Final Capstone
Susanna Vinoba

Final Report: Instagram Likes Calculator

Contents

1. Problem Statement

In the modern landscape of social media, understanding user engagement is vital for content creators, marketers, and social media managers. This project proposes the development of an Instagram Likes Calculator that analyzes various factors affecting post engagement. By considering media type (video, photo, album), caption classification, timestamp, number of tags, and user follower count, this tool aims to provide insights into the factors influencing likes on Instagram posts.

The scope of this project begins with ten sports clothing and apparel brands Nike, Adidas, Puma, New Balance, Vans, Skechers, Under Armour, Reebok, Champion and Converse. The success of this endeavor will be measured by a 20% increase in likes generated for each post by utilizing the calculator and then making tweaks to the post as needed such as posting at a different time or changing the type of post - from video to carousel.

2. The Data

The data was scraped from Instagram using the instagrapi package. The data was scraped August 2023. The features scraped include the username, follower count, post ID, media type, caption, likes, comments, timestamp, tags. A total of 100 posts were scraped from each of the ten accounts creating a dataframe of 1000 rows.

The feature set used for both EDA and the machine learning algorithm include the following features/ columns: follower count, media type, tag, clean captions, hour, day of week, and month. Note the only features that were scraped that were not used in EDA or the machine learning model were username, post id, and comments.

Post ID was not used as it is irrelevant. Username was not used as it would be collinear to follower count and follower count would be more predictive to unseen data than username would. Lastly comments were not used as comments are not a feature that can be controlled by the post creators. The comment count was included to add more insight in exploring the data if needed in the future.

Data cleaning:

The captions and time data were cleaned. A column called clean captions was created which tokenized every caption, removed punctuation, removed capital letters, and removed stop words.

The captions were also cleaned a secondary way through collocation. Wherever words that occurred frequently in the corpus together within 2 n-grams were grouped together with an underscore. The collocated captions were not used in modeling as EDA revealed that the tokenized words without the collocations had more stand out terms rather than having terms with very similar frequency.

The time data was cleaned by pulling out the hour, day of week and month from the timestamp data.

3. EDA and Statistical Analysis

Exploratory Data Analysis (EDA) was performed to gain insight into the dataset.

Captions:

The word cloud for the tokenized clean caption is as follows:

This shows the terms link and bio are used commonly. These make sense as it is common for posts to contain these to redirect users to a link in their bio. Other terms are alluding to campaigns such as a Skechers and Snoop Dogg collaboration.

Time of Posting:

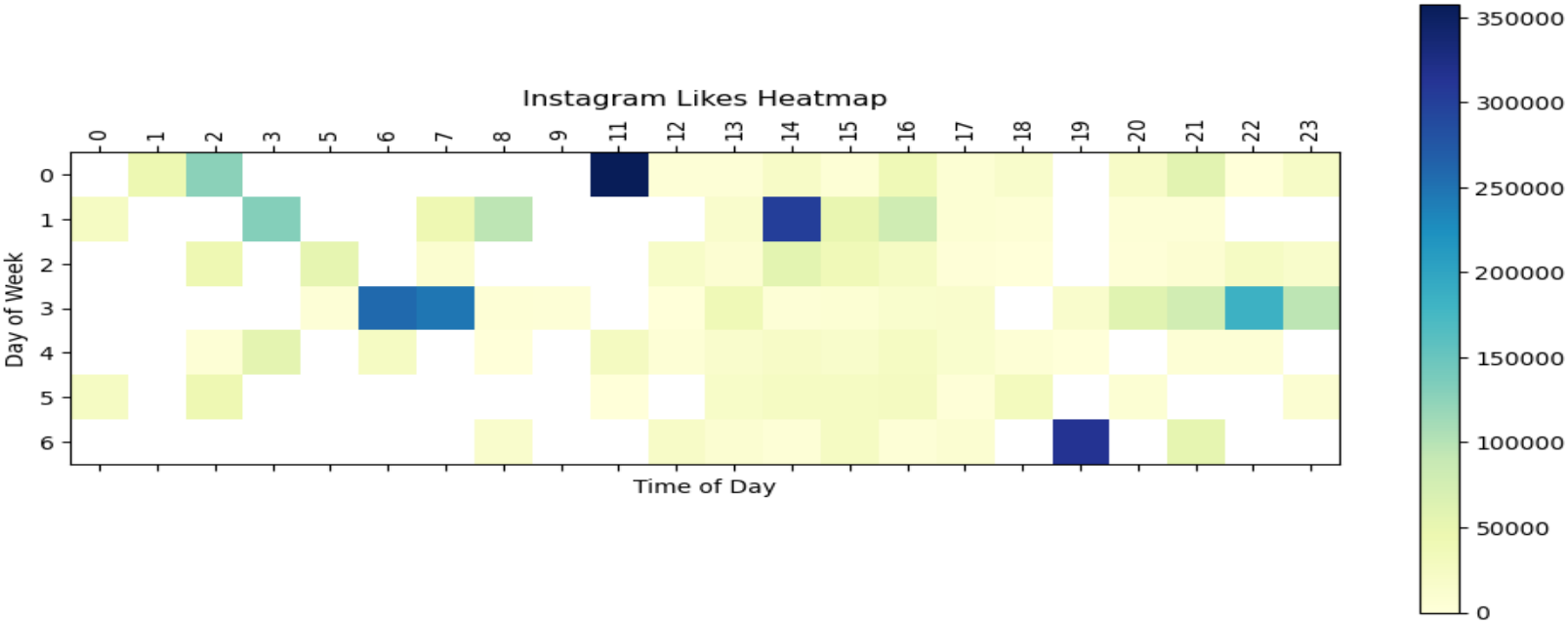
Please note outliers have been removed for time analysis to 3 standard deviations from the mean.

