Less is More: Enhancing Model Performance in Supreme Court Decisions with Minimal Data

Anonymous

1 Introduction

The United States Supreme Court cases offer a unique opportunity to test machine learning models in the context of complex and nuanced decision-making. This study focuses on developing predictive models for Supreme Court rulings using a comprehensive dataset of over 5,000 cases. The primary objective is to explore the effectiveness of various machine learning algorithms in predicting whether the Supreme Court will reverse or affirm lower court decisions.

My paper addresses three key research questions:

1. How do different machine learning models compare in their ability to predict?
2. Can I improve model performance by considering features that become available only after or at decision time?
3. How can I reduce the number of training instances while maintaining or improving model performance?

2 Literature Review

The role of partisanship in judicial decision-making has been widely studied. For instance, Fang et al. (2023a) analyzed the Super-SCOTUS dataset to investigate how a judge's political affiliation influences courtroom language. Contrary to the expectation of political neutrality, they found a strong correlation between language use and justices' political leanings, evident both at the court and individual levels over time. This paper employs BERT NER (Devlin et al., 2019) to analyze language patterns.

Cohen and Yang (2019) further examine sentencing disparities, revealing that a judge's political affiliation impacts outcomes based on defendants' race and gender. Their study showed that Republican-appointed judges imposed harsher sentences on Black defendants compared to non-Black ones, while offering more lenient sentences to female defendants versus males. These disparities were consistent from the fact-finding stage to sentencing, even when controlling for other judicial and court characteristics, using interaction regression.

These studies collectively underscore the importance of considering political affiliation, directly or indirectly, when predicting the outcome of legal decision making.

3 Method

3.1 Data

This study modifies a subset of the Super-SCOTUS dataset (Fang et al., 2023b), which includes metadata and transcripts of U.S. Supreme Court cases between 1955 and 2019.

Several features relating to state, category, issue area and Chief Justice are one-hot encoded. In the post-decision dataset, a new feature named *deliberation_length*, which is the difference between the argument date and decision date, is included. The court hearing transcript is vectorized as embeddings using the Sentence Transformer introduced by Reimers and Gurevych (2019). All features are scaled before fitting to improve convergence in logistic regression and allows for accurate distance calculation in k-nearest neighbours (used in ensemble model).

Identifying features such as case id and case title were excluded. Feature selection involves running three random forest models at different maximum depths and compare feature importance scores. Random forest is chosen for its ability to handle missing data effectively (Hayes et al., 2024), especially for features relating to state in this dataset. Features relating to the court hearing embeddings consistently ranked as the most important. After approximately top twenty features, the remaining features rankings are no longer consistent across different maximum depths. Observing the top twenty features, beside court hearing, for each maximum depth, I picked the eight common features between them (Table 1). I note that *issue_area_UNKNOWN* is one of the highest scorers but because of the broad hidden values behind it, I have chosen to exclude this feature.

Selected Features
majority_ratio
respondent_category_United States
issue_area_Judicial Power
argument_date_num
court_hearing_length
utterances_number
petitioner_category_United States
year

Table 1: List of features

Three categories of dataset are created, corresponding to the three research question: dataset with features before decision, dataset with features after decision, and dataset with features before decision but excluding noisy features (data-efficient dataset).

3.2 Evaluation

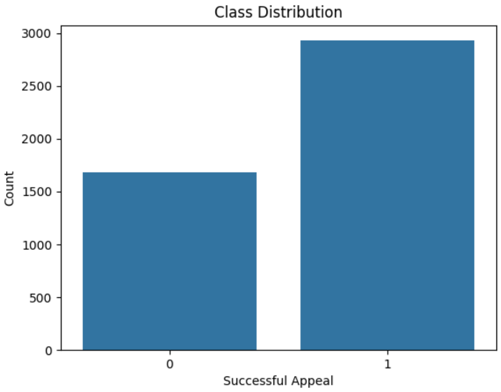

Figure 1: Target class distribution

Due to the class imbalance in the dataset (Figure 1), the F1 score is prioritized over accuracy. Similar to judicial decision-making, where false positives are more detrimental than false negatives and positive outcomes are the majority class, a weighted-average approach is necessary. Precision will also be considered alongside F1 score.

