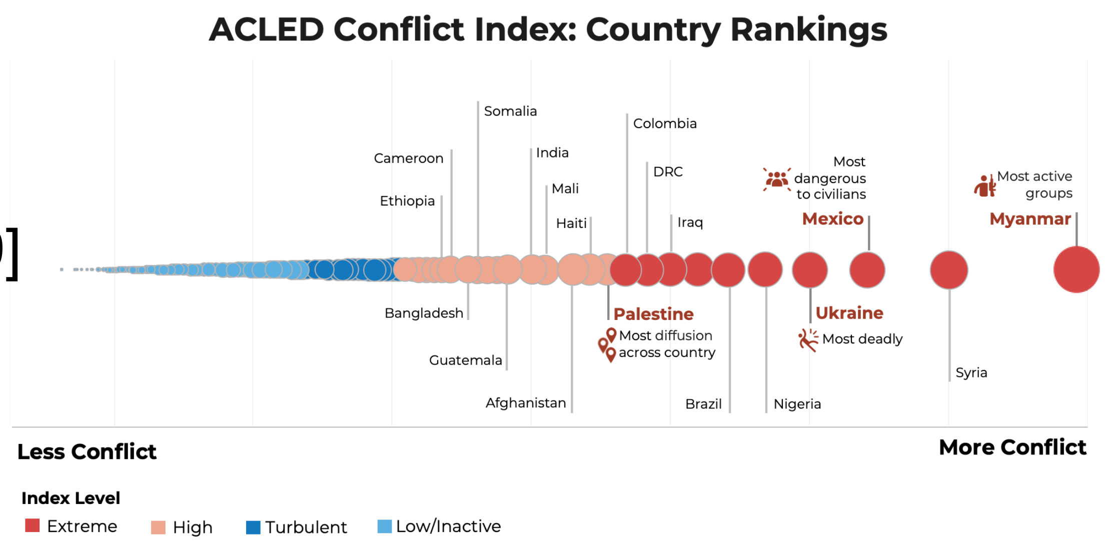
On A Machine Learning Framework for Studying Imbalanced Spatio-Temporal Data

Snigdhansu (Ansu) Chatterjee,
Sinha E-nnovate Endowed Chair Professor,
Department of Mathematics and Statistics,
University of Maryland, Baltimore County
snigchat@umbc.edu

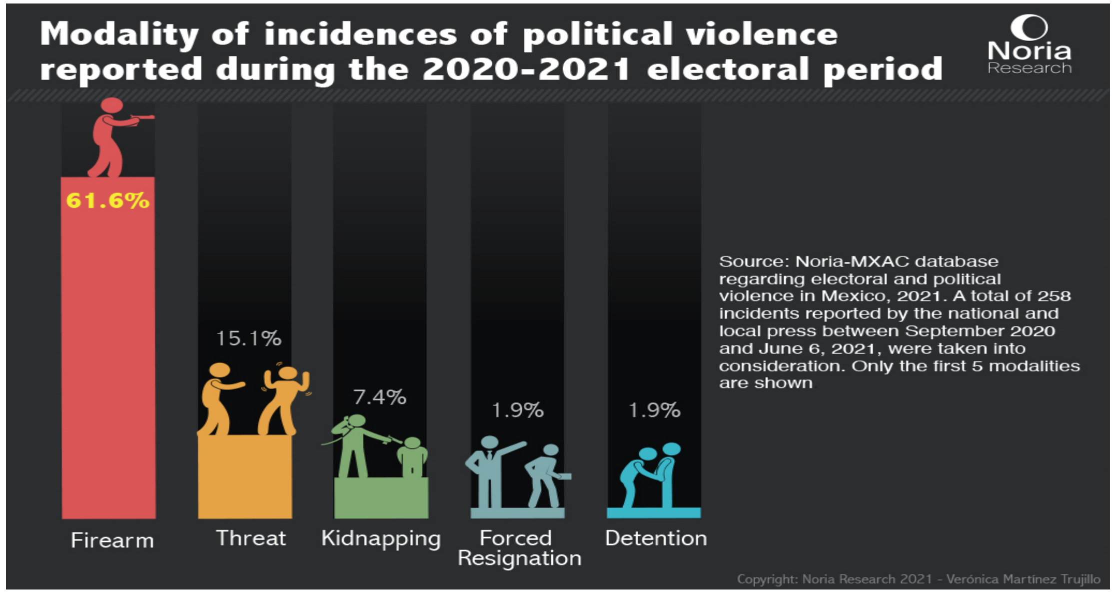
This presentation is primarily based on the MS Thesis work of V. Subedi at the University of Minnesota.

Introduction

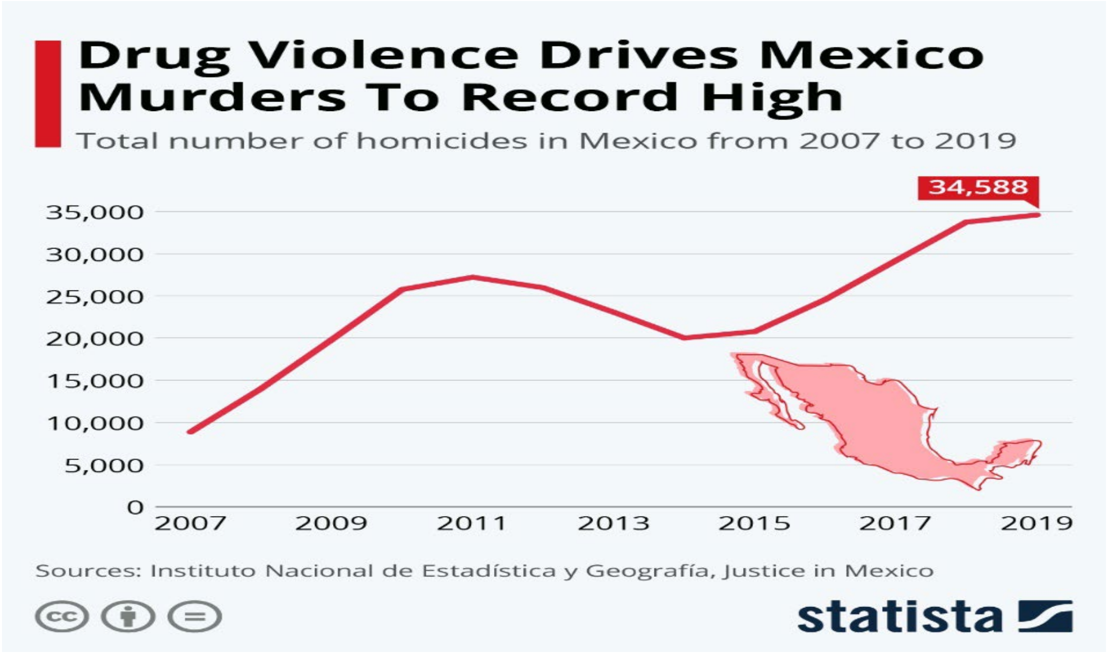
[21, 39]



[40]



[41]



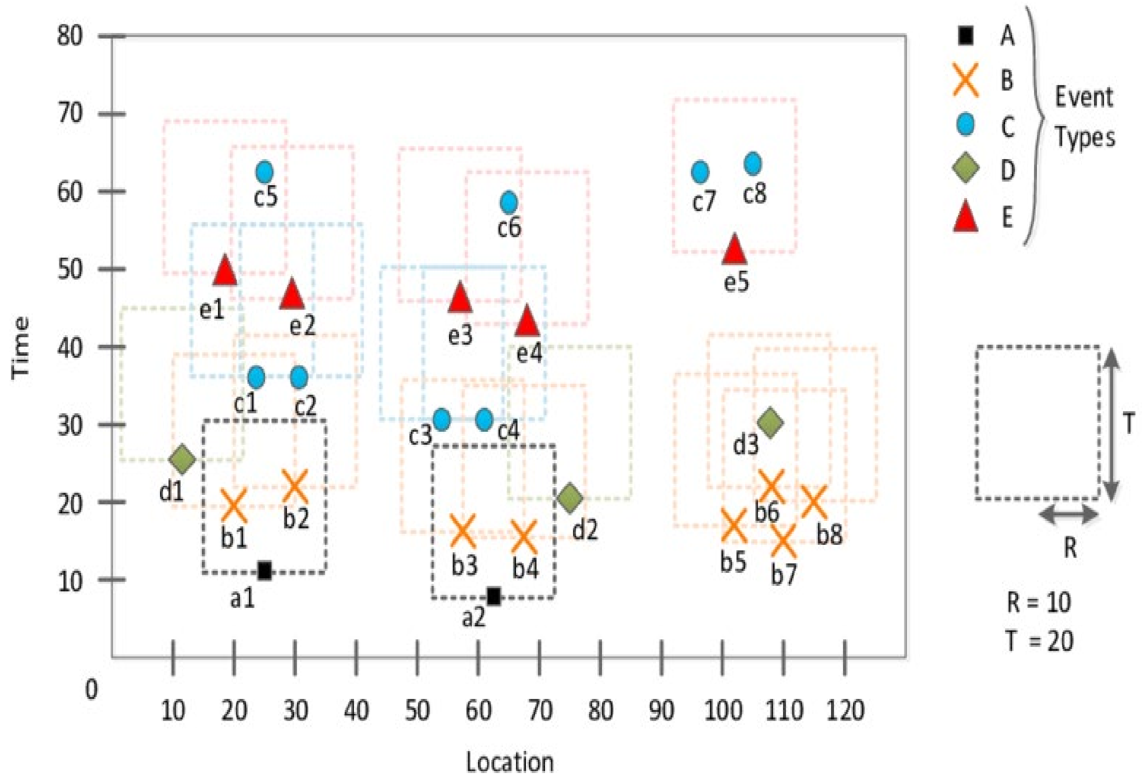
[1, 42]

