

Instagram Dashboard and Feature Prediction

Suvajit Dulal Jana
3001336

Submitted in partial fulfilment for the degree of
Master of Science in Big Data Management and Analytics.

Griffith College Dublin
September 2019

Under the supervision of Mr Osama Abushama.

Disclaimer

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the Degree of Master of Science in Big Data Management and Analytics at Griffith College Dublin, is entirely my work and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.

Signed: _____**Date:** _____

Acknowledgements

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of ceaseless people or cooperation made it possible, whose constant guidance and encourage crown all efforts with success.

I am grateful to my project guide Mr Osama Abushama for the guidance, inspiration and constructive suggestions that helped us in preparing this project. I am also grateful my program director Dr Waseem Akhtar who taught us how to prepare us for a project, present our project and prepare documentation. And also, Mr Aqeel Kazmi who taught us how to perform EDA and build models.

I also want to thank Mr Ashish (Data Analyst) who have helped me a lot by suggesting me the algorithms and libraries that can use to build models for prediction and a Graphical user interface for the project.

I also thank my colleagues who have helped in the successful completion of the project.

Table of Contents

Acknowledgements	i
List of Figures	iii
List of Tables	iii
Abstract	iv
Chapter 1. Introduction	1
1.1 Prediction	2
1.2 Goals (or Something Similar)	4
1.3 Overview of Approach (or something similar)	4
1.4 Document Structure	5
Chapter 2. Background	6
2.1 Literature Review	6
2.2 Related Work	8
Chapter 3. Methodology	13
Business Understanding	14
Data Understanding:	14
Data Preparation	14
Modelling	15
Evaluation	15
Deployment	15
Chapter 4. System Design and Specifications	24
4.1 Hardware Specification	24
4.2 Technologies Used	24
Chapter 5. Implementation	26
5.2 Problem Statement	27
Chapter 6. Testing and Evaluation	28
Chapter 7. Conclusion and Future Work	29
References	30

List of Figures

Figure 1 Aesthetic Appeal of Instagram Posts.....	8
Figure 2 Proportion of Categories.....	11
Figure 3 Proportion of users w.r.t content categories.	11
Figure 4 Clustering users based on the categories	11
Figure 5 CRISP-DM	13
Figure 6 Dataset Division	15
Figure 7 Main GUI.....	17
Figure 8 Dataset	18
Figure 9 Pair Plot	18
Figure 10 Heatmap.....	19
Figure 11 Like Distribution for all username in the dataset.....	19
Figure 12 Probability plot of Like Distribution for all username in the dataset	20
Figure 13 Followers Distribution for all username in the dataset	20
Figure 14 Probability plot of Followers Distribution for all username in the dataset.....	21
Figure 15 Linear model.....	21
Figure 16 Posts Distribution for all username in the dataset	22
Figure 17 Probability plot of Posts Distribution for all username in the dataset	22
Figure 18 Following Distribution for all username in the dataset	23
Figure 19 Probability Plot Following Distribution for all username in the dataset	23
Figure 20 JSON Data Collected from the Crawler	27

List of Tables

Table 1 Instagram post Category	10
---------------------------------------	----

Abstract

Instagram is one of the biggest social media platforms owned by Facebook. It is a platform where everyone can share post like photos, video, stories, etc. and follow other user's feed. They can make their account private where only the users who follow the account can see the feed. The platform also has a explore feature where user can look at the popular posts taken in his/her nearby location or the posts which they may have interested in.

In the business of all money, time is the most important asset of any organisation is to save them inefficient and cleverly will lead the org from making a profit as well as popularity. This application will help most organization, influencer and content creators to know how their account progress and growth. Instagram doesn't allow a normal user to get the statistic of their account. The user has to upgrade their account to a business account to see the statistic and analytic of the account. And because of novel coronavirus pandemic, the verification process is been paused for a long time to upgrade the normal account into a business account. This application will work even for a normal user.

When the project was started the idea was to get the data from the Instagram API, clean and explore the data, and then build a model to predict the data. But the problem was to get the API needed to fetch the data needed verification which was paused due to coronavirus pandemic and also it may take at least a month to get API token from Instagram.

So, to collect, explore and visualise the data I have used libraries like selenium, Pandas, Numpy, Matplotlib, Scipy, and Streamlit. And Developed a web application in which the user can collect new and finish data from the internet and build a new prediction model on that data. Under the application, the user can also visualise the data as well.

Chapter 1. Introduction

Instagram is a huge platform where many creators and influencers have been sharing their creative content, experience with others. There are many athletes and celebrities on this platform. A user can find these athletes and celebrities and also find their friends and family on the platform to see the post and feed that can be photos and video. This platform provides users to capture and share some special moment in life with their friends and family. A user can decide to keep the account private that is the only user who follows them can see their post. The platform also allows the user to edit photos or videos before post the content on their account/profile.

Now the popular business entities are using social media for advertising or promoting the product. Instagram nowadays is like an online portfolio. Businesses like hotels and restaurants use this platform to show their rooms, ambience, foods, etc. to show/attract some new customer on their premises. Some creative showcase their arts and skill and many more. But not only small businesses even the larger businesses are in this game like automobile manufacturer, airlines, tours and travel, sellers, etc. The more the individual/businesses get connected to people the more they earn goodwill and popularity in the market.

But gaining popularity is not like just post anything on the internet. To gain popularity one individual should know their audience and which audience should be the target to promote the business. Through this, an individual will get to know about what to post at what time and whom and which hashtags to be included on the post because this is the only thing to get your profile and post into the viral in social media.

To get the information about the audience and get the statistics for a profile. The user has to spend his time and money to get all the detail from Instagram. The user first has to register the profile to business and go through all the process of verification. Then the individual has to get the Application programming interface (API) Key from Instagram. Instagram provides the basic API for free which get a JSON format of data about profile name and Id of a particular user. To get the other detail of a user the individual has to verify his/her account first or a business which may take 2 weeks to a month and charges may include. This will get all the information about the users and their all post on their feed. This includes likes, comments, Number of followers, Number of following, Sponsor details of every post, etc. GraphQL is used to query and manipulate the data from the JSON format. This language is developed by Facebook to access, query and manipulate the data extracted through any API and available in many languages like JavaScript, Perl, Python, Ruby, Java, C++, C#, PHP, R, etc.

1.1 Prediction

As to collect the live dataset from Instagram API is required. But if an individual doesn't want to register for API from the internet, he/she can scrap the internet to collect the data from Instagram. The dataset must contain the following attributes to train the models for prediction

- Username
- Post
- Followers
- Following
- Like
- Is Account Verified
- Is Account Private

Collecting Dataset

The dataset must contain at least more than 200 user's data containing to predict the attribute for particular data. The dataset must include both kinds of users common and influencer so the data won't be biased. After getting the dataset Exploratory Data Analysis (EDA) has been performed. Where dataset is to be summarized the main characteristic of data, discover a pattern, process the null values from the dataset that is, discards the null row or performing some imputation technique to fill the null values. The EDA process also includes plotting the dataset on the graph to get an idea to of relationships between the columns. To use linear regression for building the model, it is necessary to remove the correlated variable to get better accuracy with the built model. We can visualise this with the heatmap which will give us the idea of the confusion matrix. We can also plot the distribution graph to check the skew in the dataset. As some of the datasets was not showing a good skewed feature on the graph so we have applied $\log(1+x)$ to all the attributes of the dataset. At this point, we have applied the EDA needed to the dataset.

