INDEX

1) Introduction

In the realm of social media marketing, understanding how far your content reaches is paramount for evaluating its effectiveness. Instagram, a leading platform for visual content sharing, provides a wealth of data ripe for analysis to gauge post-performance. This project is centered around conducting a thorough examination of Instagram reach using Python programming language.

With an active user base exceeding a billion, Instagram stands as a formidable tool for brands, influencers, and individuals to engage with their audience. Yet, merely posting content isn't sufficient; comprehending the extent of post reach and engagement is pivotal in devising a successful social media strategy.

The primary aim of this project is to dive into Instagram's API data and scrutinize various metrics to evaluate post reach. These metrics may encompass likes, comments, shares, impressions, and reach, among others. By leveraging Python, a versatile and widely utilized programming language, we can automate the data collection and analysis processes, rendering them more efficient and scalable.

Through this analysis, we endeavor to address key queries such as: What content resonates most with our audience? When is the optimal time for posting to maximize reach and engagement? Are there discernible patterns or trends in our audience's behavior that we can exploit to refine our content strategy?

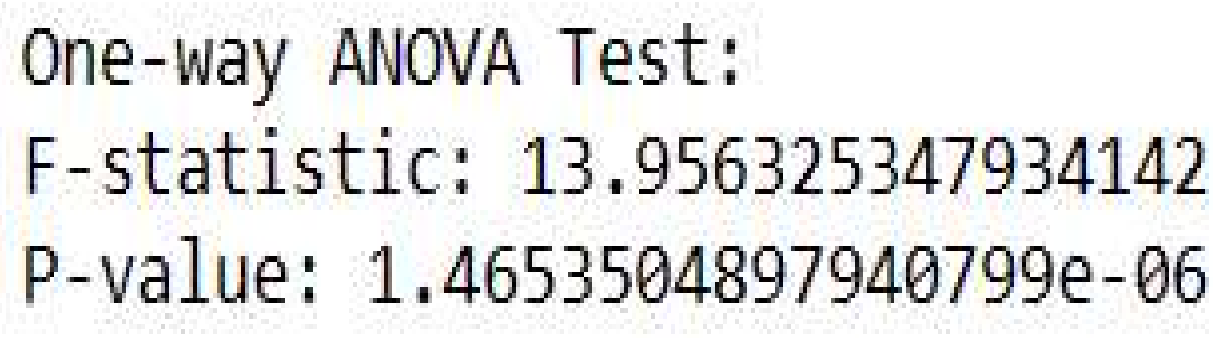
Upon project completion, we anticipate not only garnering valuable insights into our Instagram reach but also cultivating practical skills in Python-based data manipulation, visualization, and interpretation. These acquired proficiencies are transferrable not only to Instagram but also to other social media platforms and data analytics endeavors, empowering us to make informed decisions and elevate outcomes in our digital marketing endeavors.

2) Description about the data

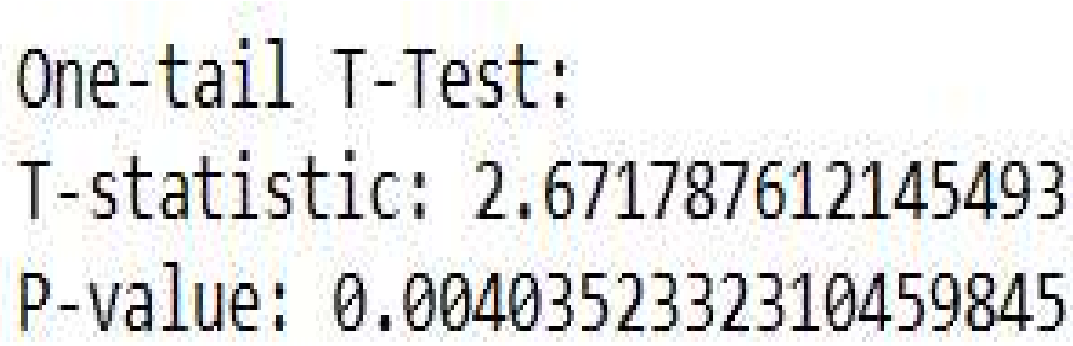
- Each row of the spreadsheet explains data for each individual posts.
- Impressions: The total number of times a post has been viewed.
- From Home: Number of impressions from users' home feed.
- From Hashtags: Number of impressions from users discovering the post through hashtags.
- From Explore: Number of impressions from users discovering the post through the Explore page.
- From Other: Impressions from sources not explicitly categorized (e.g., external shares or direct links).
- Saves: The number of times users saved the post.
- Comments: Number of comments on the post.
- Shares: Number of times users shared the post.
- Likes: The total number of likes received on the post.
- Profile Visits: The number of visits to the profile associated with the post.
- Follows: The number of new followers gained through the post.
- Caption: Text accompanying the post, providing context or additional information.
- Hashtags: List of hashtags associated with the post, contributing to its discoverability.

