Predicting Medium Article Engagement

Mei Zhu

Highlights

- Built machine learning models to predict Medium article engagement,, using number of claps as a proxy.
- Applied Random Forest for classification (high vs. low engagement) and Gradient Boosting for regression (predicting number of claps).
- Key content features such as author popularity, sentiment, and article length were found to influence engagement the most.

Background

As content discovery becomes increasingly algorithm-driven, with AI-powered search and recommendation systems shaping what readers see, understanding and predicting engagement is more valuable than ever.

During my internship at a media and publishing company, I observed a growing need to optimize content visibility, particularly for evergreen articles.

This sparked the central question of my project: Can we use machine learning to predict the engagement of online articles before they are published?

Data Preparation

- Text Processing
  - Combined title + subtitle and extracted word features using TF-IDF
  - Ran VADER sentiment analysis on article text
  - Calculated text length and subtitle length
- Metadata Engineering
  - Computed author average claps to represent reputation
  - Extracted temporal features: month, weekday, weekend
  - One-hot encoded tags (e.g. “Machine Learning,” “Big Data”)
- Target Variables
  - Classification: articles in the top 25% of claps labelled as “high engagement”
  - Regression: Used log-transformed calps to reduce skewness
- Data Cleaning
  - Removed rows with missing values in key columns
  - Final dataset size: 13,728 articles with full features

Model Selection

Classification (High vs. Low Engagement)

- Model: Random Forest Classifier (balanced class weights)
- Input Features: TF-IDF, sentiment, author avg. claps, reading time, tag, text length, data features
- Performance
  - Accuracy: 91%
  - ROC AUC: 0.875
  - Precision/Recall (high engagement): 0.81
- Key Features
  - Author average claps
  - Article length
  - Sentiment score

Regression

- Model: Gradient Boosting Regressor
- Target: Log-transformed number of claps
- Performance
  - MAE: 1.25
  - RMSE: 1.55
  - R2: 0.32
- Key Features
  - Reading time
  - Response count
  - Subtitle length

Conclusions

This project shows that machine learning can meaningfully forecast article engagement, with real applications across the content lifecycle:

- Editors can flag promising articles for promotion before they are published.
- Writers can better understand which article features correlate with higher engagement, helping them tailor their writing accordingly.
- Platforms like Medium could integrate these models into recommendation algorithms to improve user satisfaction and reading time.
- Advertisers can target high-engagement content to maximize ad visibility and ROI.