After this, we have split the data into two-part at 9:1 ratio. The large part is to train the model to predict the attribute of the dataset. And another part is for testing the model to check the accuracy of the models. We have created 5 models as follows

1. To predict like
2. To predict Followers Count
3. To predict Comments Count
4. To predict the account is verified or not
5. To predict the account is private or not

Splitting data and Training models

To predict like the like column was removed from the training dataset and only like column was provided in the testing dataset. The basic linear Regression Equation is used for the prediction that is ' $Y = a + bX$ ', where ' X ' is the explanatory variable and ' Y ' is the dependent variable. The ' b ' is represented as a slope of the regression line and ' a ' is the point of interception on the graph. R Squared statistical measurement is used to check the accuracy of the data. It is used to check how close the data has been predicted to form the regression line on the graph. The closure the point is from the regression line the high the accuracy. If the R-squared value is 0% it means the model has explained none of the variability. That is the model has predicted a point far from the regression line.

Evaluation

To verify the result and accuracy of the model got trained we have taken the actual data and predicted data and convert them to normal form by calculating the exponential values as we have used $\log(1+x)$ to get skew before. And getting the exponential values of both data, compare both and check the distance between them. The higher the distance less the accuracy.

The same approach has been taken to predict the follower and comments count. For predicting the account is private or not and Verified or not, we have not used $\log(1+x)$ for skew and the attributes were already in the good shape as this is the Boolean values that are true or false. Rest everything is same as previous models just we don't have to calculate the exponential values.

Result

After all, models have been saved, we have used real-time data to predict likes, comments count, Followers, Private and verified account using those models. This data can be used by the individual and decide what to post by comparing their account with other famous influencer or get the trends from the data.

1.2 Goals (or Something Similar)

Goals of this project were to develop a classification model with high accuracy and predict some attributes. The main motive of this project to analyse the live data from the internet and predict the attributes on the real-time data. As Instagram doesn't have a feature to analyse the data and show real-time statistic on the account. This can help individual or business account to get an idea of how their account will grow in the future. And the real-time data plays an important role in getting the real-world scenario on current trends and posts or hashtags and test how another account has been impacted with the trends on social media. And according to that information and the individual users or business can post content on their feed by hitting the target audience on the platform. This may also result in gathering more target audience to the account without promoting anything and without paying anything to pay for promoting the account.

1.3 Overview of Approach (or something similar)

This application was built using the python language and many libraries have been used to give the application some major functionalities. The libraries like NumPy, Pandas, Sklearn, Instagram-Explore, Selenium, Seaborn, Streamlit, Scipy and matplotlib has been used. At the initial stage of the project, we were using Instagram-Explore to collect the Instagram data. This library was extracting the JSON data from the Instagram by using Pandas the required data was then saved into a data frame and saved them in a Comma Separated Values (CSV) file. But at the stage the Instagram-Explore library got some error due to constant changes on Instagram data policies and a new feature called reels is been deployed to the platform. Therefore, to overcome we the problem and gather the real-time data, we have to use selenium web driver to build a crawler that will scrap the internet and collect the recent data that is last 10 – 13 post of every account. A list of usernames has been supplied to the crawler. The list includes all types of user that can be normal individual, influencers and business. The collected data is then saved to the CSV file.

After this, Exploratory Data Analysis (EDA) is performed on the dataset and the dataset is been divided into the training set and testing set. Using this, multiple models are been made. To train the model linear regression is been used. Other algorithms were also used like Gradient Boosting Regressor, LGBM Regressor and Cat Boost Regressor but the desired accuracy was not achieved. The accuracy achieved is for predicting likes is 93%, comment count is 86%, follower count is 90%, a Verified account is 53% and Private 40%

1.4 Document Structure

Chapter 1 gives a brief introduction about the project, Chapter 2 is the Background and literature work done before starting the project. Chapter 3 is an explanation of the methodology used to build the project. Chapter 4 is about all technical, hardware and software specification of the. Chapter 5 gives the brief of the implementation and deployment of the project, Chapter 6 define all the testing made to the application and their result. Chapter 7 Concludes the project with all the enhancement that can be made to the application.

Chapter 2. Background

2.1 Literature Review

The growing audience is important. We determine them as a group. A quad closure is a group of four people who are following each other. Based on the triadic closure method that is whom they follow what they like, where they comment. There are many factors to be considered while determining the group like information about the users, what they like, whom they tag, which tags they use most, sentimental analysis, etc. this makes it complex to use today in recommendation system in Instagram. For example, you follow your 4 friends on Instagram. If two of your friends like photos which have tags for cars then the other friends in the group will receive some car-related feed on their recommendation. Same for following if 4 of your friends follow the same account then you may get the same account for recommendation to follow. This technique is very helpful to gather more audience. [1] Online Social media can be used to investigate real-world human behaviour. Instagram shares major features like sharing media, tagging and sharing media, Like contents, etc. The main point of interaction is following, like a post, comment, post, etc. The investigation takes 3 major aspects. [2]

1. The group they follow or following and made a heterogeneous interaction network, to unveil the emergence of self-organization and topically-induce community structure.
2. The engagement they do on the platform like which post they like, where they have commented, whom they follow to understand global trends and popular users emerge.
3. The behaviour of the user's context, caption label, hashtags, to determine how the user has attracted the attention and explore the variety of topic he/she is interested.

These 3 aspects will give us the idea and mechanisms of the user can gain popularity on the platform.

When it to promote business online social media is the best way to promote something. In a book called Instagram Power [3], the author asks to take a closer look in Instagram a photo and video sharing application Described by Miles as “the breakout social network of the iPhone revolution,” Throughout the book author Stresses the point that while the online world was once computer-based, and now smartphones are taking place. Earlier the promotion used to make on a social media platform like Twitter and Facebook. “Instagram provides an opportunity for you to bring your company into the new mobile revolution without complexity or drama,” Miles writes. Miles describes major points to get successful on Instagram.

The points are as follows [4]:

- How to use (and How Not to use) hashtags
- The power of copywriting
- Six Catalysts for Growth
- Tips for branding and selling on Instagram
- How to integrate Instagram into your online Market

The book also devotes to describe the skills or product needed for Instagram. Miles doesn't feel obliged to devote far more space to over-describing the need for a skill or product. Getting down to the brass tacks. Instagram Power makes a concise case for social media marketing before. Getting down to practical, step-by-step instructions on how to promote a business using Instagram.

“Using Instagram as a tool to jump into the new mobile format can be a simple step toward your eventual success in this exciting new space,” __Jason G. Miles.

To provide a working knowledge of Instagram, the current research seeks to explore the structural dimensions of the intentions of end-users to use Instagram and to explore the relationships between identified reasonings and key variables of norms and behaviour purpose. A thorough questionnaire was undertaken in which a total of 212 Instagram users assessed their Instagram inspiration, production factors, intention to use and perception. The results suggest that there are five primary social and psychological intentions for Instagram users: social interaction, archiving, self-expression, escapism and peeking.

2.2 Related Work

Popularity prediction of images and videos on Instagram

People share moments on the platform with their friends and family. Today everyone is using Instagram. The posted content has attribute like caption, like and comments. Today many individual and large companies want to attract more audience. The problem is not all post is seen by all the follower/audiences therefore it is important to know what to post and when to post. So that audience will get to know about the post. [5] Studying the dataset from Instagram will help us to understand the best way to post the content on the platform. In this study, they have collected the like images and video their caption, comments and like information on every post. They ran an experiment with 10-Fold Cross-Validation, where they got results od popularity score prediction with 0.002 in RMSE and Popularity Class prediction with 90.77% accuracy and their dataset was limited to the Iranian popular influencers on Instagram.

Instagram Likes for Architectural Photos can be Predicted by Quantitative Balance Measures and Curvature.

