Analytics in Style: Predicting Instagram Likes and Clickbait Posts

MQM Team A09: *Amy Chen, Hao Cheng, Mike Liu, Taneesha Mahajan, Yichi Zhang*

I. Business Understanding

Instagram has emerged as a valuable social media platform for fashion brands where they can engage with their consumers directly. The platform's emphasis on visual content pairs well with fashion and style, so Instagram has naturally explored into becoming a product discovery option for style and clothing. The business problem we are addressing is how to improve user engagement on Instagram, particularly in the context of fashion brands. A very simple yet effective engagement measure is the number of Likes a post can generate. Instagram's core business, advertising through sponsored posts, relies heavily on user engagement. By predicting the number of Likes for sponsored posts, Instagram can gain insight into how to work with brand partners to structure the posts in order to attract positive engagement for both the advertisement partner as well as the platform itself.

Clickbait posts are an issue that all social media platforms are tackling. Although they can stimulate user engagement to a certain extent, they negatively affect the overall user experience and can interfere with the product discovery process on Instagram[1]. Therefore, we are interested in developing a predictive model to identify Clickbait posts. The model can help Instagram remove Clickbait posts that prevent users from finding the products they are interested in, thus improving user engagement as well as the recommendation algorithm and enabling Instagram to evolve into an ideal product search platform for Style and fashion.

[1] https://techcrunch.com/2019/04/10/instagram-borderline/

II. Data Preparation and Understanding

- **Data source**

The dataset in this analysis is from the Harvard database. It was created for an analysis that focused on classifying clickbait posts using machine learning algorithms and image recognition tools[2].

Datapoints are obtained from the fashion subset of Instagram, as this is an attractive platform for many clickbaiters. Each post is characterised according to various features such as Smile, Face, Logo, ProductOnly, etc. and also contains information about the number of Likes, Followers, and Hashtags. Lastly, posts are classified as whether or not they are a clickbait.

- **Data cleaning**

We removed columns containing variables of X1-X1024 (by products of image recognition) because they were not required in our analysis. We also removed the Hashtags and Captions columns because their lengths and other critical information were extracted into other columns. We reclassified the 62 brands in the 'SearchedTag' variable into four categories: Mega couture, Small couture, Designer and High street[3]. We only focused on posts with Likes less than 3000, as we considered the rest of them as outliers that comprised less than 1% of our total observations. We also removed observations with NA values.

- **Key variables explanation**

Clickbait post is a post that has an obvious discrepancy between the post's content, captions, and hashtags. In our dataset, whether a post is a Clickbait or not is manually labeled by the creators of the dataset.

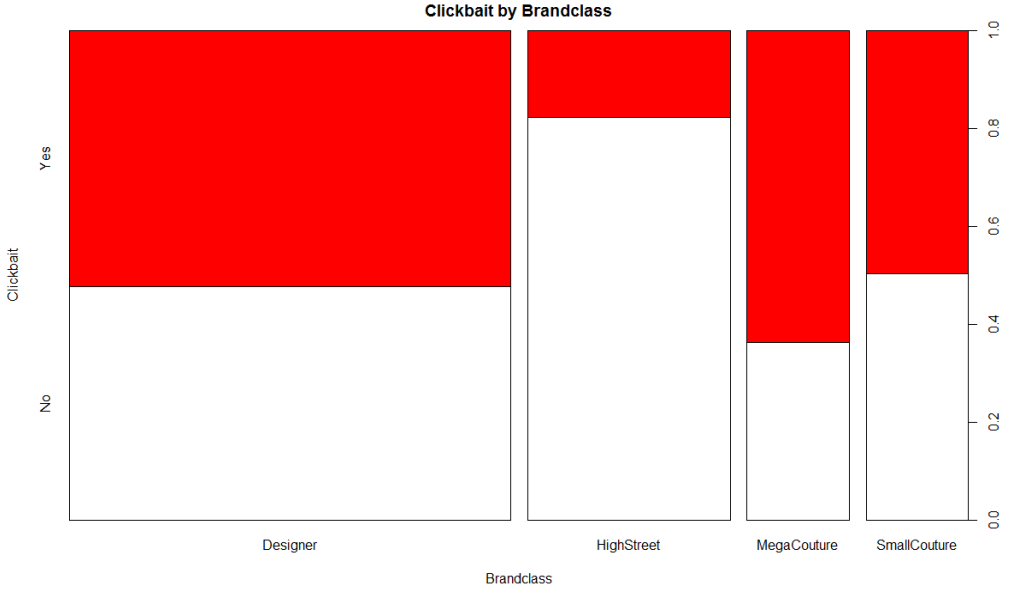
The last ten variables in this dataset show various aspects of the image features. The dataset providers analyzed Instagram images and quantified their features from 0-1 in respect to probability. Since we were not confident in deciding the threshold that these probabilities should be converted to 1s or 0s, we did not turn them into factors.

