

[4] Sean M. Gerrish and David M. Blei. "Predicting Legislative Roll Calls from Text". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Omnipress, 2011, pp. 489–496. ISBN: 9781450306195.

[5] Richard L. Hall. "Participation and Purpose in Committee Decision Making". In: *American Political Science Review* 81.1 (1987), pp. 105–127. DOI: 10.2307/1960781.

[6] Smith, Noah A. Yano, Tae and Wilkerson, John D. "Textual Predictors of Bill Survival in Congressional Committees". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 793–802.

12 Appendix

12.1 Definition of label

For each bill, the label is positive if the bill get passed within a year, otherwise the label is negative. In more detail, we first get the time interval between the introduction date and the final status date from database table `bill_progress`. We use both the time interval and the final status from `bill_progress` to decide the label. The deciding rule is
If time interval <= 365 days and final status = "Failed" then negative
else if time interval <= 365 and final status = "Passed" then positive
else negative

The distribution of bill labels in Illinois can be seen in figure 24.

12.2 List of all features generated

feature type	feature names	description
no_text	introducing_party	Contains the bill attribute features and the bill state demographic features (according to the categorisation in the report) The total length of the feature vector was 36
	introducing_party_ratio_in_power_in_senate	
	introducing_party_ratio_in_power_in_house	
	house_introduced	
	time_since_session_start	
	time_till_session_end	
	sponsor_pass_rate	
	sponsor_count	
	num_rep(resentative)_sponsors	
	num_sen_sponsors	
	state_demographics_median_age	
	state_demographics_total_population	
	state_demographics_median_income	
	state_demographics_gender_ratio	
	state_demographics_poverty_ratio	
	state_demographics_ethnicity_distribution	
	state_demographics_education_distribution	
	state_demographics_employed_ratio	
description_only	bill_description_tfidf	The total length of feature vector was 10654
description_text	bill_description_tfidf	The total length of feature vector was 49664
	bill_text_tfidf	

12.3 Model grid used

Model	Parameters
Decision Tree	"criterion": ["gini", "entropy"]
	"max_depth": [3,10,50,100,500,1000,10000,None]
	"max_features": ["sqrt", "log2"]
	"class_weight": ["balanced", None]
Random Forrest	"n_estimators": [100,500,1000],
	"criterion": ["gini", "entropy"]
	"max_depth": [100,1000,10000,None]
	"max_features": ["sqrt", "log2"]
	"class_weight": ["balanced", None]
SVM	"C": [0.01, 0.1, 1, 10, 100]
	"kernel": ["linear", "rbf", "poly", "sigmoid"]
	"class_weight": ["balanced", None]
Logistic Regression	"penalty": ['none', 'l1', 'l2']
	"C": [0.01, 0.1, 1, 10, 100]
	"class_weight": ["balanced", None]

Table 2: Model Grid Used

12.4 List of train/validation sets

	train	validation
1	2009-2010	2012
2	2009-2012	2014
3	2009-2014	2016
4	2009-2016	2018

12.5 Temporal graphs

This section provides the temporal plots for all the feature types and temporal splitting types that were tested.

1. Type II temporal splitting
   (a) `no_text` feature type: as shown in figure 11 (This is the same plot as was included in the report)
   (b) `description_only` feature type : as shown in figure 12
   (c) `description_text` feature type: as shown in figure 13
2. Type I temporal splitting
   (a) `no_text` feature type: the experiments were ran on subset of the provided no text features, which are not directly comparable to the other models. So, we are not presenting those results. (These were the first models that were run, results from which were provided in the weekly update). Since we eventually shifted to using Type 2 splits exclusively, these features were never run in there most recent form on type I splits.
   (b) `description_only` feature type : as shown in figure 14
   (c) `description_text` feature type: as shown in figure 15

We propose that the performance of the text based features was not at par with the no text features since the passing of a bill is very highly governed by the socio-political landscape of the bill, rather than being solely controlled by what the bill holds. We thinks intelligently combining the no text features with the text based features should prove beneficial to the model, but that is part of our proposed future work.