The aim of this research was in the field of experimental aesthetics, to investigate how to influence daily decision

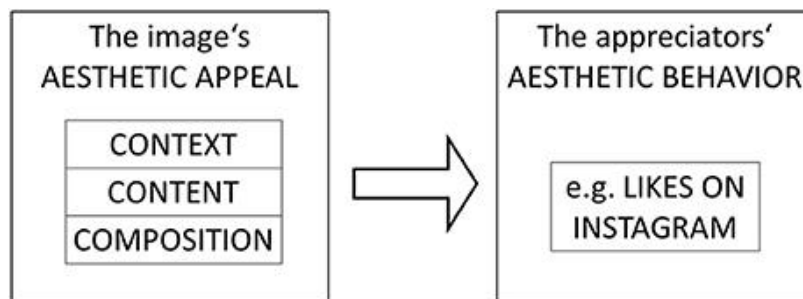


Figure 1 Aesthetic Appeal of Instagram Posts

The holds for both consumers—when they buy artfully designed products, enjoy visits to museums, galleries and exhibitions, or even search for an attractive partner—and for producers—when artists create artworks, advertisers design campaigns, researchers visualize data, or ordinary people arrange their flats, take photos, or do handicrafts. We all strive to put things together in a visually pleasing way because what is beautiful is usually considered as good. Extensive research on the “beautiful is good” phenomenon has shown that it applies to persons—“who is beautiful has more socially desirable personality traits and leads a better life” [6] [7] [8] [9] [10]

The data generated on Instagram by photos and video posted by users on Instagram can gather and used for further analysis. The final dataset contains 700 posts with likes count generated on Instagram. [6] Their first aim is to validate the like count, the second aim is to make some distribution plot to visualise the data. By considering two visualise the balance and the preference of curvature over angularity. Also, compared 2D and 3D plot. The comparison shows that the predicted curve was more complex for dynamic images gives more meaning to like prediction. The likes for other images was not much complex therefore the predicted likes were less.

How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention



Data Mining on Social media has been in trending between data scientists and psychologists. However, according to the research, the data is not suitable to determine for a single user because the attributes of this user are fixed. The other data can only determine whether the user is famous/popular or not. Or how much popularity user can achieve. Therefore, the researcher collects his data and divide the data into multiple environments according to the user. Based on the relevant data, they devise a novel dual-attention model to incorporate an image, Caption, and user environment. The dual-attention model considers two parts one is an environment and other is the feature of the image. A hierarchical structure followed to predict the information for the account. The classification result shows that the model outperforms the baseline and a statistical analysis identifies what kind of picture or caption can help the user achieve a relatively higher like count and increase the engagement of the post and higher the engagement the high number of audience can be achieved. [11]

What We Instagram: A First Analysis of Instagram Photo Content and User Types

Instagram had seen rapid growth in the number of users, daily user and user interaction on the platform since 2010. [12] It is the most popular photos and video capturing and sharing platform. In this research, they have present qualitative and quantitative analysis on Instagram. They have also used computer vision to examine the photo content. Based on that they divide the post into the following categories. The result of the research reveals several insist about Instagram those are as following:

- Eight popular photos categories
- Five distinct types of Instagram users in terms of their posted photos
- A user's audience (number of followers) is independent of his/her shared photos on Instagram.

Table 1 Instagram post Category

Category	Exemplary Photos
Friends (users posing with others friends; At least two human faces are in the photo)	
Food (food, recipes, cakes, drinks, etc.)	
Gadget (electronic goods, tools, motorbikes, cars, etc.)	
Captioned Photo (pictures with embed text, memes, and so on)	
Pet (animals like cats and dogs which are the main objects in the picture)	
Activity (both outdoor & indoor activities, places where activities happen, e.g., concert, landmarks)	
Selfie (self-portraits; only one human face is present in the photo)	
Fashion (shoes, costumes, makeup, personal belongings, etc.)	

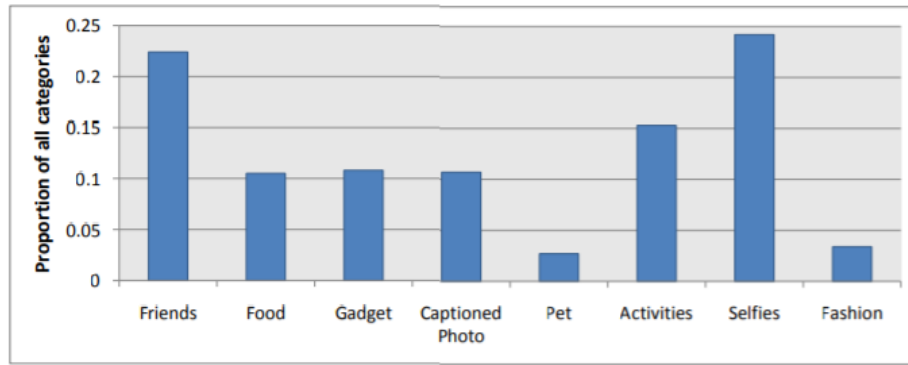


Figure 2 Proportion of Categories

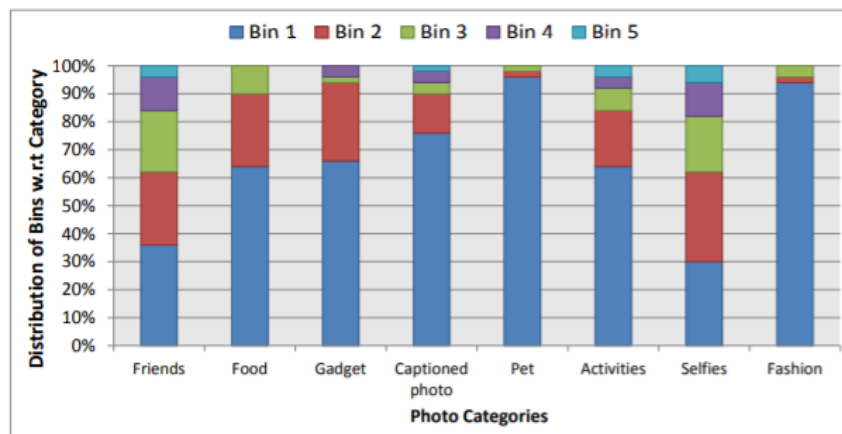


Figure 3 Proportion of users w.r.t content categories.

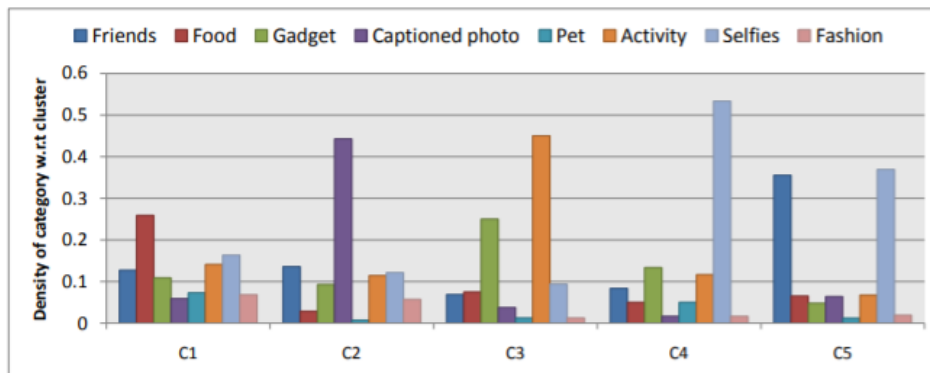


Figure 4 Clustering users based on the categories

The interaction of Instagram Followers in the fast fashion sector: The case of Hennes and Mauritz (H&M)

This research aimed to analyse the interaction between fashion brands and their followers on social media. The research was done on the relationship between the H&M a fast-fashion company and their customers/followers on Instagram. [13] A thorough analysis is been done that how company engage with their customers and followers and what is the impact on their business. The research answers questions like what their follows likes the most so the company can design more that kind of products or design on the market. This also helps them to improve and gather more audience on social media. Three categories were used to analyze the data. The categories were considering the point of view of the content of the message communication, company's strategies, formal aspect and product posted. By this result, other business can also follow this to promote the product on online social media to gather more audience and gain more customers.