[2] http://www.sscnet.ucla.edu/comm/jjoo/web/icwsm18-clickbait-instagram.pdf.
[3] https://www.researchgate.net/publication/316098452_Fashion_Conversation_Data_on_Instagram

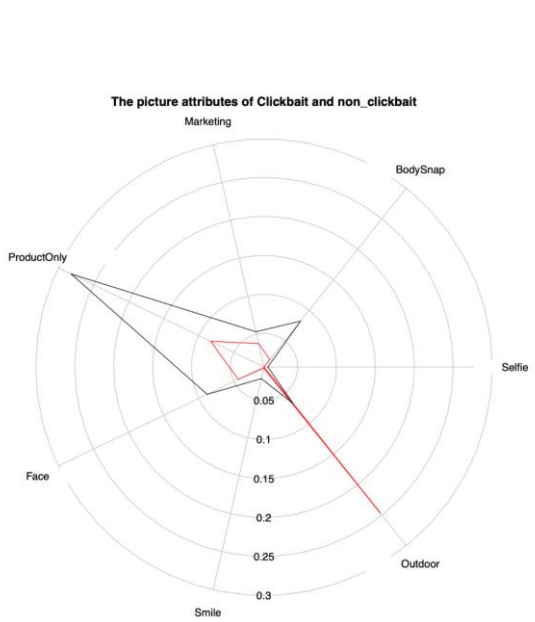
Selfie	Whether it contains a face that occupies 50% of the height
Body_snap	Whether it contains any body part
Marketing	Whether it contains runway or ceremony scenes
Product only	Whether it contains products without people
Non fashion	Whether it missing fashion products
Face	Whether it contains frontal or side faces
Logo	Whether it contains any brand logo
Brand Logo	Whether it contains specific brand logo
Smile	Whether it contains any smiling faces
Outdoor	Whether it contains outdoor background

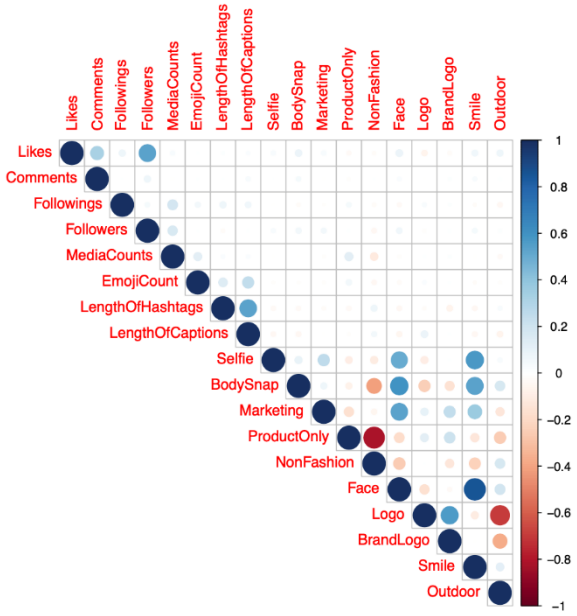
- **Exploratory Analysis**

We first explored the average number of clickbait posts associated with each fashion brand class. MegaCouture has the highest clickbait rate because it is a common tactic for clickbait posts to include luxury brand hashtags to increase exposure through the recommendation algorithm. On the other hand, high-street brands have relatively low clickbait rate since their brand popularity is lower.



