Predicting Medium Article Engagement

Mei Zhu

Project Overview & Objectives

Research Question

To what extent can machine learning models predict the engagement level of Medium articles using features such as textual content, author history, and publishing metadata?

Additionally, how accurately can we estimate the actual number of claps an article will receive, and can this prediction help us classify its engagement level?

Motivation

Understanding and predicting article engagement can benefit multiple stakeholders in the content ecosystem:

Writers

Optimize content features to boost interaction

Editors

Plan publishing schedules and identify promising content

Platforms

Recommend highly engaging articles to improve UX

Advertisers

Place ads strategically for maximum visibility

Dual Modeling Approach

Classification Task

Identify whether an article belongs to the top 25% most engaging articles (binary: high vs. low)

Regression Task

Predict the actual number of claps an article will receive (log-transformed to handle distribution skewness)

Data Preparation & Feature Engineering

Text Features

- Combined title and subtitle text
- TF-IDF feature extraction
- VADER sentiment analysis scores

Metadata Features

- Author popularity (average claps)
- Categorical tags (one-hot encoded)
- Temporal features: month, weekday, weekend flag
- Reading time and response counts

Classification: Identifying High-Engagement Articles

Approach

- Random Forest classifier with balanced class weights
- Top 25% of articles labeled as "high engagement" (1)
- Remaining 75% labeled as "low engagement" (0)
- 80/20 train-test split with stratification

Performance Metrics

Overall accuracy: 92%

ROC AUC Score: 0.8949

Low Engagement (Class 0)

