Barcelona School of Economics

Homework 1

Forecasting and Nowcasting with Text as Data

DSDM - BSE

Noemi Lucchi

Pablo Fernández

May 16, 2025

Part 1

Think of a (onset/incidence) variable to forecast. Optimally, this variable provides you with a large panel (long and many units), i.e. if you are using yearly data try to get many countries/units. Design this well with part 2 in mind.

For this project, we decided to focus on forecasting the occurrence of major protest events across the United States. Specifically, we forecast protest incidence at the state-month level for four sociopolitical issues that have relevant in the last years:

1. **Healthcare access and policy**
2. **Gun control and criminal justice reform**
3. **Housing affordability and homelessness**
4. **Racism and immigration**

These four domains represent critical topics in American society that frequently catalyze public demonstrations, and additionally they offer several advantages as forecasting targets:

- They represent recurring issues with varying intensity across time and geography;
- As a consequence of the point above, they should be well-documented in news media, allowing us to use this source to feed our models;
- They represent politically significant events that decision-makers might want to anticipate.

The data has been obtained from the Crowd Counting Consortium, a joint project from Harvard University and the University of Connecticut which has been recording protests at a detailed level since 2017. After merging, cleaning and preprocessing their datasets, the relevant features of this data can be summarized as follows, where each row is a protest:

- The date and the US state where the protest happened. These features constitute the time and unit dimensions, respectively. Consequently, and since the protest record started in January 2017 and the last recorded month was in April 2025, the panel that is created from this dataset (which is explained in detail in part 3) is wide (several time periods, one per month since January 2017) and relatively long (as many units as US states).
- The issues that can be associated to the protest (e.g., "healthcare", "housing", etc.), where more than one issue can be associated to the same protest.
- The participants and organizers of the protest (text feature).
- The main claims of the protest; e.g., "for banning nuclear weapons", "against the Dakota Access Pipeline". This is another text feature.
- The size of the crowd that attended the protest (though the data is very incomplete).
- Binary indicators for whether there have been any arrests in the protest, injuries among the crowd or among the police and property damage.

We defined a "major protest" threshold for each state based on historical patterns, marking a month as having protest incidence (value = 1) when the number of protests for a specific issue exceeds the state's historical monthly average for that category. This approach accounts for differences in baseline protest activity across states.

Part 2

Merge other time series to this outcome variable. One needs to be a text- based measure. Ideas: Google trends data, GDELT data, our topics from conflictforecast.org download section. You do not need to come up with your own text mining pipeline.

Two additional time series datasets have been added as features:

- LDA topic share *stocks* for each US state from January 2017 to March 2025.
- Google Trends indexes for some common n-grams of claims for each type of protest issue (e.g., "rent" for the "housing" issue).

LDA topic share stocks

Regarding the first feature dataset, the original data consisted of raw LDA topic shares per article and month, where the articles were linked to a specific US county (ADM2 territorial level). By applying basic geospatial operations, the articles were linked to one US state (which represents our unit dimension in this project). However, and for the features to be useful, we had to aggregate the topic shares from the article level to the US state level. There were mainly 2 possibilities for performing this aggregation:

- By computing the mean topic share per US state and month;
- By computing the stock of LDA topic shares per US state and month.

Given that we wanted to reduce the impact of noise and minimize volatility in the data, the aggregation method that was used was the stock for LDA topic shares. As explained in detail below, this measure keeps some "memory" of the amount of text (and its topic) that has been generated. To aggregate following this method, first we computed the stock of tokens:

$$W_{t=T} = \sum_{t=1}^{T} \delta^{T-t} w_t$$

where:

- w_t: number of tokens in all documents of a specific state at month t (i.e., how much text is being produced in one month).
- $\delta = 0.8$: decay factor, which enables that information is carried through time, but with recent months weighing more than previous ones in the stock of tokens.
- W_t: weighted sum of all past token counts up to month T.

Once we had the calculation for the stock of tokens, we then proceeded with the stock for topic shares:

$$X_{k,t=T} = \frac{\sum_{t=1}^{T} \delta^{T-t} w_t x_{k,t}}{W_T}$$

where:

- $x_{k,t}$: share of topic k across all documents of a specific US state at month $t \rightarrow$.
- $X_{k,t}$: present value of the flow of tokens for each topic and state.
- $\delta^{T-t} w_t x_{k,t}$: discounted contribution of that month's topic tokens.