A hold-out strategy was used, utilizing the training dataset for model fitting, transformation, and tuning, while reserving the development dataset solely for evaluation. I acknowledge that relying on a single evaluation set lacks robustness, as model performance may vary with different target class distributions.

3.3 Machine Learning Models

Logistic regression is chosen due to its simplicity and the interpretability of its coefficients.

An ensemble model was utilized, which combines five different classifiers: a logistic regression model, a Gaussian Naive Bayes model, a decision tree model, a k-nearest neighbors (KNN) model, and a random forest model. These models are stacked together, and a simple voting mechanism is employed as the meta-classifier. A limitation of ensemble learning is the lack of interpretability, especially when applied to the legal context.

Hyperparameter tuning uses 5-fold cross-validation. Table 2 indicates the hyperparameters considered when tuning for each model.

Model	Hyperparameter
Tree	max_depth, min_samples_split, min_samples_leaf, criterion
KNN	n_neighbours, weight, metric
Logistic	C, solver, max_iter
NB	var_smoothing
RF	n_estimators, max_depth, min_samples_split, min_samples_leaf, bootstrap

Table 2: Model hyperparameters

3.4 Noisy Data Identification

This study employs three different methods to exclude noisy data. Active learning is applied separately on logistic regression model, decision tree model, Gaussian Naive Bayes and k-nearest neigbours model. Logistic regression model used earlier is modified to exclude instances where the probability of prediction is between 45-55% (herein refers as 'boundary logistic model') (Figure 2). Ensemble model is also modified to exclude instances that have two out of five disagreeing votes (herein refers as 'ensemble disagree model') (Figure 3).

4 Result

All models perform better than the baseline in terms of F1 score. Logistic model, despite being

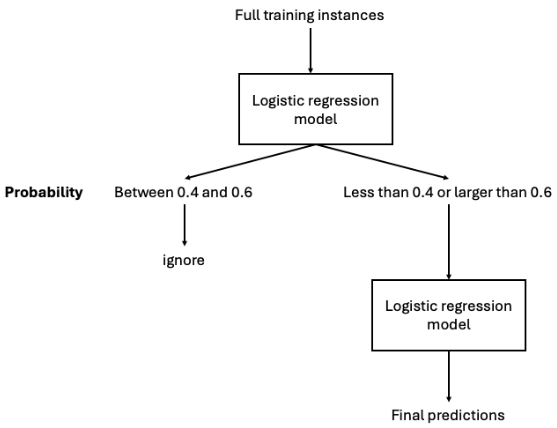
Full training instances
Logistic regression model
Probability
Between 0.4 and 0.6
Less than 0.4 or larger than 0.6
ignore
Logistic regression model
Final predictions

Figure 2: Logistic boundary model

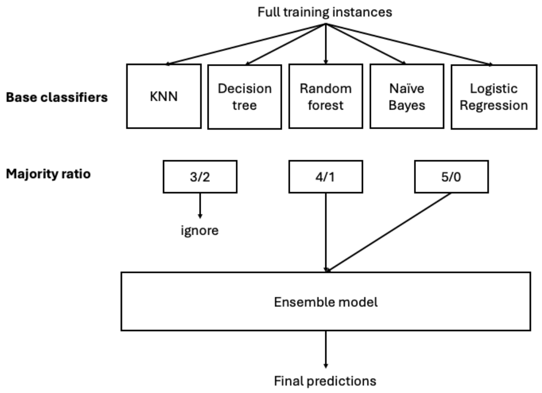

Figure 3: Ensemble disagree model

significantly simpler, performs on-par with ensemble model in the pre-decision dataset. However, in both the post-decision dataset and data-efficient dataset, ensemble model and its variant prove significantly superior to logistic model and its variant. Including the post-decision features improves the performance of both models.