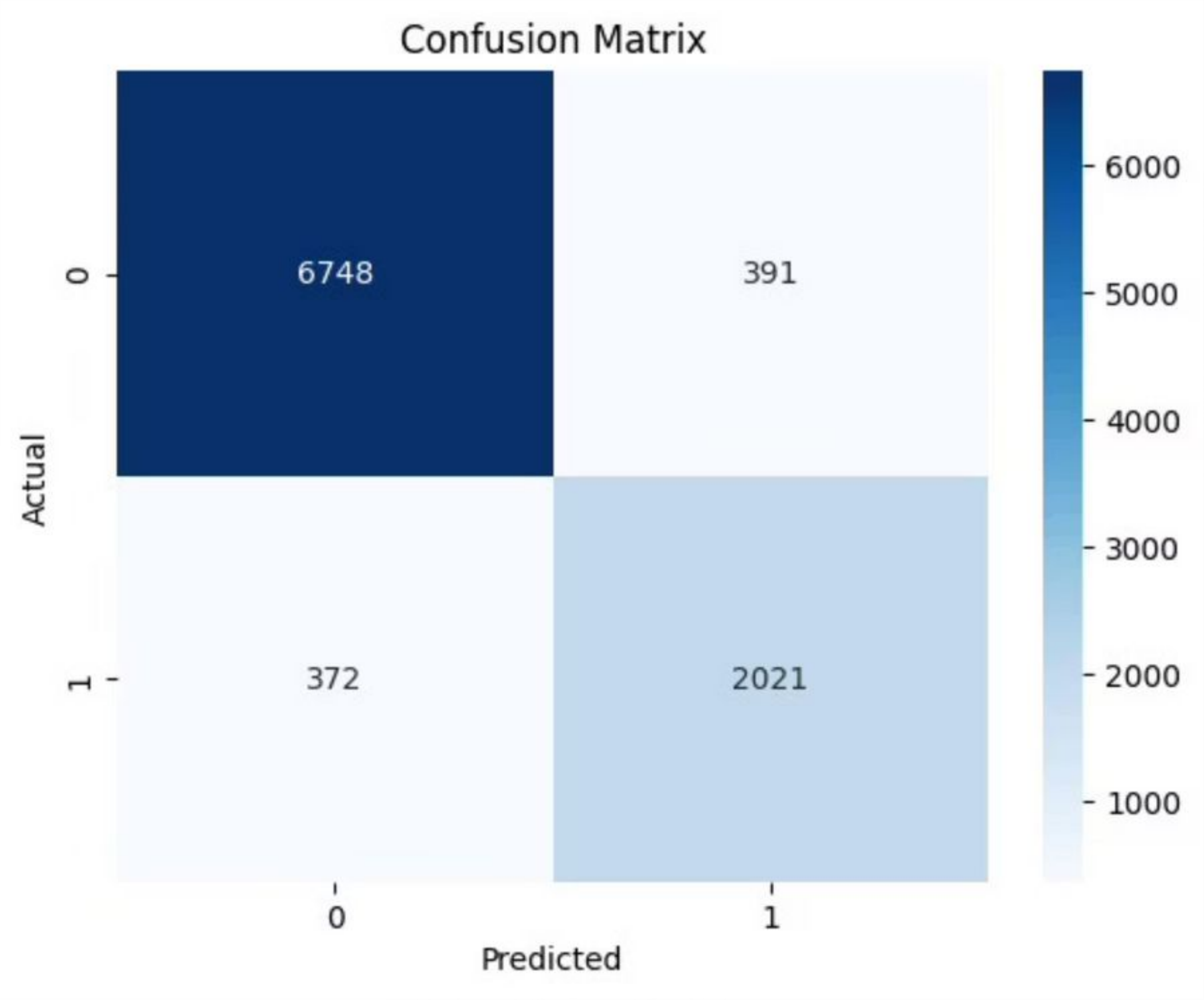
Precision: 0.95

Recall: 0.95

High Engagement (Class 1)

Precision: 0.84

Recall: 0.84



Confusion matrix showing strong performance in both classes, with slightly lower performance on the minority high-engagement class.

Classification Performance Analysis

Key Insights

1 Strong Discrimination Ability

ROC AUC near 0.9 indicates excellent separation between high and low engagement articles, making the model reliable for editorial decisions

2 Primary Predictive Features

Author historical performance (author_avg_claps), text length, and sentiment were the most influential features, suggesting quality and positivity drive engagement more than timing