For the day of week the numbers represent the following days:

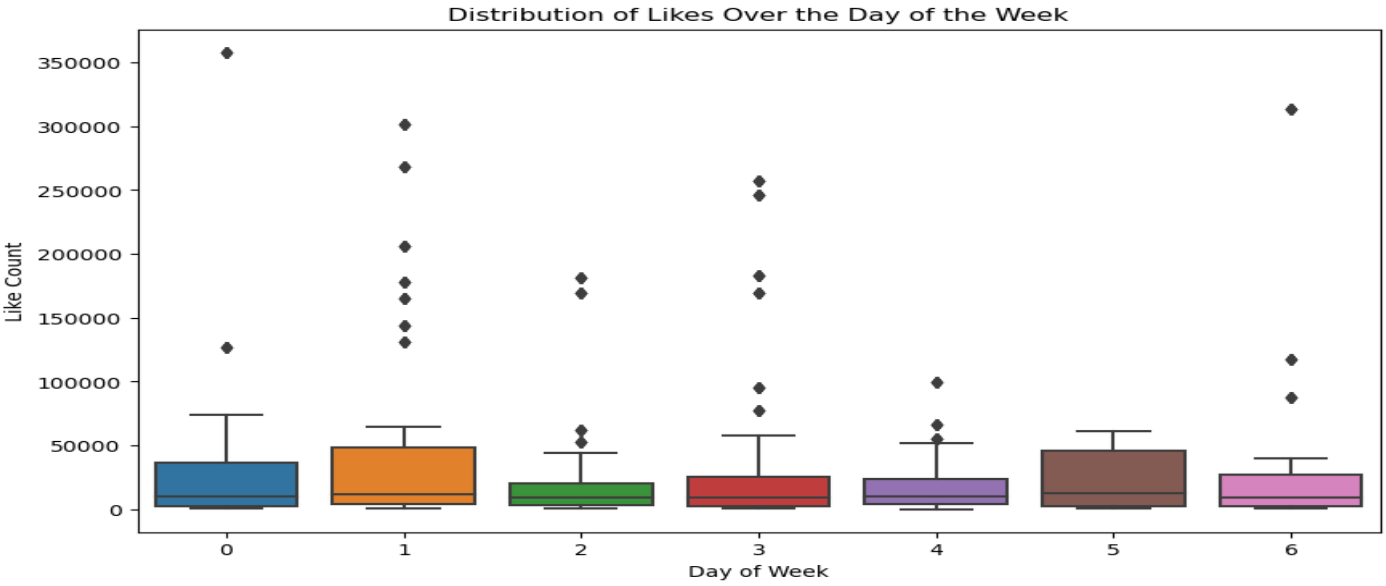
- Monday: 0
- Tuesday: 1
- Wednesday: 2
- Thursday: 3
- Friday: 4
- Saturday: 5
- Sunday: 6

For hour of day 0 corresponds to midnight (12:00 AM), 1 corresponds to 1:00 AM, and so on, up to 23 for 11:00 PM.

