**MINI PROJECT**

**Topic:**

**Instagram Post Reach Analysis**

Abstract:

Instagram is one of the most popular social networks for marketing. Predicting the popularity of a post on Instagram is important to determine the influence of a user for marketing purposes. There were studies on popularity prediction on Instagram using various features and datasets. However, they haven't fully addressed the challenge of data variability of the global dataset, where they either used local datasets or discrete output. This research compared several regression techniques to predict the Accuracy Rate of posts using a global dataset. The prediction model, coupled with the results of the popularity trend analysis, will have more utility for a larger audience compared to existing studies. The features were extracted from hashtags, image analysis, and user history. It was found that image quality, posting time, and type of image highly impact Accuracy Rate. The prediction accuracy reached up to 79.1% using the Gradient Boosting Regressor.

Methodology:

* Loading the dataset
* Prepare a function to examine the word cloud to understand the maximum type of words used in Instagram captions and hashtags
* Understand the number of likes we get based on the number of followers we have and the time elapsed since our post
* Gradient Boosting Regressor (train, test)
* Linear regression (train, test)
* Prediction on 100, 200, 500, 1000 followers

Data Sample

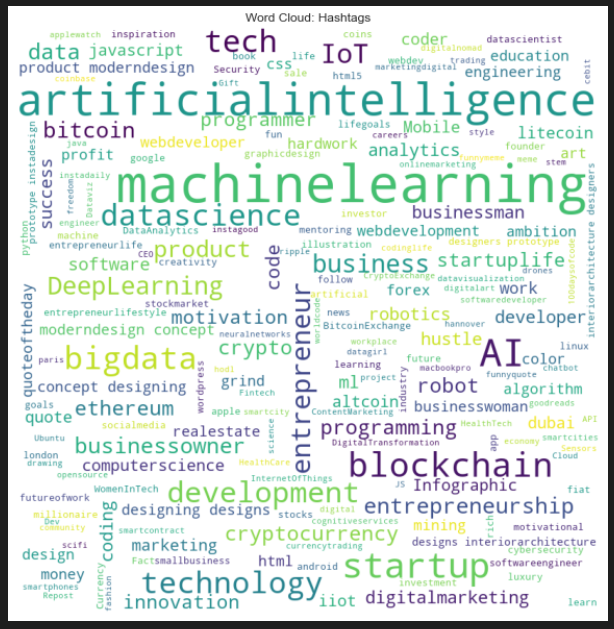


**Dataset Used:** [**instagram\_reach.csv**](https://drive.google.com/file/d/1B1mEPdf4eKfrUzePNSv1gCzxAxzma2P5/view?usp=sharing)