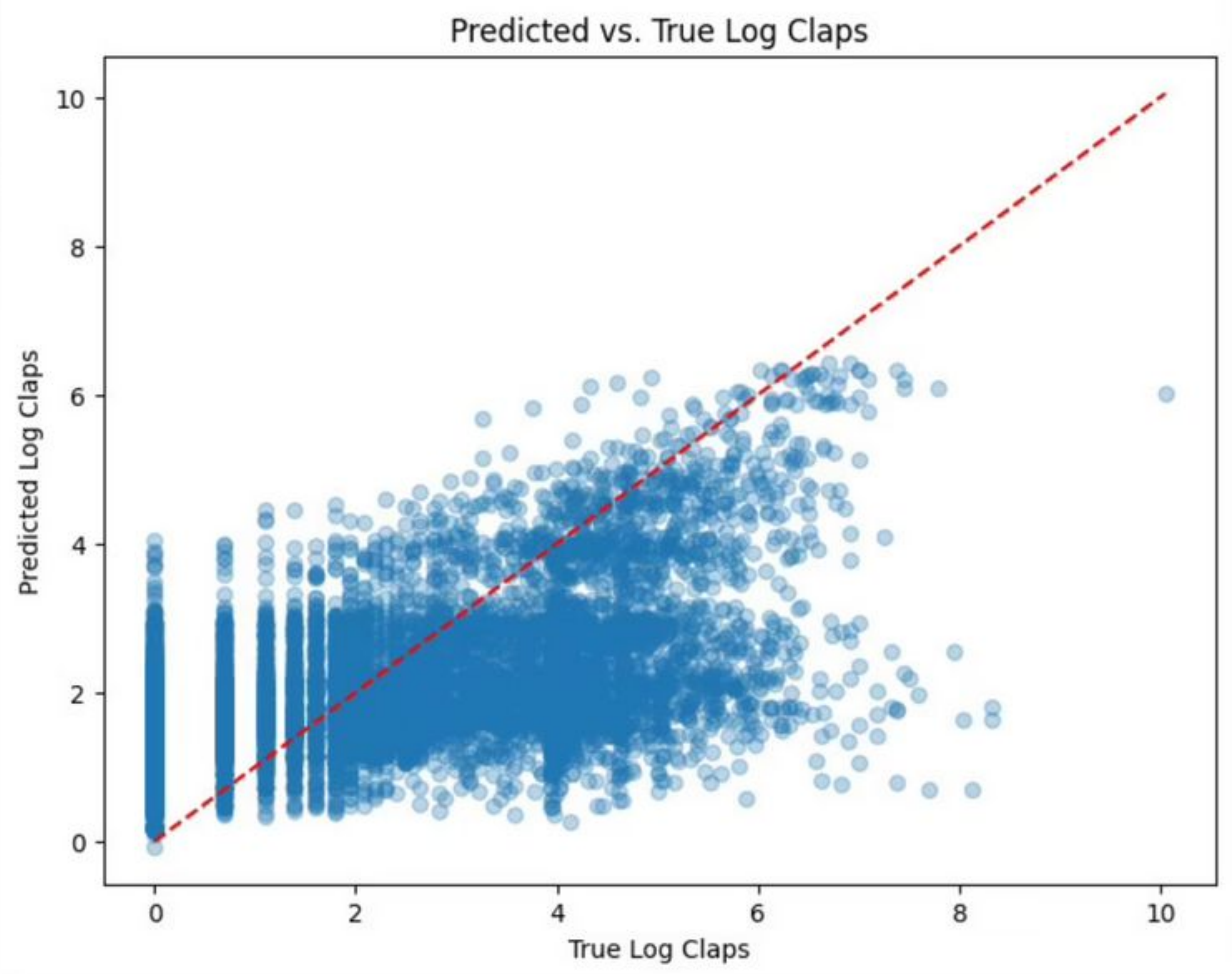
3 Practical Applications

The classifier can effectively flag potentially high-performing content for prioritization in recommendation systems or promotional campaigns

Regression: Predicting Exact Clap Counts

Approach

- Log-transformed target variable to normalize skewed clap distribution
- Gradient Boosting Regressor model
- Features: TF-IDF vectors, one-hot encoded tags, reading time, response count, text lengths
- 80/20 train-test split



Performance Metrics

1.25	1.55	0.32
MAE	RMSE	R^2
Mean Absolute Error (log scale)	Root Mean Square Error (log scale)	Coefficient of determination

Regression Performance Analysis

Key Insights

1 Model Performance & Nuance

- The model captures general trends but struggles with extreme values (very high or low engagement)
- R^2 of 0.32 indicates the model explains about one-third of variance in clap counts
- MAE of 1.25 in log scale represents reasonable prediction accuracy given engagement's inherent variability

2 Distinct Feature Drivers

- Most influential features differed from classification model:
  - Response count
  - Reading time
  - Subtitle length

3 Granular Prediction Capability

- These features better predict exact engagement levels rather than just high/low classification

The regression approach provides more nuanced insights into expected engagement levels, enabling granular content performance analysis and personalized recommendations.