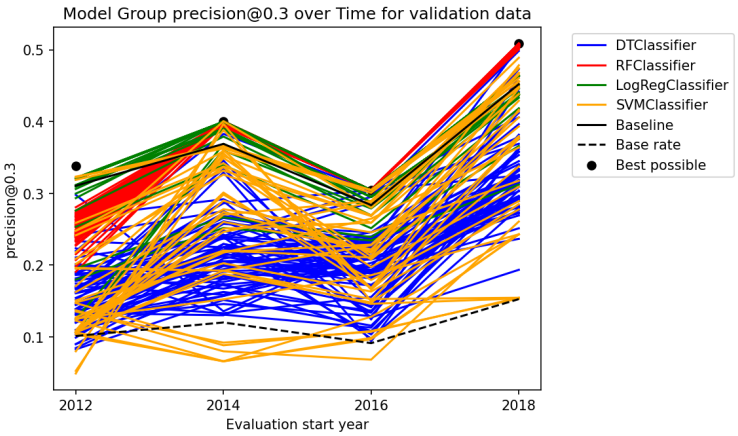


Figure 11: Type II Temporal graph for `no_text` features

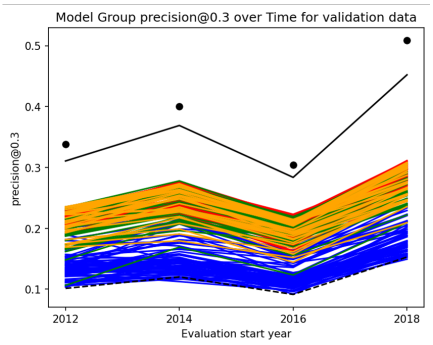


Figure 12: Type II Temporal graph for `description_only`

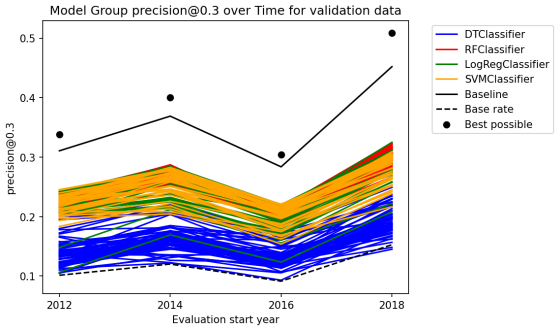


Figure 13: Type II Temporal graph for `description_text`

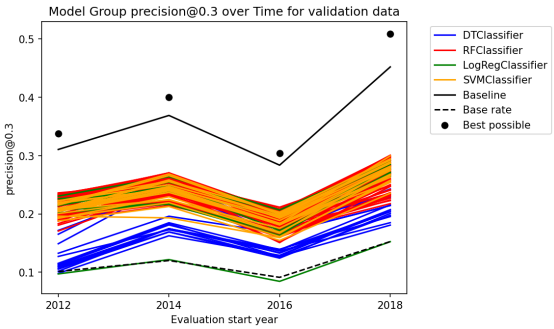


Figure 14: Type I Temporal graph for `description_only`

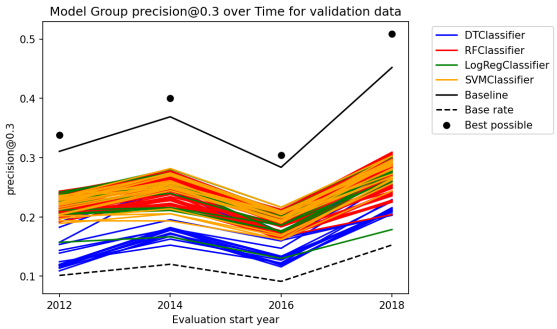


Figure 15: Type I Temporal graph for `description_text`

12.6 Criteria used to select top models

The criteria for selecting top models is the average precision@30% across 4 temporal block validations.