5 Discussion

5.1 Pre-decision Dataset

The logistic regression model exhibits a slight tendency toward overfitting (Figure 4). Various strategies were employed to enhance its performance, including the removal of specific features such as state, category, and year, as well as the omission of hyperparameter tuning. Regardless, the overfitting issue does not improve.

The current ensemble model with five tuned base models indicates severe overfitting (Figure 5). Removing the tuning improves the performance but the overfitting issue persists, albeit to a lesser extent (Table 5 and Table 6). I further reduced the complexity of the model by eliminating one classifier at a time (Table 4), and the two lowest performers (Naïve Bayes and logistic regression) together. However, there is minimal improvement to the issue of overfitting and performance suffers further.

Model/Dataset	Precision	F1 Score
Baseline	.64	.49
1. Pre-decision		
Logistic	.54	.55
Ensemble	.56	.53
2. Post-decision		
Logistic	.61	.58
Ensemble	.66	.56
3. Data-efficient		
Active (logistic)	.56	.55
Boundary logistic	.55	.56
Disagree ensemble	.62	.55

Table 3: Evaluation of main models

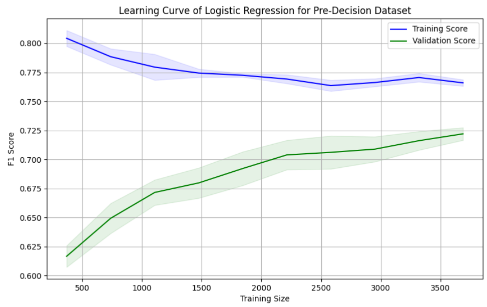

Figure 4: Logistic regression model with hyperparameter tunings

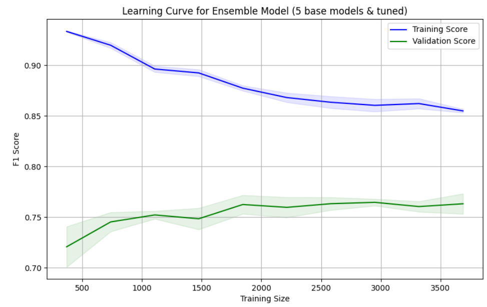

Figure 5: Ensemble model with 5 base classifiers that are tuned

Although both models achieve similar

Removed	RF	KNN	Tree	NB	Logistic
Precision	0.57	0.57	0.57	0.61	0.60
F1 score	0.57	0.57	0.56	0.55	0.57

Table 4: Reducing Base Models

	3 base models	5 base models
Not tuned	.64	.65
Tuned	.64	.62

Table 5: Precision of various ensemble models

weighted-average recall, their recall compositions differ significantly. The ensemble model correctly identifies nearly all positive cases (0.93) but only 0.09 for negative cases. Because of the overlap in base classifiers' behaviours, a positive prediction is more likely to be the final vote (Figure 6). In fact, due to the class imbalance and the relative simplicity of the models, both k-nearest neighbours and Naïve Bayes return positive predictions on all cases of the development set, exacerbating the ensemble model's performance in recognizing negative cases.