Chapter 3. Methodology

This project was aimed to make an application for an individual or business. In the business of all money, time is the most important asset of any organisation is to save them inefficient and cleverly will lead the org from making a profit as well as popularity. This application will help most organization, influencer and content creators to know how their account progress and growth. Instagram doesn't allow a normal user to get the statistic of their account. The user has to upgrade their account to a business account to see the statistic and analytic of the account. And because of novel coronavirus pandemic, the verification process is been paused for a long time to upgrade the normal account into a business account. This application will work even for a normal user.

Therefore, the Cross-Industry Standard Process for Data Mining (CRIP DM) methodology is been used to develop this project.

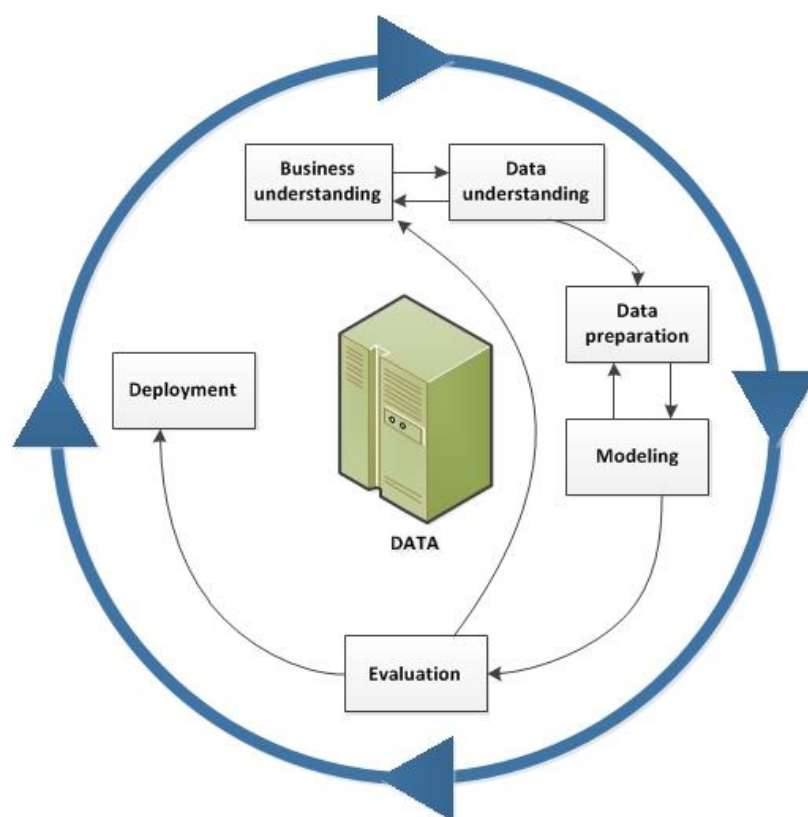


Figure 5 CRISP-DM

Business Understanding

Nowadays the popular business entity is using social media for advertising or promoting the product. Instagram nowadays is like an online portfolio. Businesses like hotels and restaurants use this platform to show their rooms, ambience, foods, etc. to show/attract some new customer on their premises. Some creative showcase their arts and skill and many more. But not only small businesses even the larger businesses are in this game like automobile manufacturer, airlines, tours and travel, sellers, etc. The more the individual/businesses get connected to people the more they earn goodwill and popularity in the market.

Data Understanding:

Understanding data plays an important role in build a model. Because without any knowledge about built a model will be useless. For this project, we have used a dataset with more the 5000 variable and 21 attributes. To collect the data, we have built a crawler who crawls through the internet and scrape the data and prepare one CSV file for further process. The crawler gathers the JSON data for each user extract top 10-13 post and each user. And the users whose account are private for them the crawler collect basic information but no for the user and store them in the CSV file. The crawler also extracts the time of post, like count and comment count for every post.

Data Preparation

To collect the data, the crawler crawls through the internet and scrape the data and prepare one CSV file for further process. The crawler gathers the JSON data for each user extract top 10-13 post and each user. Once the CSV file is prepared the dataset is then converted into a data frame for easy manipulation and visualisation. After the data gets a load on the program, we built a new data frame where we use group by clause to group the post for all the users and average function is used for getting likes, followers, followings, comments count, and tags on the post of that particular user. We have also plot heatmap and pair plot of all the variables to get an overview of the relationship between the attributes in the data. The data set is then divided into 2 part set one to train the model and one to test the model.

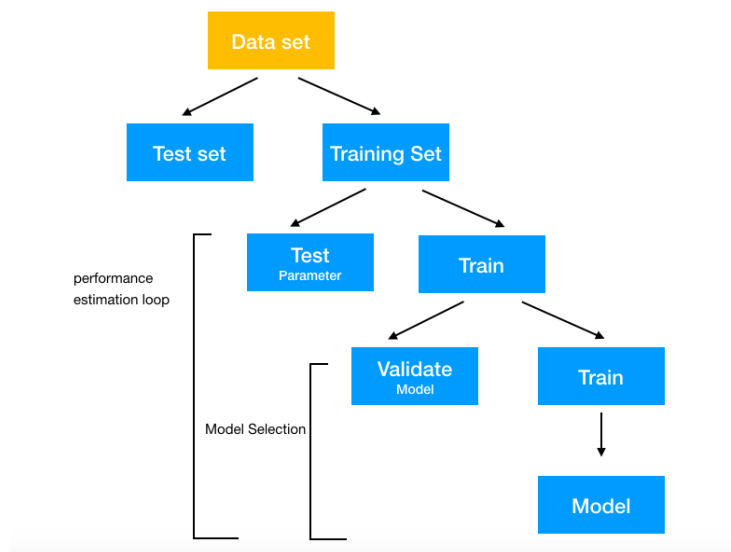


Figure 6 Dataset Division

Modelling

The whole data is divided into two part that is training set and testing set into 9:1 ratio the random data is select to make both sets. The attributes on both data set go through a generalisation factor $\log(1+x)$ to genialise the dataset. From the training set, the attribute which is to predict is been remove and supplied to the model for training. To increase the accuracy of the model all the other useless columns where drop. Later this training model has tested the testing dataset and calculate the accuracy. The same training set is given to multiple models but only one model gave us the higher accuracy.

Evaluation

Here at this stage, the trained model is then tested here to check the accuracy of the model. The testing dataset is used here to calculate the accuracy. The check and compare the results we have to calculate the exponential value of all data coming from the model as well as on the test model. The higher the accuracy the better the model. To calculate the accuracy, we have to use the R-Square score.

Deployment

The dataset must contain at least more than 200 user's data containing to predict the attribute for particular data. The dataset must include both kinds of users common and influencer so the data won't be biased. After getting the dataset Exploratory Data Analysis (EDA) has been performed. Where dataset is to be summarized the main characteristic of data, discover a pattern, process the null values from the dataset that is, discards the null row or performing some imputation technique to fill the null values. The EDA process also includes plotting the dataset on the graph to get an idea to of relationships between the columns. To use linear regression for building the model, it is necessary to remove the correlated variable to get better

accuracy with the built model. We can visualise this with the heatmap which will give us the idea of the confusion matrix. We can also plot the distribution graph to check the skew in the dataset. As some of the datasets was not showing a good skewed feature on the graph so we have applied $\log(1+x)$ to all the attributes of the dataset. At this point, we have applied the EDA needed to the dataset.