By following this aggregation at the US state level we were able to obtain smoother LDA topic shares than with simple mean aggregation, as shown in Figure 1.

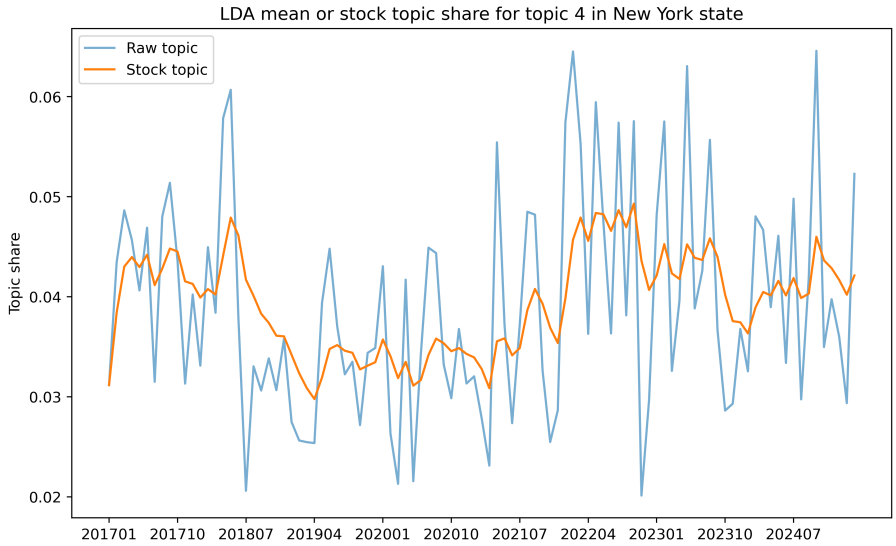

Figure 1: LDA mean vs stock topic share for topic 4 in New York state.

Google Trends

Regarding the Google Trends dataset, we tried to extensively scrape the Google Trends index for the most relevant n-grams associated with each of the considered issues. To do so, we applied TF-IDF to the claims data and computed the term frequencies of the actors (either organizers or participants), dividing the data by one text corpus per type of issue and with one protest constituting a document. Furthermore, to avoid data leakage, we did not consider those terms or actors that made their first appearance after the starting period of the validation set, as including them would imply leaking knowledge from the future to a present in which these terms should not be known.

Table 1 shows some examples of the results that we obtained after ranking the top claims and the top actors.

Nevertheless, and after trying different approaches, we were not able to avoid being blocked from doing more requests after a few were done. Therefore, in the end we only had indexes (for each US state and from January 2015 to May 2025) for the following terms: *health*, *healthcare*, *rent*, *criminal justice* and *immigrants*. However, we still expect these indexes to provide meaningful insights, with an increase in the index expected to imply an increase worry related to the topic that the term is linked to.

Issue	Sample of top claims	Sample of top actors
Healthcare	healthcare, health insurance, vaccination mandates, abortion, hospital benefits	nurses; healthcare workers; National Nurses United; Massachusetts Nurses Association; United Food & Commercial Workers Local 7
Racism & Immigration	police brutality, immigration, antiracism, white supremacy, black lives matter	NAACP; ICE; Families Belong Together; Patriot Front; students
Housing	affordable housing, evictions, homelessness, rent, homeless encampments	tenants; community members; advocates; local unhoused people; students
Guns & Criminal Justice	gun control, gun violence, school safety, legislation, police	Moms Demand Action; March for Our Lives; prisoners; elected officials; families belong together

Table 1: Sample of 5 queries and actors for each issue, as identified by TF-IDF.

Part 3

Define an onset or incidence variable based on your data in 1. Carefully explain why forecasting this is useful. Who could be the decision-makers that can make use of this forecast and how? Explain the prediction policy problem they face.

Our incidence variable is the outbreak of major protests related to specific topics. To define our target variable, we first computed the monthly, state-level average number of protests for each protest type. We then assigned a value of 1 to the target variable if the number of protests in a given month exceeded this average. Given this definition, the incidence variable is equal to 1 if the number of protests in that state exceeds this historical average within our forecast horizon (that has been set to 6).

The purpose of our model is to forecast major unrest on specific topics, enabling more targeted and effective policy interventions that respond to people's needs and concerns. It is reasonable to assume that no government, regardless of wealth, can fully satisfy the demands of all its citizens. Grievances are therefore expected to persist, especially in the presence of limited resources and competing interests across different social groups. A certain level of protest is natural and does not always warrant intervention.