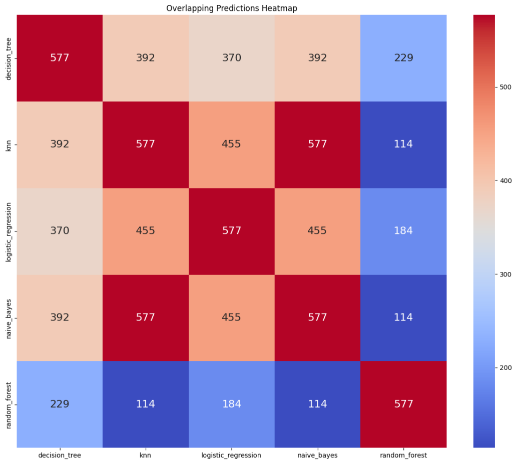

Figure 6: Overlapping matrix

	3 base models	5 base models
Not tuned	.51	.56
Tuned	.50	.53

Table 6: F1 score of various ensemble models

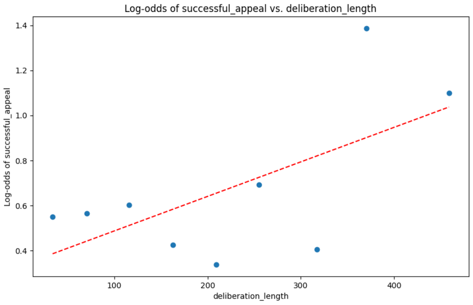

Figure 7: Log-odds of deliberation length

5.2 Post-decision Dataset

Table 3 shows that the ensemble model effectively utilizes the newly introduced majority ratio and deliberation features. This is likely because neither the deliberation length (Figure 7) nor the majority ratio (Figure 8) has a linear relationship with the log-odds of the target variable. This non-linearity may explain why the logistic regression model, which assumes linearity, did not benefit as much from the post-decision data. In contrast, the ensemble model leverages this information more effectively as it captures the non-linear relationships and feature interactions better. However, the ensemble model's tendency to favour positive outcomes, especially in a development dataset dominated by positives, may also contribute to its superior performance in this context.

5.3 Data-Efficiency

Active learning is typically used when labeled data is scarce. However, even with sufficient labeled data, reducing the dataset to 1,200 instances (about 25% of the full training set) yields comparable performance for active learning model using logistic regression, decision trees, and Naïve Bayes. Interestingly, active learning with k-nearest neighbors (KNN) struggles significantly in precision and F1 score. This is because active learning relies on assumptions about data distribution, while KNN, a non-parametric classifier, makes no strong assumptions, resulting in poor performance (Ritter and Chaudhuri, 2023). Furthermore, due to class imbalance, the active learning model using KNN fails to predict any negative instances.

In contrast to active learning, which focuses on boundary instances, Joshi et al. (2024) suggest that in Contrastive Language-Image Pre-Training, targeting images with clear corresponding captions yields superior prediction performance compared to using the entire dataset. In the legal context, cases with out-

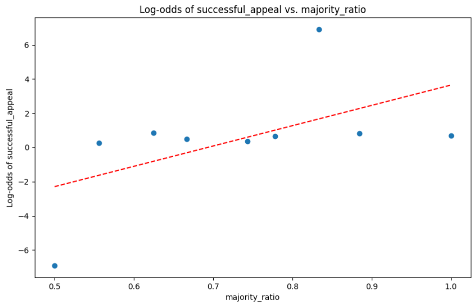

Figure 8: Log-odds of majority

	KNN	Decision Tree	NB	Logistic
Precision	0.40	0.54	0.57	0.56
F1 Score	0.49	0.55	0.57	0.57

Table 7: Active learning using different underlying classifiers

comes easily inferred from the given features are less likely to contain noise, which may stem from procedural irregularities, or other external influences unique to the instance. Consequently, they are more likely to be used as precedents for future decisions.

Both boundary logistic and ensemble disagree models removed about 25% of training instances, with approximately half of these instances overlapping. Both models show significant improvement in addressing overfitting compared to their original versions (Figure 9 and Figure 10). By eliminating ambiguous cases, the models avoid fitting to noisy data and enhance generalization.

While the boundary logistic model's performance on the development set remains unchanged, the disagree ensemble model exhibits substantial improvement, particularly in precision. The boundary model's underperformance may stem from the costs of learning case nuances, suggesting that some removed instances are not noisy. Conversely, the disagree ensemble model becomes more conservative yet accurate in classifying positive instances by leveraging its enhanced ability to identify patterns.

Overall, the active learning model using 25% of the training dataset, alongside the boundary logistic and ensemble models using 75% (Table 8), shows slightly improved performance compared to traditional logistic and ensemble models trained on the full dataset. This underscores that the quality and relevance of training data, rather than mere quantity, significantly affect model performance. However, in an environment of class imbalance in both the training and development dataset, the focus on clear-cut cases may lead to overperformance, which might not be reflected in another development set with a different class distribution.