https://drive.google.com/drive/folders/1frRumz8gMnp-JcrERGIHQ2Zysq6ARWbM
(This drive consists of the excel data file and the code)

3) One- way ANOVA and One-tail T-test

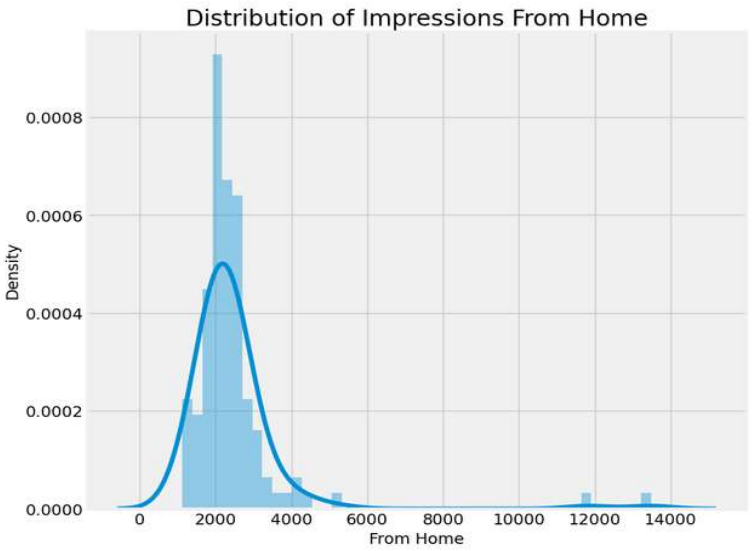


1) We have tried to find out whether there is a significant difference between the reach through home feed recommendations, Hashtags, and explorations
2) The null hypothesis for this is “There is no significant difference between the Instagram reach through the above ways”
3) Mean of impressions from home feed recommendations, hashtags and exploration are statistically different from the above test.

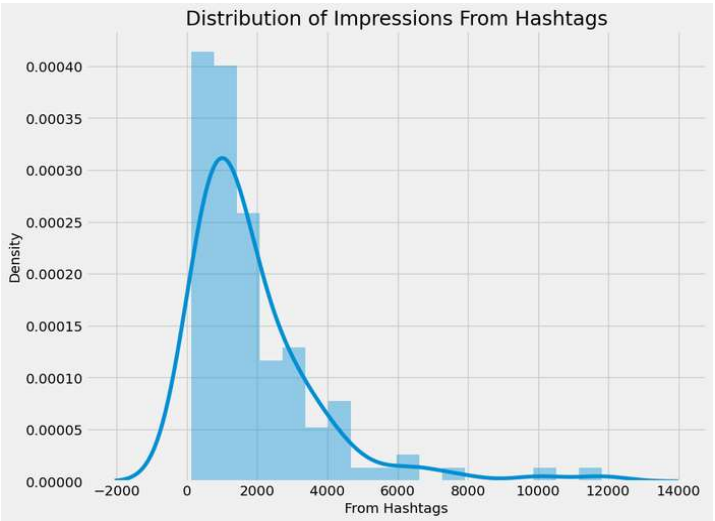


1) Next, we have tried to find out whether there is a significant difference between the reach through home feed recommendations, from Hashtags since these are the two ways most users see the posts through One-tail T-Test.
2) The null hypothesis for this is “There is no significant difference between the Instagram reach through home feed recommendations and from Hashtags ”
3) Reject the null hypothesis. The mean of 'From Home' is significantly greater than the mean of 'From Hashtags'.

