A Project Report on

Predicting Instagram Influencer Engagement Rate

Submitted in partial fulfillment of the requirements for the award
of the degree of
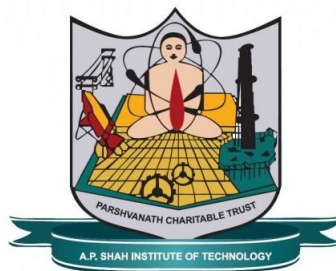
Bachelor of Engineering

in

Computer Engineering

by

Aishwarya More(18102008)
Shweta Patil(19202002)
Preksha Vora(19202004)
Moksha Shah (18102057)

Under the Guidance of

Prof. Shafaque Fatma sayed

Department of Computer Engineering
NBA Accredited
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI

Academic Year 2021-2022

Approval Sheet

This Project Report entitled *"Predicting Instagram Influencer Engagement rate"* Submitted by *"Aishwarya More"(18102008),"Shweta Patil"(19202002),"Preksha Vora"(19202004),"Moksha Shah"(18102057)* is approved for the partial fulfillment of the requirement for the award of the degree of **Bach**-elor of Engineering in Computer Engineering from University of Mumbai.

Prof. ShafaqueFatmaSyed

Guide

Prof. Sachin Malave
Head Department of Computer Engineering

Place:A.P.Shah Institute of Technology, Thane
Date:

CERTIFICATE

This is to certify that the project entitled *"Predicting Instagram Influencer Engagement rate"* submitted by *"Aishwarya More" (18102008)," Shweta Patil" (19202002),"Preksha Vora" (19202004),"Moksha Shah" (18102057)* for the partial fulfillment of the requirement for award of a degree **Bachelor of Engineering** in **Computer Engineering** ,to the University of Mumbai, is a bonafide work carried out during academic year 2021-2022.

Prof. Shafaque Fatmasyed
Guide

Prof. Sachin Malave   Dr. Uttam D.Kolekar
Head Department of Computer Engineering   Principal

External Examiner(s)

1.

2.

Place:A.P.Shah Institute of Technology, Thane
Date:

Acknowledgement

We have great pleasure in presenting the report on *Predicting Instagram Influencer Engagement rate* We take this opportunityto express our sincere thanks towards our guide **Prof. Shafaque Fatma syed** & Department of Computer, APSIT thane for providing the technical guidelines and suggestions regarding line of work. We would like to express our gratitude towards his constant encouragement, support and guidance through the development of project.

We thank **Prof. Sachin Malave** Head of Department,Computer, APSIT for his
en-couragement during progress meeting and providing guidelines to write this report.

We thank **Prof. Amol Kalugade** BE project co-ordinator, Department of Computer, APSIT for being encouraging throughout the course and for guidance.

We also thank the entire staff of APSIT for their invaluable help rendered during the course of this work. We wish to express our deep gratitude towards all our colleagues of APSIT for their encouragement.

Student Name1:Aishwarya More
Student ID1:18102008

Student Name2:Shweta Patil
Student ID2:19202002

Student Name3:Preksh vora
Student ID3:19202004

Student Name4:MokshaShah
Student ID4:18102057

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

————————————————

(Signature)

————————————————

(Aishwarya More 18102008)
(Shweta Patil 19202002)
(Preksha Vora 19202004)
(Moksha Shah 18102057)

Date:

Contents

Chapter 1

Introduction

Benefit of Machine Learning is allows the user to feed a computer algorithm an immense amount of data and have the computer analyze and make data-driven recommendations and decisions based on only the input data.The basic objective of AI (also called heuristic programming, machine intelligence, or the simulation of cognitive behavior) is to enable computers to perform such intellectual tasks as decision making, problem solving, perception,understanding human communication (in any language, and translate among them).

Basis on this , we are trying to make system that will be helpful for brands which are trying to collaborate with Instagram influences based in their engagement rate and growing and declining phases in basis of their recent activities.

Instagram influencers are regular Instagram users but with a unique ability to influence others, well-established credibility, and a considerably large audience. From a marketing perspective, an influencer is defined as a person with the power to influence potential buyers of a product or service.