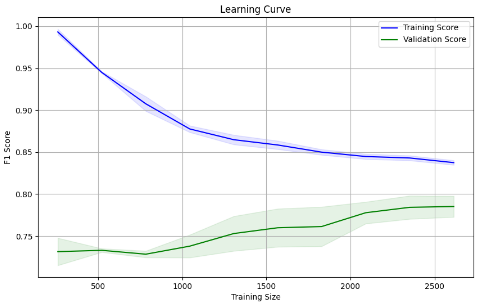

Figure 9: Boundary logistic model

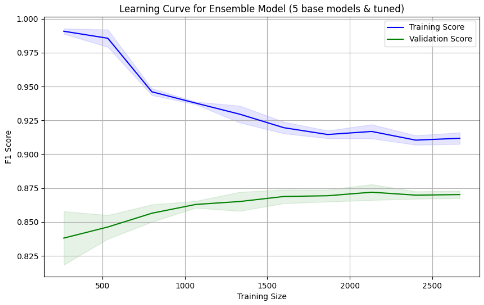

Figure 10: Ensemble disagree model

Dataset	Number of Instances
Training Dataset	4612
Active Learning	1200
Boundary Logistic	3268
Ensemble Disagree	3334

Table 8: Number of instances used in fitting the model

6 Conclusion

This study explored the effectiveness of various machine learning models in predicting Supreme Court decisions, focusing on data efficiency and the impact of temporal information. Our findings demonstrate that ensemble models generally outperform simpler logistic regression models, particularly when post-decision data is incorporated. The study also highlights the potential of data-efficient approaches, with models trained on a subset of carefully selected data performing comparably to those trained on the full dataset but with better generalisability.

Future research could explore a more mathematically grounded approach to identifying and eliminating noisy instances, as compared to the intuitive approach used in this study. An example of such is the research conducted by Mirzasoleiman et al. (2020) to identify coreset that facilitates faster convergence in logistic regression and neural network, as well as improving performance.

References

Bonica, A. and Sen, M. (2021). Estimating judicial ideology. *Journal of Economic Perspectives*, 35(1):97–118.

Cohen, A. and Yang, C. S. (2019). Judicial politics and sentencing decisions. *American Economic Journal: Economic Policy*, 11(1):160–91.

Fang, B., Cohn, T., Baldwin, T., and Frermann, L. (2023a). More than votes? voting and language based partisanship in the us supreme court. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4604–4614. Association for Computational Linguistics.

Fang, B., Cohn, T., Baldwin, T., and Frermann, L. (2023b). Super-SCOTUS: A multi-sourced dataset for the Supreme Court of the US. In Preoțiuc-Pietro, D., Goanta, C., Chalkidis, I., Barrett, L., Spanakis, G., and Aletras, N., editors, *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 202–214, Singapore. Association for Computational Linguistics.

Hayes, T., Baraldi, A. N., and Coxe, S. (2024). Random forest analysis and lasso regression outperform traditional methods in identifying missing data auxiliary variables when the mar mechanism is nonlinear (p.s. stop using little's mcar test). *Behavior Research Methods*, null(null):1–32.

Joshi, S., Jain, A., Payani, A., and Mirzasoleiman, B. (2024). Data-efficient contrastive language-image pretraining: Prioritizing data quality over quantity. *International Conference on Artificial Intelligence and Statistics (AISTATS).*

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. (2020). Coresets for data-efficient training of machine learning models. *International Conference on Machine Learning (ICML)*, pages 6950–6960.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rittler, N. and Chaudhuri, K. (2023). A two-stage active learning algorithm for k-nearest neighbors.

Satoła, A.; Satoła, K. (2024). Performance comparison of machine learning models used for predicting subclinical mastitis in dairy cows: Bagging, boosting, stacking, and super-learner ensembles versus single machine learning models. *Journal of Dairy Science*, 107(6):3959–3972.