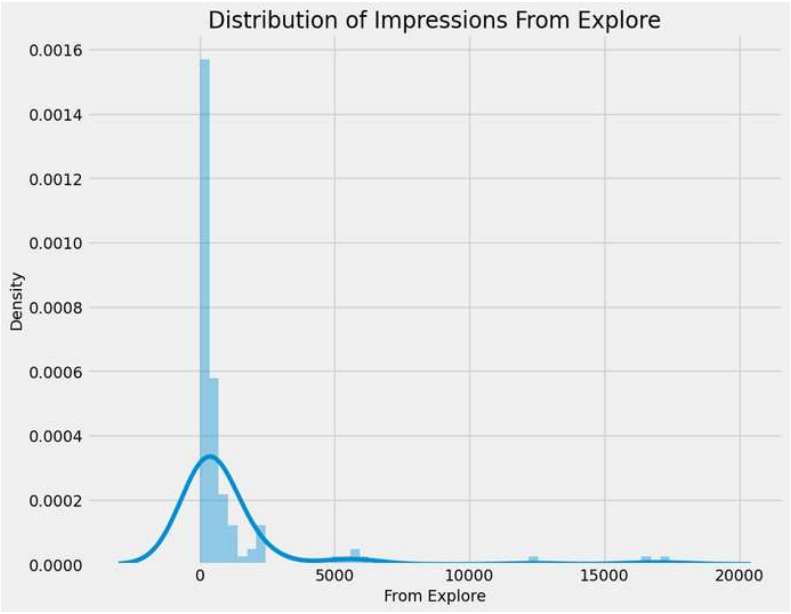
4) Distribution curves



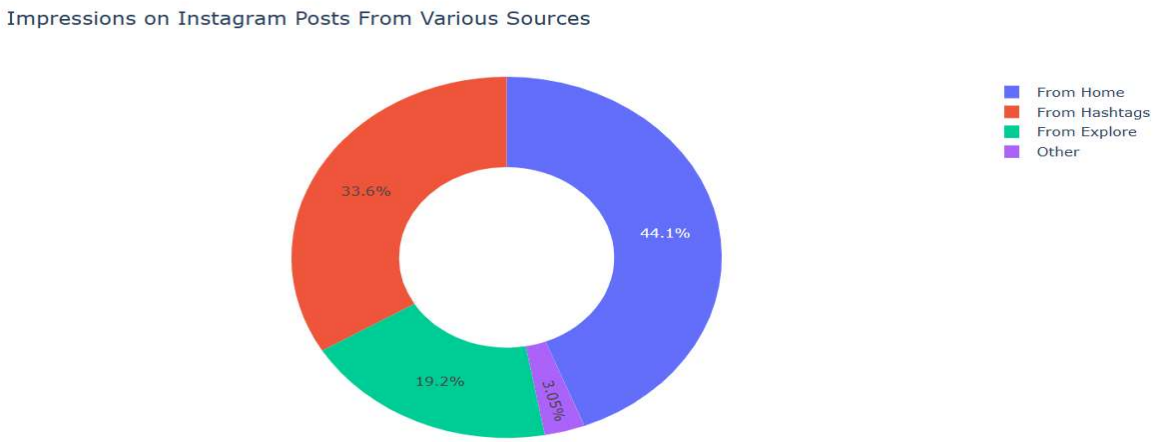
- The impressions received from the Instagram home section indicate the extent to which the posts reached by the followers.
- Assessing these home impressions reveals the challenge of consistently reaching all followers on a daily basis.



- Hashtags are tools we use to categorize our posts on Instagram to reach more people based on the kind of content we are creating.
- Looking at hashtag impressions shows that not all posts can be reached using hashtags, but many new users can be reached from hashtags.



- The explore section of Instagram recommends posts to the users based on their preferences and interests.
- By looking at the impressions we have received from the explore section, we can say that Instagram only recommends our posts a little to the users.
- Some posts have received a good reach from the explore section, but it's still very low compared to the reach of hashtags.