Motivation

- Spatiotemporal data => Samples dependent spatially and temporally
- Sparse Data
- High dimensional feature space
- Imbalanced class distribution

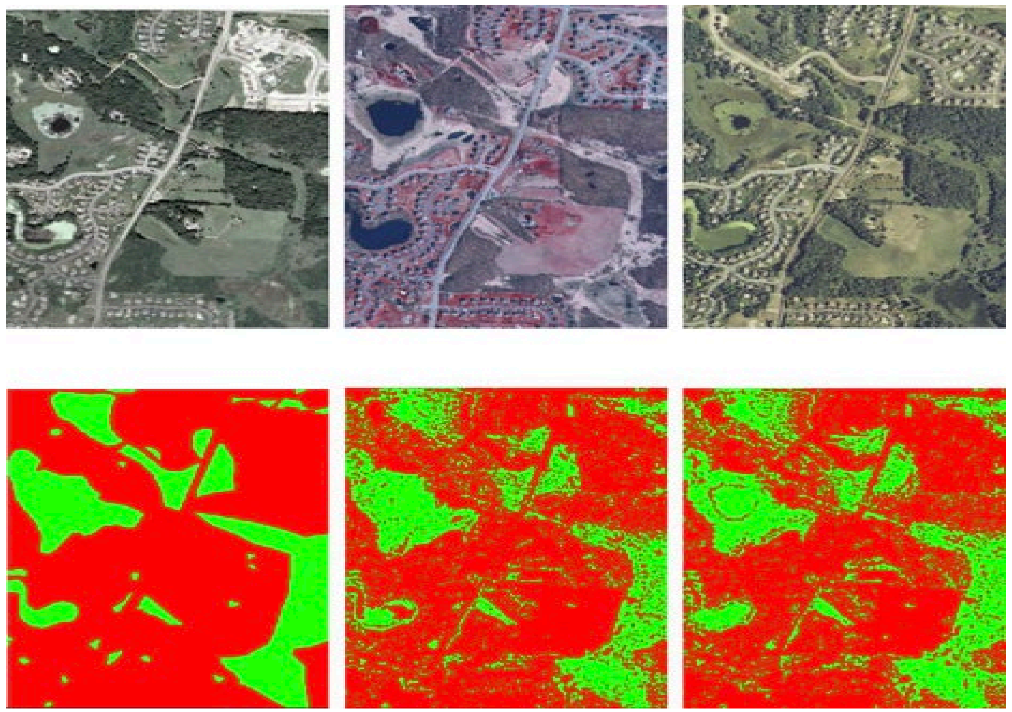
Challenges

Neighboring samples correlated spatially



[43]

Classical machine learning algorithms fail!

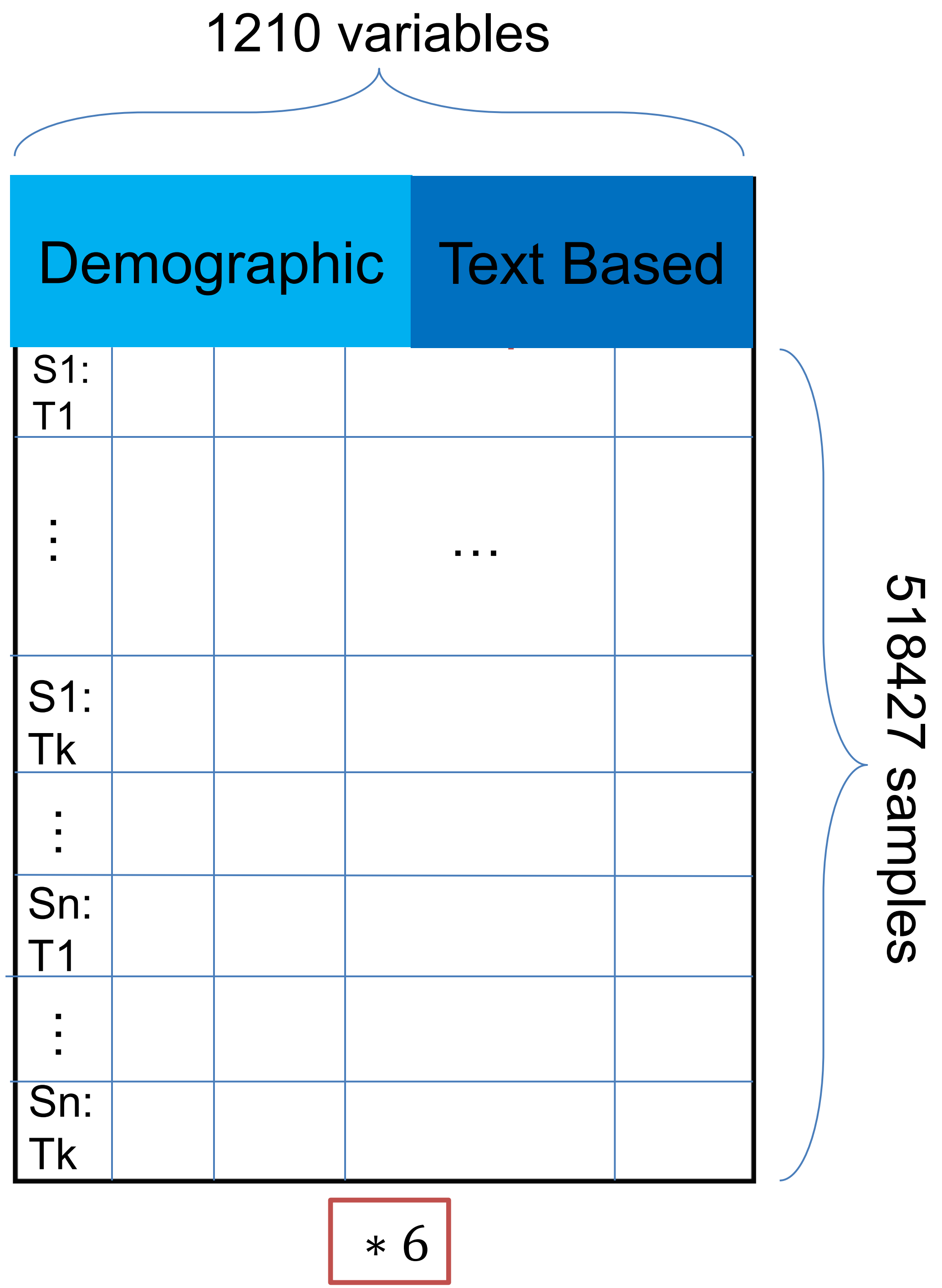


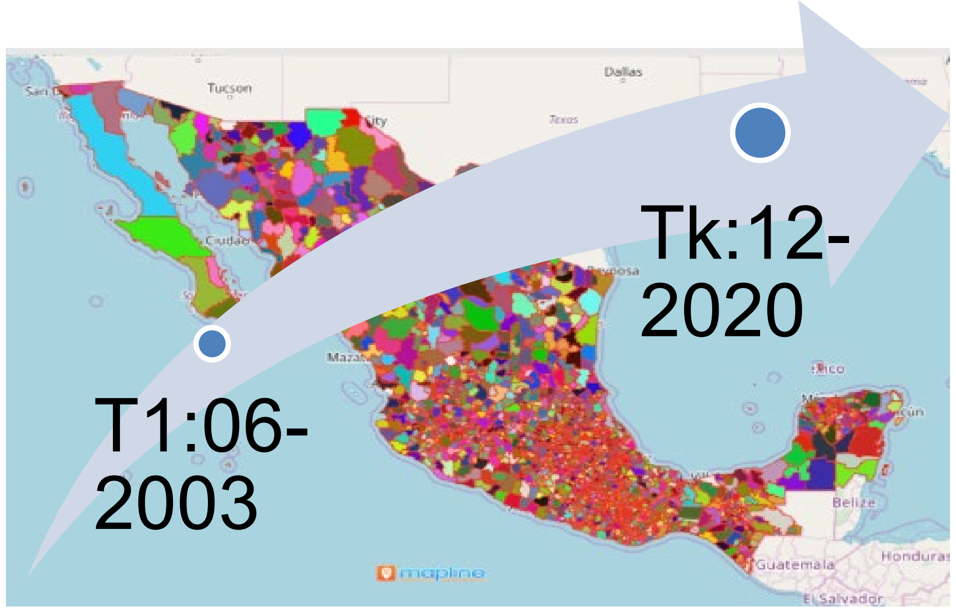
[2]