12.7 Results of top 5 models

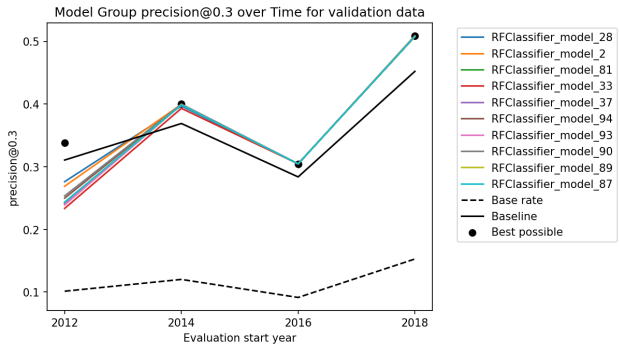


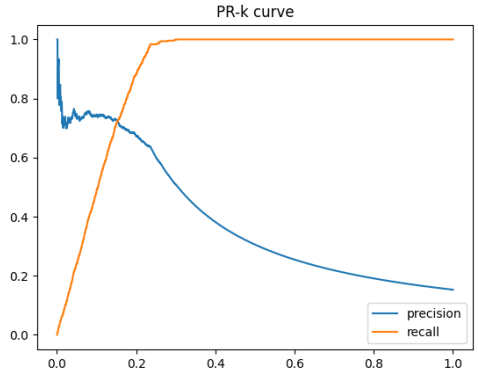
Figure 16: Temporal plot with our best 10 models

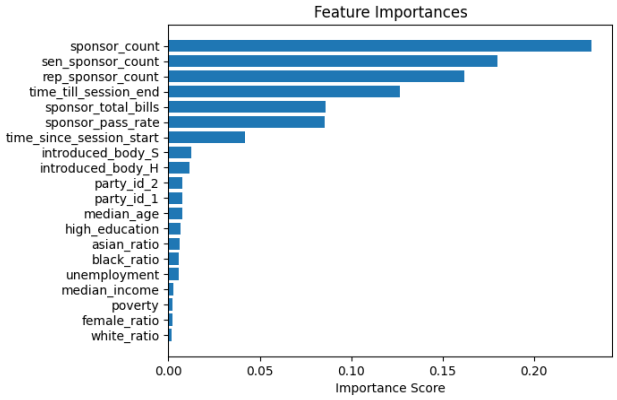
12.7.1 1st model

model type	Random Forest
parameters	n_estimators: 100
	criterion: "entropy"
	max_depth: 10000
	max_features: "log2"
	class_weight: null
Bias metrics	
TPR disparity (African American)	1
TPR disparity(Asian)	1

cross tabs

feature_names	top_vals	bottom_vals
rep_sponsor_count	6.008	1.058
sponsor_count	9.824	1.756
sen_sponsor_count	3.816	0.698
introduced_body_S	0.533	0.394
introduced_body_H	0.467	0.606
time_till_session_end	107.907	95.927
time_since_session_start	397.093	409.073
party_id_2	0.449	0.453
party_id_1	0.551	0.547
sponsor_pass_rate	0.236	0.236



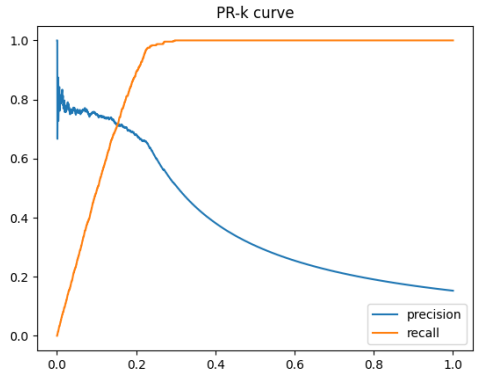


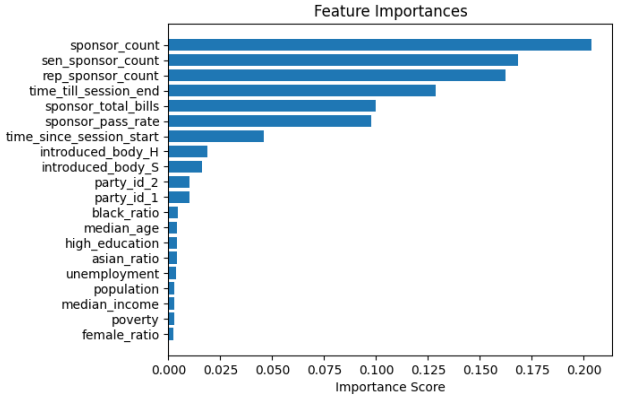
12.7.2 2nd model

model type	Random Forest
parameters	n_estimators: 100
	criterion: "gini"
	max_depth: 100
	max_features: "sqrt"
	class_weight: null
Bias metrics	
TPR disparity (African American)	1
TPR disparity(Asian)	1

cross tabs

feature_names	top_vals	bottom_vals
rep_sponsor_count	6.008	1.058
sponsor_count	9.82	1.758
sen_sponsor_count	3.812	0.699
introduced_body_S	0.533	0.394
introduced_body_H	0.467	0.606
time_till_session_end	107.843	95.954
time_since_session_start	397.157	409.046
party_id_2	0.449	0.453
party_id_1	0.551	0.547
sponsor_pass_rate	0.236	0.236



