Instagram Likes Calculator

Problem statement

Background: In the modern landscape of social media, understanding user engagement is vital.We propose an Instagram Likes Calculator to analyze factors affecting post engagement. Aims to provide insights into factors influencing Instagram post likes.

Features: media type, captions, timestamp, tags, and follower count.

Scope: 10 sports clothing brands

Success measured by a 20% increase in likes per post after analysis with the calculator.

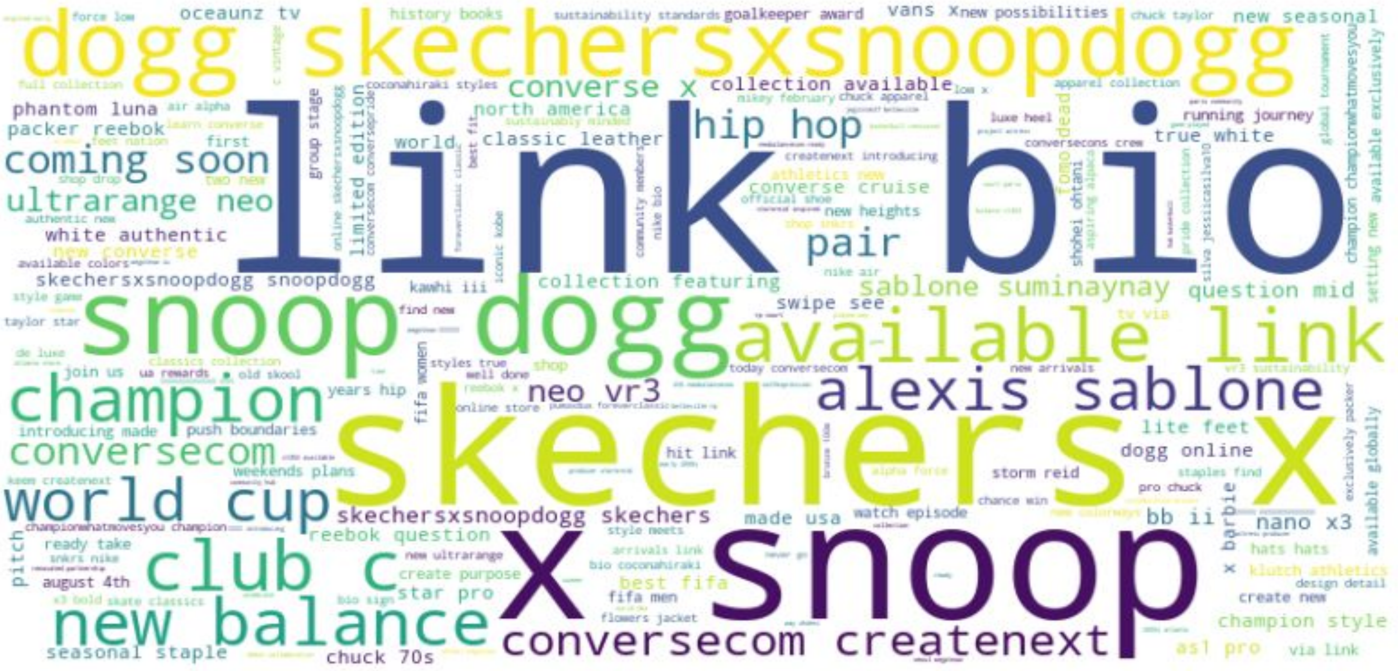
The data

- Data was scraped using instagrapi in August 2023.
- 100 posts from each of 10 accounts, creating a 1000-row dataframe.
- Features: username, follower count, post ID, media type, caption, likes, comments, timestamp, tags.
- Features used in modelling: follower count, media type, tags, clean captions, hour, day of week, and month.
- Data cleaning:
    a. Cleaned captions
    b. extracted time components.