Below is a heat map of the distribution of likes over the week and hour of day:



Times of posting that show the consistency in the likes given are around mid day most likely around a lunch break and late in the day as people are going to bed in the EST time period. The best day to post seems to be Thursday either earlier in the day or later in the day. Some of these results can be explained by outliers even with the outlier removal since trending campaigns could have influenced these results.

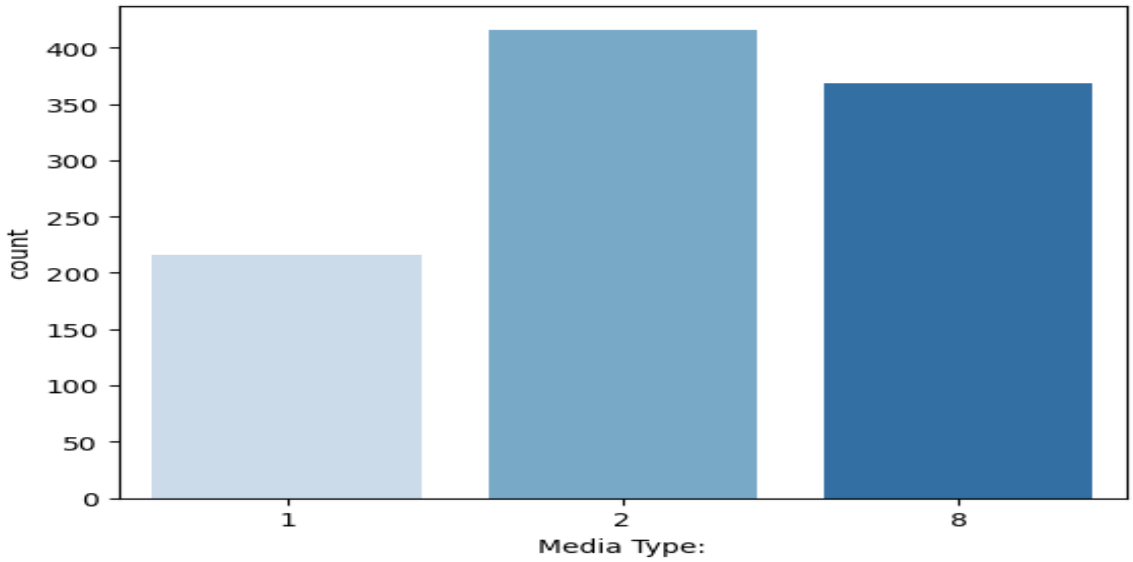


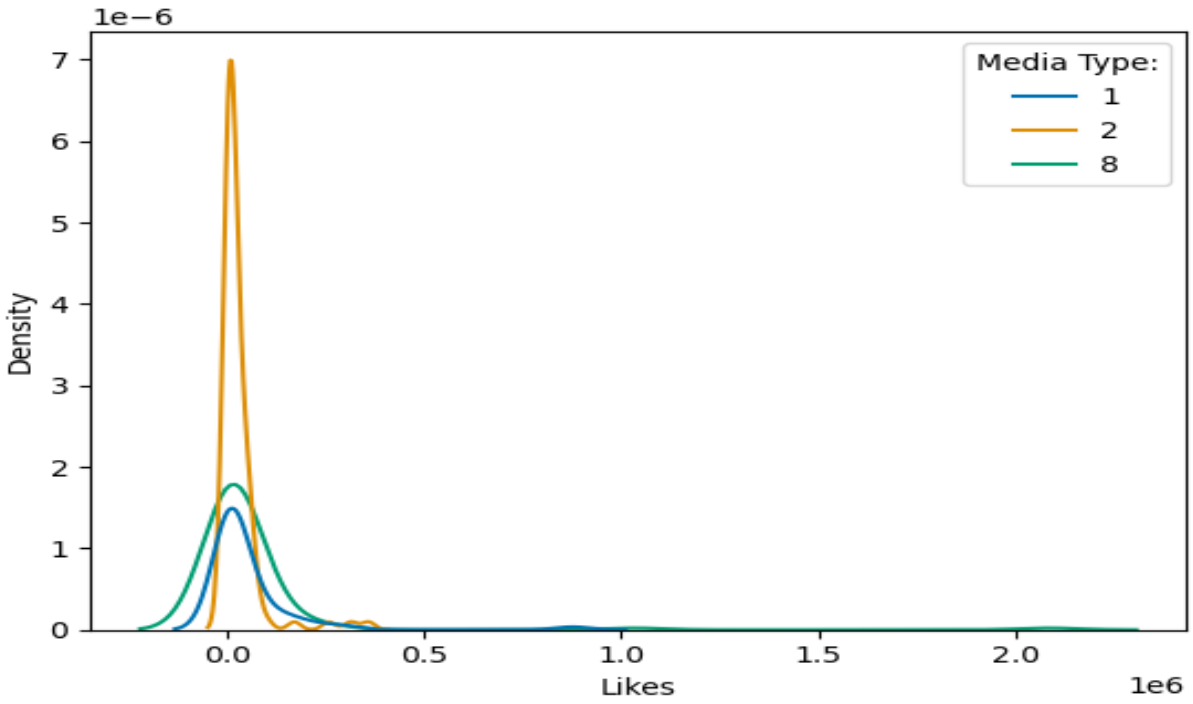
The distribution of likes shows that Tuesday is the best day to post and Saturdays follow that. Tuesday seems to be a day where posts tend to go viral as it captures plenty of outliers.