Given these structural constraints, our model is designed to act as a filter: it helps identify those situations in which intervention is necessary to prevent the escalation of unrest into larger crises.

We emphasize that our notion of intervention is strictly preventive—not repressive. It does not involve suppressing protests through force, but rather anticipating the underlying issues and working to solve them before discontent increases excessively. In a large and diverse country like the United States, unrest can arise for many reasons. To maintain focus and interpretability, we chose to concentrate on four protest topics that we found to be especially prominent.

Just as it is not useful to forecast protests that fall within normal levels, it is also not meaningful to predict major protests over issues that do not reflect widely shared concerns. Our model is therefore calibrated to detect only those outbreaks that are both exceptional in scale and relevant in substance.

Part 4

Do some feature engineering on the right-hand-side data if necessary. Explain the reasons for the feature engineering carefully.

To enhance our model's ability to forecast protest activity, we used the `FeatureEngineer` class which provides methods for creating temporal features that capture the historical patterns in our data.

Rolling Features

Following the course materials, we implemented weighted rolling means to capture temporal trends in protest activity. We created weighted rolling means of protest counts using window sizes of 1, 3, and 12 months, which allowed our model to learn from both recent and longer-term protest patterns.

For the weighting scheme, we used an exponential decay with $\alpha = 0.8$, which assigns higher importance to more recent observations while still accounting for activity in earlier months. This approach aligns with the intuition that more recent protest activity is more predictive of future protests, but historical patterns still provide valuable context.

Lastly, we maintained the original values (rather than log-transforming them) to preserve the interpretability of the features, particularly since protest counts rarely reach extreme values that would need log transformation.

Discrete Features

To capture specific temporal dynamics related to protest activity thresholds, we implemented the since and the ongoing variables.

For the since variables, we calculated the number of months since protest activity exceeded specific thresholds (2, 5, and 10 protests). This feature helps the model identify states that have been peaceful for extended periods, capturing the temporal distance from the last significant protest activity. By using multiple thresholds, we enable the model to distinguish between different intensities of past activity, recognizing that the predictive value of past protests may depend on their scale.

For the ongoing variables, we tracked consecutive months of protest activity above the same thresholds, resetting to zero when activity falls below the threshold. This feature captures the tendency for sustained protest movements to continue. These variables are especially valuable for distinguishing between isolated protest events and sustained social movements.

Lagged Variables

We incorporated direct lags of the protest count variable at 1, 2, and 3 months. These lags provide the model with the most recent protest counts, capturing short-term autocorrelation in protest activity. Unlike rolling means which aggregate information, these lagged values preserve the exact protest counts from specific prior months. Finally, the inclusion of multiple lags allows the model to detect various temporal patterns, such as monthly or quarterly cycles in protest activity.

LDA Topic Stocks

As text-based features, we incorporated LDA topic stocks derived from news articles (Mueller and Rauh, 2022). These features represent the discounted accumulation of topic shares over time, capturing the evolving narrative around each state in the news media. The stocks smooth out short-term fluctuations in media coverage while maintaining sensitivity to meaningful shifts in discourse. By merging these topic stocks with our protest data, we enable the model to capture patterns between specific narrative themes and upcoming protests.

Google Trends Data

Another text-based feature incorporated into the dataset is Google Trends data for specific keywords associated with each protest category. This data serves as a proxy for public interest and awareness, under the assumption that higher search volumes may reflect heightened concern or attention to a particular issue. As with the previously discussed features, including this variable in the model may help uncover patterns linking public interest to protest activity.

After analyzing each feature specifically, we briefly discuss some commonalities of this step:

- We ensured absence of data leakage by creating features in a way that use only past information that would be available at prediction time.
- We incorporated both features that capture short-term fluctuations (lags) and those that capture longer patterns and the historical trend (rolling mean and since/ongoing variables).
- We integrated diverse data sources to capture different aspects of the phenomenon (raw statistics about previous protests, media narrative and public attention).

When discussing the features we created, it is also crucial to analyze their contribution to the model's predictions. In Figure 2, we present the feature importance scores, broken down by model type, since we employ a binary classifier for each protest category. Overall, the most influential features are the 12-month rolling mean of protest counts and the variable indicating the number of months since the last occurrence of more than 10 protests. These are followed by the Google Trends search volume and, subsequently, the LDA-based stock topic share.