We made a radial plot to better understand the visual attributes of both clickbait and non-clickbait posts. Clickbait posts have a high likelihood to show pictures with only Products, Body pictures, and Faces. Non-clickbait posts are more likely to have Outdoor pictures.





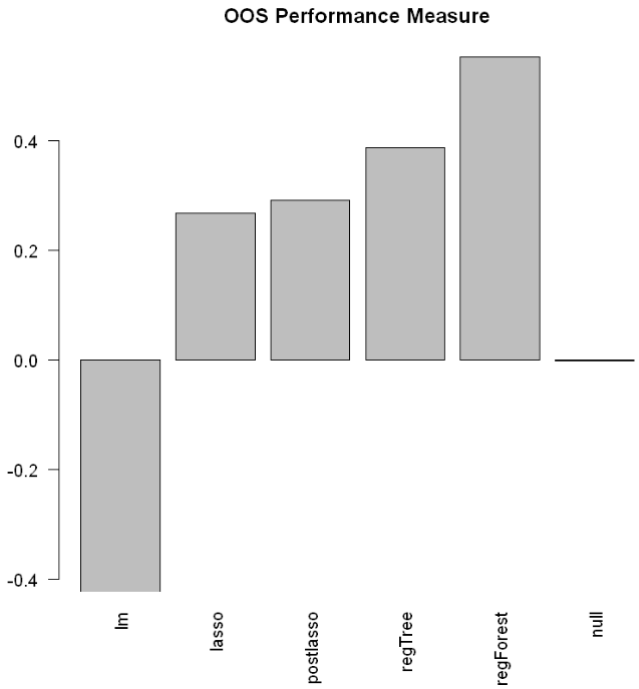
Correlation matrix is a good way to understand the relationship between different variables. As shown above, Likes has a high positive correlation with the number of followers. This explains why fashion brands tend to involve influencers with massive followings in their promotion campaigns. Posts that include a lot of hashtags also tend to have long captions. In terms of pictures attributes, we can find certain attributes are associated with each other. For example, Selfie pictures are more likely to have a smile and Outdoor photos are less likely to have brand logos in them.

III. Modelling & Evaluation

- **Predicting Likes**

We utilized various models to predict the number of Likes for a post, including linear regression, lasso regression, post-lasso linear regression, regression tree, and random forest, with the first three including interactions. We first split the data randomly into a train and test sample, then performed a 10-fold cross validation of all models on the train sample. We evaluated the models based on out-of-sample R^2. With a negative OOS R^2, the simple linear model was very prone to overfitting

while the best model was the random forest with OOS R^2 of 0.53, meaning that the random forest model can accurately explain 53% of the variation in Instagram Likes. While it is not very high, it does offer a significant increase over the baseline null model (average likes per post across the dataset) or even other models.



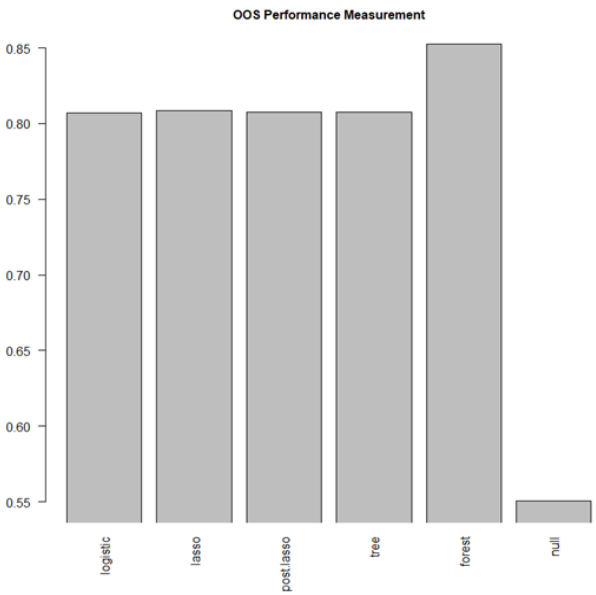
The post-lasso linear regression has a lower R^2 (0.29) but has better interpretability compared to the random forest model. Out of all the variables, lasso selected only 8 useful ones, which are all related to Followers, Comments, Location Existence. The highest coefficient is an interaction between Comments and Marketing, which shows that having comments and having a runway or ceremony scene in the photo can generate a lot of likes. Brands can utilize this information to photograph new products in runway settings to attract more engagement.