Objectives

- Developing a generalized methodology to model imbalanced event type spatiotemporal data using a subset of high dimensional feature space.
- Analyzing spatiotemporal patterns in political conflicts.
- Find the set of predictors that are important in classifying the labels (the predictors of political conflicts).

Dataset

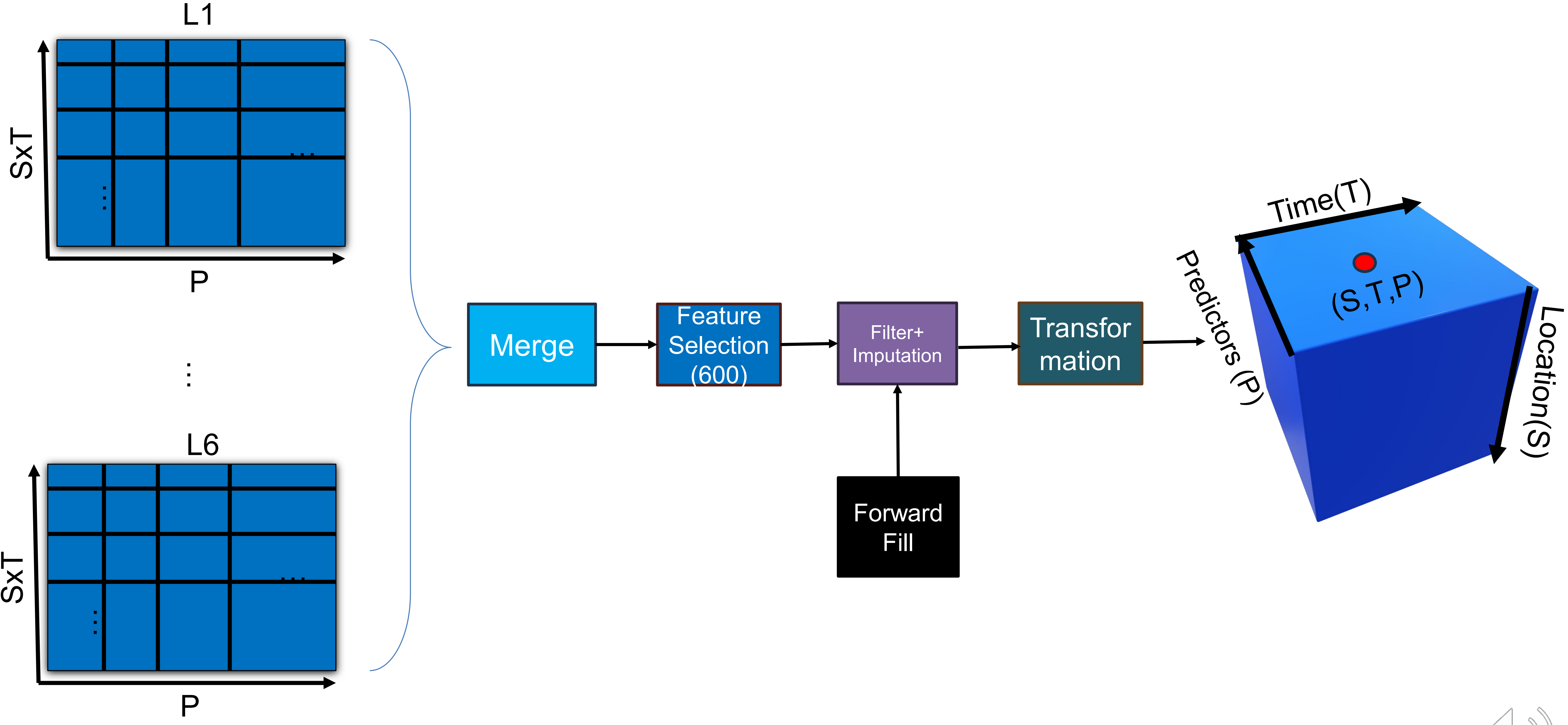




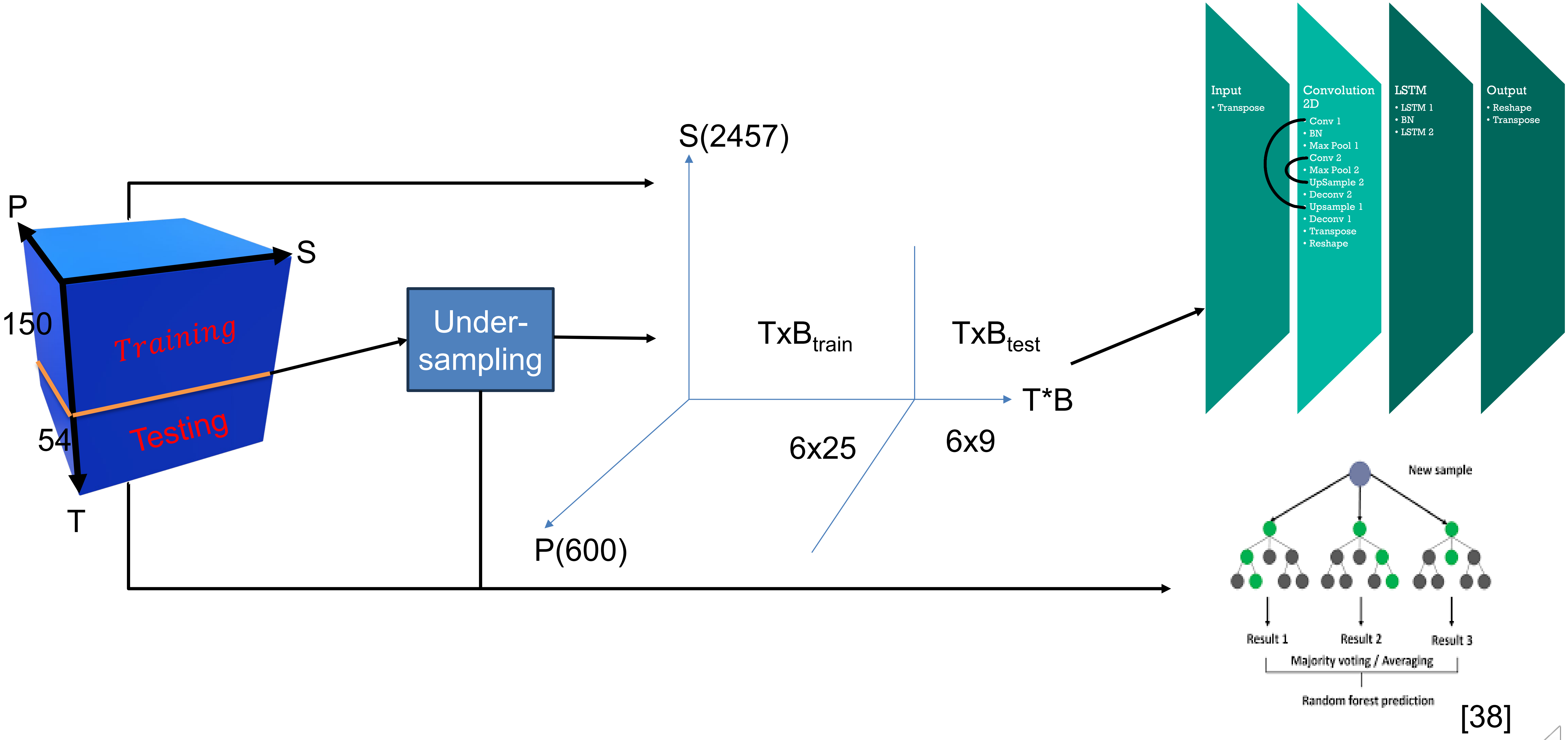
Homicide, Accident, Suicide, Population by gender, Material conflicts, Verbal conflicts