These findings are consistent with the underlying nature of protest dynamics. Given that protests are inherently population-driven events, it is reasonable to assume that early signals may emerge from public sentiment (reflected in online search behavior captured by Google Trends) as well as from historical protest patterns, such as the frequency and recentness of large-scale events. Much like in conflict studies, regions with a history of recurring protests are more likely to experience similar events again, compared to regions that have been relatively quiet for a prolonged period (e.g., the last 10 months).

In this context, it is also coherent that the LDA topic share ranks third in predictive power. This variable reflects media narratives, which may lag behind or selectively ignore public concerns. As such, while informative, media coverage may only partially capture the early signals of emerging unrest.

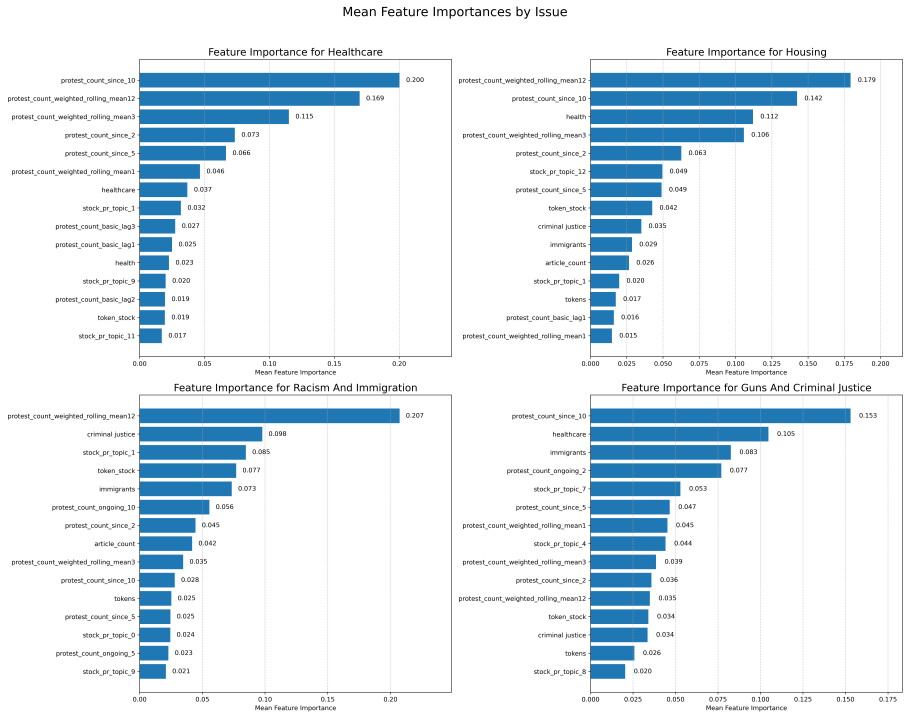

Figure 2

Part 5

Adapt the rolling forecast code that you should by now all be familiar with. Make sure you implement this correctly. Don't write your own code but adapt Ben's.

We start by discussing the choice of the parameters:

- **Horizon:** We selected a 6-month forecast horizon, allowing our model to predict protest activity several months in advance, which provides sufficient time to design and implement interventions.
- **Gap:** We set the as `horizon - 1`. This configuration ensures that we avoid data leakage by properly accounting for the temporal structure of our target variable, which uses information from t to $t + 5$ for each forecast.
- **Test size:** We used `test_size = 1`, meaning that each train-test split uses a single month for testing.
- **Number of splits:** We dynamically determined `n_splits` based on the number of test periods available in our dataset from October 2022 onwards. This approach ensures comprehensive evaluation of our model's performance across a full year of data.

This configuration creates a series of train-test splits that ensures that each forecast is made using only information that would have been available at prediction time.

We implemented the `PanelSplit` class in the following way:

- For each of our four protest categories (healthcare, guns/criminal justice, housing, and racism/immigration), we initialized a separate `PanelSplit` object.
- Each `PanelSplit` object was configured with the periods from our dataset, ensuring that

the temporal structure of our panel data was properly maintained.

- The resulting split objects contain a sequence of train-test indices that define which observations are used for training and testing in each iteration of our rolling forecast.

This implementation ensures that our forecasting procedure properly respects the panel structure of our data, with states as the cross-sectional units and months as the time dimension.