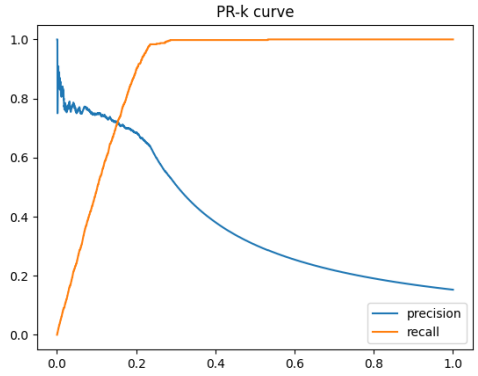


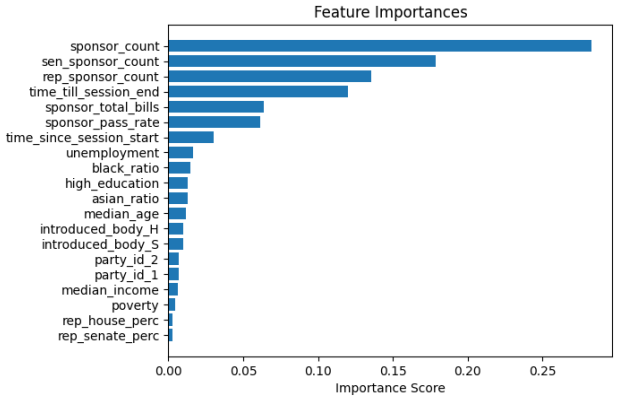
12.7.3 3rd model

model type	Random Forest
parameters	n_estimators: 1000
	criterion: "entropy"
	max_depth: 100
	max_features: "sqrt"
	class_weight: "balanced"
Bias metrics	
TPR disparity (African American)	1
TPR disparity(Asian)	1

cross tabs

feature_names	top_vals	bottom_vals
rep_sponsor_count	6.008	1.058
sponsor_count	9.86	1.74
sen_sponsor_count	3.852	0.682
introduced_body_S	0.533	0.394
introduced_body_H	0.467	0.606
time_till_session_end	107.678	96.025
time_since_session_start	397.322	408.975
party_id_2	0.449	0.453
party_id_1	0.551	0.547
sponsor_pass_rate	0.236	0.236



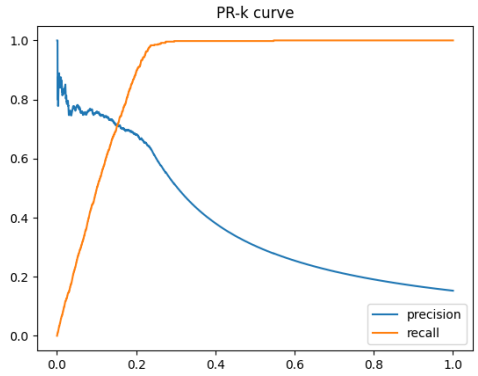


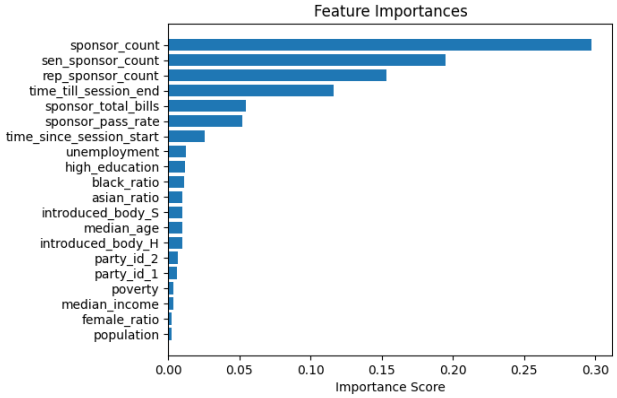
12.7.4 4th model

model type	Random Forest
parameters	n_estimators: 500
	criterion: "gini"
	max_depth: 100
	max_features: "sqrt"
	class_weight: "balanced"
Bias metrics	
TPR disparity (African American)	1
TPR disparity(Asian)	1