Categories:

Please note outliers have been removed to 3 standard deviations from the mean.

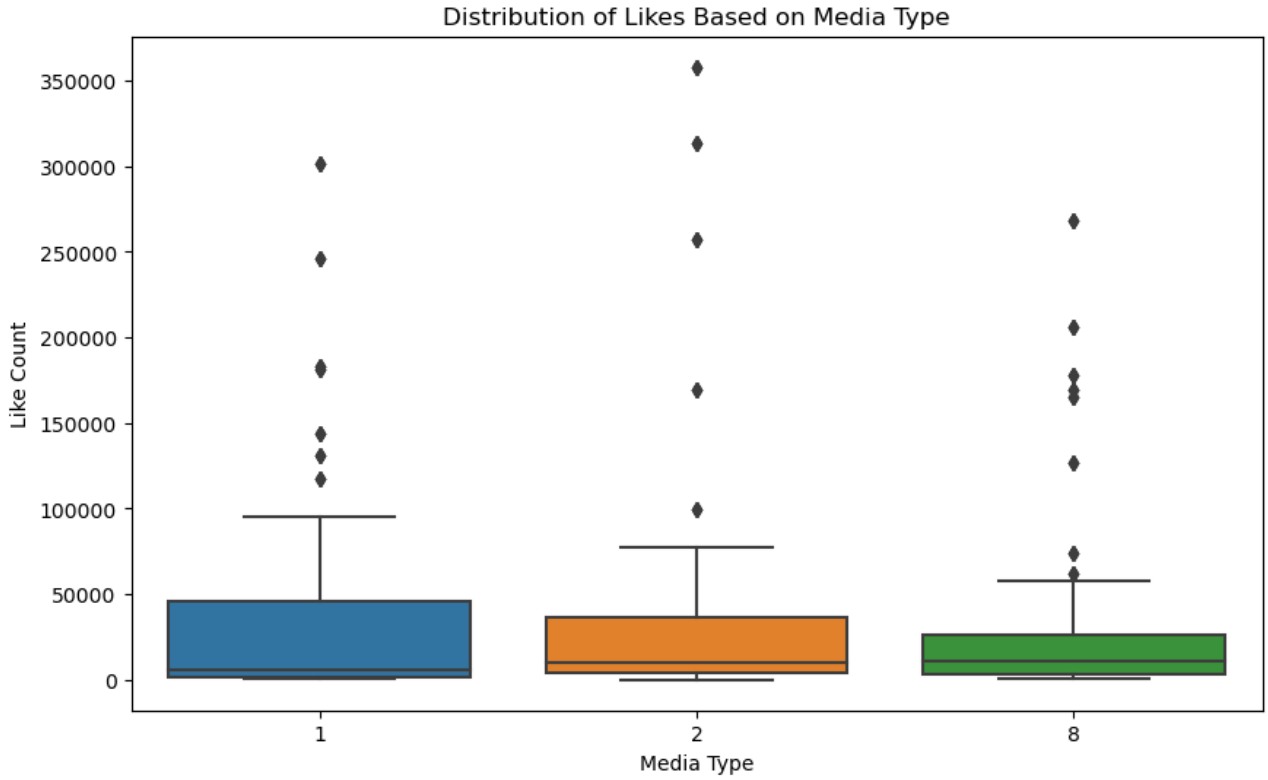
- Photo - When media_type=1
- Video - When media_type=2
- Album - When media_type=8