**Features:** Caption, Followers, Hashtags, Time Since posted, Likes

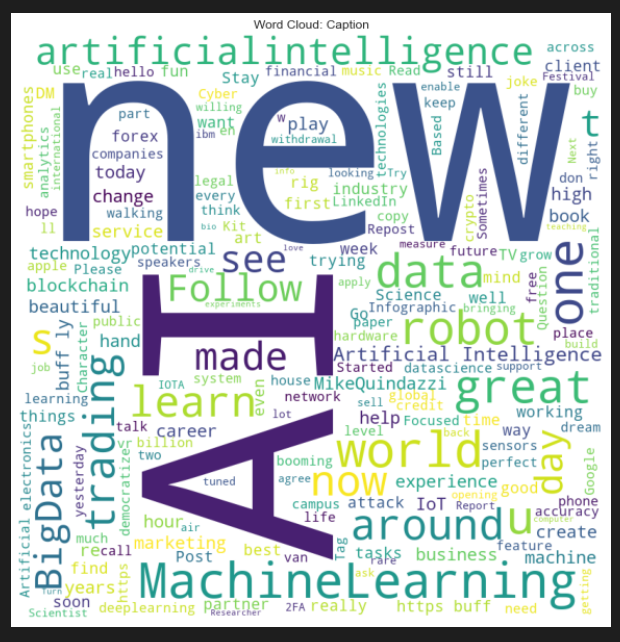
**Datapoints:** 100

Word Cloud: Hashtags

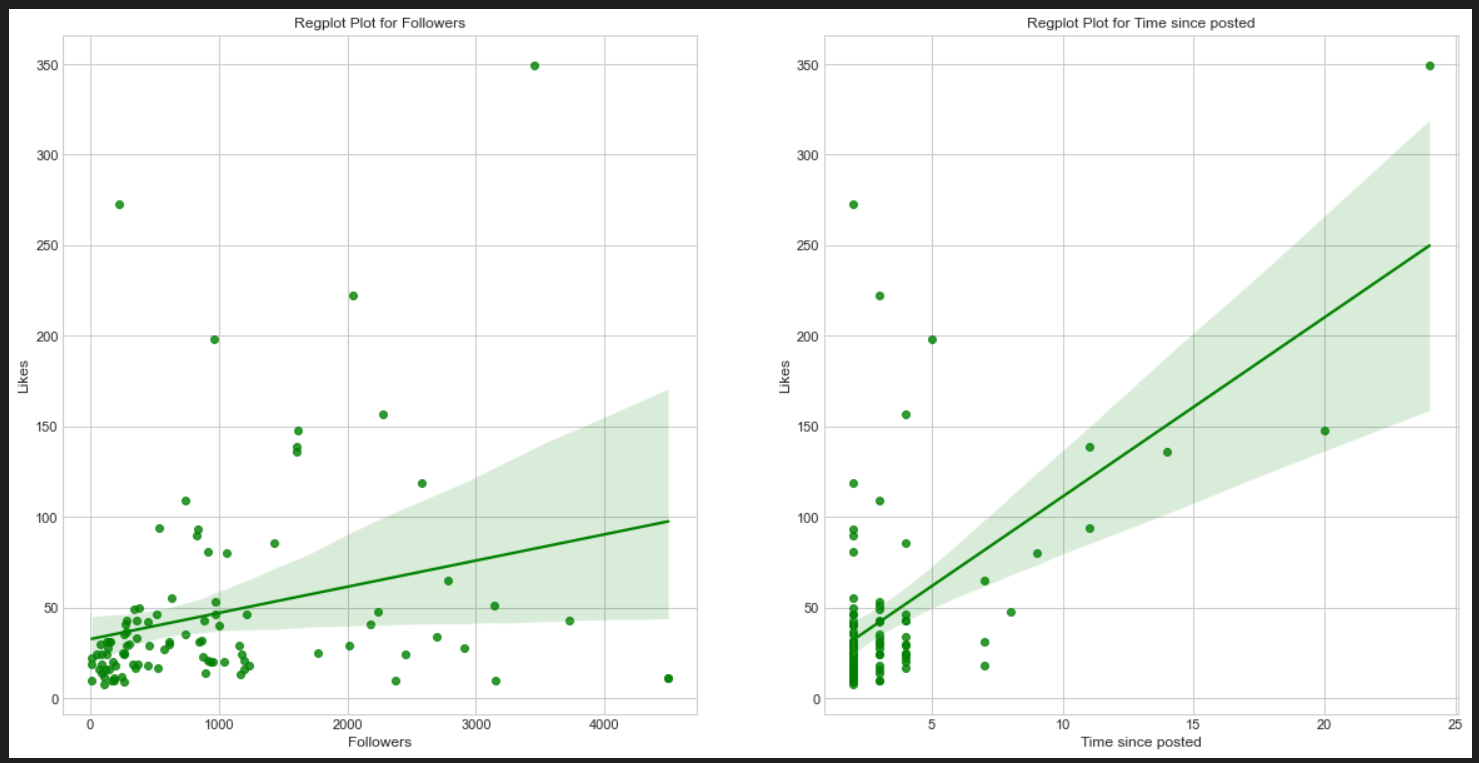


Word Count: Caption

Word Cloud: Captions



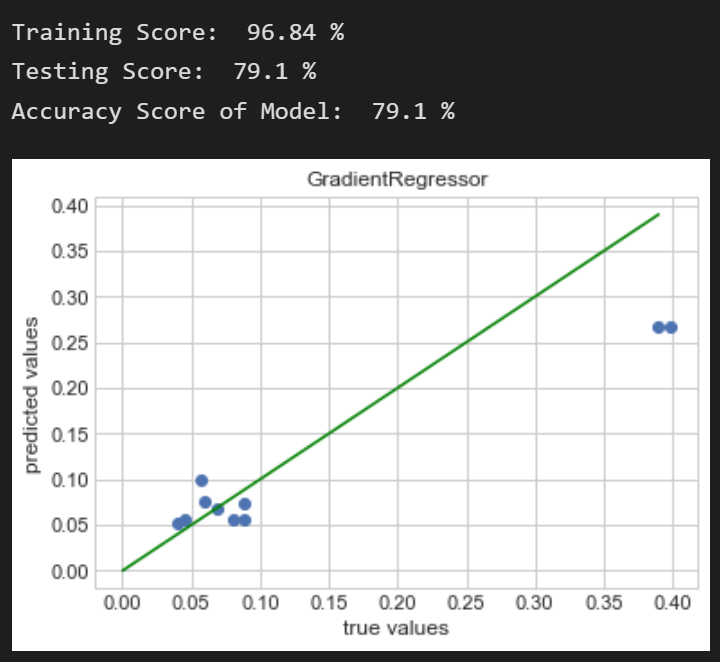
Plots for Likes vs Followers and Likes vs Time since posted