This application was built using the python language and many libraries have been used to give the application some major functionalities. The libraries like NumPy, Pandas, Sklearn, Instagram-Explore, Selenium, Seaborn, Streamlit, Scipy and matplotlib has been used. At the initial stage of the project, we were using Instagram-Explore to collect the Instagram data. This library was extracting the JSON data from the Instagram by using Pandas the required data was then saved into a data frame and saved them in a Comma Separated Values (CSV) file. But at the stage the Instagram-Explore library got some error due to constant changes on Instagram data policies and a new feature called reels is been deployed to the platform. Therefore, to overcome we the problem and gather the real-time data, we have to use selenium web driver to build a crawler that will scrap the internet and collect the recent data that is last 10 – 13 post of every account. A list of usernames has been supplied to the crawler. The list includes all types of user that can be normal individual, influencers and business. The collected data is then saved to the CSV file.

After this, we have split the data into two-part at 9:1 ratio. The large part is to train the model to predict the attribute of the dataset. And another part is for testing the model to check the accuracy of the models. We have created 5 models as follows

1. To predict like
2. To predict Followers Count
3. To predict Comments Count
4. To predict the account is verified or not

To predict the account is private or not.

After this, Exploratory Data Analysis (EDA) is performed on the dataset and the dataset is been divided into the training set and testing set. Using this, multiple models are been made. To train the model linear regression is been used. Other algorithms were also used like Gradient Boosting Regressor, LGBM Regressor and Cat Boost Regressor but the desired accuracy was not achieved. The accuracy achieved is for predicting likes is 93%, comment count is 86%, a follower is 90%, a Verified account is 53% and Private 40%

Instagram Dashboard

Collect fresh New Data

Collect

Train New Models

Train

Enter a IG Username

Submit

Dataset

Show Data

Dataset Overview

Show

Figure 7 Main GUI

Dataset

Show Data

	No. of Post	following	followers	Likes	No of Tags	C
eatfamous	4387	1377	275746	7.0852	4.6667	
eatingnyc	3461	711	317582	7.8619	3	
eddpinto_	617	2891	264849	5.9236	2.5000	
edkashi	1671	851	388285	6.9840	0	
edehumor	5174	273	179529	7.4860	1.0833	
el_kilombo	21562	1550	3817641	10.9189	0	
elaine_yiu	2247	838	816720	9.4169	0.5833	
elenscarriere	1825	645	488431	8.6388	0.8333	
elensham	2426	326	193744	8.7649	0	
elisabethrioux	842	482	1828673	11.7969	1.4167	
elizabethgadd	1814	228	179165	9.0611	0	