EDA: Captions

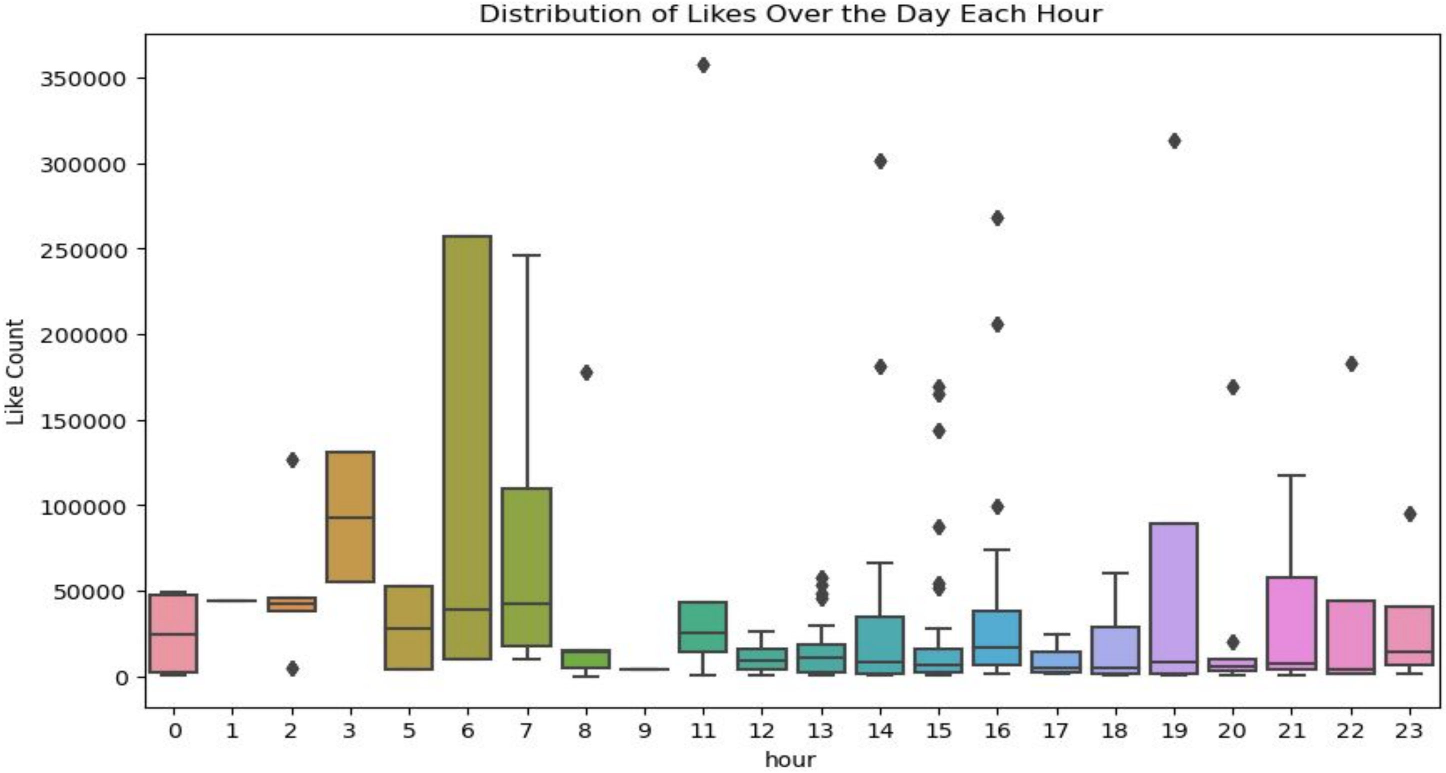
This shows the terms link and bio are used commonly. These make sense as it is common for posts to contain these to redirect users to a link in their bio.

Other terms are alluding to campaigns such as a Skechers and Snoop Dogg collaboration.