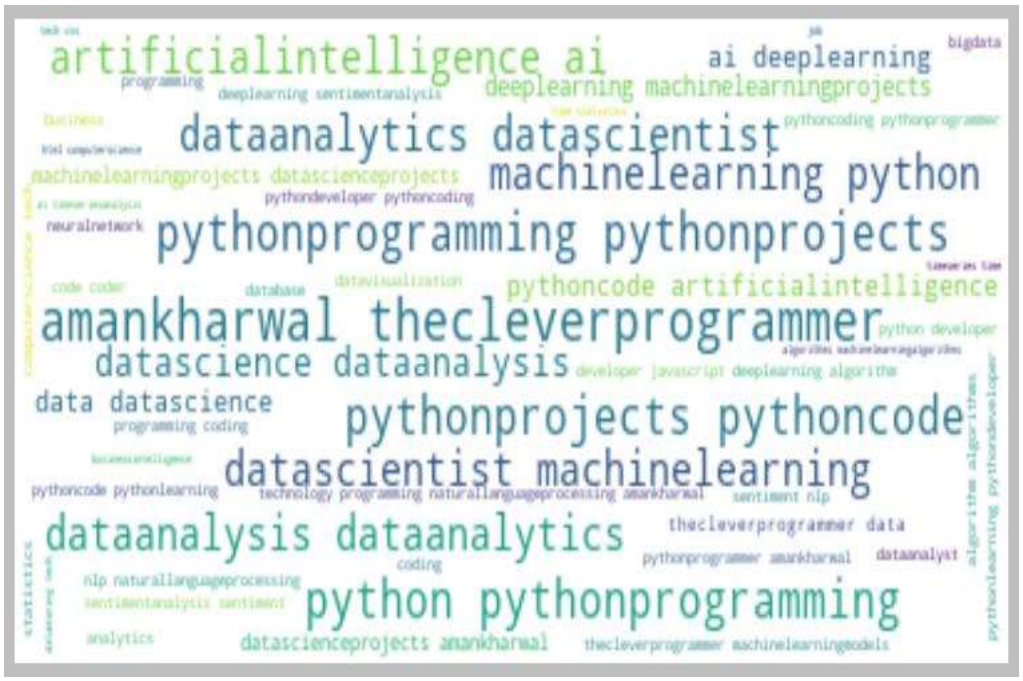
Percentage of impressions we get from various sources on Instagram

1) From the pie chart we can say Instagram reach "From Home" is highest among all.
2) We also proved that there is a statistical difference between "From Home" and other modes by ANOVA test and t-tail test.
3) Hence it can be concluded that Instagram reach for any post is reached maximum from home feed of any user.

5) Text Analytics

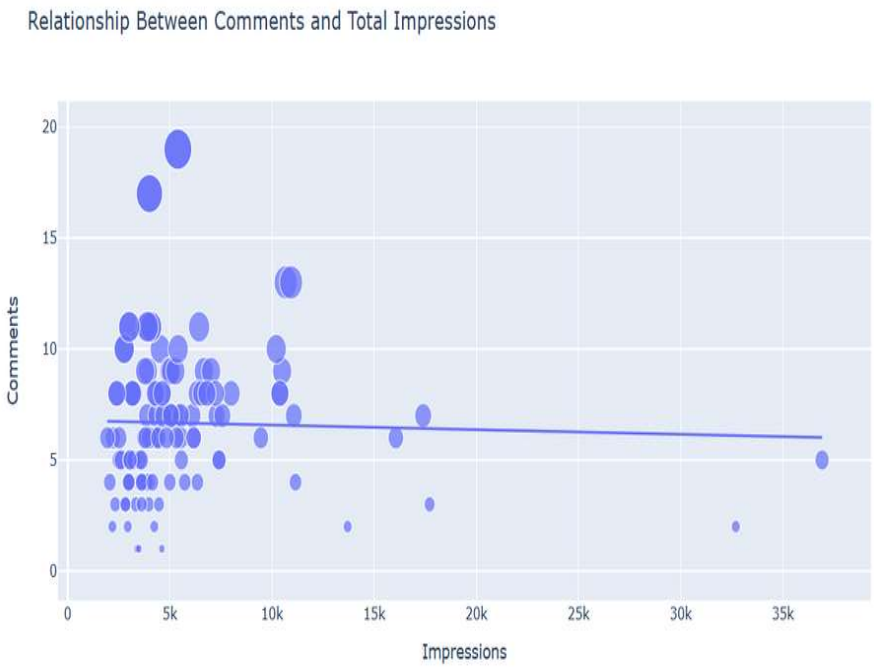
1) We have tried to apply text analytics and create Word cloud for "Captions" and "Hashtags"
2) The word Cloud for "Captions" indicates that words like 'Data Science', 'Machine Learning' have the highest frequencies.
3) The word Cloud for "Hashtags" indicates that words like 'PythonProgramming', 'ThecleverProgrammer' have the highest frequencies.

