Conflict and Independence: A Machine-Learning, Time-Series Analytic Approach to Forecasting Ethnopolitical Conflict in Postcolonial Sub-Saharan Africa

MDSC Capstone by Claire O'Shaughnessy

Advisor: Peter Mucha

Abstract

Since European decolonization in the mid-20th century, postcolonial Sub-Saharan nations have been plagued by recurrent periods of intrastate violence. In fact, more than two-thirds of the nations in Sub-Saharan Africa[2]—ninety percent of which were colonized by European nations before 1886[3]—have experienced civil conflict since 1960, resulting in millions of deaths and immense political instability. Using a machine learning approach, this paper leverages Paine's 2019 dataset[6] to 1) compare the performance of tree-based machine learning models against logistic regression in accurately classifying instances of conflict, and 2) simulate conflict predictions. The results of my first inquiry demonstrate the superior performance of decision tree and random forest models in classifying conflict using a variety of historical and present-day ethnic group characteristics. The results of my second inquiry suggest that historical data, including static characteristics such as an ethnic group's precolonial power structure, neolithic timing, and slave exports, can be used to improve present-day conflict forecasting. Further, the random forest model's relatively even distribution of feature importance values suggests that this model is the most practical for forecasting conflict, particularly when some feature values cannot be accurately extrapolated. However, these findings also highlight the need for improved methods for forecasting dynamic, high-importance feature values that follow more complex underlying patterns, as well as the importance of continuing to collect data that will make predictive civil war models more robust to error.

Data

This report leverages the dataset from Paine's 2019 *Ethnic Violence in Africa* paper. The group classifications in the dataset were compiled from the Ethnic Power Relations database (EPR), which provides data on politically relevant ethnic groups and their access to central power.[6]

Key characteristics

- Uses ethnic group-years (years in which an ethnic group existed) as its unit of analysis
- Contains 8,567 entries associated with 204 groups from 37 Sub-Saharan African countries
- Includes majority static and historical variables (e.g. approximate date of a group's Neolithic transition, local geography, slave exports)
- Classifies groups by their degree of organization in the precolonial era, finding these categories to be statistically significant predictors of a group's involvement in civil war