EDA: Hour

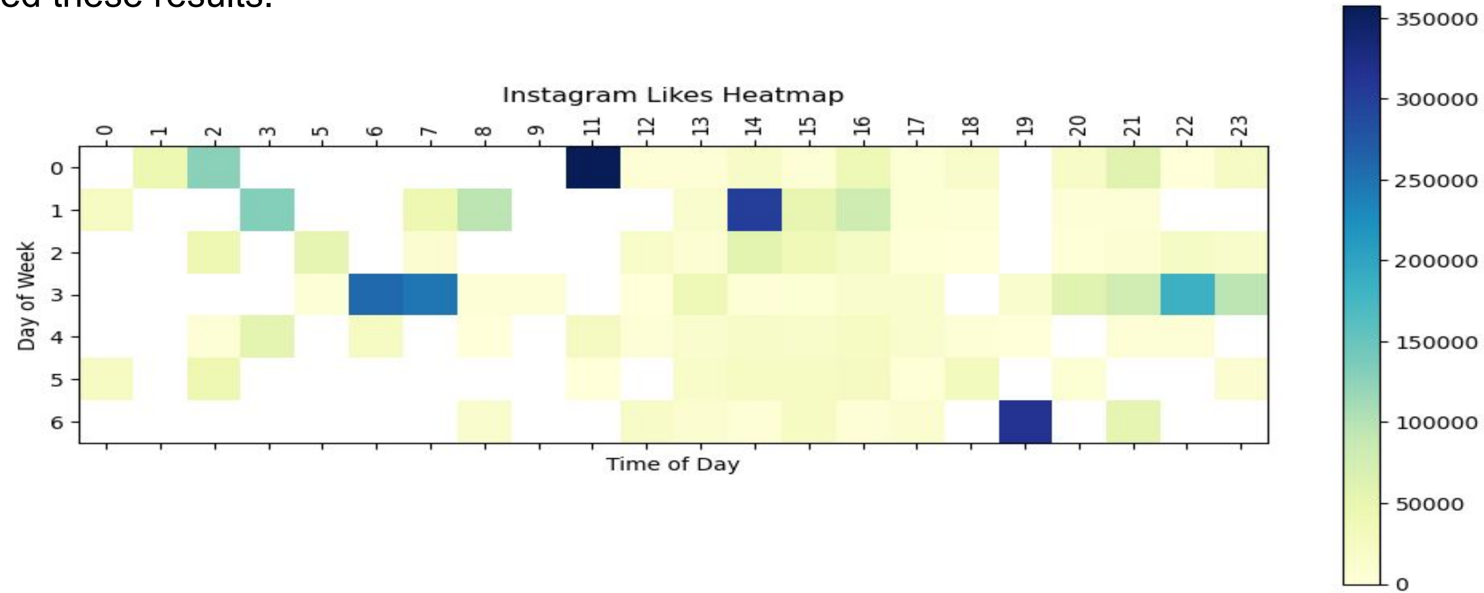
The distribution shows that 3am, 6am and 7am are the best times to post



EDA: Hour and week

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.