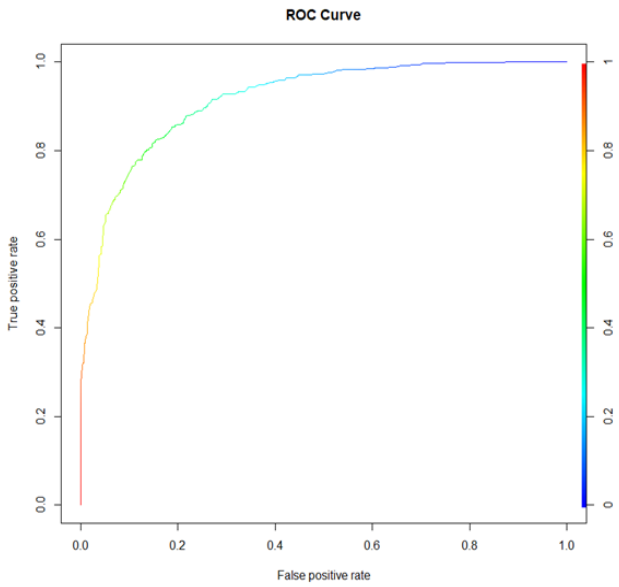
R^2 is chosen as the model evaluation metric because advertisement posts often rely on the number of likes as a fundamental engagement metric and both Instagram, as well as partnered brands, care about how well the model can predict the number for promoted posts. That being said, only around half of the variation in Likes is captured by the model, so the prediction will likely be very different from the actual number of likes. More variables will have to be included to improve the

performance of the model, for example, more specific content of the picture, frequency of posts, or whether the post contains a product link.

- **Predicting Clickbaits**

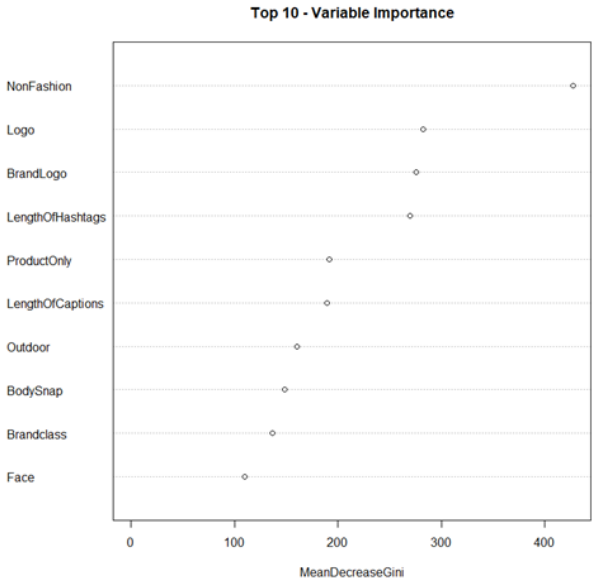
To predict if posts are Clickbait posts or not, we utilized classification tools, such as logistic regression, lasso regression, post-lasso logistic regression, classification tree, and random forest models. The best model, random forest, was found based on the accuracy of predicting 1s or 0s based on a threshold of 0.5 and through the 10-fold cross validation technique used previously. Compared to the null model (average number of Clickbait posts in the dataset), the random forest model greatly improves predictive capabilities. Afterwards, we further tested it by using it to predict Clickbaits in the test sample. By plotting a ROC curve that calculated all of the thresholds, we found the best threshold that generated the highest TPR with the lowest FPR was approximately 0.7 (seen from the curve below). The area under the curve was 0.92, which is quite high and a slight improvement to the 0.88 AUC under the logistic model. Therefore, the random forest model is very good at correctly predicting clickbait posts.