Gradient Boosting Regressor

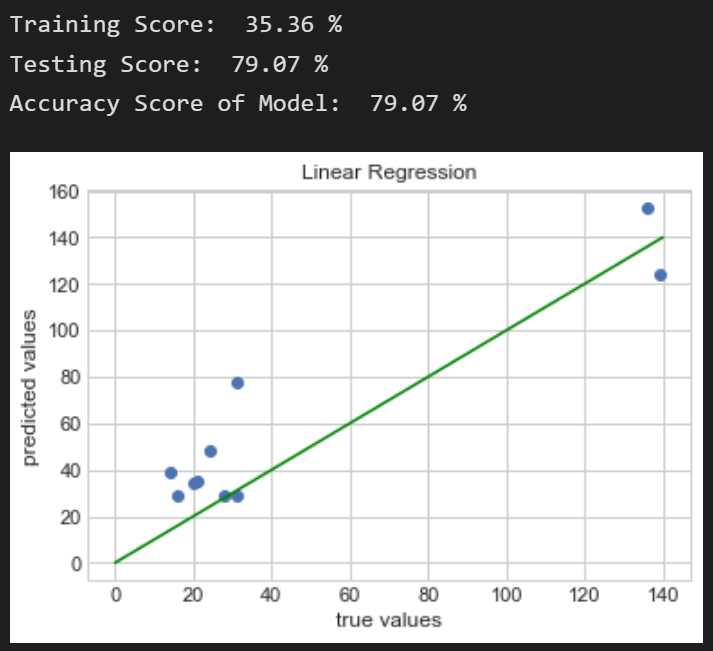
**Gradient Boosting** is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor’s error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels.

There is a technique called the **Gradient Boosted Trees** whose base learner is CART (Classification and Regression Trees).