Document data: Counts of occurrences of unique violent/abusive words between citizens

Phase I: Pre-Processing



Phase I: Training



Phase I: Results

Original Data

Class	Precision	Recall	F1
0	1.00	1.00	1.00
1	0.53	0.09	0.15

Random Forest

Class	Precision	Recall	F1
0	1.00	0.78	0.88
1	0.00	0.31	0.01

CNN2D+LSTM

Under-sampled Data

Class	Precision	Recall	F1
0	0.99	1.00	1.00
1	0.63	0.16	0.25

Random Forest

Class	Precision	Recall	F1
0	0.99	0.64	0.77
1	0.01	0.38	0.02

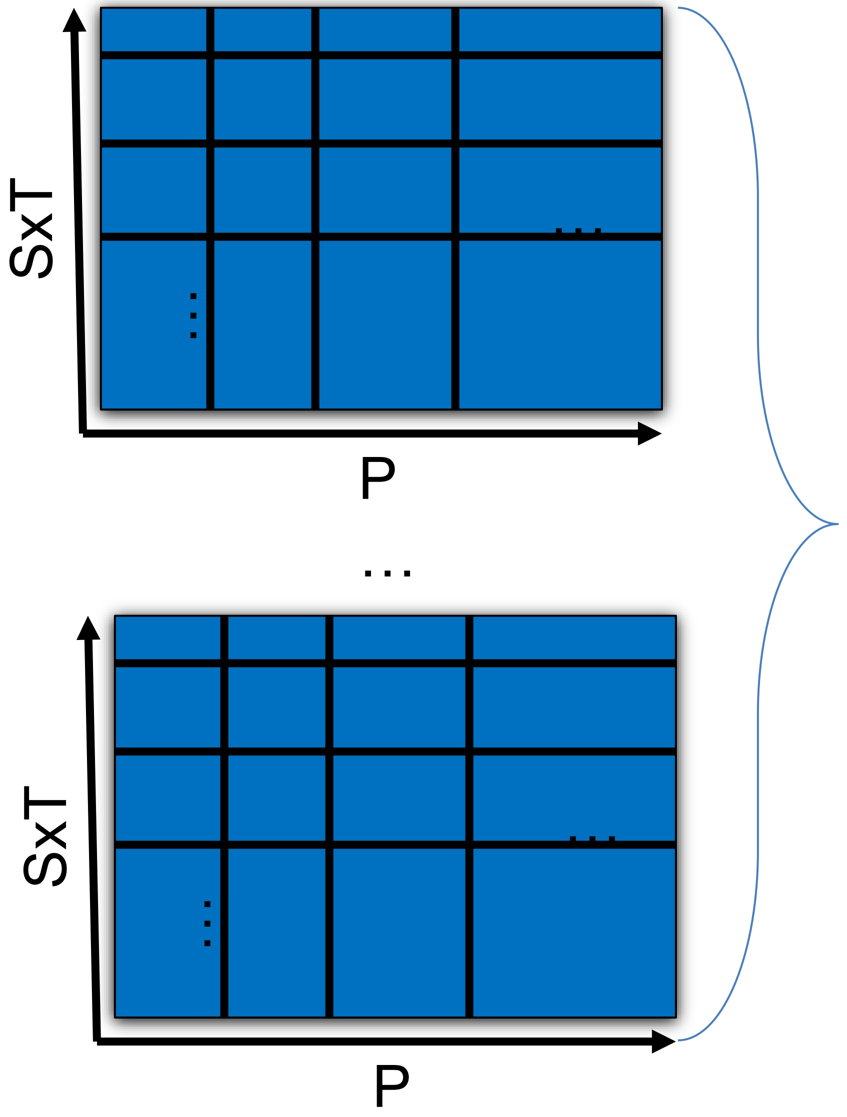
CNN2D+LSTM

Concatenation

Class	Precision	Recall	F1