Figure 8 Dataset

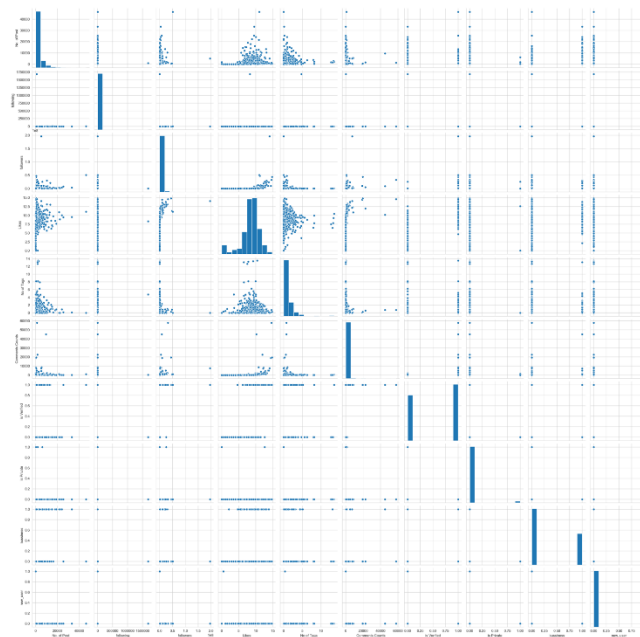


Figure 9 Pair Plot

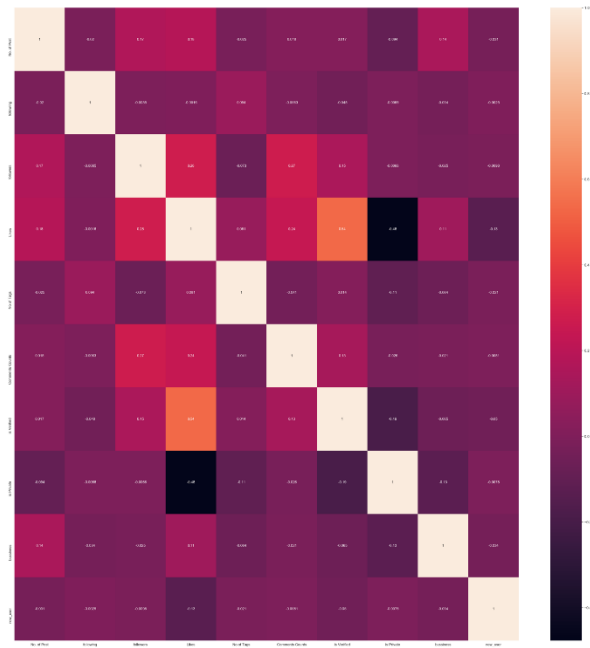


Figure 10 Heatmap

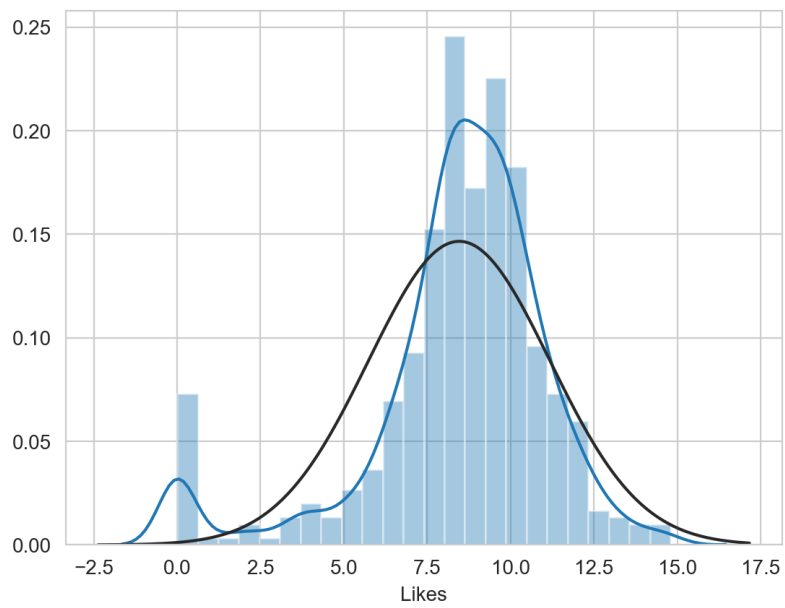


Figure 11 Like Distribution for all username in the dataset

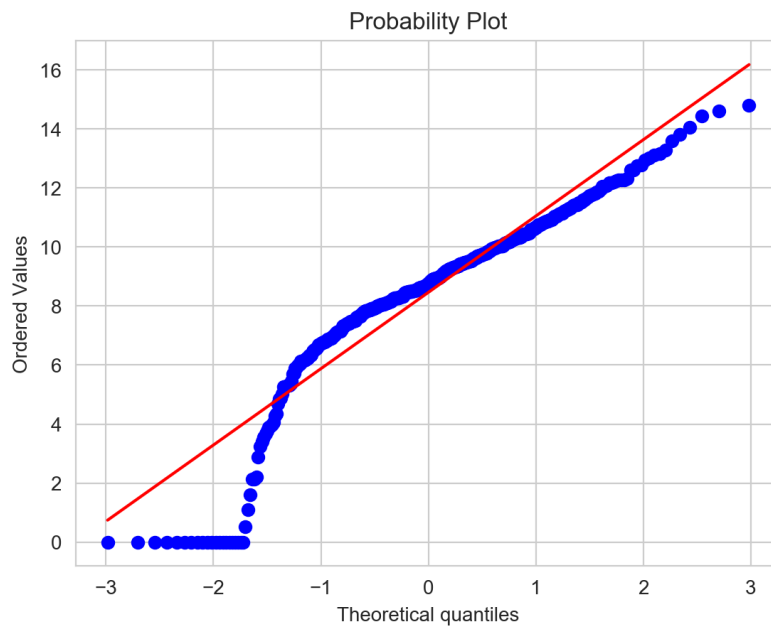


Figure 12 Probability plot of Like Distribution for all username in the dataset

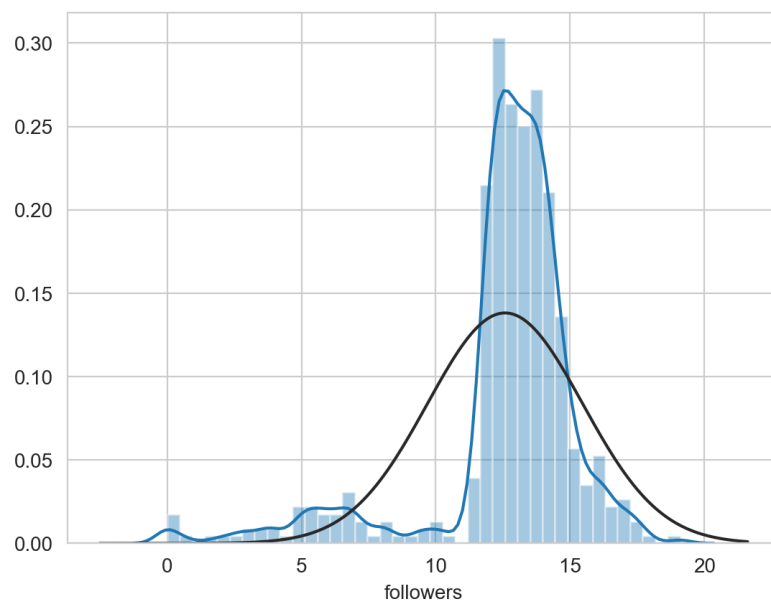


Figure 13 Followers Distribution for all username in the dataset

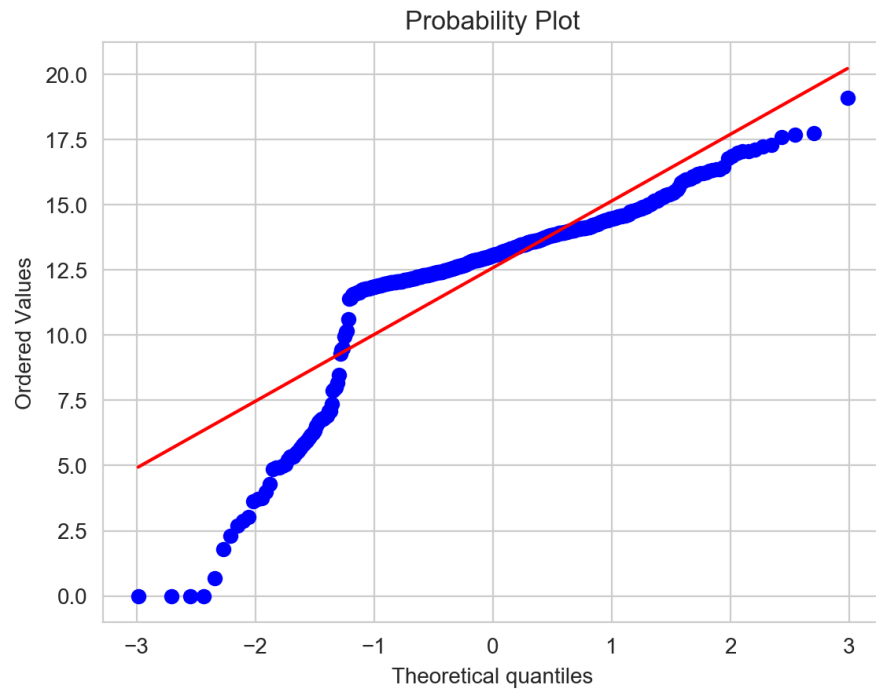


Figure 14 Probability plot of Followers Distribution for all username in the dataset