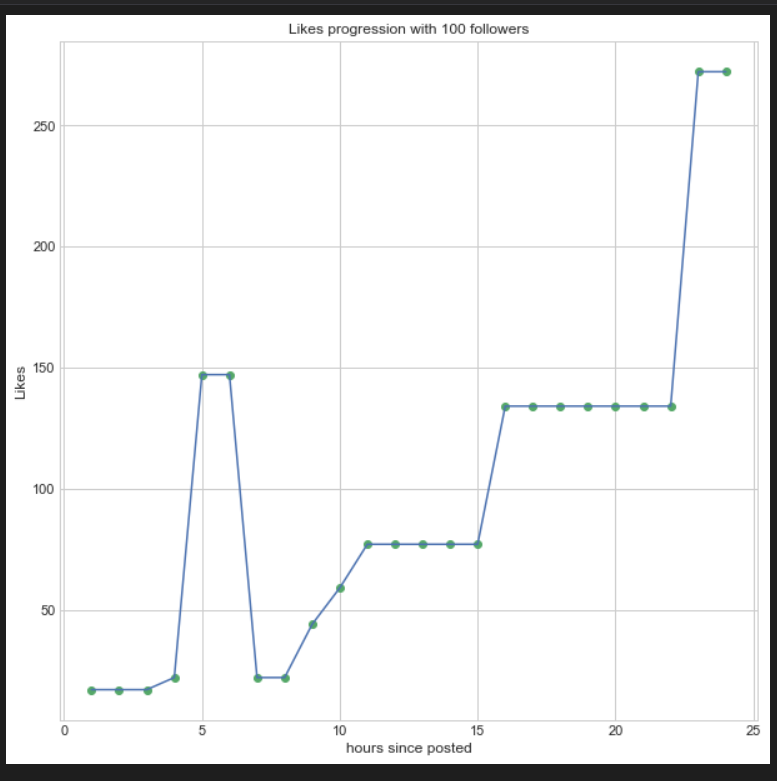


Linear regression

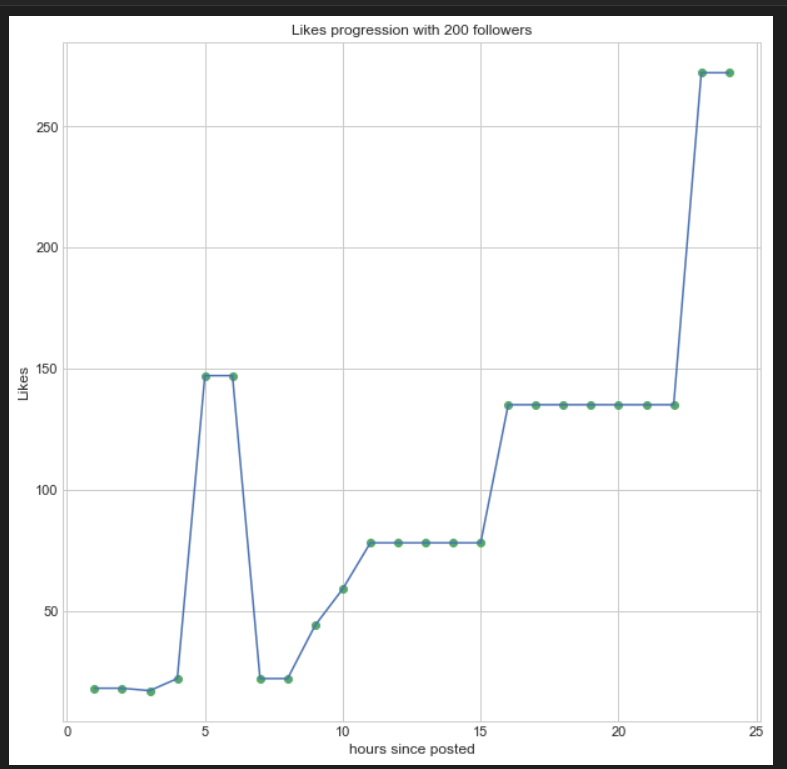
**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