Captions

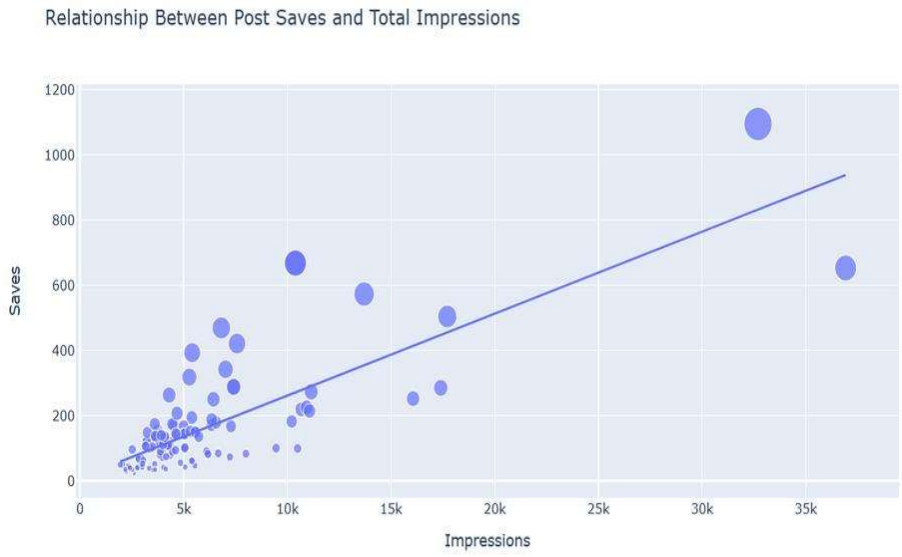


Hashtags

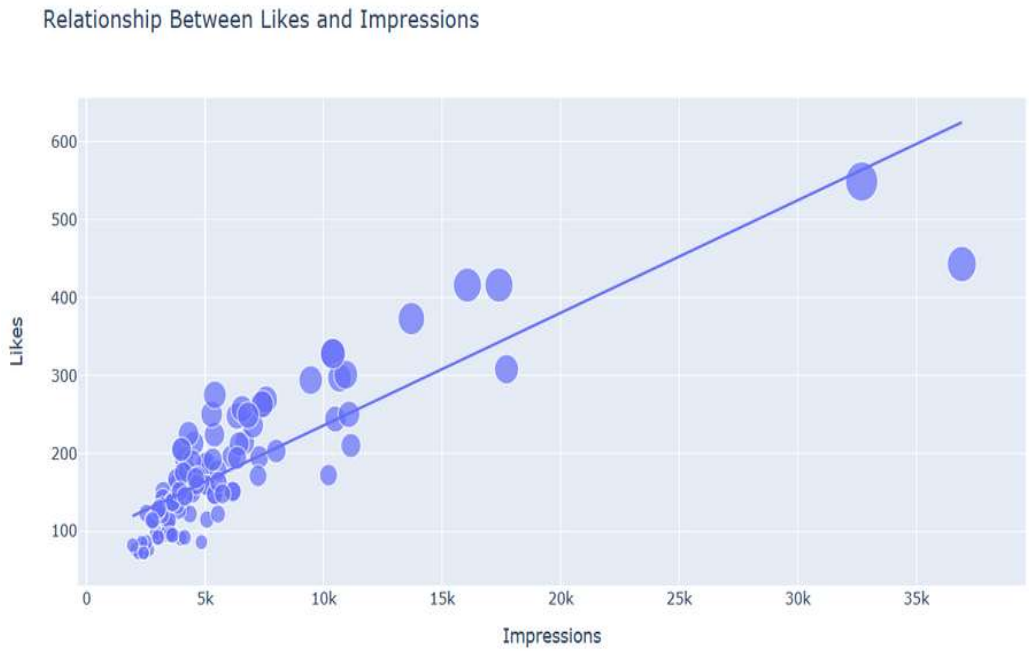
6) Scatter Plots



From this Scatter Plot we can conclude that impressions and number of comments do not have a proper linear relationship, it cannot be said that with increase in number of impressions the comments are not increasing, they are saturated at one cluster.



From this Scatter Plot we can conclude that impressions and number of saves have a more proper linear relationship than Comments, it can be said that with increase in number of impressions the comments are not increasing, they are saturated at one cluster.



From this Scatter Plot we can conclude that impressions and number of likes do not have a proper linear relationship, it cannot be said that with increase in number of impressions the likes are not increasing, they are saturated at one cluster.