For each protest category, we prepared the features and the target variable in the following way:

- As features, we included all engineered features previously described (i.e., rolling means, since and ongoing variables, lags, topic stocks, and Google Trends data).
- As target variable, we used the incidence variable `inc_anymajorprotest_th0_h6`, which indicates whether protest activity exceeded the historical state-level average in the 6-month horizon.

We then configured a Random Forest classifier as our model, which has been used to make predictions with the four different datasets:

- We carefully chose hyperparameter (`max_depth=4`, `max_features=0.2`, and `min_samples_leaf=50`), since we where initially overfitting, which was highly probable given the class unbalanceness.
- The `min_samples_leaf` parameter ensures that terminal nodes contain a substantial number of observations, preventing the model from capturing noise in the training data.

Finally, we implemented our rolling forecast using the `cross_val_fit_predict` function from the `panelsplit` package:

- For each protest category, we trained and evaluated our model across all train-test splits defined by the corresponding `PanelSplit` object.
- We used the `predict_proba` method to obtain calibrated probability estimates rather than binary predictions.

Part 6

Plot ROC curve and precision recall curve of the merged fitted values. Discuss these curves in detail.

The ROC curves (Figures 7-10) plot the True Positive Rate (TPR) against the False Positive Rate (FPR) across different classification thresholds, providing insight into the model's ability to discriminate between months in which there will be high protests activity and months in which there won't be.

Our models demonstrate strong discriminative ability across all four protest categories, with the AUC spanning from 0.74 in the racism and immigration model, to 0.85 in the housing one. These AUC values substantially exceed the 0.5 baseline of a random classifier, indicating that our models capture meaningful signals predictive of protest activity.

The shape of the ROC curves reveals important characteristics of our forecasts:

- All curves show a steep initial rise, indicating that our models achieve substantial true positive detection with minimal false positives at high-confidence thresholds.

- The curves for healthcare and housing show particularly strong performance in the low FPR region (0-0.2), suggesting these models are especially valuable when high precision is required.
- The flatter curve sections at higher FPR values suggest diminishing returns when attempting to capture every protest event, as expected in real-world prediction tasks, and considering also nature of the prediction task which is not trivial.

While ROC curves are standard in classification evaluation, Precision-Recall curves (Figures 3-6) are particularly informative for imbalanced classes like protest events, focusing on the trade-off between precision and recall .

Our models achieve remarkably high average precision scores, spanning from 0.87 for the guns and criminal justice dataset, to 0.95 for both the healthcare and the housing datasets.

These high values suggest that our models maintain high precision even at substantial recall levels, which is a desired characteristic in real-world applications where false positive are costly.

The Precision-Recall curves reveal several notable patterns:

- All four categories maintain precision above 0.8 even at recall levels approaching 0.8, indicating robust performance across a wide range of operating thresholds.
- The healthcare and housing models maintain particularly high precision (¿0.9) even at high recall levels, only dropping below 0.9 precision near 0.9 recall.
- The guns and criminal justice model shows a more gradual decline in precision as recall increases, indicating a more challenging prediction task for this category.
- All curves show a characteristic drop in precision at very high recall values (¿0.9), reflecting the difficulty of capturing the most challenging major protest outbreaks without generating false positives.

Analyzing these evaluation metrics together, we can draw the following conclusion about our models' performance:

- Both evaluation approaches rank healthcare and housing as the most predictable categories, with guns and criminal justice and racism and immigration showing relatively lower performance.
- The notably high average precision scores suggest that our models are well-calibrated for forecasting our target variable, since they maintain high precision despite class imbalance.
- The fact that precision is high across different recall ranges indicates that our forecasts would generate false positives while still capturing the majority of true positives, which is a desirable property for an implementation like this one.