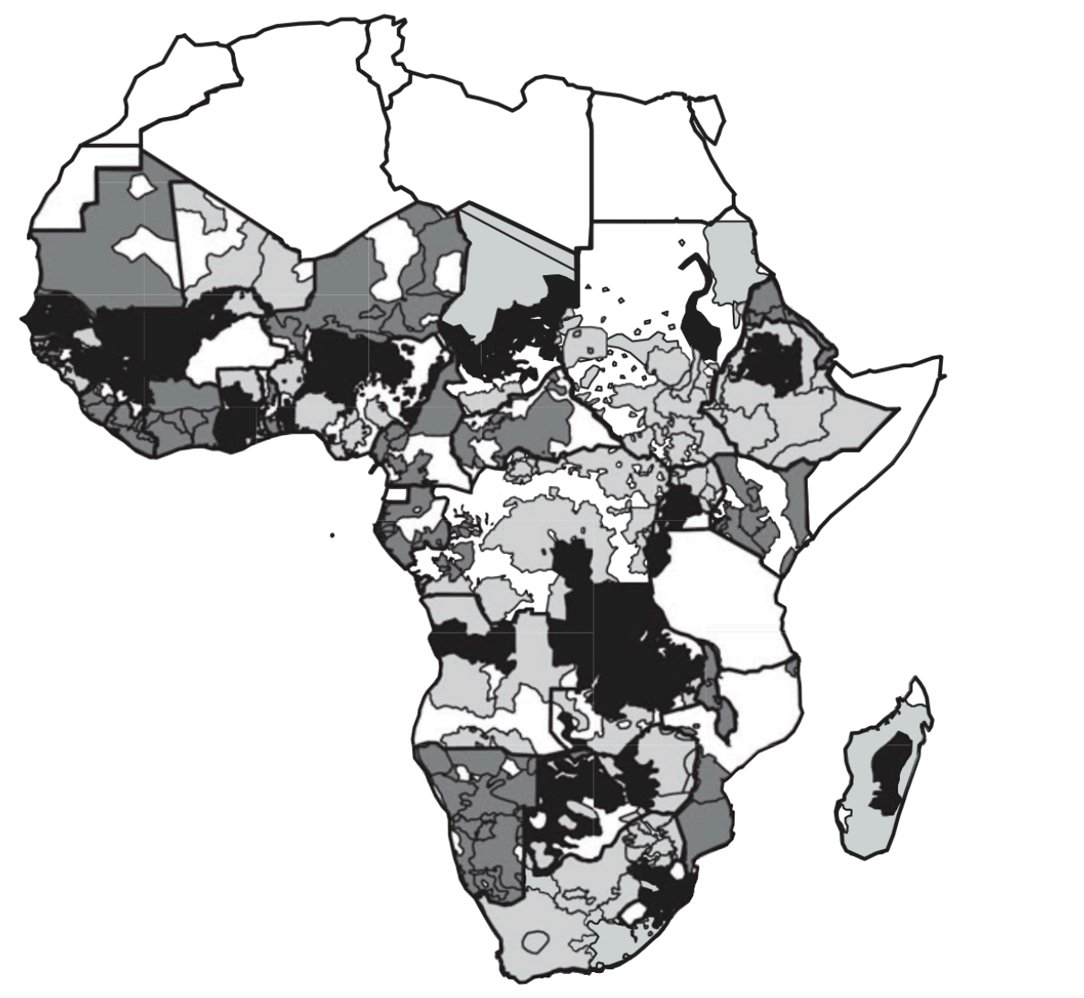

***Fig.1 : Geographic representation of the territories of ethnic groups included in Paine's ethnic group coding system**. PCS groups (black territories): those that were well-established in the pre-colonial era and advantaged under colonialism. PCS states: those in which PCS groups reside. SLPCS groups (gray territories): those that reside in PCS states but were not well-established in the precolonial period or advantaged under colonialism. SL states (white territories): those that lacked well-established groups in the precolonial era. SL groups: those residing within SL states.[6]*

Methods

To use Paine's dataset to model conflict prediction, we make key modifications in response to two primary considerations:

Consideration 1: *High autocorrelation characteristic of time-series data and predictive power of past conflict.* Paine's dataset is organized by years, and thus his explanatory model controls for prior conflict. However, when creating an accurate predictive model, incorporating past conflict variables is crucial, as such features often exhibit strong predictive power.

Solution: *To compensate for this, we use time-lagged predictors*, which capture temporal dependencies in the data and allow the model to leverage information from prior time points to improve forecast accuracy.

Consideration 2: *Temporal leakage and train test split.* If we randomly allocate group-years to training and testing sets, there will be group-years in the test set that come chronologically before group-years in the training set, meaning that the model is trained on data that should be unavailable in a prediction context.

Solution: *To reduce train-test leakage, we use a year-threshold train-test split.* This restrains the model from learning information that should be unavailable during the training phase and enables the model to mimic future conflict prediction.

Initial Model Comparison

Methods Compared

- Decision tree with grid search hyperparameter tuning
- Random forest with grid search hyperparameter tuning
- Logistic regression with grid search hyperparameter tuning
- Logistic regression without regularization or hyperparameter tuning *(control)*
- Glmnet regression with LASSO regularization and lambda optimization

Results & Discussion

- Decision tree and random forest models consistently outperform all three logistic regression model variations.
- Tree-based modes are typically 1) able to more reliably discriminate between positive and negative classes and 2) able to achieve high precision and a high precision-recall balance at the default threshold of 0.5.

Pseudo-Predictive Time-Lagged Models

Methods Compared

- Decision tree with grid search hyperparameter tuning
- Random forest with grid search hyperparameter tuning
- Logistic regression with grid search hyperparameter tuning
- Logistic regression without regularization or hyperparameter tuning *(control)*
- Glmnet regression with LASSO regularization and lambda optimization

Results & Discussion

- Random forest models perform better than the decision tree models across all time lags.
- F1 scores do not consistently increase or decrease across lags.
- Decision tree models achieve higher AUC values as lagging increases, but random forest models do the opposite. This is likely due to some tradeoff between two factors: 1) time-lagging decreases the effects of autocorrelation and allows the model to learn temporal dependencies in the data and 2) reduced lagging preserves data points.

