# Ancient Texts NLP & Classification Project

## Results of Model Building, Evaluating, and Testing

*All code at https://github.com/izsolnay/Ancient_NLP*

## ISSUE / PROBLEM

Discover if there are quantifiable differences between hymns (or praise texts) and other textual genres

**Objective**: Build a machine learning model that can classify a text as a hymn out of all texts

**Steps**: Build two models, evaluate, and test on winning model

## RESPONSE

- Created new features: 'top_word_freq', 'divine_power_count', and 'divine_power_percent'
- Conducted collinearity analyses using VIF and a Pearson correlation coefficient matrix
- Developed Random Forest and XGBoost models using recall as metric
- Validated models on val data set
- Generated confusion matrices
- Evaluated with ROC curves and AUCs
- Tested winning XGBoost model
- Extracted feature importances

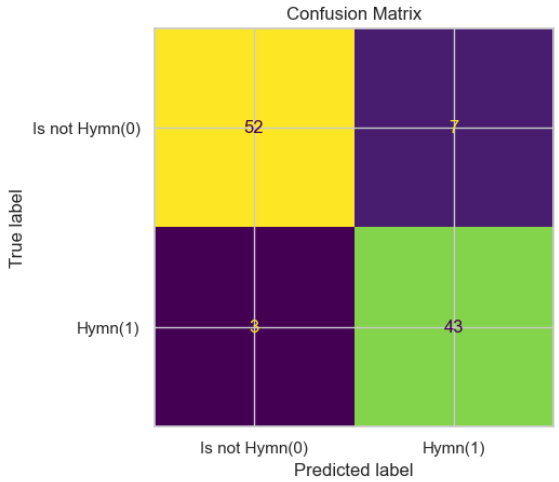
## IMPACT

The winning XGBoost model exceeded expectations achieving a recall score of 93% on the test data.

| Model | Precision | Recall | Accuracy | F1 |
|---|---|---|---|---|
| XGB TEST | 0.86 | 0.93 | 0.90 | 0.90 |

*Recall measures the model's efficacy in classifying whether a text is a hymn rather than not a hymn.*

## MODEL EVALUATION: VERY ROBUST

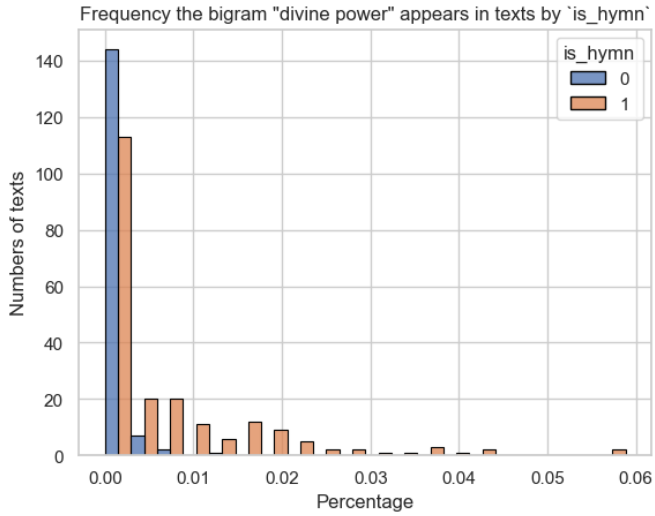



- True Negatives = 52
- True Positives = 43
- False Positives = 7
- False Negatives = 3

*Upper left: number of texts accurately classified as not hymns. Lower right: number of texts accurately classified as hymns. Upper right: number of texts inaccurately classified as hymns. Lower left: number of texts inaccurately classified as not hymns.*

## DIVINE POWER

The frequency the bigram "divine power" is observed in a 'hymn = 1' text is far greater than the frequency it is found in a 'hymn = 0' text




## KEY INSIGHTS

The most important features for classification:

- Whole text subjectivity scores
- Word counts
- Whole text polarity scores

The count, percentage, and bigram of "divine power" were also salient.

Other important features were the bigrams:

- 'heaven earth'
- 'foreign land'
- 'holy inana'