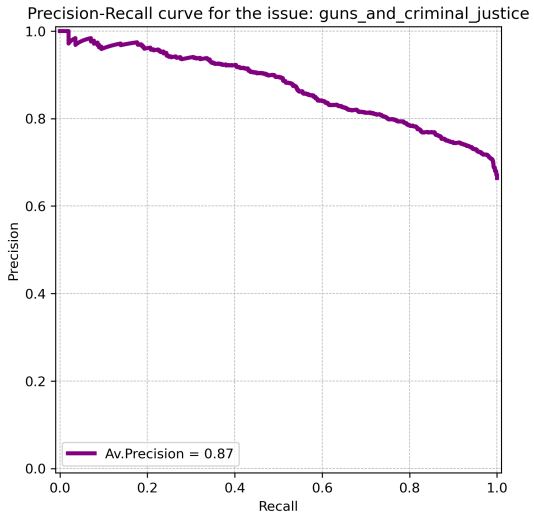

Figure 3: Precision-Recall: Guns and Criminal Justice

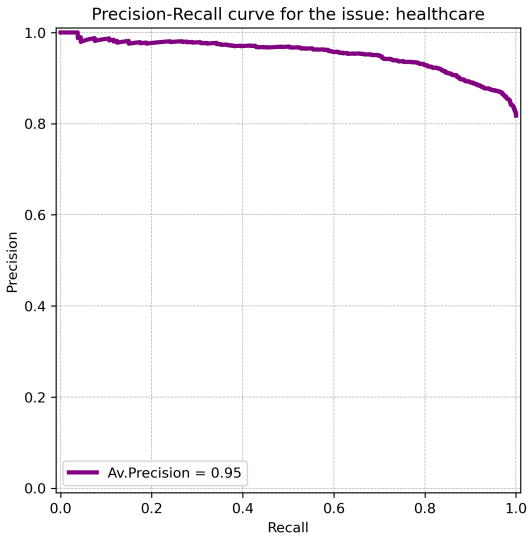

Figure 4: Precision-Recall: Healthcare

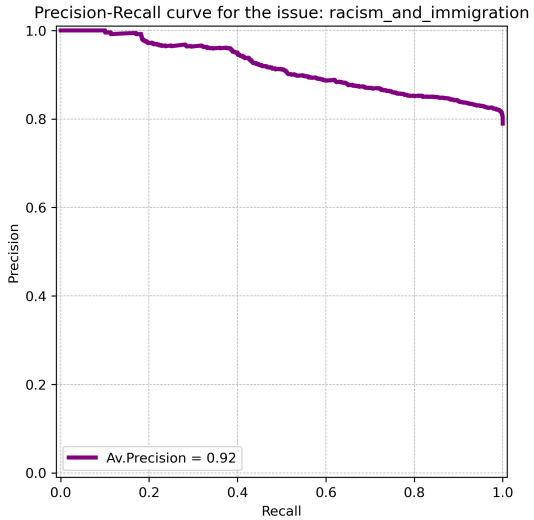

Figure 5: Precision-Recall: Racism and Immigration

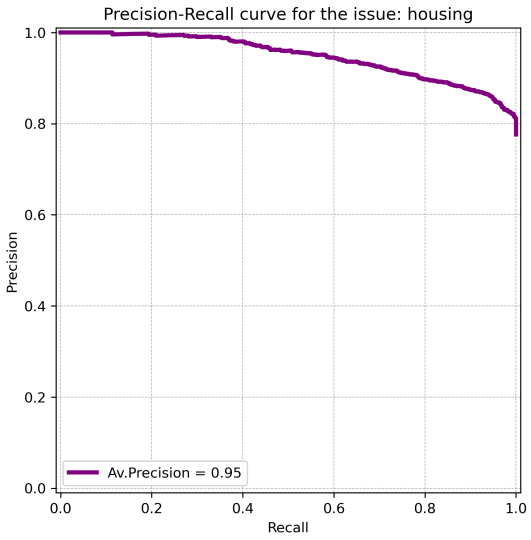

Figure 6: Precision-Recall: Housing

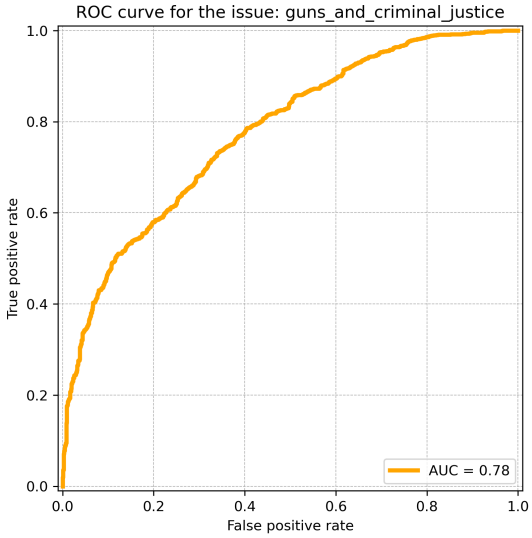

Figure 7: ROC: Guns and Criminal Justice

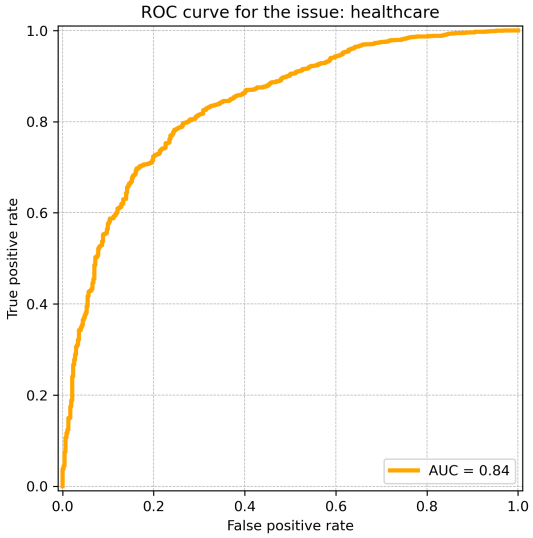

Figure 8: ROC: Healthcare

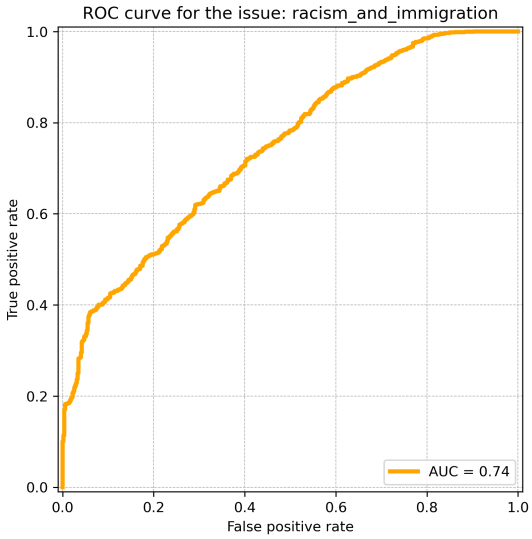

Figure 9: ROC: Racism and Immigration

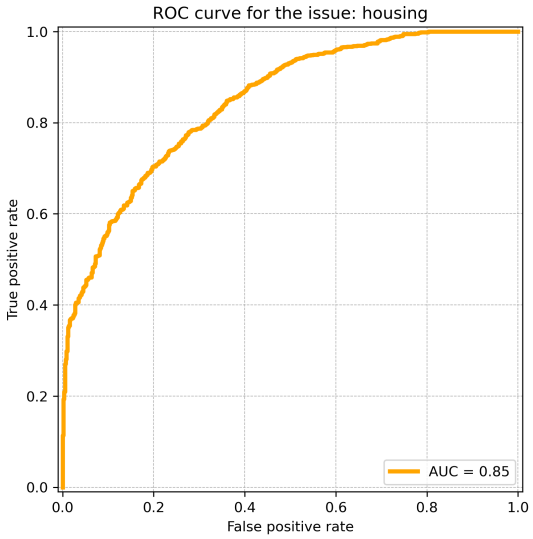

Figure 10: ROC: Housing

Bonus

Adapt the cost model to your purpose. Use it to find the optimal cut- off under assumptions on the costs on TP, FP, TN and FN. In order to argue for the reasonability of your assumptions it will help to think of what FP and FN in particular mean for your decision-maker from question 3). What decisions could be triggered by the forecast?

As previously mentioned, in this specific context, intervention does not entail preventing protests from occurring altogether. Rather, it involves implementing measures aimed at improving the underlying conditions that drive people to protest in the first place. Since we selected topics that are particularly relevant within the American socio-political landscape, we believe that government authorities (at either the state or federal level) might be incentivized to intervene if certain issues appear to be widespread.

This approach aligns with the idea of addressing the root causes of discontent, rather than passively waiting for unrest to escalate and subsequently attempting to suppress it. Protests can cause significant disruptions in urban areas and, in some cases, may involve episodes of violence. Being able to anticipate the issues that lead to public grievances—and to take action before they evolve into large-scale unrest—is precisely the objective of our model.