0	0.99	0.90	0.94
1	0.03	0.34	0.05

CNN2D+LSTM

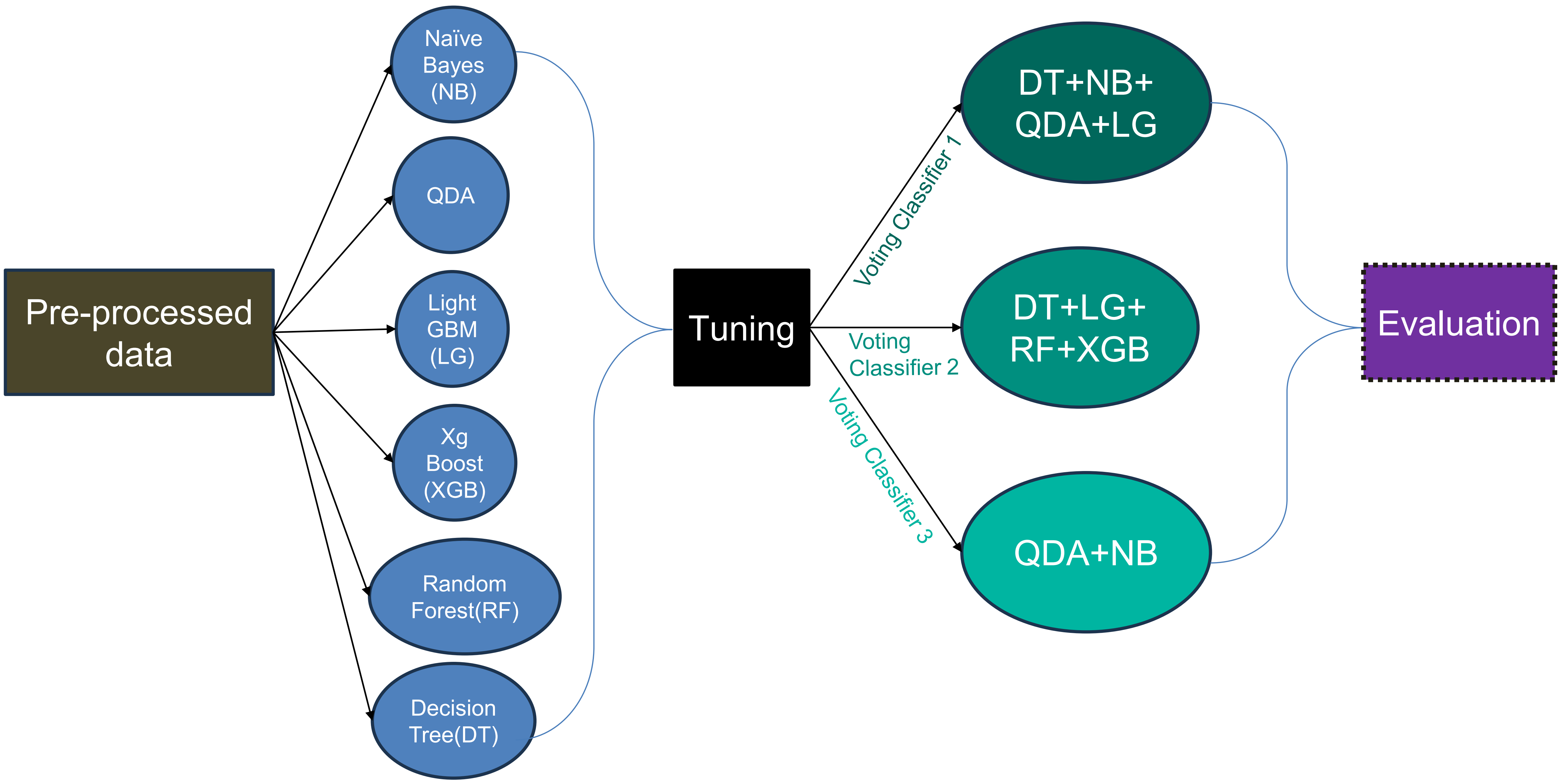
Phase II: Motivation



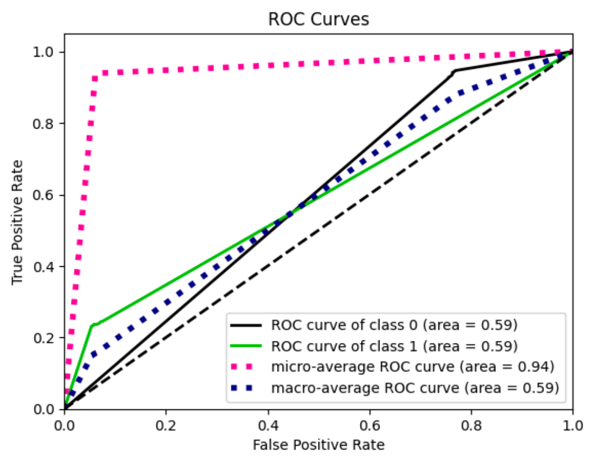
features = 600 → High dimensionality

Worth adding all the lags? → Analyze temporal signals

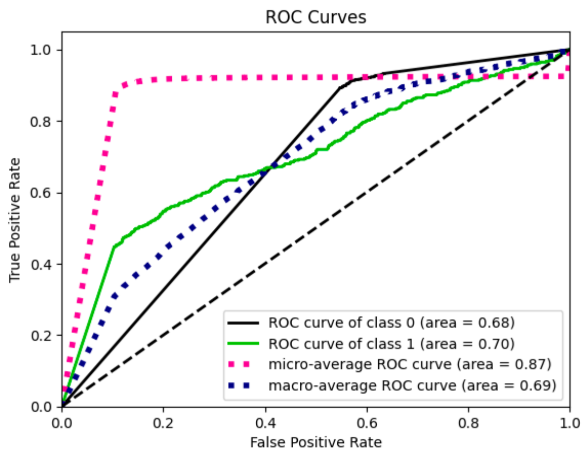
Phase II: Training



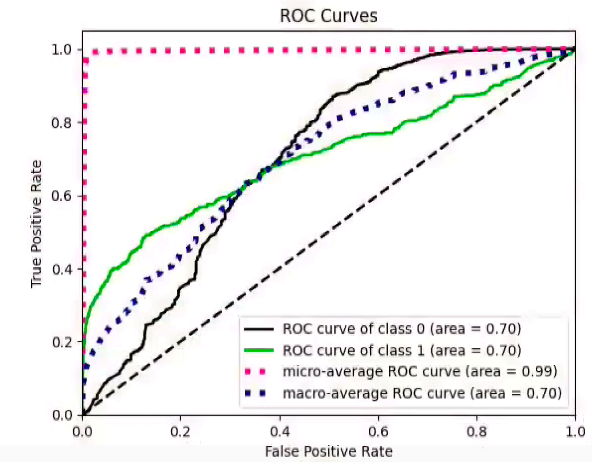
Phase II: Results



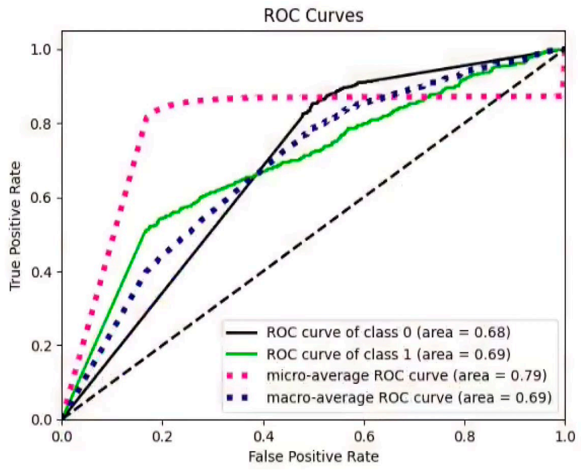
Decision tree



Naïve Bayes



Light GBM



QDA

Voting Classifier 1

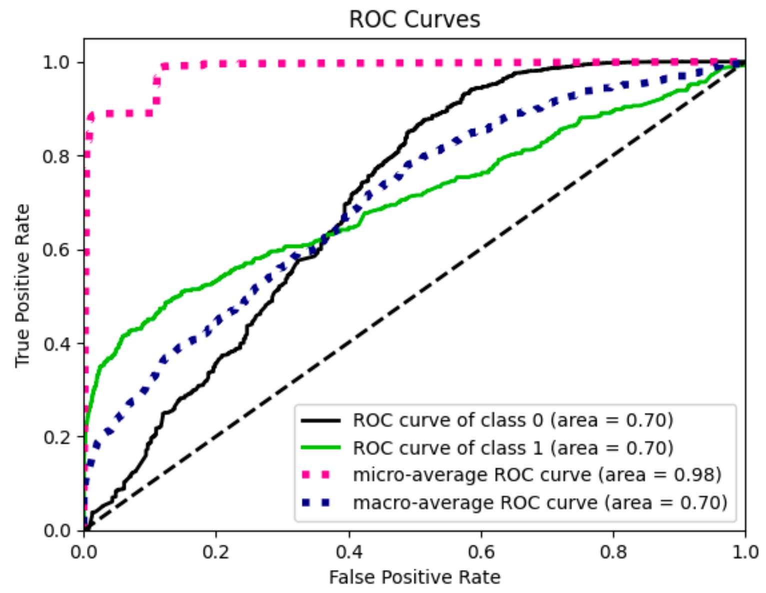
Class	Precision	Recall	F1
0	0.99	1	0.99
1	0.34	0.21	0.26