7) STEPS TAKEN TO ANALYSE THE DATA

a) Data Extraction:

Extracted engagement metrics such as Likes, Saves, Comments, Shares, Profile Visits, and Follows from the dataset, ensuring a comprehensive understanding of user interactions with the content. By collecting these diverse metrics, we capture a holistic view of audience engagement, which is crucial for analyzing post-performance2 effectively.

b) Data Array Creation:

Organized the extracted data into a NumPy array, facilitating efficient handling and analysis of the dataset. By structuring the data into a NumPy array, we streamline data manipulation processes, enabling seamless integration with various analytical tools and algorithms. This structured format enhances readability and accessibility, laying a solid foundation for subsequent analyses.

c) Target Variable Isolation:

Identified "Impressions" as the target variable, the focal point for prediction in the analysis. By isolating Impressions as the target variable, we establish a clear objective for our modeling efforts, directing our focus towards understanding and predicting the reach of posts on Instagram. This targeted approach enables us to tailor our modeling strategies to specifically address the dynamics of Impressions.

d) Data Splitting:

Employed the train_test_split function to partition the dataset into training and testing sets, allocating 50% of the data for testing. This systematic approach ensures a fair evaluation of model performance by separating data used for training from data used for testing. By reserving a significant portion of the dataset for testing, we can accurately assess the generalization capabilities of our model and identify potential overfitting issues.

Ensured reproducibility in the split by setting a random seed (random state=42), guaranteeing consistency in results for future iterations. By fixing the random seed, we ensure that the dataset is split in a consistent manner across different runs of the analysis. This reproducible process enhances the reliability of our findings and facilitates comparison across multiple experiments.

e) Model Evaluation Metrics:

Evaluated model performance using key metrics: R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE). These metrics provide quantitative measures of the model's predictive accuracy and explanatory power, allowing us to assess its overall effectiveness in capturing the variability in Impressions. By examining multiple metrics, we gain a comprehensive understanding of the strengths and weaknesses of the model.

f) Interpretation:

R-squared value of 0.83 suggests a robust fit, explaining approximately 82.64% of the variance in Impressions. This high R-squared value indicates that a significant portion of the variability in Impressions can be attributed to the features included in the model. Consequently, the model provides a reliable framework for understanding the factors influencing post reach on Instagram.

MSE of 3,524,089.47 and MAE of 1,328.91 provide quantifiable measures of prediction accuracy, highlighting the model's effectiveness in estimating Impressions. Despite some level of prediction error, as indicated by non-zero MSE and MAE values, the model's predictions are generally close to the actual values of Impressions. This suggests that the model is capable of making accurate estimations of post reach on Instagram.

g) Model Strengths:

High R-squared value indicates a substantial proportion of explained variance, demonstrating the model's capability to capture underlying patterns in the data. This high explanatory power signifies that the features included in the model effectively account for the variability in Impressions, allowing us to make reliable predictions of post reach.

Despite non-zero MSE and MAE, the model yields accurate predictions, reflecting its reliability in estimating Impressions. While there may be some level of prediction error inherent in the model, the magnitude of these errors is relatively small compared to the range of Impressions observed in the dataset. As a result, the model's predictions can be considered trustworthy for practical applications in social media marketing.

h) Potential Reasons for High Errors:

The complexity of the system suggests that factors beyond the included features may influence Impressions, potentially contributing to prediction errors. Instagram's algorithm for determining post reach is influenced by various dynamic factors, such as user behavior, content relevance, and platform updates, which may not be fully captured by the features included in the model. Consequently, there may be inherent limitations to the model's ability to accurately predict Impressions under certain conditions.

8) Conclusion:

The model demonstrates strong explanatory power, as evidenced by the high R-squared value, indicating its effectiveness in explaining the variance in Impressions. This high explanatory power underscores the utility of the model as a valuable tool for understanding and predicting post reach on Instagram, providing actionable insights for optimizing content strategy and audience engagement.

Ongoing efforts to refine the model include further exploration of feature engineering and outlier management techniques to enhance predictive performance and reliability. By continually refining and iterating upon the model, we can improve its accuracy and effectiveness in capturing the dynamics of user engagement on Instagram, empowering social media marketers to make data-driven decisions and achieve better outcomes in their marketing campaigns.

References and links for further study

1. Search | Kaggle
2. Instagram Scraping | Kaggle
3. The Amazing Ways Instagram Uses Big Data And Artificial Intelligence | Bernard Marr
4. Social Media Scraping
5. How web scraping impact social media
6. Pattern Recognition
7. Pattern Basic | Geek for Geeks