Predictagram

@NasheathAhmed, @MatanGans, @IshanHasan, @LeonJiang

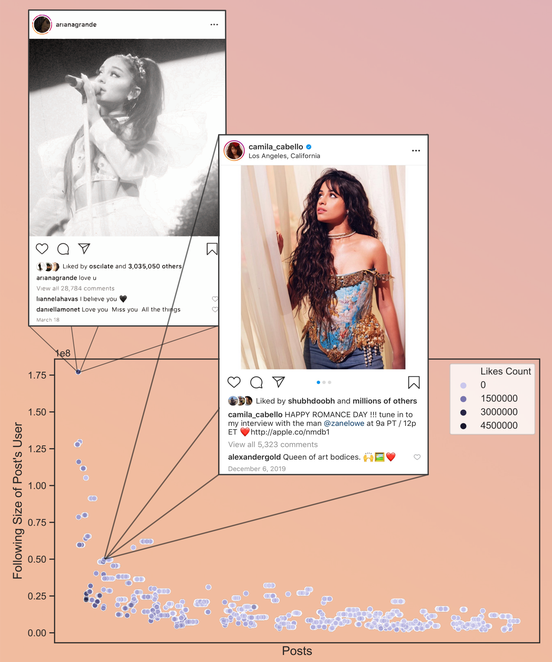
Introduction

For musicians on Instagram, the number of likes on a post serves not only as a metric of their fame and popularity, but as a key indicator of marketability. In the wake of Instagram's shift to potentially remove the likes count on posts, the need for a tool to accurately gauge content impact may arise among managers, record labels, and anyone else who needs this metric to understand artists' popularity trends and fanbase engagement. Therefore, we set out to create a model that attempts to accurately predict the number of likes a musician's post will accumulate, given data on Instagram posts made by musicians in the same year.

Data Collection

42,165 Instagram posts from 323 unique users

Our data consists of posts in the past year from a manually-filtered list of Instagram musicians, originally amounting to 47,737 posts from 356 users. Cleaning the data involved trimming away outliers that could potentially skew our prediction model. These outliers were posts with anomalously high or low numbers of likes, posts from users that over- or under-represented our data, and posts from unverified accounts. Our final data set contained a total of 42,165 Instagram posts from 323 unique users.



Analysis and Results

1. Our full model (including all features) did not outperform our baseline model (which predicts a post's number of likes to be the average number of likes for a post by the same user).

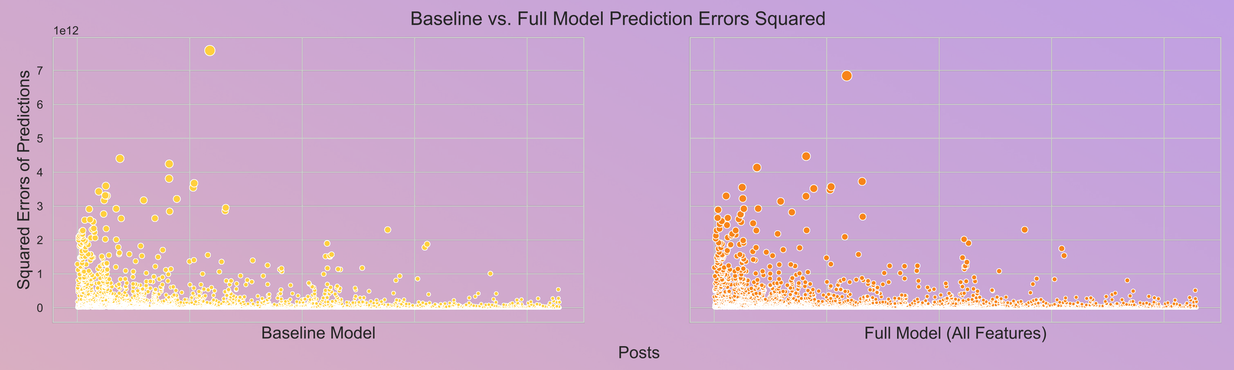


Figure 1: The larger dots in the graph of the full model indicate a larger error between our predictions and the actual value of likes, but the differences are very slight and we cannot determine whether the discrepancy is significant.

2. Most of the predictive power in our models comes solely from the posts' "temporal likes" feature. This is given by the linear regression at the post's timestamp.