Voting Classifier 2

Class	Precision	Recall	F1
0	0.99	1	0.99
1	0.38	0.19	0.25

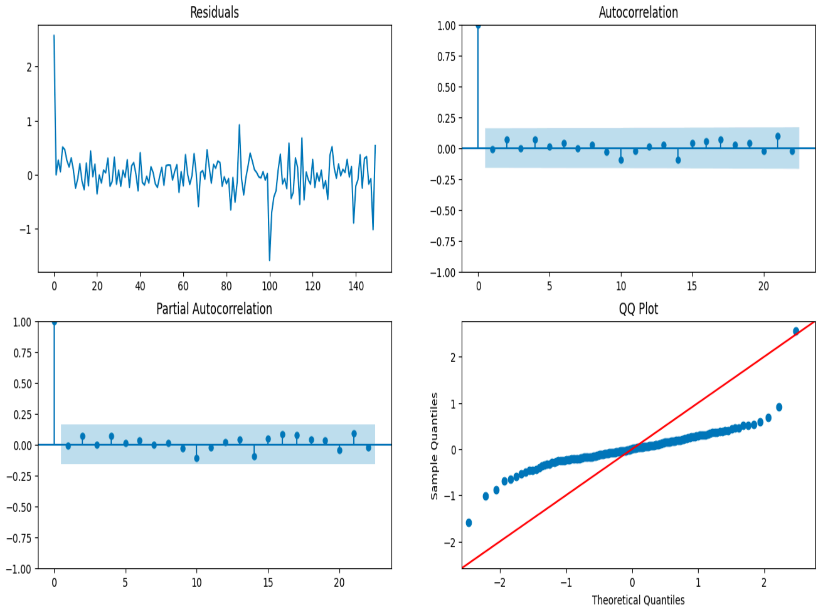
Voting Classifier 3

Class	Precision	Recall	F1
0	0.99	0.89	0.94
1	0.04	0.45	0.07



VC 1 Soft Voting

Phase III: Motivation



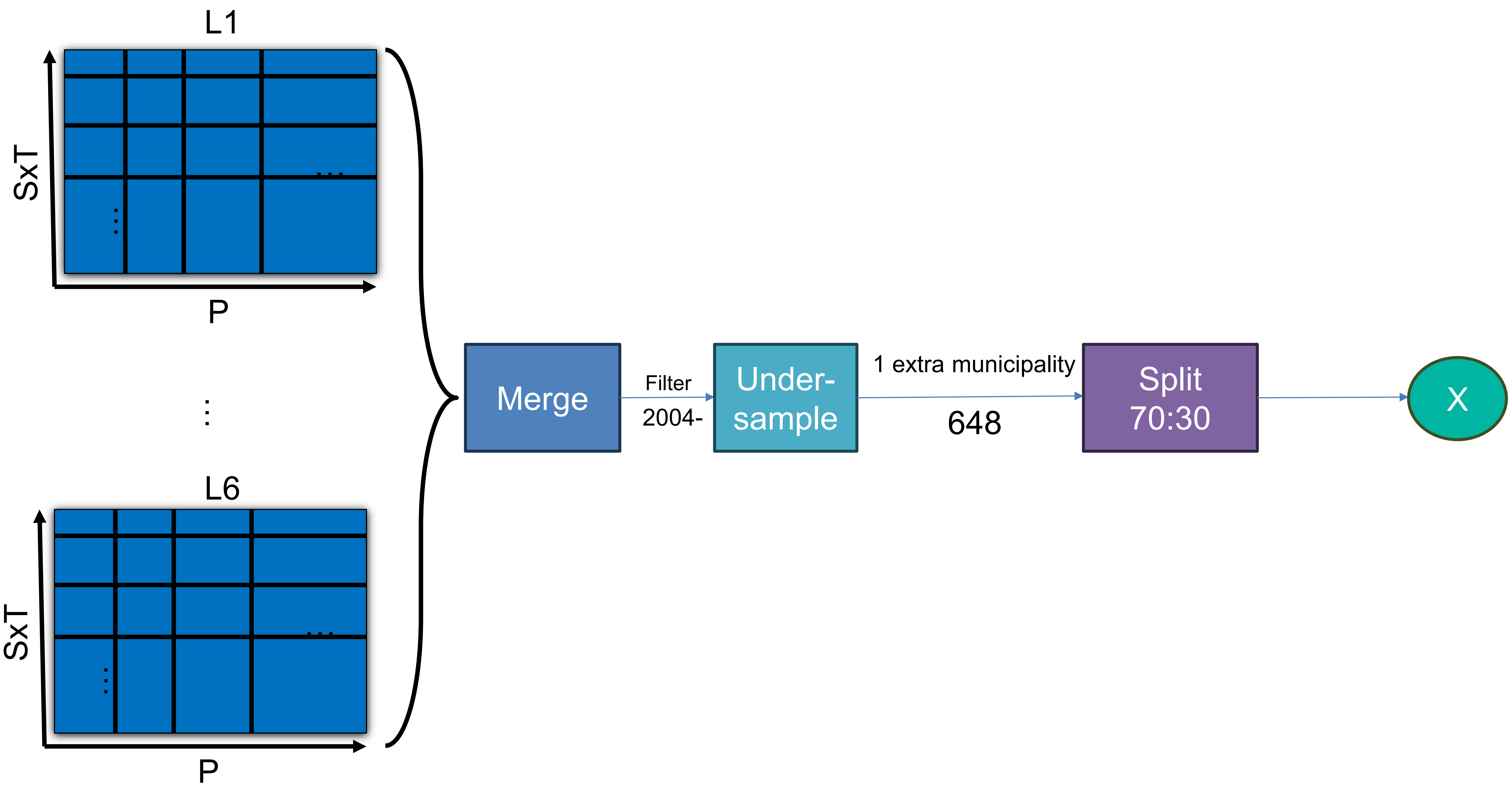
- Model fit is not very reliable.
- Also need to focus on the important predictors

Voting Classifier 1:
DT+NB+QDA+LG

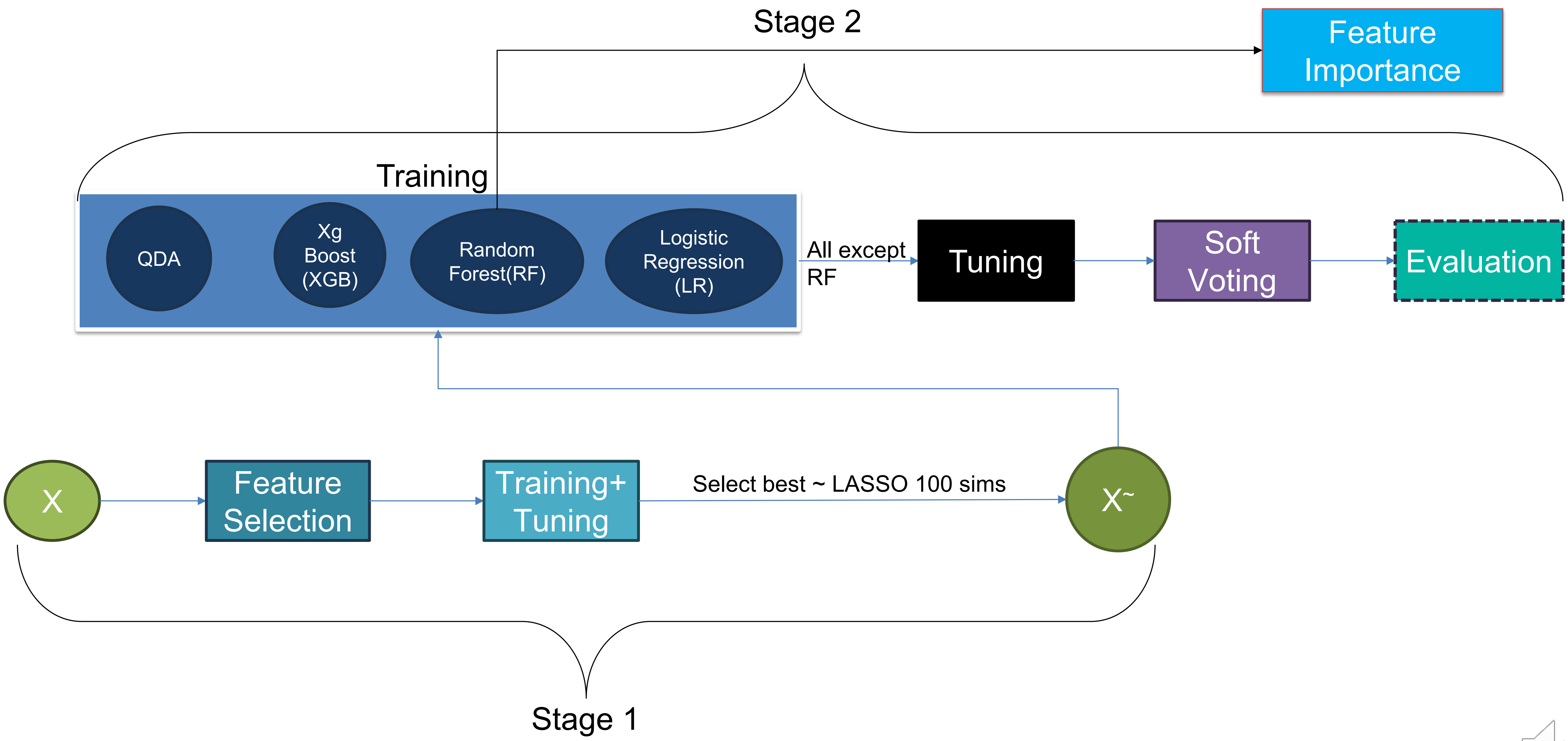
- Large feature space ~ 600 features
- Model is too complex
- Low Bias High Variance (Overfitting)

Feature Selection

Phase III: Pre-Processing



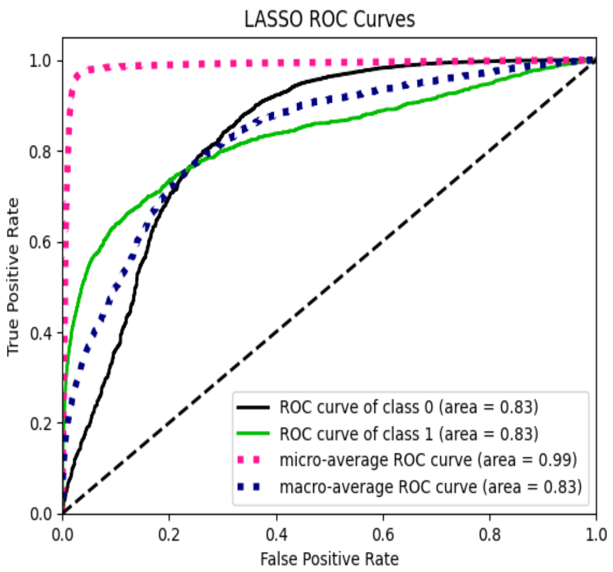
Phase III: Training



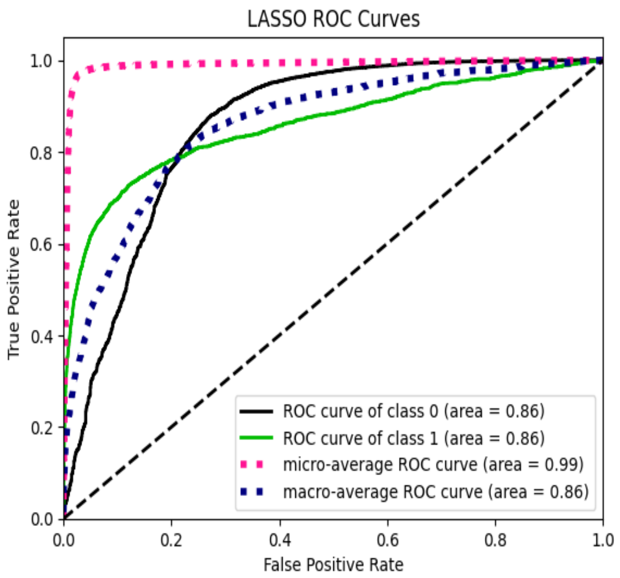
Phase III: Results

Method(fea--tures selected)	Class	Precision	Recall	F1
Mutual Information(50)	1	0.61	0.17	0.27
Forward Stepwise(77)	1	0.37	0.18	0.24
RFE(3616)	1	0.79	0.13	0.23
Forward Λ RFE(49)	1	0.36	0.18	0.24
Forward *U* Mutual Information(117)	1	0.67	0.17	0.27
LASSO(479)	1	0.24	0.66	0.35
PCC(2213)	1	0.17	0.50	0.26
LASSO 100 sims(82)	1	0.71	0.22	0.34