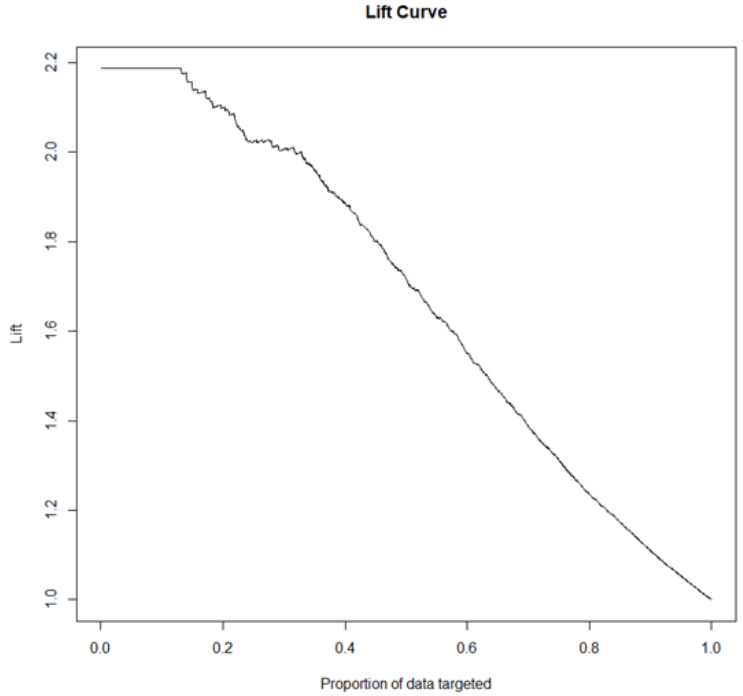
The random forest model provides the most accurate predictions, but a major flaw is that it cannot provide insight into whether a variable is positively or negatively correlated with the

dependent variable. However, it does show which variables are important for prediction. The following top-10 variable graph shows the most important variables used for predicting clickbaits, with a higher MeanDecreaseGini representing greater importance. It is evident that image data (NonFashion, BrandLogo, Logo, Outdoor, Face) and length of hashtag/captions are very helpful for classifying. This echoes our Instagram experience that clickbait posts often include many hashtags/captions in order to cheat the algorithm. The logistic regression also shows that these factors have the largest coefficients: NonFashion has the largest negative coefficient while Logo has the largest positive coefficient.



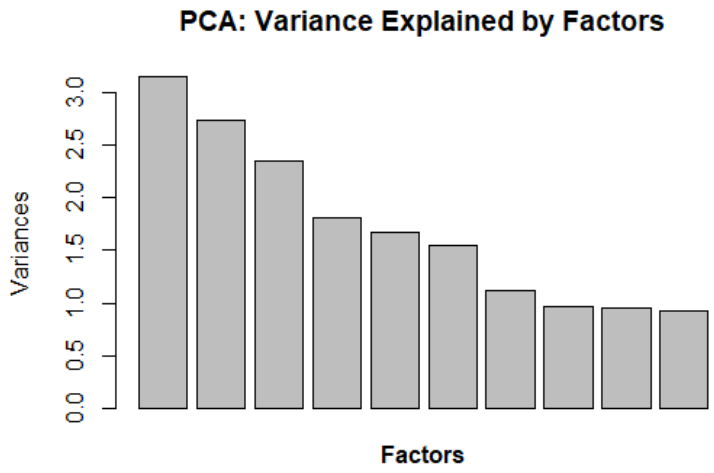
Accuracy is chosen as the evaluation measure because it assesses the risk that incorrectly labelling posts as Clickbait can cause the post to stay hidden to other users. It can potentially damage the poster's income if the post is a paid advertisement that relies on engagement metrics. It will also be a frustrating user experience that can lead to lower engagement on the platform or even abandonment altogether. Considering the pros and cons, we believe the model should prioritize having a low FPR until the model is validated against more data. Therefore, we recommend choosing a threshold equal to or higher than the 0.7 determined from the ROC curve to satisfy this condition.

