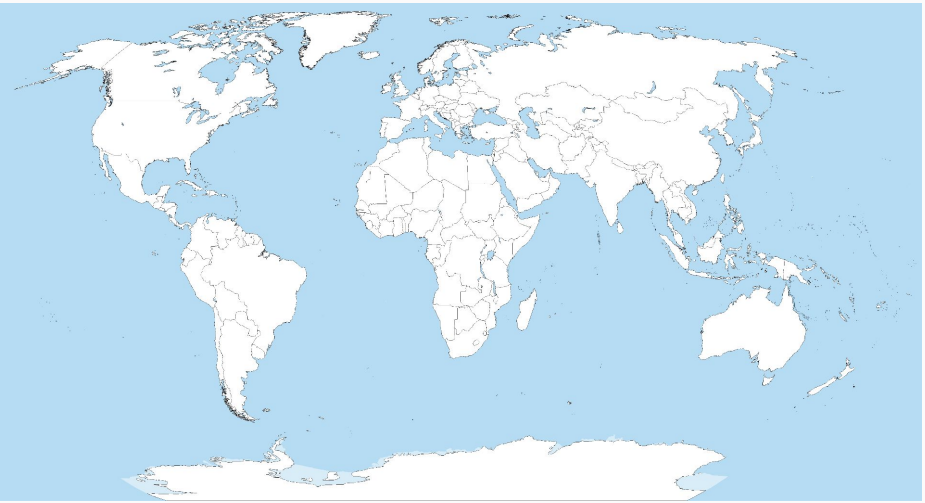


Predicting Language Endangerment

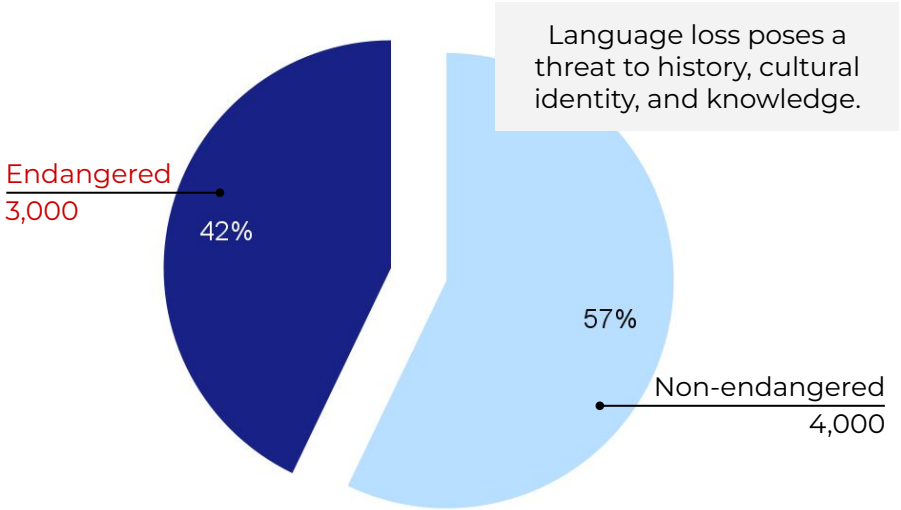
DATASCI 207 Section 5

Jordan Andersen, Courtney Chen,
Brian Woods, Helin Yilmaz

Question & Motivation

What makes a language endangered, and can we predict this before it's too late?

Languages worldwide are are disappearing at alarming rates.



Our approach uses ML to predict endangerment.