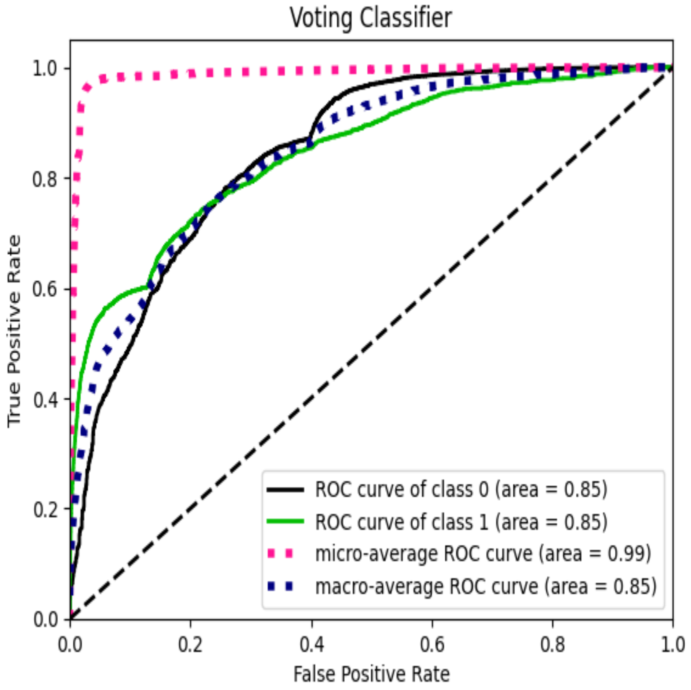
Tuned LASSO



Tuned LASSO 100 sims



Model	Class	Precision	Recall	F1
Logistic Regression	1	0.68	0.27	0.39
Random Forest	1	0.68	0.20	0.31
Xg Boost	1	0.64	0.25	0.36
QDA	1	0.15	0.60	0.23
Voting Classifier	1 0	0.97 0.60	0.99 0.31	0.98 0.41



Model Comparison

Model	Precision(1)	Recall(1)	F1(1)	Training Time (sec)	Evaluation Time (sec)	Support distribution
CNN + LSTM (Concatenation with Undersampling)	0.01	**0.65**	0.02	105	1	42375(0) vs 393(1)
CNN + LSTM (2 Independent Datasets with Undersampling)	0.03	0.36	0.05	50	1	42375(0) vs 393(1)
Logistic Regression	**0.67**	0.27	0.38	44.68	1.43	38181(0) vs 1416(1)
Decision Tree	0.27	0.30	0.29	557.99	1.67	38181(0) vs 1416(1
Random Forest (Unweighted)	0.82	0.17	0.28	497.22	4.97	38181(0) vs 1416(1
KNN	0.55	0.11	0.18	17.22	**420.84**	38181(0) vs 1416(1
Naive Bayes	0.15	0.34	0.20	14.34	5.15	38181(0) vs 1416(1
Gradient Boosting	0.66	0.27	0.39	**1120.34**	2.69	38181(0) vs 1416(1)
Voting Classifier with LASSO 100 sims (Soft)	0.60	0.31	0.41	34.24	1.39	38181(0) vs 1416(1)
Voting Classifier with LASSO complete (Soft)	0.63	0.33	**0.43**	111.96	2.90	38181(0) vs 1416(1)

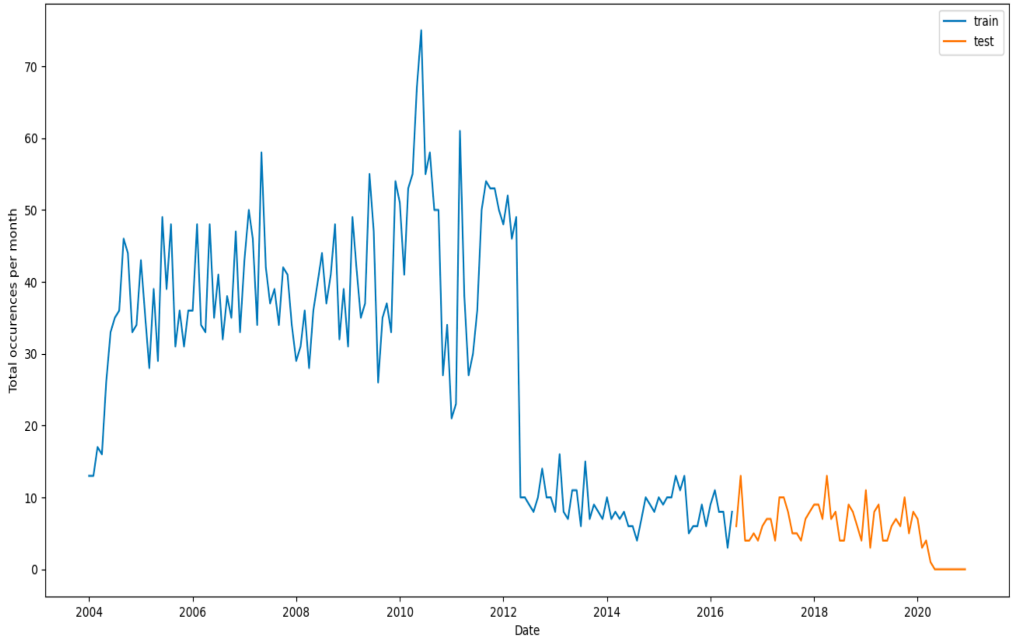
- Model Complexity
- Feature Engineering
- Scalability
- Generalization

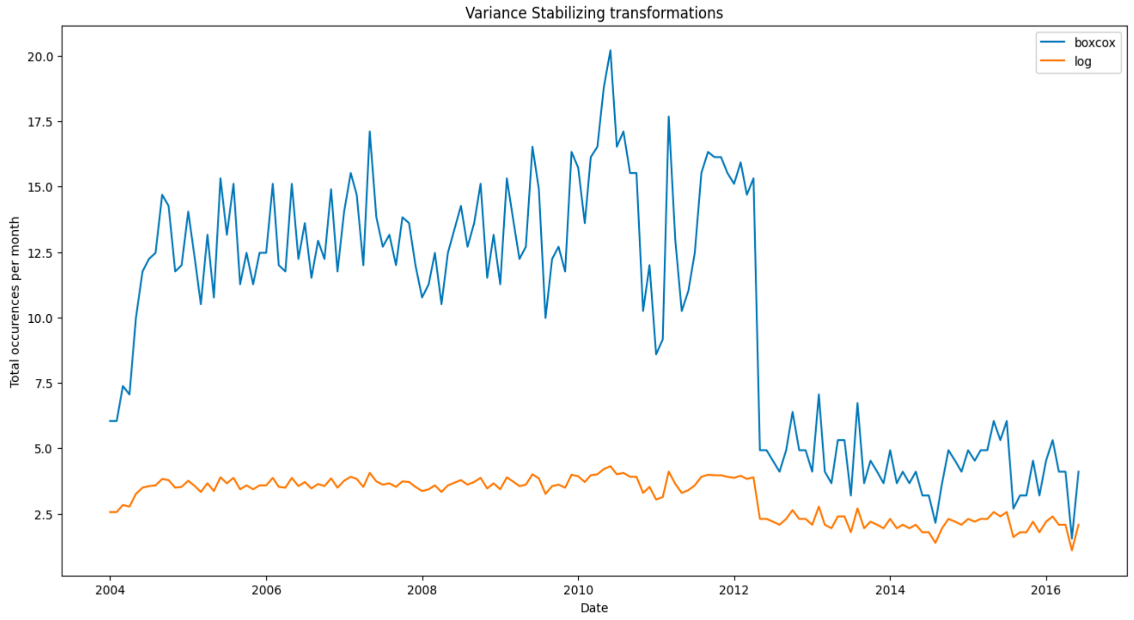
The classical methods are trained on whole data with 647 municipalities with 7232 predictors having both the majority and minority class. The neural networks and the voting classifier with LASSO 100 are trained on 82 predictors from the union of predictors from 100 independent LASSO simulations. The voting classifier with LASSO complete is trained using 479 predictors.