In our framework, the cost of intervention is represented by the reforms and structural changes that governments would need to implement to mitigate the unrest. On the other hand, the cost of a false negative (i.e.,failing to predict an upcoming surge in protests) includes both the immediate consequences of the protests (e.g., injuries, property damage, traffic disruptions) and the longer-term risks associated with growing, unresolved social tensions.

It is important to clarify that our model is not designed to predict whether a specific protest will occur. Rather, it forecasts whether the number of protests on a given topic will exceed the historical average in a given state over a forecast horizon of h months. Predicting isolated events is not informative in assessing the broader social climate and could even be misleading: for instance, a protest organized by a radical fringe group might not reflect any meaningful or widespread societal concern that warrants government intervention.

Given this perspective, we represented our target variable as a binary indicator of whether, in the next h months, the number of protests on a specific topic will be significantly higher than the historical average for that topic in the given state. To achieve this, we identified four major themes that are common causes of unrest in the United States and built separate forecasting models (one for each topic) to predict potential surges in protest activity.

The minimization of the total cost, which is given by the major protest outbreaks and the intervention itself, is represented by the following function:

$$\min_c \mathbb{E}[\text{Cost}_c] = \text{Cost}_{TP} \times \text{TP}_c + \text{Cost}_{FP} \times \text{FP}_c + \text{Cost}_{FN} \times \text{FN}_c + \text{Cost}_{TN} \times \text{TN}_c \quad (1)$$

Cost Components

True Positive Cost

When correctly predicting future protests and implementing policy changes, the cost is:

$$\text{Cost}_{TP} = I + (1 - p)V_D - S \quad (2)$$

Where:

- I = cost of policy intervention
- p = effectiveness of policy intervention in solving underlying problems (0 to 1)
- V_D = cost of protests if no intervention
- S = social benefit from addressing legitimate concerns

We still add the term $(1 - p)V_D$ with a positive sign since it represents the potential partial protest costs that still occur despite intervention, if concerns are not addressed properly. Also, this includes protests that happen for other reasons. Indeed, the policy is specifically suited to solve one specific issue, so it generally will be ineffective in solving any other issues, and this will also determine costs associated to protests.

False Positive Cost

When predicting the outbreak of major protests, which won't actually happen, and so implementing some policy reforms, the cost is:

$$\text{Cost}_{FP} = I - \alpha S \tag{3}$$

Where α represents a discount factor (0 to 1) for social benefits when addressing issues that weren't urgent enough to cause major protests, but still people might appreciate the policy reform that has been implemented and so this reduce the total cost.

False Negative Cost

When forecasting that there won't be any major protests, but there will actually be, the cost is given by the cost of protests themselves:

$$\text{Cost}_{FN} = V_D \tag{4}$$

True Negative Cost

When correctly predicting no major protest, there might still be a cost, which is given by the protests that normally occur and are considered the benchmark level:

$$\text{Cost}_{TN} = \beta V_D \tag{5}$$

Where β represents a discount factor (0 to 1) since the cost of the usual level of protest is not as sever as the one caused by major protests.

Parameters defined above highly effect the results, and they are also specific to the topic:

- **Higher effectiveness (p)**: As policy interventions becomes more effective in addressing root causes, the intervention threshold lowers, because
- **Higher social benefit (S)**: As policy reforms create more social value, intervention becomes justified at lower risk thresholds
- **Varying intervention costs (I)**: Cheaper interventions justify actions at lower risk thresholds
- **Discount factor** α: If the social benefits determined when addressing issues that weren't urgent enough to cause major protests are high, then intervention becomes justified at lower risk thresholds
  item **Discount factor** β: An high cost of usual level of protest justifies actions at lower risk thresholds since the state is already experiencing high costs due to the usual level of protests, and it can't face the more severe costs of major protests

Assuming moderately high levels of effectiveness, social benefits and parameters α and β, and that the required interventions are not prohibitively expensive, our cost model suggests that governments should adopt a relatively proactive stance in addressing potential sources of discontent, rather than waiting for protests to materialize.

Even when an intervention turns out to be unnecessary with respect to the specific topic under consideration, it is likely to still yield positive externalities. This is due to the persistent baseline level of protest activity, as well as the potential for overlapping or interconnected issues that may benefit from the same reform.

Under these assumptions, the cost of a missed protest prediction (i.e., a false negative) is greater than the cost of an unnecessary intervention (i.e., a false positive). As a result, the optimal decision-making strategy involves intervening at relatively low probability thresholds, thereby favoring early and preventive action over reactive measures.