The following lift curve indicates how the random forest model predictions compared to random guessing predictions given a certain portion of samples. For instance, when targeting approximately 35% of our test set, our random forest model performs twice as good as random guessing.



By being able to classify clickbait posts automatically, Instagram's algorithm can proactively adjust the likelihood of recommending such posts to other users prior to the deterioration of user engagement. Additionally, by providing more accurate and useful search results, Instagram can evolve into a better product discovery and product search platform in the fashion space.

In addition to the random forest model, we observed from the PCA clustering output several factors that posts with a high clickbait component score exhibit. The 1st component is high in Face, Smile, BodySnap, and Selfie variables, interpreted as posts with a face or body picture. The 2nd component is high in ProductOnly and BrandLogo, low in Outdoor and NonFashion, interpreted as fashion products that highlight the brand logo. The 3rd component is high in Top100HashOfInsta, ProductOnly, LengthOfHashtags and low in NonFashion, interpreted as excessive use of irrelevant hashtags that are not associated with the product in the picture.



IV. Deployment

After implementing the likes model, Instagram can better price their advertisement services depending on the number of Likes a promoted post is expected to generate. It can also add a new service that advises brands on how to structure posts to attract more user engagement. Click-through-rate will be a key metric for promoted posts since they often contain product links. If more likes on promoted posts lead to higher click-through-rate, then accurately predicting likes will greatly improve Instagram's ability to generate advertising revenue from fashion brands. However, as mentioned before, more variables should be included in the proposed model to more accurately predict the number of likes.

After implementing the clickbait model and removing clickbait posts from the recommended explore section, Instagram can measure the average time spent looking at a post to quantify whether user engagement has increased. Since clickbait posts are often misleading, users are likely to quickly stop looking at that post. Useful posts, on the other hand, invite users to stay for longer and more likely to lead to interest and even purchase.

As discussed previously, incorrectly labelling clickbait posts could damage the social network users have built or even their monetization abilities. Therefore, Instagram should be more careful when choosing thresholds for the model. We advise that they start with a small batch of style posts while tracking the accuracy of predictions, and as more data is collected update the model and slowly

ramp up the percentage of posts the model examines. Since users cannot see the algorithm and attribute any decrease in engagement directly to the recommendation algorithm, minor mistakes are likely to go unnoticed so Instagram has a significant error of margin when implementing the model.

There are some limitations to our analysis. Firstly, we somewhat disagreed with the method to determine whether the posts are clickbait or not based on going into the posts and judging it ourselves. But for the purposes of this analysis, we assumed they have all been accurately determined. Secondly, Likes is currently the best approximation for user engagement, but it is not the most ideal metric. Comments and Shares (not included in the dataset) require more effort and commitment on the part of users, making them higher quality forms of engagement. These metrics should not be weighted the same during engagement calculations. Additionally, there are also talks of Instagram removing Likes completely from the platform, at which point other metrics will be prioritized to determine user engagement. Until such time comes, Likes will remain at the core of user engagement on Instagram. Thirdly, extensive use of clickbait posts will likely cause a decline in followers, but this dataset does not contain that information. With time series data of users, we can better understand the impact of clickbaits on sustained user engagement.

V. Conclusion

Instagram's attempt at becoming a product discovery channel is a natural extension of its business. In order for its users to treat it as such, the recommendation algorithm must eliminate the irrelevant clickbait posts and offer the most helpful and useful content. Our model for predicting Likes serves as a good tool to help the company partner with brands and improve its promoted posts capabilities. Our model for predicting clickbaits can help Instagram remove the insignificant content that harms user experience. Despite limitations of the dataset, the predictive models provide significant value that Instagram can benefit from upon implementation.