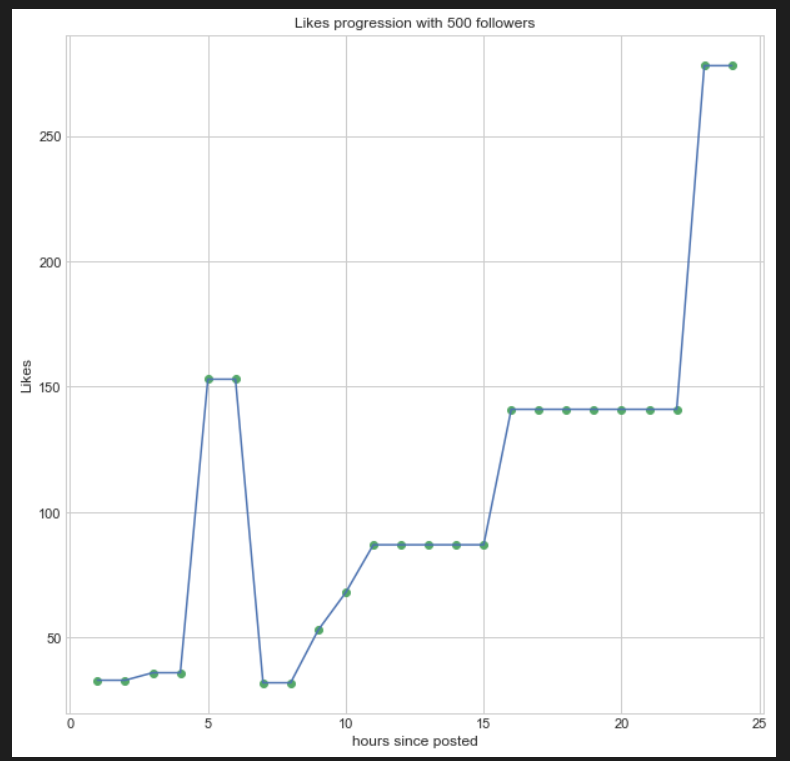
Prediction on 100 followers



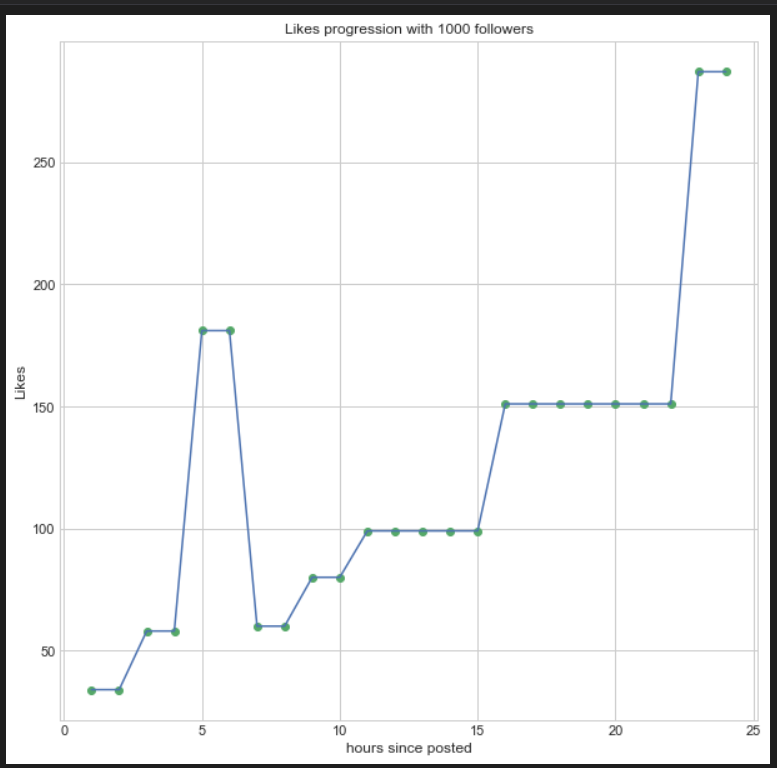
Prediction on 200 followers



Prediction on 500 followers



Prediction on 1000 followers



**Conclusion:**

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Gradient Boosting Regressor** | **Linear Regression** |
| **Training Score** | **96.84%** | **35.36%** |
| **Testing Score** | **79.1%** | **79.07%** |
| **Model Accuracy** | **79.1%** | **79.07%** |

Based on the predictions above, we can observe one thing: if you have a higher number of followers, your post is more likely to get more likes in attendance for life. But the maximum likes will not increase more than 20 times the number of followers only contributes to an increase of about 20% in likes.