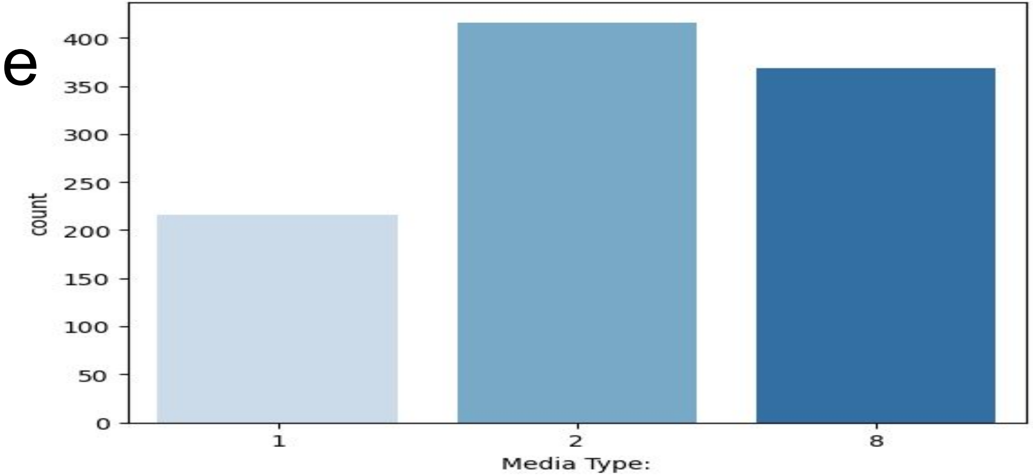


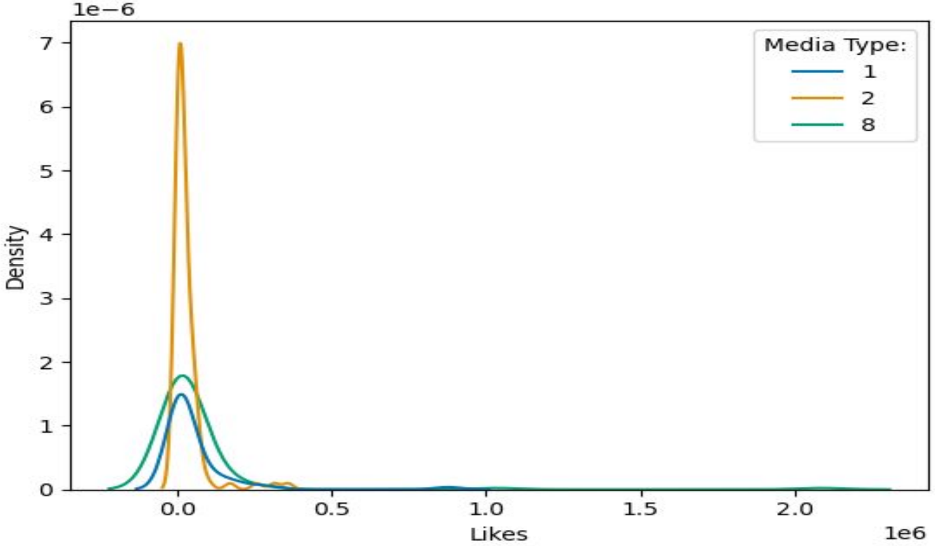
EDA: Categories/ Post type

- 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.

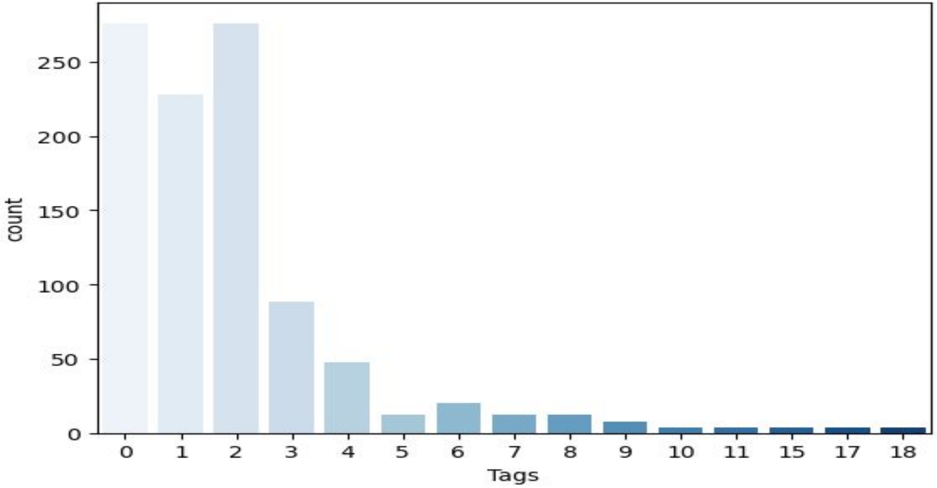
Videos are posted the most and they have a taller distribution meaning they perform pretty consistently. The bumps at the right tail are showing the videos that performed beyond the expected (positive outliers).