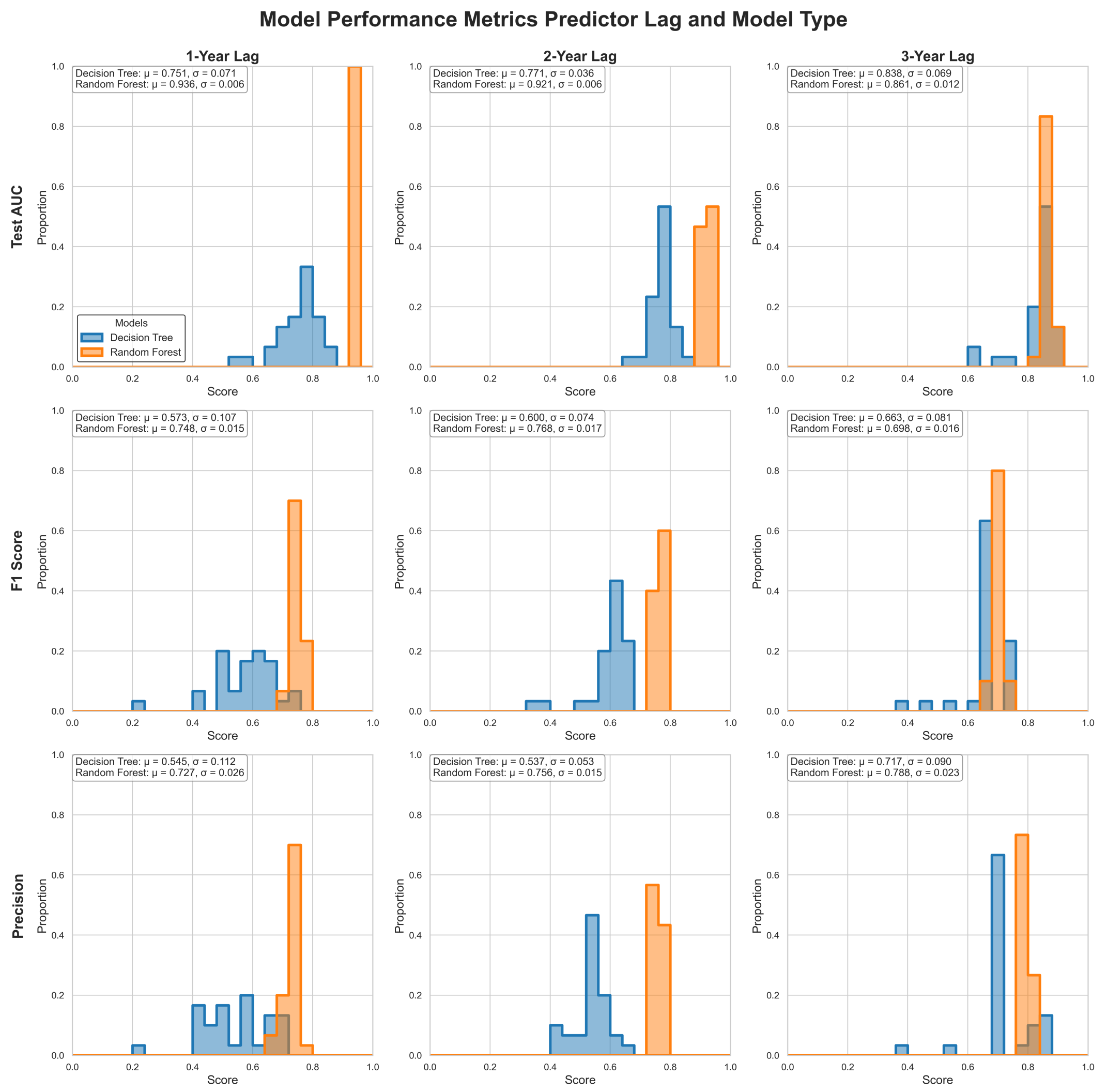

***Fig. 2: Histogram distributions of performance metrics for time-lagged models.** Thirty trials were conducted for each model and associated time lag. Mean and standard deviations of the performance metric distributions for each model type and corresponding lag are shown in the top right corner of each subplot.*