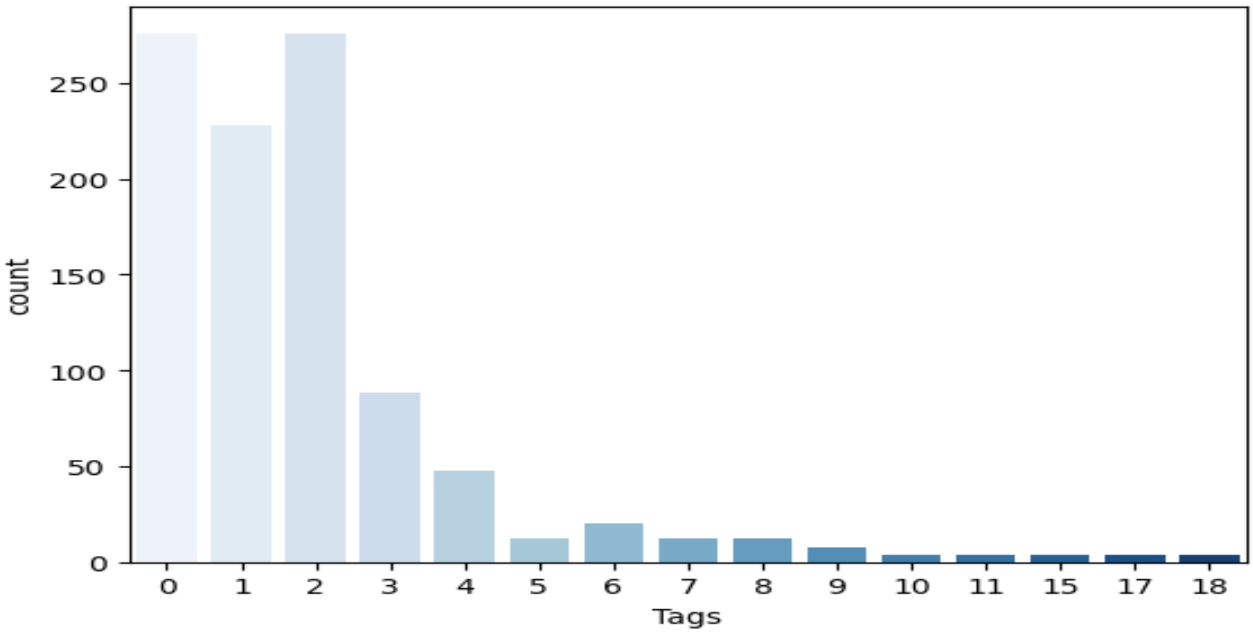
Photos and albums follow a similar distribution shape. The main difference is where there are more albums posted than photos and that is why it is taller. Photos start at a higher level of likes showing that they have less likelihood of performing poorly. Albums seem to have a longer tail meaning they have outliers that are performing extremely well.

Although I wouldn't have assumed it Videos are posted the most and they have a denser distribution meaning they perform pretty consistently. The bumps at the right tail are showing the videos that performed beyond the expected (positive outliers).