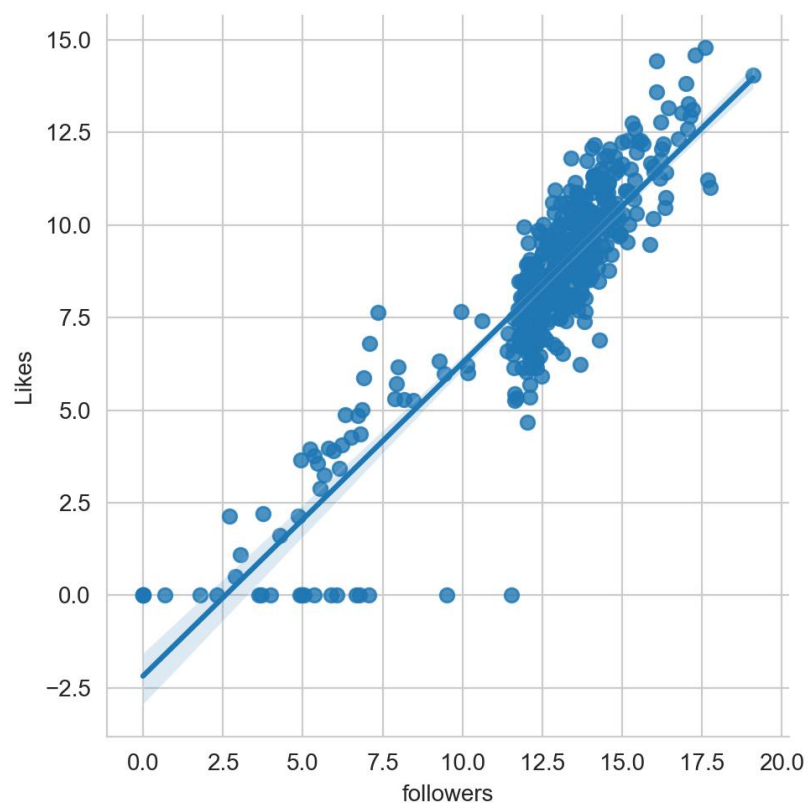


Figure 15 Linear model

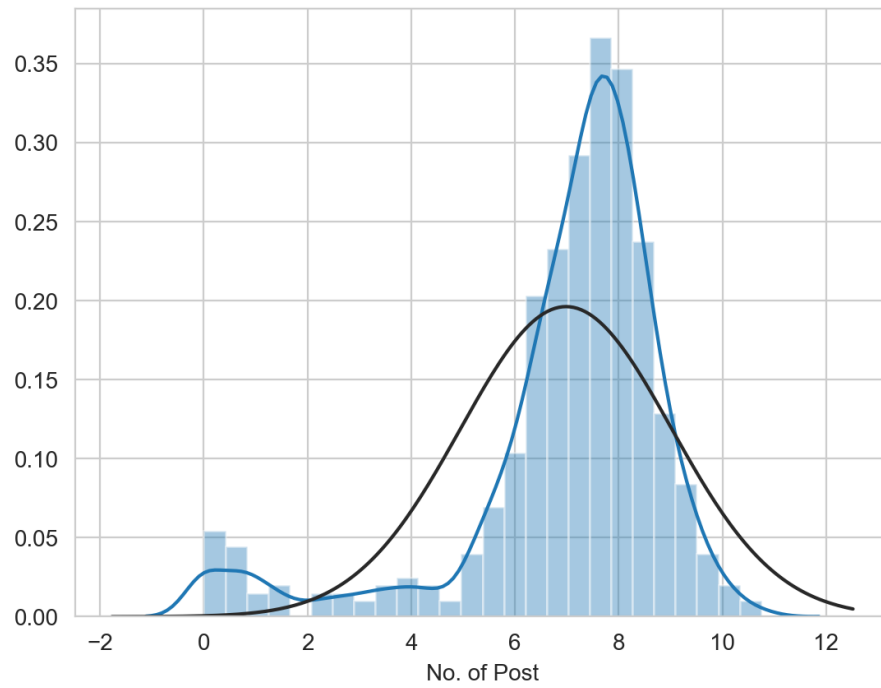


Figure 16 Posts Distribution for all username in the dataset

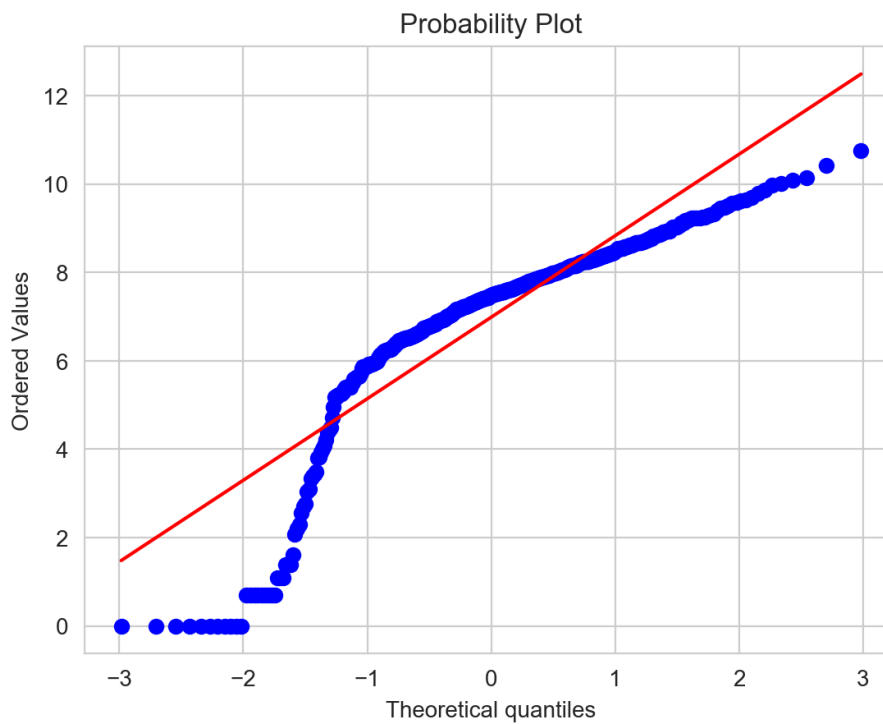


Figure 17 Probability plot of Posts Distribution for all username in the dataset

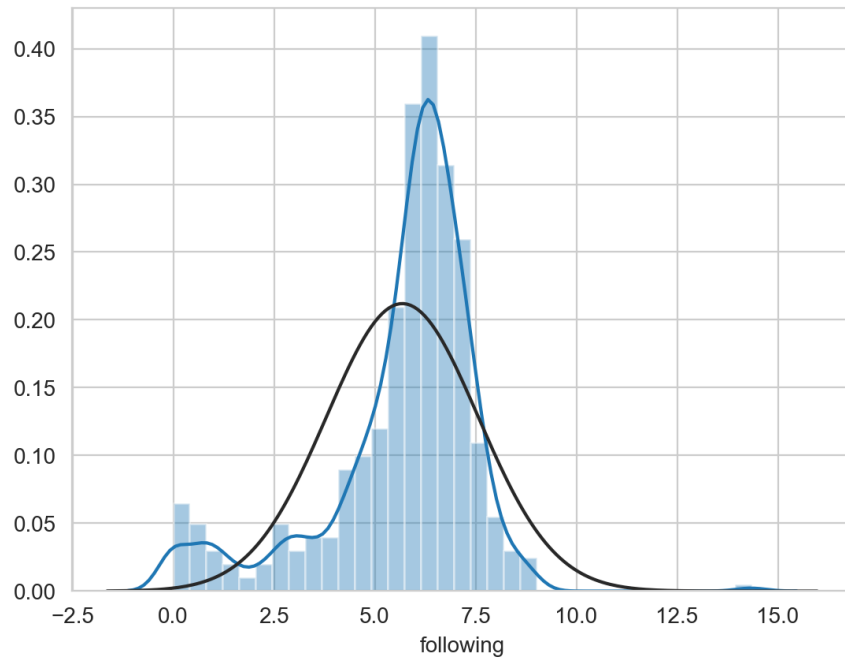


Figure 18 Following Distribution for all username in the dataset

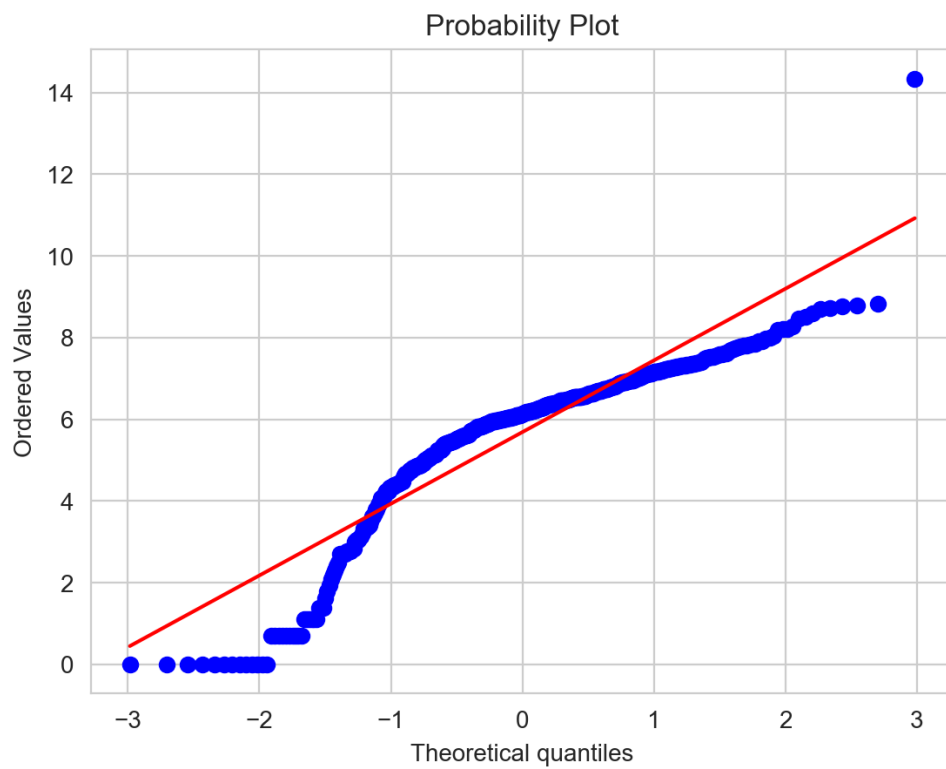


Figure 19 Probability Plot Following Distribution for all username in the dataset