Decade-Specific Tree-Based Models

Methods Compared

- Decision tree with predictors lagged one year, tested on 1980s, 1990s, and 2000s data (cumulative and non-cumulative models)
- Random forest with predictors lagged one year, tested on 1980s, 1990s, and 2000s data (cumulative and non-cumulative models)

Results & Discussion

- Models generally perform best on the 2000s test set.
- Model performance varies more significantly across decade-specific models than between cumulative and non-cumulative models.
- Cumulative models do not consistently perform better across any decades or metrics.
- Random forest models perform better than decision tree models, with the exception of the 1980s dataset which saw mixed performance across metrics.

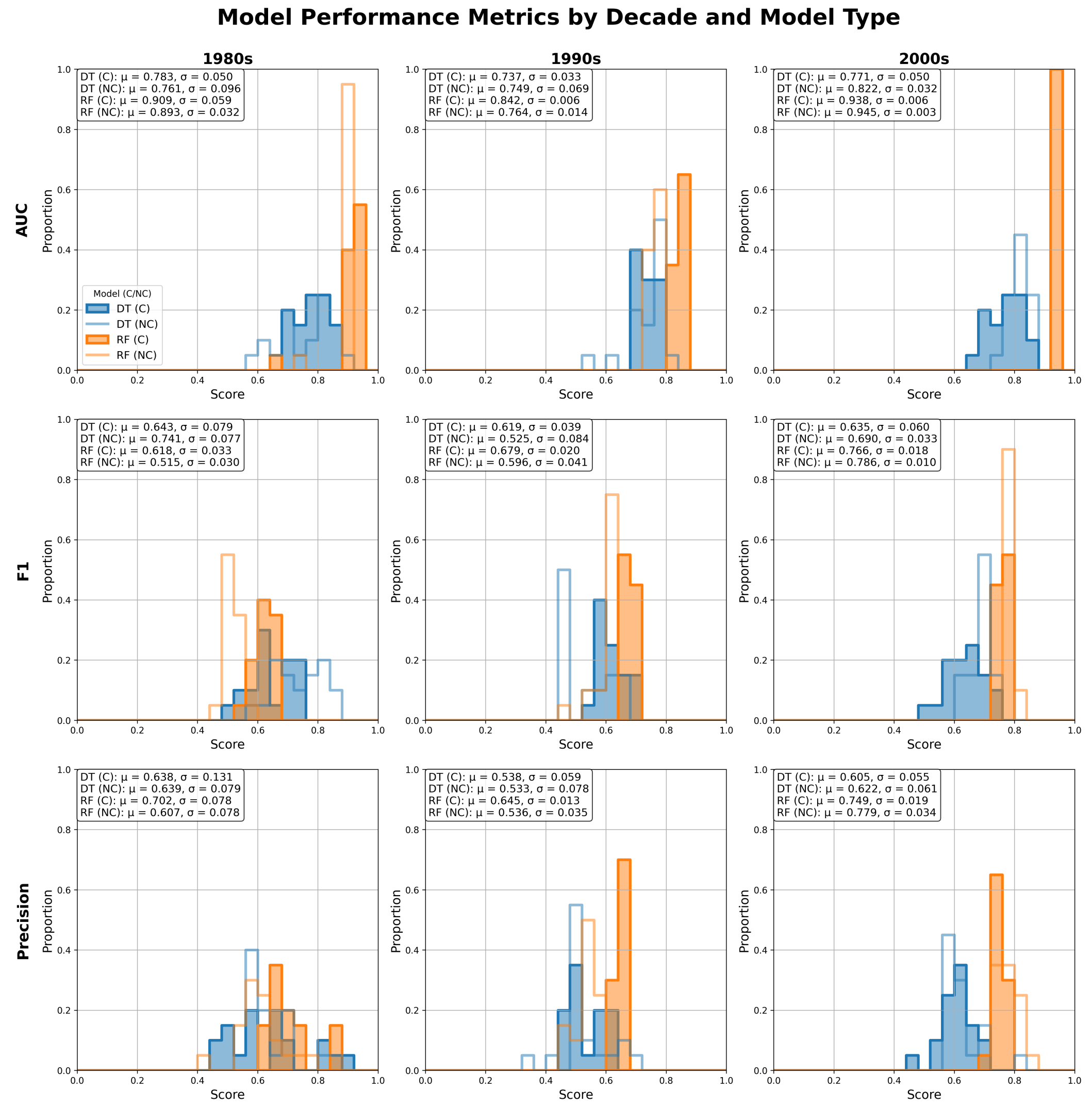

***Fig. 3: Histogram distributions of performance metrics for decade-specific models with predictors lagged by one year.** Twenty trials were conducted for each model and associated decade and cumulativity classification. Mean and standard deviations of the performance metric distributions for each model type and corresponding decade and cumulativity classification are shown in the top right corner of each subplot.*

References

1. Fearon, J. D., & Laitin, D. D. (2003). Ethnicity, Insurgency, and Civil War. *The American Political Science Review*, 97(1), 75–90. http://www.jstor.org/stable/3118222
2. Burke, M. B., Miguel, E., Satyanath, S., Dykema, J. A., & Lobell, D. B. (2009). Warming increases the risk of civil war in Africa. *Proceedings of the National Academy of Sciences*, 106(49), 20670–20674. https://doi.org/10.1073/pnas.0907998106
3. GPF Team. (2016, June 3). Colonial Powers in Sub-Saharan Africa | Geopolitical Futures. Geopolitical Futures. https://geopoliticalfutures.com/colonial-powers-in-sub-saharan-africa/