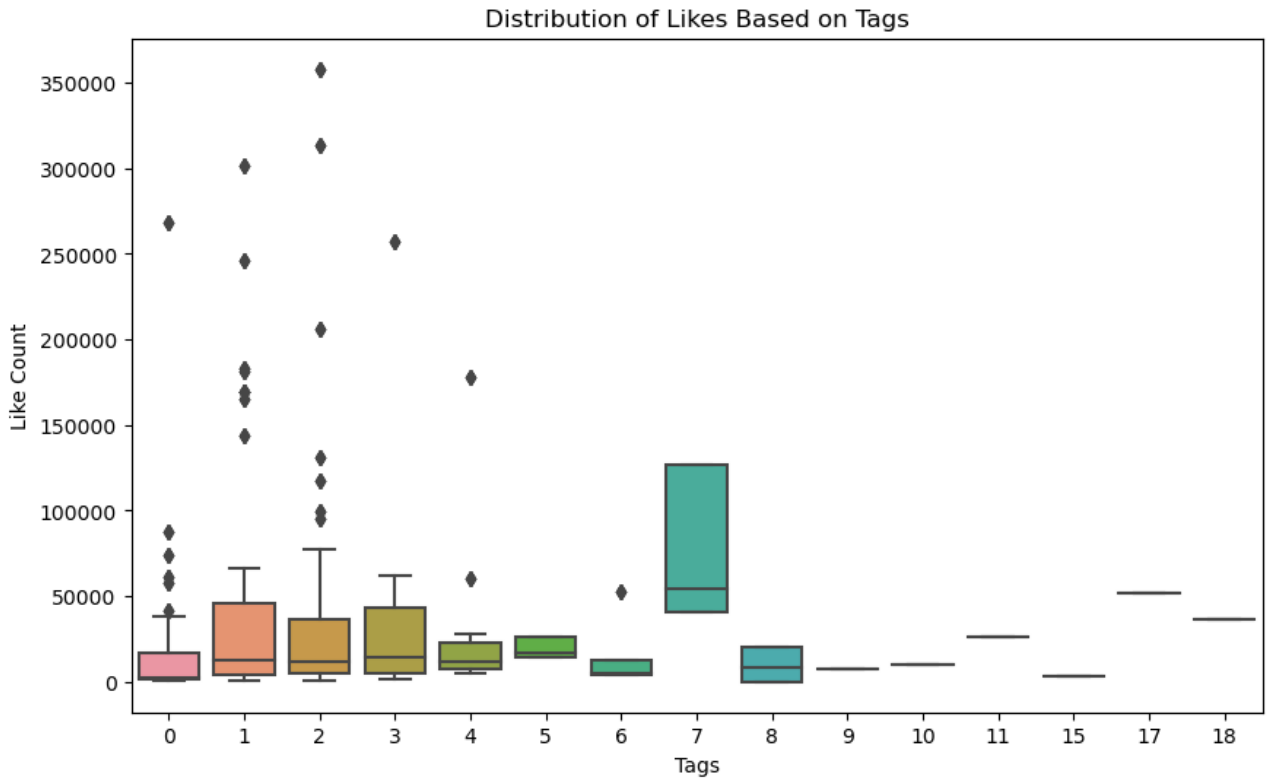


Tags:

Please note outliers have been removed to 3 standard deviations from the mean.



The most common number of tags on a post are between 0-2 and the count of posts reduces as there are more tags.



The most interesting thing here is wherever 7 tags seem to perform well for the less than 50 posts. After checking the data the seven people who were tagged were a part of campaigns. Other than that, one and three perform similarly and two has a slightly longer tail with more outliers.

4. Application of Machine Learning

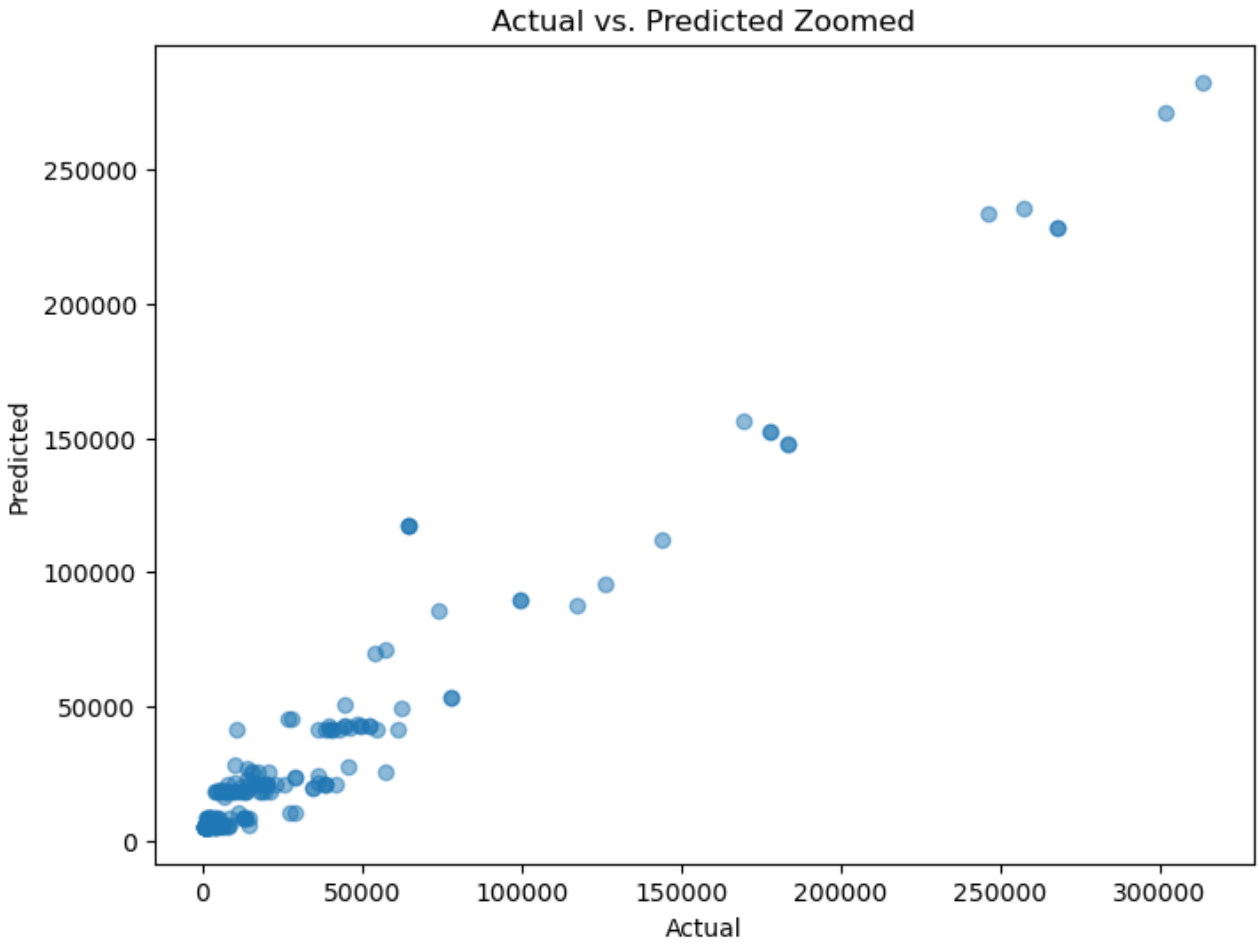
Model selection: Regression boosting was chosen as the model. The target variable is 'Likes,' which is a numerical value. Predicting a continuous numerical value is a regression task. Regression boosting models can also effectively handle a mix of feature types, offering a robust solution for predicting likes accurately.