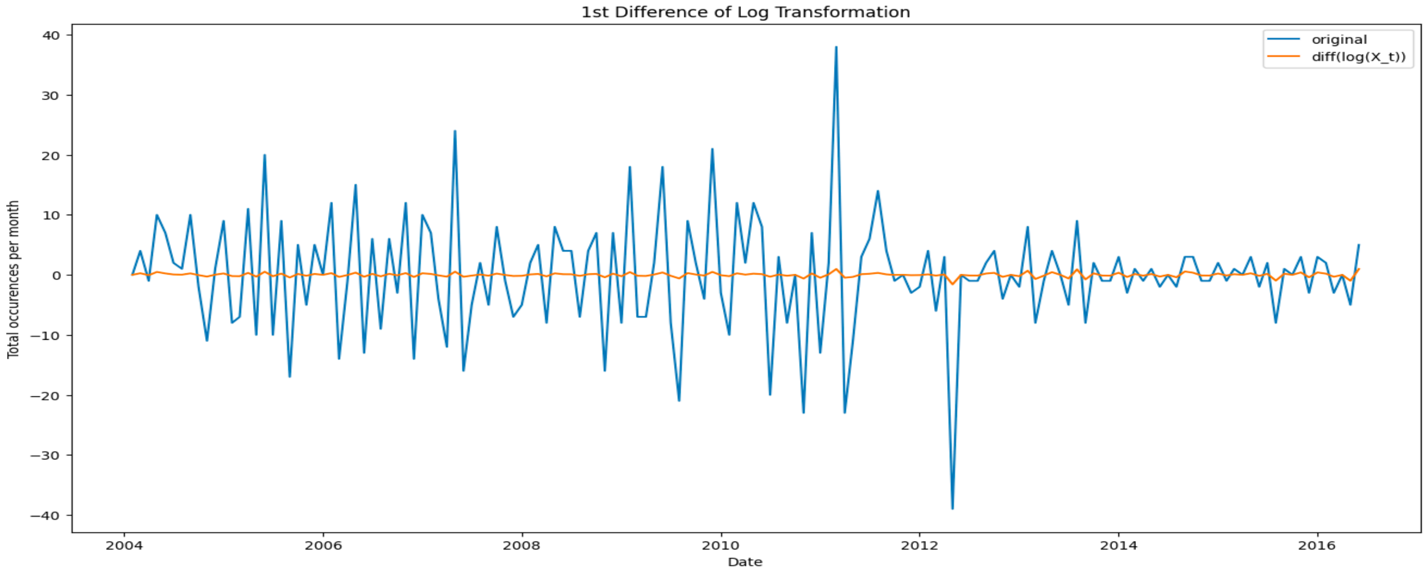
Conclusion

- Classical methods outperform deep neural network models for tabular dataset.
- Univariate time series analysis revealed previous 1 lag dependency using which voting classifier was trained which performed the best.
- Feature selection using LASSO 100 sims selected just 82 predictors and gave a F1 score of 41%.
- Year, month, accident, material conflicts and presence of positive sentiments turn out as the important predictors that drive the political conflicts

Time series of violence







This research is partially supported by the US National Science Foundation (NSF) under grants # DMS-1737918, # OAC-1939916, # DMR-1939956 and # DMS 2436549, and a grant from Cisco Inc.

Thank you

References

1. Muchlinski, D., Yang, X., Birch, S., Macdonald, C., & Ounis, I. (2021). We need to go deeper: Measuring electoral violence using convolutional neural networks and social media. Political Science Research and Methods, 9(1), 122-139. doi:10.1017/psrm.2020.3
2. Zhe Jiang, Shashi Shekhar, Xun Zhou, Joseph Knight, and Jennifer Corcoran. Focal-test based spatial decision tree learning. IEEE Transactions on Knowledge and Data Engineering, 27(6):1547–1559, 2014.
3. Peter Hart. The condensed nearest neighbor rule (corresp.). IEEE transactions on information theory, 14(3):515–516, 1968.
4. Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, (3):408–421, 1972.
5. Ivan Tomek. A generalization of the k-nn rule. IEEE Transactions on Systems, Man, and Cybernetics, (2):121–126, 1976.
6. Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets, volume 126, pages 1–7. ICML, 2003
7. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.
8. Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pages 1322–1328. Ieee, 2008.
9. Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1):20–29, 2004.
10. Harold Hotelling. Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6):417, 1933.

References

11. Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 2(11):559–572, 1901
12. Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3):37–52, 1987.
13. Benjamin Fruchter. Introduction to factor analysis. 1954.
14. Peter J Huber. Projection pursuit. The annals of Statistics, pages 435–475, 1985.
15. Kenji Kira and Larry A Rendell. A practical approach to feature selection. In Machine learning proceedings 1992, pages 249–256. Elsevier, 1992.
16. Laura Elena Raileanu and Kilian Stoffel. Theoretical comparison between the gini index and information gain criteria. Annals of Mathematics and Artificial Intelligence, 41:77–93, 2004.
17. Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. Machine learning, 46:389–422, 2002.
18. Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1):267–288, 1996.
19. Susan D Hyde and Nikolay Marinov. Which elections can be lost? Political analysis, 20(2):191–210, 2012.
20. Paul R Brass. Theft of an idol: Text and context in the representation of collective violence, volume 8.
21. Clionadh Raleigh, rew Linke, H̊avard Hegre, and Joakim Karlsen. Introducing acled: An armed conflict location and event dataset. Journal of peace research, 47(5):651–660, 2010.
22. Elizabeth Boschee, Jennifer Lautenschlager, Sean O’Brien, Steve Shellman, James Starz, and Michael Ward. Icews coded event data. Harvard Dataverse, 12, 2015.

References

23. Cullen S Hendrix and Idean Salehyan. No news is good news: Mark and recapture for event data when reporting probabilities are less than one. International Interactions, 41(2):392–406, 2015.
24. Ursula E Daxecker. The cost of exposing cheating: International election monitoring, fraud, and post-election violence in africa. Journal of Peace Research, 49(4):503–516, 2012.
25. Nils B Weidmann. A closer look at reporting bias in conflict event data. American Journal of Political Science, 60(1):206–218, 2016.
26. Daniel Archambault, Fabio Celli, Elizabeth M Daly, Ingrid Erickson, Werner Geyer, Germaine Halegoua, Brian Keegan, David R Millen, Raz Schwartz, and N Sadat Shami. Reports on the 2013 workshop program of the seventh international aaai conference on weblogs and social media. AI Magazine, 34(4):116–118, 2013.
27. Idean Salehyan, Cullen S Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. Social conflict in africa: A new database. International Interactions, 38(4):503–511, 2012.
28. Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping ak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836, 2016.
29. Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In Proceedings of the 2nd International Conference on Neural Information Processing Systems, NIPS'89, page 396–404, Cambridge, MA, USA, 1989. MIT Press.
30. Sepp Hochreiter and JÅNurgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.
31. https://d2l.ai/chapter_appendix-mathematics-for-deep-learning/information-theory.html
32. https://en.wikipedia.org/wiki/Partial_correlation.html
33. https://www.analyticsvidhya.com/blog/2023/05/recursive-feature-elimination.html
34. https://quantifyinghealth.com/stepwise-selection.html

References

35. https://ilyasbinsalih.medium.com/what-is-hard-and-soft-voting-in-machine-learning-2652676b6a32.html
36. https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn.html
37. https://goldinlocks.github.io/Multivariate-time-series-models/
38. https://medium.com/@roiyeho/random-forests-98892261dc49
39. https://acleddata.com/acled-conflict-index-mid-year-update/
40. https://noria-research.com/data-on-electoral-violence-mexico-2020-2021/
41. https://www.statista.com/chart/12635/drug-violence-drives-mexico-murders-to-record-high/
42. https://pianalytix.com/violence-detection-using-ml/
43. Maciąg, Piotr & Kryszkiewicz, Marzena & Robert, Bembenik. (2019). Discovery of closed spatio-temporal sequential patterns from event data. Procedia Computer Science. 159. 707-716. 10.1016/j.procs.2019.09.226.
44. https://stats.stackexchange.com/questions/351638/random-sampling-methods-for-handling-class-imbalance