cross tabs

feature_names	top_vals	bottom_vals
sen_sponsor_count	3.885	0.668
sponsor_count	9.894	1.726
rep_sponsor_count	6.008	1.058
party_id_1	0.683	0.49
party_id_2	0.317	0.51
introduced_body_S	0.533	0.394
introduced_body_H	0.467	0.606
time_till_session_end	107.833	95.959
time_since_session_start	397.167	409.041
sponsor_pass_rate	0.236	0.236



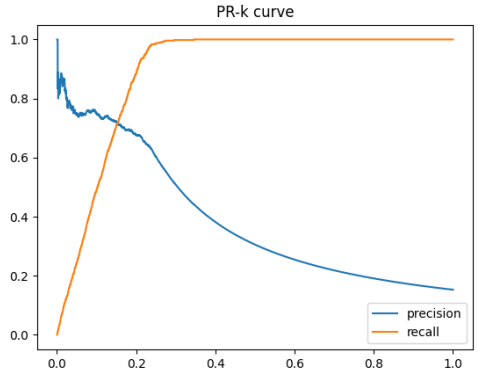


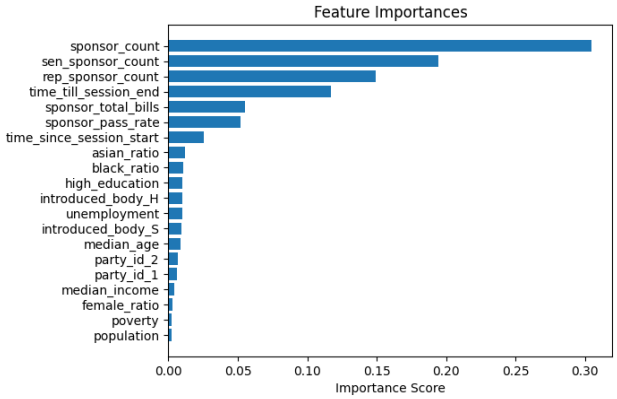
12.7.5 5th model

model type	Random Forest
parameters	n_estimators: 500
	criterion: "gini"
	max_depth: 1000
	max_features: "sqrt"
	class_weight: "balanced"
Bias metrics	
TPR disparity (African American)	1
TPR disparity(Asian)	1

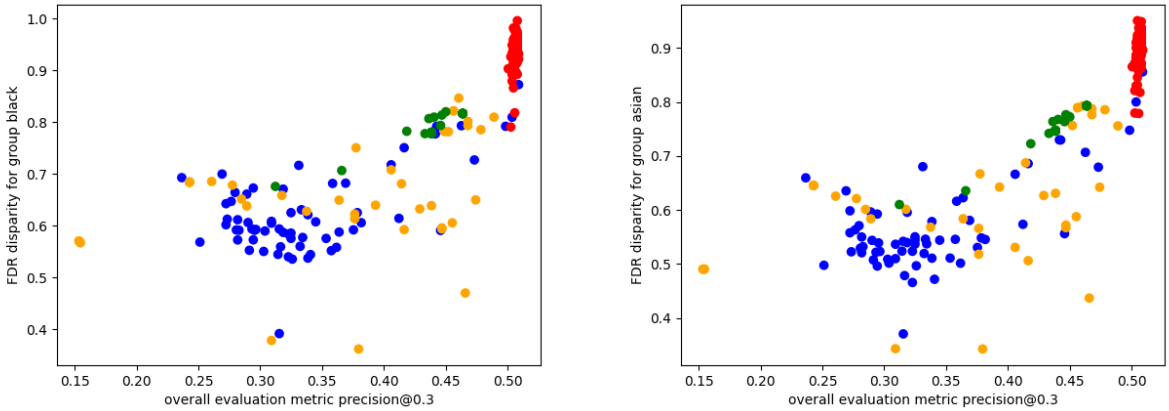
cross tabs

feature_names	top_vals	bottom_vals
sen_sponsor_count	3.864	0.677
sponsor_count	9.873	1.735
rep_sponsor_count	6.008	1.058
introduced_body_S	0.533	0.394
introduced_body_H	0.467	0.606
time_till_session_end	107.727	96.004
time_since_session_start	397.273	408.996
party_id_2	0.449	0.453
party_id_1	0.551	0.547
sponsor_pass_rate	0.236	0.236