For caption processing count vectorizer was used instead of TFIDF as the length of the captions is so short that repeated words would still have meaning. For example link and bio although it can be used in various campaigns it is still helpful to know that if people are lead to a link in a bio what the performance of the post is.

For tags and follower count although they could have been categorical features seeing as follower count had 10 unique values and tags 18 a standard scaler was used instead so the model can better interpret unseen/ new data that fall just outside the scale such as 20 tags or 16.

Model:

Please note outliers have been removed to 3 standard deviations from the mean (anything above 400,000 likes were removed)



The model showed great accuracy especially with the posts that do well.

Feature importance (top 15):

1. caption__pumaxdua: 0.18343067070373784
2. caption__lorenzoposocco: 0.10013655522916509
3. caption__lookbook: 0.08900465236821718
4. caption__red: 0.07669657615410763
5. caption__jennaortega: 0.07374358563978418
6. caption__theo123456: 0.07294197496597013
7. caption__violets: 0.05877856504258899
8. caption__sneak: 0.05310462225649454
9. caption__shot: 0.042723294009601565
10. caption__offwhite: 0.03284812400604805
11. remainder__month: 0.026665220278516505
12. remainder__dayofweek: 0.02557527074510758
13. caption__classics: 0.015211053759076599
14. numeric__Follower count:: 0.012524641083407223
15. caption__wegotnow: 0.011822709330479513

The model had an easy time predicting the likes for posts that were a part of influencer campaigns or that featured influencers such as the Dua Lupa campaign with Puma. The other words that are not an influencer are repeatedly used in captions as part of the influencer campaign. Surprisingly month had a pretty high feature importance. It may be possible that use of apps change with the

weather. Day of week makes sense as during working days there will be different usage than the weekends. Follower count also shows up, I would have assumed it would have more importance but the model recognizing a campaign seems to have precedence.

5. Conclusions and Future Work

Conclusions: A model was made to estimate the number of likes a post will receive based off of features such as the caption, time of posting, type of post, and number of tags.

Future work: There is plenty of future work that can be done with this data. Work can be done around the comments field to understand the important features that will generate comments and see if comments are directly related to the number of likes as it is assumed. Also, work can be done into influencer marketing across various brands to see which influencers have the greatest effect on engagement. Lastly this project was only focused on the sports clothing industry. This study can be expanded to other industries to create a more comprehensive Instagram likes calculator.