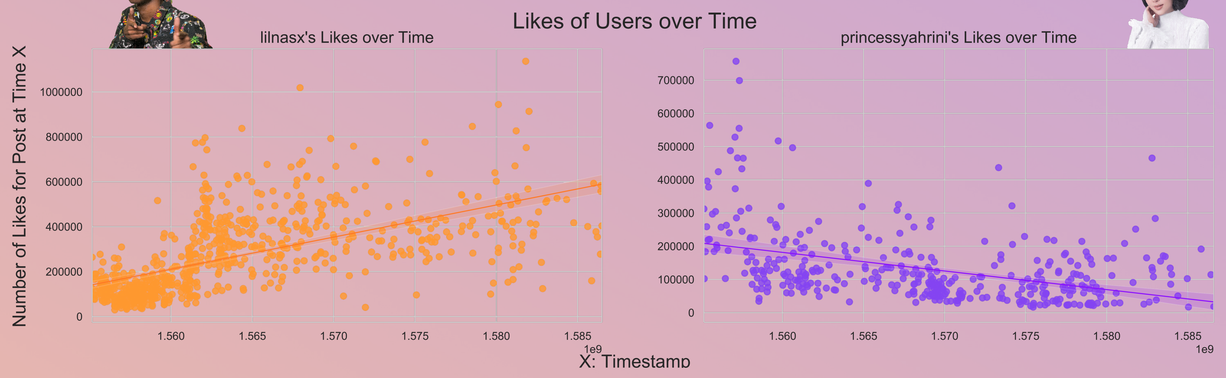


Figure 2: These regressions illuminate trends in likes over time. Positive (lilnasx), negative (prinessyahrini), and neutral (not depicted) trajectories were discovered for individual users, motivating the addition of "temporal likes" as a feature in our model.

3. We observed a higher MSE and lower R-squared when including NLP features in the model.

Model	Test RMSE	R-squared
Baseline	287967.1590	0.6042
Temporal Likes Only	280061.7537	0.6256
Full Model w/o NLP	279767.9131	0.6264
Full Model (all features)	288843.1832	0.6018

Table 1: When using only the temporal likes feature, we outperform the baseline. Using all features but NLP ones slightly decreases RMSE and increases R-squared. Using all features (including NLP) does not outperform the baseline. We cannot determine whether these changes are significant

Model feat. Features

With an 80/20 train-test split, we maintained the same distribution of users across train and test sets. We trained our model using a ridge regression using a mix of scraped, extracted, and generated features, and evaluated its predictive power using the reported root mean-squared error and R-squared value. Below are examples of some of the features that were included:

Time of Post

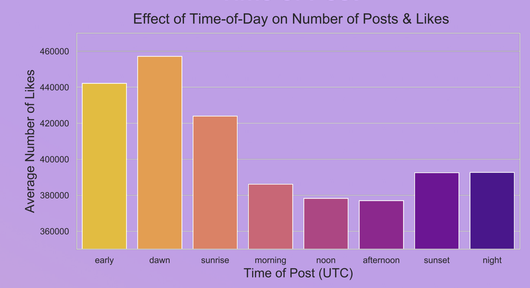


Figure 3: We believed that time of post and similarly extracted features could help predict likes received.

Caption Length

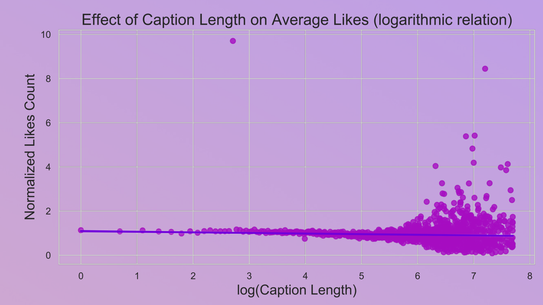


Figure 4: We used the caption length feature due to a visible polynomial relation and linear logarithmic relation.

Looking back, and Looking Ahead

In comparing our models, we found evidence to support the notion that the strongest feature for making accurate predictions was the predicted number of likes derived from users' individual trend lines (Fig. 2). Impactful features beyond this are difficult to pinpoint, and likely harder to generate, given the wide range of possible factors. We believe that a better predictive model would factor in more user-specific trends, or even predict for users individually, but this would require a lot more data to make it generalizable. With more time and data, we could try features more specific to the content of the post, whether it be image classification or more meaningful NLP features including sentiment analysis, as well as joining the data with news analysis that gives more information about our users' popularity in other media. If we could generalize this model to more diverse users, we could see it having applications in fields such as advertising and marketing, where influencers are paid to post product-related content on Instagram