12.8 FDR disparity



(a) African American vs White group

(b) Asian vs White group

Figure 18: FDR Disparity

12.9 Creating racial groups

The following telling words were present in bill descriptions from the African-American group's bills but not those from the reference group.

	Word	Number of Occurrences
1	Landlord	72
2	Eviction	71
3	Sterilization	63
4	Distressed	60
5	Post-partum	42
6	Smoking	41
7	Parole	40
8	Arrest	35
9	Officer-Involved	36
10	Detained	32

The following telling words were present in bill descriptions from the Asian-American group's bills but not those from the reference group.

	Word	Number of Occurrences
1	Sterilization	60
2	Eviction	46
3	Landlord	36
4	Asian	32
5	Dealer	21

12.10 Exploratory data analysis plots

	state_id	state_abbreviation	count	rank
0	32	NY	100575	1
1	13	IL	71185	2
2	43	TX	56456	3
3	30	NJ	47002	4
4	11	HI	45191	5
5	23	MN	41990	6
6	21	MA	36870	7
7	36	OK	33359	8
8	24	MS	32268	9
9	46	VA	32076	10
10	38	PA	28573	11
11	5	CA	27953	12

Figure 19: Number of bills introduced per state

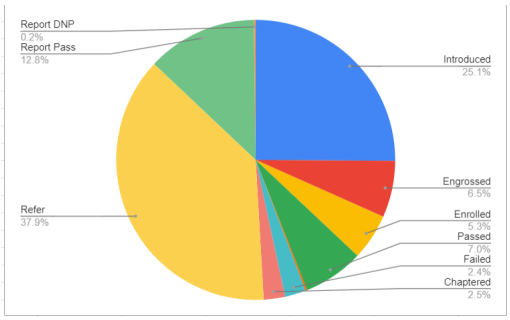


Figure 20: Distribution of `bill_status` labels across all data

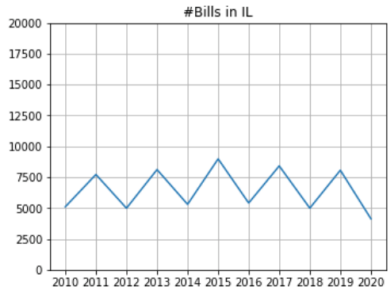


Figure 21: Bill introduced per year in IL state

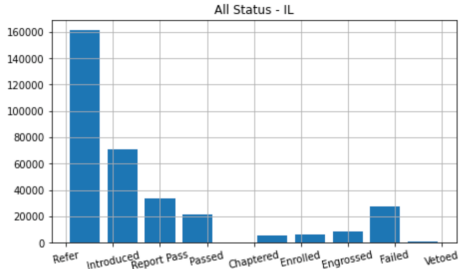


Figure 22: The distribution of labels in IL state bills

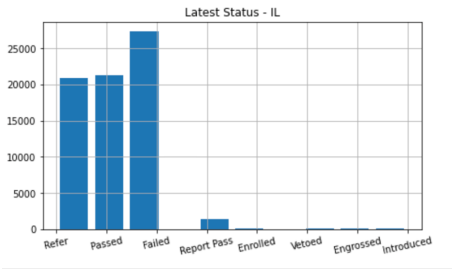


Figure 23: Distribution of the final status label in IL

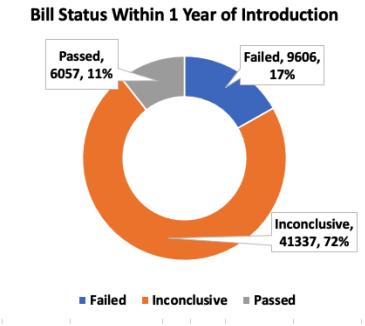


Figure 24: Bill final status label within a year in IL