Chapter 4. System Design and Specifications

The project was built on the local machine. The performance implies on the thread the CPU can handle and how much RAM is present on the machine to store the data temporarily on the RA

4.1 Hardware Specification

The hardware specification of the local machine used to build the project are as follows

Model: MSI GV62 8RE

Processor: Intel® Core™ I7-8750H GPU @ 2.20GHz (12 CPUs), ~2.2GHz

Memory: 16GB

Storage: 256GB SSD,
1TB HDD

GPU: Intel® UHD Graphic 630 4GB
NVIDIA GeForce GTX 1060 6GB

Operating System: Windows 10 Home Single Language

4.2 Technologies Used

We used the following libraries

- 1) OS: No Need to install it comes preinstalled with python. This OS is required to read the system file path.
- 2) NumPy: To use this library, we need to install it first and then we can use it. (Install it via pip install NumPy via command prompt/terminal). This is used to perform linear algebraic functions.
- 3) Pandas: To use this library, we need to install it first and then we can use it. (Install it via pip install pandas via command prompt/terminal). This is used to perform various functions, including reading a huge CSV file.
- 4) Matplotlib: To plot charts
- 5) Sklearn: To use this library, we need to install it first and then we can use it. (Install it via pip install sklearn via command prompt/terminal). This is used to perform various functions such as creating a confusing matrix, implementing models splitting data into test and train data, model reports Calculating Precision, Recall, F1 Score.
- 6) Seaborn: To use this library, we need to install it first and then we can use it. (Install it via pip install seaborn via command prompt/terminal). This is used to plot various interactive graphs.

- 7) Streamlit: To use this library, we need to install it first and then we can use it. (Install it via pip install streamlit via command prompt/terminal). This is used to show all work on a web browser. We have used Streamlit framework so that we can interact with our prediction data quite a user appealing and looks user friendly.
- 8) Instagram-explore library: To use this library, we need to install it first and then we can use it. This is used to gather Instagram data in JSON format without any difficulties.

Chapter 5. Implementation

The dataset must contain at least more than 200 user's data containing to predict the attribute for particular data. The dataset must include both kinds of users common and influencer so the data won't be biased. After getting the dataset Exploratory Data Analysis (EDA) has been performed. Where dataset is to be summarized the main characteristic of data, discover a pattern, process the null values from the dataset that is, discards the null row or performing some imputation technique to fill the null values. The EDA process also includes plotting the dataset on the graph to get an idea to of relationships between the columns. To use linear regression for building the model, it is necessary to remove the correlated variable to get better accuracy with the built model. We can visualise this with the heatmap which will give us the idea of the confusion matrix. We can also plot the distribution graph to check the skew in the dataset. As some of the datasets was not showing a good skewed feature on the graph so we have applied $\log(1+x)$ to all the attributes of the dataset. At this point, we have applied the EDA needed to the dataset.

This application was built using the python language and many libraries have been used to give the application some major functionalities. The libraries like NumPy, Pandas, Sklearn, Instagram-Explore, Selenium, Seaborn, Streamlit, Scipy and matplotlib has been used. At the initial stage of the project, we were using Instagram-Explore to collect the Instagram data. This library was extracting the JSON data from the Instagram by using Pandas the required data was then saved into a data frame and saved them in a Comma Separated Values (CSV) file. But at a stage the Instagram-Explore library got some error due to constant changes on Instagram data policies and a new feature called reels is been deployed to the platform. Therefore, to overcome we the problem and gather the real-time data, we have to use selenium web driver to build a crawler that will scrap the internet and collect the recent data that is last 10 – 13 post of every account. A list of usernames has been supplied to the crawler. The list includes all types of user that can be normal individual, influencers and business. The collected data is then saved to the CSV file.

Now for the real challenge: scraping the Instagram profiles of these users. This involves reading the metadata from the profile (number of followers/following, number of posts, description of the profile), crawling the 10 latest posts from the users (image in JPG format, number of likes, number of comments, timestamp, description text). The final result is a JSON file for each user that looks like this:



Suvajit Jana

Chapter 6. Testing and Evaluation

There were a few test cases while implementing the application.

1. What does the crawler do if the account is private?

The Crawler will successfully add the data of particular user registering the follower count, the following count, etc. all metadata will be register but the information about the post will null.

2. What does the crawler do if there any changes in the script while loading the scripts?

We have implemented the case so if the crawler does not find the script the crawler will find and gather information from different Script.

3. How the data is stored in the CSV file?

The crawler gathers the data in one data frame and appends the data frame into the CSV file according to the column name

4. How much response and crawler handles?

The Crawler can send on 1000 request to Instagram to collect the data

5. Does GUI response as per the expectations?

Yes, the GUI responsive and working as expected. As we have used streamlit library which makes it useful to build an attractive and responsive GUI

Chapter 7. Conclusion and Future Work

In this project enterprises, multi-million companies, etc. can use this extract data information for the commercial purpose. This application work as a third party where an organisation can gather information like count follower, followings, trends, tags, etc. this gather information helps them to choose a suitable celebrity, person, individual for their advertising and marketing purpose. Even this help gather an audience and promote their product or services on a suitable page using trendy tags.

For future work, I want to make this application more appealing and want to enhance the accuracy of output by selective more suitable models for this application which can help the application to predict the more accurate output. I also want to use trends in the future to enhance the performance of the application.

References

- [1] M. K. Ahmet Anil Mungen, "Mining quad closure patterns in Instagram," in *ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2016.
- [2] R. I. A. T. Emilio Ferrara, "Online popularity and topical interests through the lens of Instagram," 2014.
- [3] J. G. Miles, *Instagram Power*, McGraw-Hill, 2014.
- [4] The Editors of SUCCESS Magazine, "Book Review Instagram Power by Jason G. Miles," March 2014. [Online]. Available: http://videoplus.vo.llnwd.net/o23/digitalsuccess/SUCCESS%20Book%20Summaries/2014%20March%20SBS/InstagramPower_Review.pdf.
- [5] H. S. A. Y. Alireza Zohourian, "Popularity prediction of images and videos on Instagram," IEEE, 2018.
- [6] K. T. a. R. Hübner, "Instagram Likes for Architectural Photos Can Be Predicted by Quantitative Balance Measures and Curvature," *frontiersin*, 2018.
- [7] K. B. E. a. W. E. Dion, "What is beautiful is good.," *Psychol*, 1972.
- [8] A. H. A. R. D. M. M. G. a. L. L. C. Eagly, "What is beautiful is good, but? A meta-analytic review of research on the physical attractiveness stereotype.," *Psychol. Bull.*, 1991.
- [9] N. K. A. a. I. D. Tractinsky, "What is beautiful is usable. *Interact.*," *Comput.*, 2000.
- [10] V. W. (. Krohn, "Die ästhetischen Dimensionen der Wissenschaft," *Ästhetik in der Wissenschaft*, 2006.
- [11] T. C. Z. Z. J. I. Zhongping Zhang, "How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention," IEEE, 2018.
- [12] L. M. S. K. Yuheng Hu, "What We Instagram: A First Analysis of Instagram Photo Content and User Types".
- [13] J. L. d. O. A. & D. A. María Del Rocío Bonilla, "he interaction of Instagram followers in the fast fashion sector: The case of Hennes and Mauritz (H&M)," *Journal of Global Fashion Marketing*, 2019.