Feature Importance Analysis

Classification Model

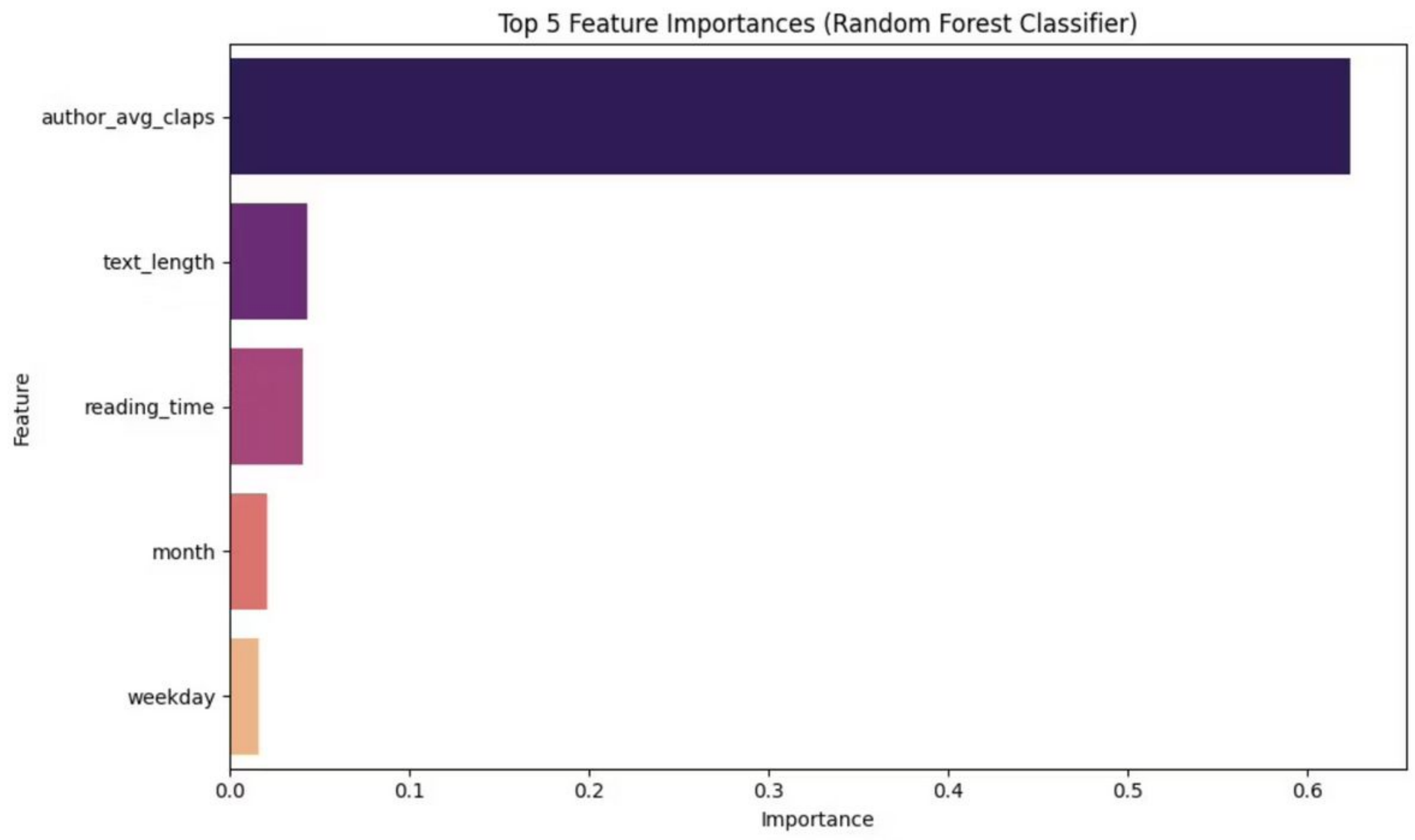
Top predictors of high vs. low engagement:

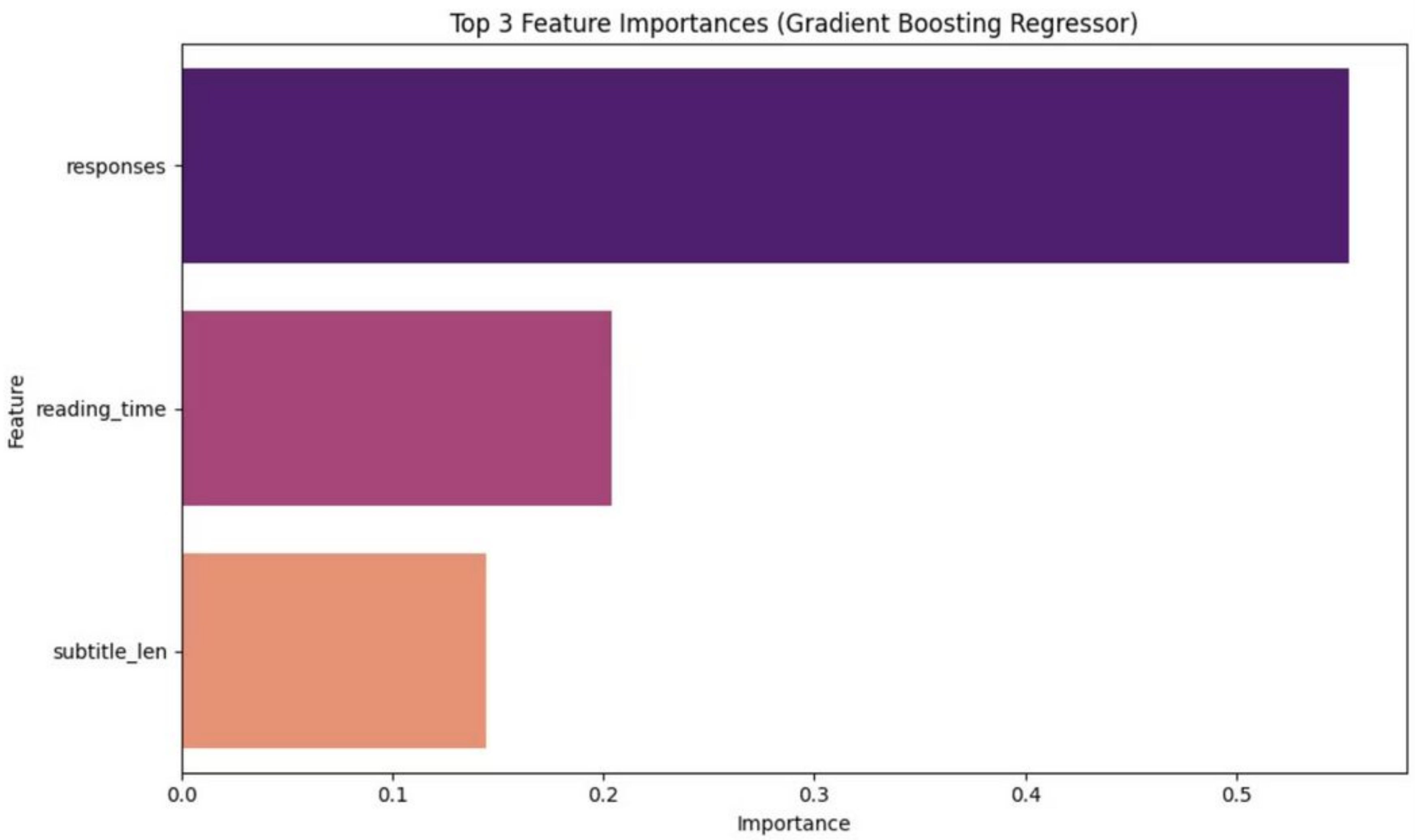
1. **Author average claps** - historical performance is highly predictive
2. **Text length** - comprehensive articles tend to engage more
3. **Sentiment score** - positive content generally performs better
4. **Reading time** - mid-length articles optimize engagement