Model Inputs (Numeric)

Language presence: Speakers, countries, legal status
Socioeconomic context: Urbanization rate, internet use

Models Used

Gradient Boost, Ensemble, Neural Networks, Extra Trees
(Baseline Majority Class Predictor: ~46% test accuracy)

Model Output (Categorical)

Endangerment level *(Not Endangered, At Risk, Endangered, Highly Endangered, Extinct)*

Related Work

Journal	Source	Model	Strengths	Limitations
Global Predictors of Language Endangerment[1]	Nature Ecology & Evolution	Ordinal probit model analyzing 51 predictor variables targeting language “maintenance”.	Statistically rigorous model designed for ordinal predictions.	Current predictors may not capture historical patterns associated with endangerment.
Predicting Language Endangerment[2]	Pankaj Dwivedi	Simple linear regression predicting extinction year for Indian languages.	Focuses on the timeline of extinction, which is critical for language revitalization efforts.	Regional rather than global scope.

1. Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., & Hua, X. (2022). *Global predictors of language endangerment and the future of linguistic diversity. Nature Ecology & Evolution, 6*(2), 163–173. https://doi.org/10.1038/s41559-021-01604-y
2. Dwivedi, P. (2020). *Predicting language endangerment: A machine learning approach.* In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 147–153). https://doi.org/10.1109/ICCCNT49239.2020.9225576

Description of the Data

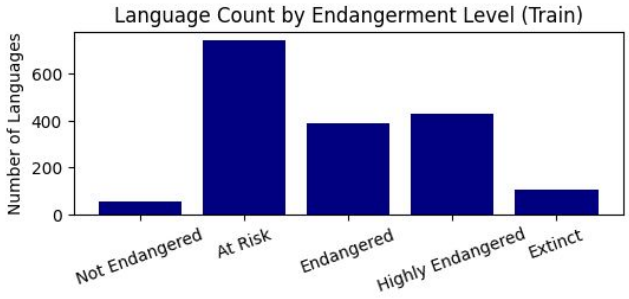
Datasets

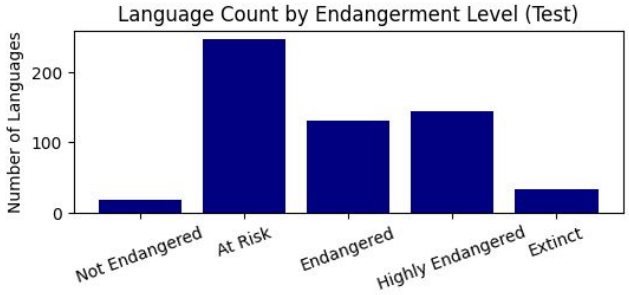
- Endangered Languages[1]
- Official Languages by Country[1]
- Most Spoken Languages[2]
- Rate of Urbanization[3]
- Internet Use[3]

Data Preparation

1. Missing value drops and imputations
2. RegEx formatting for consistency
3. Log transformations on speakers
4. Data balancing with SMOTE
5. Interaction features for joint patterns

Distribution of the Data





1. Endangered Languages Project
2. Ethnologue
3. World Bank Indicators

Gradient Boost: Experiments & Results

Gradient Boost is an ensemble model that builds sequential trees, with each tree improving upon the previous one.