4. Henderson, E. A., & Singer, J. D. (2000). Civil War in the Post-Colonial World, 1946-92. *Journal of Peace Research*, 37(3), 275–299. http://www.jstor.org/stable/425346
5. Harkness, K. A. (2016). The Ethnic Army and the State: Explaining Coup Traps and the Difficulties of Democratization in Africa. *The Journal of Conflict Resolution*, 60(4), 587–616. http://www.jstor.org/stable/24755887
6. Paine, J. (2019). Ethnic Violence in Africa: Destructive Legacies of Pre-Colonial States. *International Organization*, 73(3), 645–683. https://www.jstor.org/stable/26758061
7. Musumba, M., Fatema, N., & Kibriya, S. (2021). Prevention is better than cure: Machine learning approach to conflict prediction in sub-Saharan Africa. *Sustainability*, 13(13), 7366. https://doi.org/10.3390/su13137366
8. Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J., Love, B. C., Raimondo, F., & Patil, K. R. (2024). On Leakage in Machine Learning Pipelines (No. arXiv:2311.04179). *arXiv*. https://doi.org/10.48550/arXiv.2311.04179
9. Soni, B. (2023, March 15). Why random forests outperform decision trees: A powerful tool for complex data analysis. *Medium*. https://medium.com/@brijesh_soni/why-random-forests-outperform-decision-trees-a-powerful-tool-for-complex-data-analysis-47f96d9062e7
10. Willig, G. (2023, January 16). Decision tree vs logistic regression. *Medium*. https://gustavwillig.medium.com/decision-tree-vs-logistic-regression-1a40c58307d0
11. Meek, C., Chickering, D. M., & Heckerman, D. (2002). Autoregressive Tree Models for Time-Series Analysis. In *Proceedings of the 2002 SIAM International Conference on Data Mining* (pp. 229–244). Society for Industrial and Applied Mathematics. https://epubs.siam.org/doi/10.1137/1.9781611972726.14

Feature Importance

Feature importance: how significantly a feature contributes to the predictive power of the model.