Regression Model

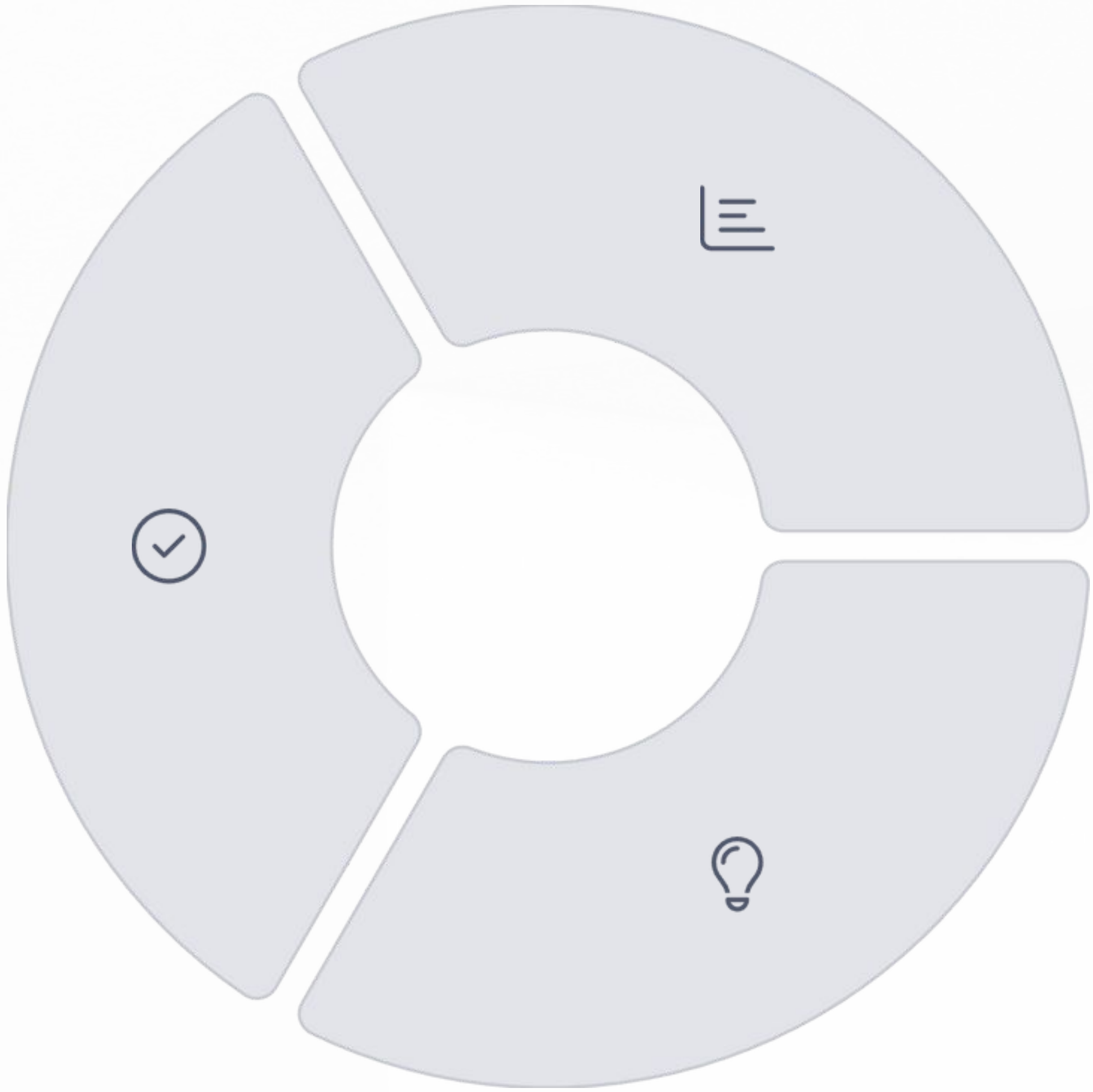
Top predictors of exact clap counts:

1. **Response count** - articles generating discussion receive more claps
2. **Reading time** - more influential for predicting exact engagement
3. **Subtitle length** - detailed subtitles signal content quality
4. **Temporal features** - publishing timing affects exact engagement levels





Summary of Results



Regression Insights

MAE of 1.25 (log scale) with R^2 of 0.32, providing granular engagement predictions

Classification Success

92% accuracy with strong ROC AUC of 0.89, effectively identifying high-potential content

Key Findings

Author reputation, content quality and sentiment drive engagement more than temporal factors

Applications of the Dual Modeling Approach

The complementary models provide both broad classification for quick decision-making and detailed regression for nuanced content optimization, creating a comprehensive engagement prediction system.

Next Steps & Recommendations

Model Improvements

Feature Engineering

- Topic modeling of article content
- Author social media metrics integration
- More granular temporal features (time of day, holidays)

Advanced Techniques

- XGBoost and neural network architectures
- SMOTE for better class imbalance handling
- Hyperparameter optimization via cross-validation

Analytical Enhancements

Interpretability

- SHAP values for deeper feature impact analysis
- Partial dependence plots for feature relationships
- A/B testing to validate model recommendations

Deployment Strategy

- Real-time prediction API for writers
- Integration with content management systems
- Feedback loop for continuous model improvement