Hyperparameter Tuning

Estimators		
50	100	200

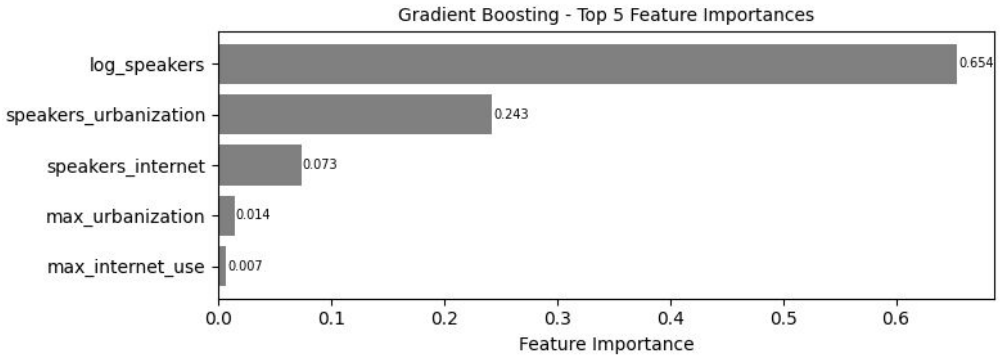
Depth		
1	3	5

Learning rate	
0.05	0.1

Subsample ratio	
0.6	0.8

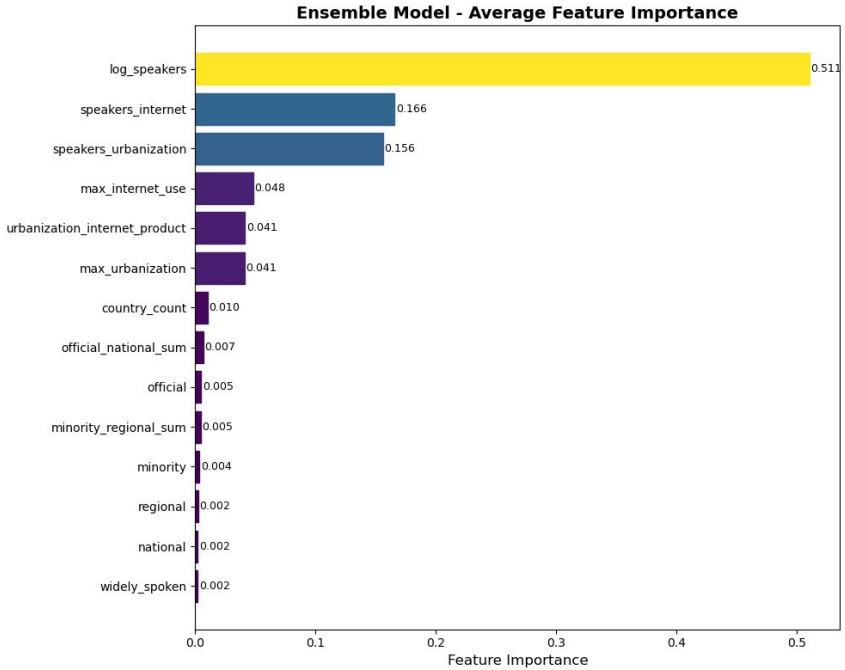
Model Evaluation

	Train Acc.	Validation Acc.	Test Acc.
Base	0.976	0.697	0.721
Final	0.918	0.901	0.883



Ensemble: Experiments & Results

Ensemble models incorporate multiple models such as random forests, gradient boosting and extra trees, then use methods like voting, bagging or boosting which aim to generate the best possible estimator for the task.



Model Evaluation

	Train Acc.	Validation Acc.	Test Acc.
Base Model	0.956	0.890	0.882
Final Model	0.911	0.889	0.883

Hyperparameter Tuning

Varied voting method (hard/soft) and class imbalance handling (None, SMOTE, SMOTETomek) along with hyperparameter tuning.

	Estimators	Max Depth	Learning Rate
Gradient Boosting	[50, 100, 200]	[3, 5, 7]	[0.01, 0.1, 0.5]
Extra Trees	[50, 100, 200]	[5, 10, 15]	N/A
Random Forest	[50, 100, 200]	[5, 10, 15]	N/A

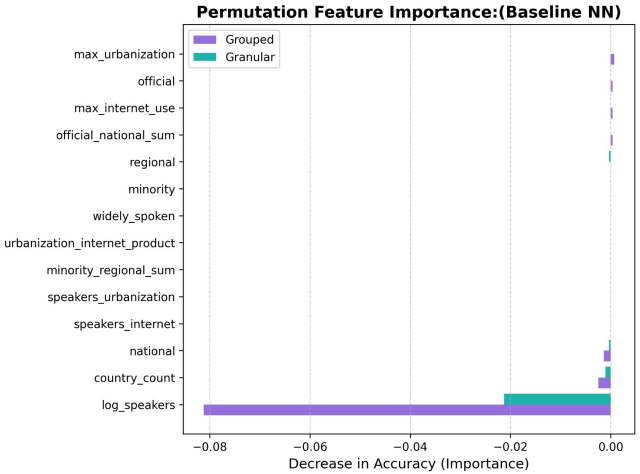
Neural Network: Experiments & Results

A Neural Network model is a layered computational model that processes input data through a series of weighted sums and non-linear activation functions, allowing it to learn complex, non-linear patterns in the data.

