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 RF model
- Extracted feature importances

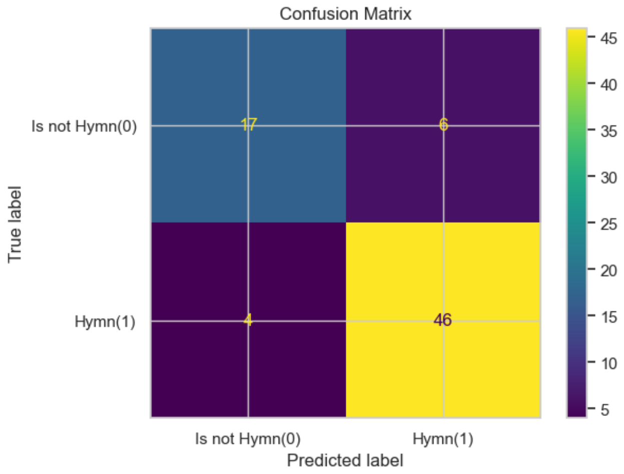
IMPACT

The winning Random Forest model exceeded expectations achieving a recall score of 92% on the test data.

Model	Precision	Recall	Accuracy	F1
RF TEST	0.885	0.920	0.863	0.901

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

MODEL EVALUATION: VERY ROBUST

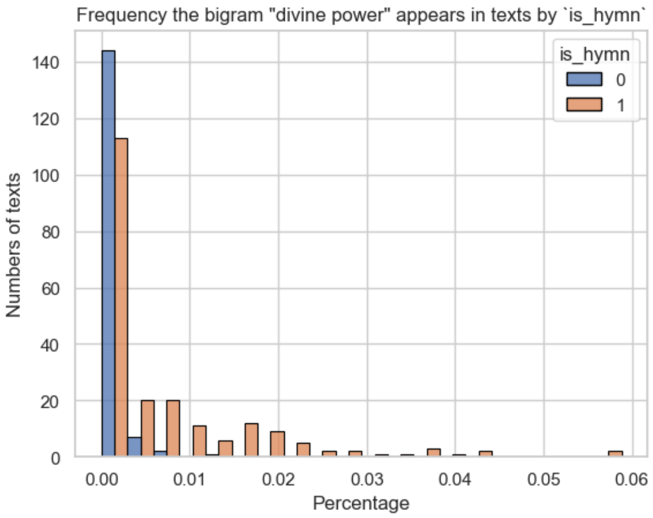


- True Negatives = 17
- True Positives = 46
- False Positives = 6
- False Negatives = 4

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



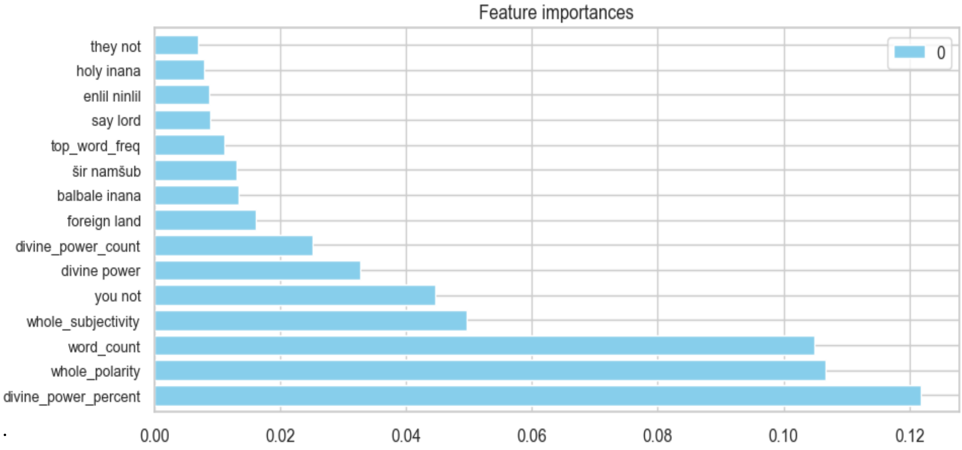
KEY INSIGHTS

The most important features for classification:

- The percentage the bigram "divine power" appeared in a text
- Polarity scores
- Word counts

The count for "divine power" was not nearly as salient as might be expected, nor was the bigram feature 'divine power'.

The least predictive of the top 15 features were bigrams: 'say lord', 'enlil ninlil', 'holy inana', and 'they not'.


Feature importances
they not
holy inana
enlil ninlil
say lord
top_word_freq
šir namšub
balbale inana
foreign land
divine_power_count
divine power
you not
whole_subjectivity
word_count
whole_polarity
divine_power_percent