Observations

- Importance was more evenly distributed across features in the tree-based models than in the logistic regression models.
- 10 predictors consistently have highest feature importance.
- 3 fundamental dynamic predictors: GDPPC, groupwise population, years since last ruling group change.

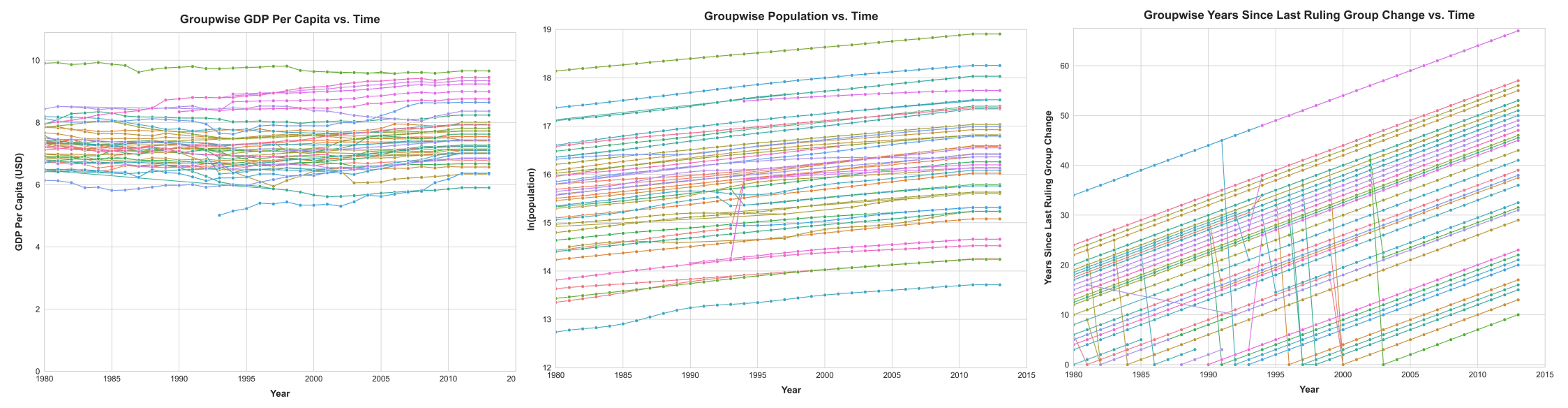

***Fig. 4 : Group-level trends over time.** Log GDP per capita (Left), Log group population (Center), Years since last ruling group change (Right). We observe that GDP per capita and population are more easily approximated by a linear model, while "peace years" is more volatile.*

Discussion & Conclusion

Tree-based models outperform logistic regression in classifying instances of conflict on Paine's sparse dataset.

Random forest models tend to outperform decision tree models in a time-lagged, pseudo-predictive context, and perform best on modern-day data when trained on only recent data instances.

Feature importance is distributed relatively evenly across features in the tree-based models, suggesting that these models are more robust to crude linear extrapolation techniques required to make real-world conflict predictions.

The random forest model with grid search hyperparameter tuning can achieve high AUC and moderately high F1 scores when tested on lagged 2000s conflict data, suggesting that:

- ***Random forest models are the best option*** for conflict prediction is Sub-Saharan Africa.
- ***Historical or static features are meaningful*** features for conflict-prediction models.

Preliminary results are optimistic, ***but underlying causes of conflict are not well-understood and may be continuously evolving*** with emerging pressures (e.g. development, climate change, technological innovation).

We should consider ***leveraging the power of tree-based models in tandem with more mathematically complex techniques.*** Autoregressive Tree (ART) models for time-series analysis might help us gauge civil war risk on a continuous 0-1 scale, while being compatible with time-series data, lagged predictors, and small, sparse datasets.[11]

Acknowledgements

Thank you to my advisor, Professor Mucha, for his ongoing holistic support throughout the term. I could not have completed this project without the accountability or expert interpretation of results that you provided. Thank you, also, to my Machine Learning Professor Yaoquing Yang for providing the foundational skills necessary to conduct this research.