Instagram Post Like Prediction Project

Data Preprocessing and Transformations

The exploratory data analysis (EDA) revealed significant skewness in key variables (likes, no_of_comments, follower_count). I applied log transformations for likes and no_of_comments and a Box-Cox transformation for follower_count, reducing skewness below |1|. Outliers in no_of_comments were removed, increasing the correlation between likes and no_of_comments from 0.69 to 0.78.

Feature Engineering

Time variables were decomposed into month, day, and hour. I applied sine and cosine transformations to the hour variable to capture cyclical patterns while retaining the month variable for temporal context. An ablation study demonstrated the critical impact of the comments feature on prediction accuracy, with R^2 scores dropping from 0.888 (with comments) to 0.749 (without).

Image-Based Modeling

I fine-tuned Vision Transformer (ViT) models, which struggled, yielding low R^2 scores (as low as -0.85 with transformed data and 0.12 without comments on untransformed data). Image embeddings from CLIP were transformed into GloVe word embeddings, with PCA applied for dimensionality reduction. This combined approach showed modest improvements when integrated with metadata.

Best Model Performance

The best results came from an XGBoost model utilizing PCA-reduced GloVe embeddings alongside metadata. This model achieved an R^2 of 0.900 with comments and 0.772 without.