Hyperparameter Tuning

Baseline vs. Tuned

Dense Layers	
2 (128 and 64)	2 (160 and 128)
Dropout	
0.3	0.3 and 0.2
Learning rate	
Default	0.001
Class Imbalance Handling	
None	SMOTE



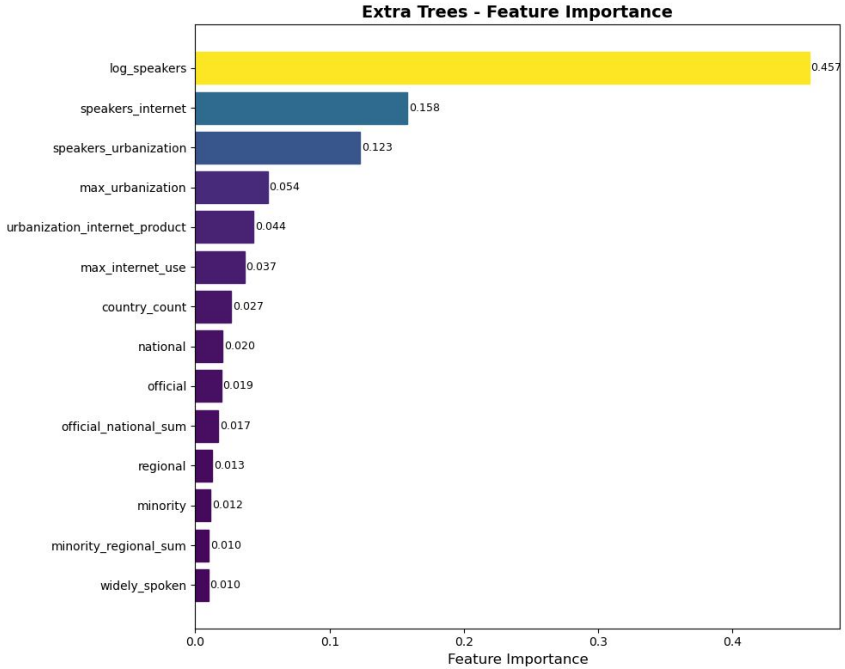
Model Evaluation

	Train Acc.	Validation Acc.	Test Acc.
Base Model (Granular)	0.754	0.720	0.728
Base Model (Grouped)	0.865	0.892	0.876
Tuned Model (Granular)	0.751	0.718	0.742
Tuned Model (Grouped)	0.862	0.887	0.857

Granular (8-class) Grouped (5-class)

Extra Trees: Experiments & Results

Extra Trees is an ensemble model that builds multiple trees in parallel, with extra randomization in feature and split selection to enhance diversity.



Model Evaluation (after being tuned)

	Train Acc.	Validation Acc.	Test Acc.
Base Model	0.8066	0.7561	0.7509
Final Model	0.9123	0.8937	0.8780

Hyperparameter Tuning

Parameter	Best	Explanation
N_Estimators	100	Best from [100, 200, 300, 500]
Max Depth	15	Best from [10, 15, 20, None]
Min Samples Split	5	Best from [2, 5, 10]
Min Samples Leaf	1	Best from [1, 2, 4]

Conclusion

Key Findings

- High predictive performance across all models, driven by the strong link between number of speakers and endangerment.
- Number of speakers consistently emerged as the most influential feature across all models.
- **Best Model:** Close tie between Gradient Boosting and Ensemble models.

Limitations & Future Work

- Historical patterns may not be captured in current data.
- Including additional socio-economic, geographic or historical factors may help push the test accuracy past the 88% threshold.