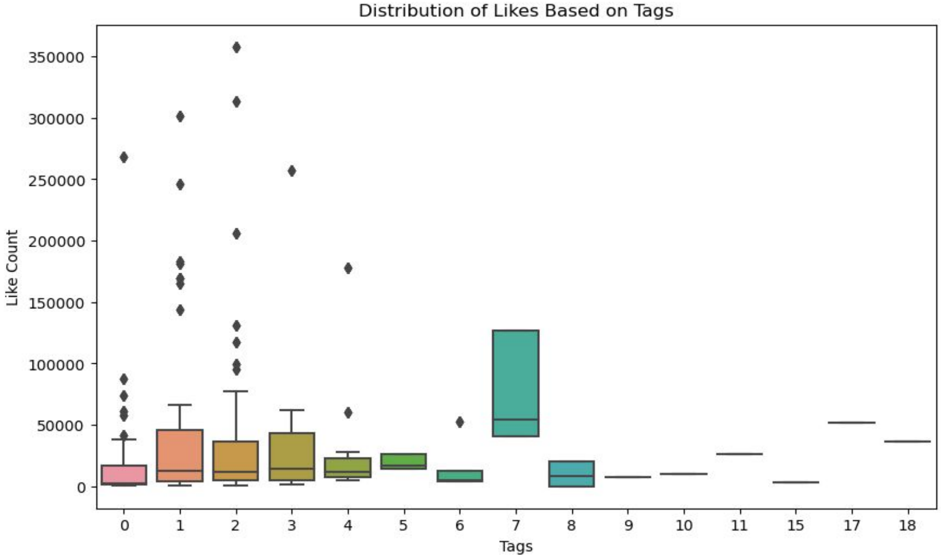




EDA: Tags

The most interesting tag count is 7 tags. It seems 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.





Model selection

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.

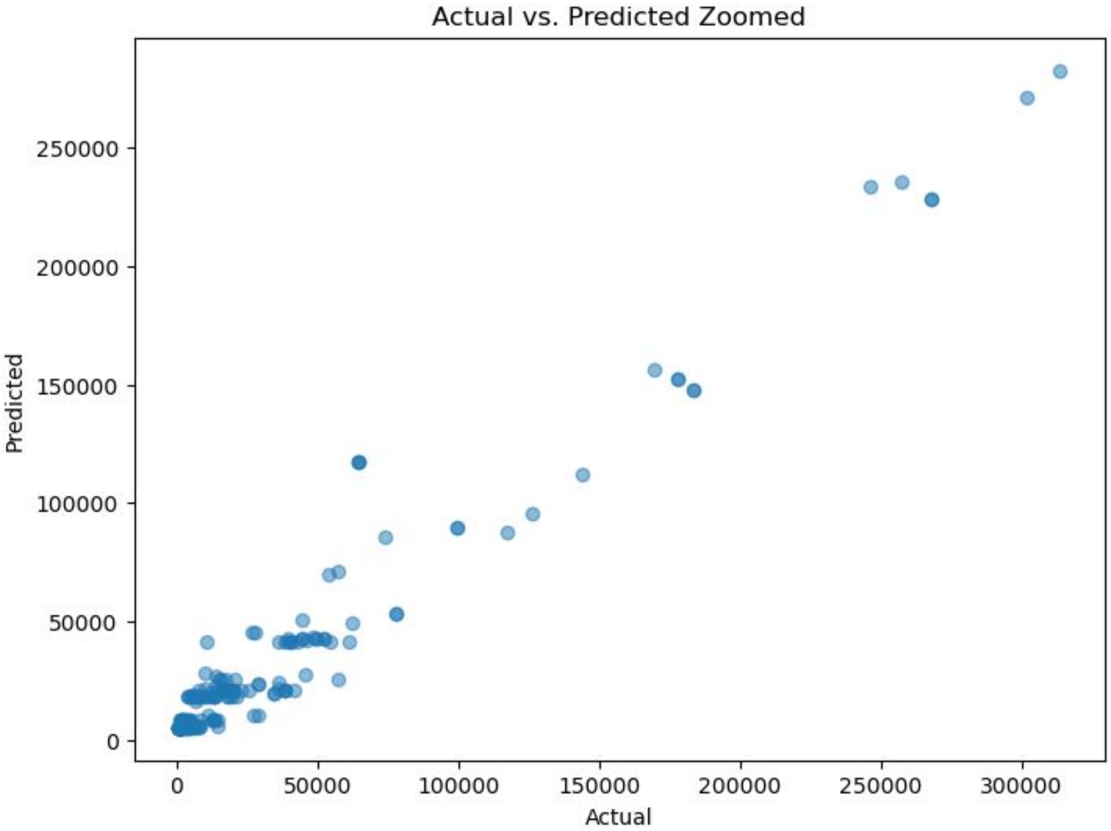
Preprocessing

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 performance

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



Feature importance

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: 0.025575270745107 58
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 if the model recognizes a campaign it seems to have precedence over the follower count.

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.