Chapter 2

ProjectConcept

2.1 Abstract

Predicting the popularity of posts on social networks has taken on significant importance in recent years, and several social media management tools now offer solutions to improve and optimize the quality of published content and to enhance the attractiveness of companies and organizations. Scientific research has recently moved in this direction, with the aim of exploiting advanced techniques such as machine learning, deep learning, natural language processing, etc., to support such tools. In light of the above, in this work we aim to address the challenge of predicting the popularity of a future post on Instagram, by defining the problem as a classification task and by proposing an original approach based on Gradient Boosting and feature engineering, which led us to promising experimental results.

2.2 Objectives

1 Study Influencer Marketing and its process as one of the emerging field of marketing in digital space.
2 Overview of Sentiment Analysis and its effects and consequence
3 Exploring the practice for deciphering the parameters of influence within the application of sentiment analysis to evaluate the effectiveness of Influencer
4 Explore the opportunities and challenges of using sentiment analysis to evaluate Influencer Marketing

2.3   Literature Review

Published Year	Author	Title	• Methodology used	Algorithm used
2020	Kanwal Zahoor; Narmeen Zakaria Bawany; Soomaiya Hamid	Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning	The focus of this paper is to analyze the customer reviews about various restaurants across Karachi. The contribution of this research is twofold. First, it performs sentiment analysis and classifies each comment as positive, negative.	Random forest algorithm (95% accuracy)
2019	Arry Akhmad Arman	Measurement of Engagement Rate in Instagram	The main objective in this study is to take an explorative engagement approach dealt with: government data processing; social networking (online); Internet; learning (artificial intelligence); mobile computing; computer aided instruction; smart cities; Big Data; educational institutions; Internet of Things	Random forest algorithm (95% accuracy)
2020	Badra Al Aufa; Wahyu Sulistiadi; Faizah Abdullah Djawas	Measuring Instagram Activity and Engagement Rate of Hospital: A Comparison Before and During COVID-19 Pandemic	The study aimed to identify the hospitals' frequency post in their Instagram account and the engagement rate before and during the COVID-19 pandemic.	Mann whitney test
2019	Yshika Agarwal; Dilip Kumar Sharma; Rahul Katarya	Sentiment/Opinion Review Analysis: Detecting Spams from the good ones!	In this article, we go through this in a step by step format of different papers and summarize for other readers how we can identify the correct emotions and differentiate between the real and fake reviews.	coupled stochastic mode
2019	S M Asiful Huda; Md Mohiuddin Shoikot; Md Anower Hossain; Ishrat Jahan Ila	An Effective Machine Learning Approach for Sentiment Analysis on Popular Restaurant Reviews in Bangladesh	In this study, we built a model using natural language processing techniques and machine learning algorithms to automate the approach of classifying a review on around 200 popular restaurants of Bangladesh as Satisfactory or Poor.	Support Vector Machine
2018	Achmad Fauzi Azmi; Indra Budi	Exploring practices and engagement of Instagram by Indonesia Government Ministries	The results of analysis of ministry Instagram data revealed that the engagement level of the people does not correlate with the frequency of Instagram ministries post.	Linear regression
2018	Roy Ling Hang Yew; Syamimi Binti Suhaidi; Prishtee Seewoochurn; Venantius Kumar Sevamalai	Social Network Influencers' Engagement Rate Algorithm Using Instagram Data	Research has been done about related case studies and existing algorithms to get a better understanding of how to use available metrics such as likes, comments and followers, to calculate the engagement rate of influencers in the most accurate way.	Xg-Boost

2.4   Problem Definition

The challenge is to recommend influencer(s) for a particular marketing campaign. Companiesmight be interested in micro influencers for a certain industry/region and they would also liketo understand which content works for the audience of a particular influencer.

After finding the right influencers they might want to start a marketing campaign with them and monitor the efficiency of such a campaign.

we built a smart search algorithm that would take into account these attributes as well as various metrics. Finally, we designed a machine learning based recommender that links content with the influencers based on their audience. This content would serve as an inspiration to the campaign manager.

2.5   Scope

1.The main idea behind this project is to develop a system which can recommendinfluencer(s)
for a particular marketing campaign.

2.Analyze the sentiment of the comment section of the post and identify the number of
  positive and negative comments.

3.Classify by% of female and male followers in profile. I.e. e.g. It would be easy for a cosmetic brand
to analyze and influence them with maximum female followers becauseprofiles maximum male
followers will not give them so much commitment.

2.6   Technology Stack

Front- End
- Html-HTML (Hypertext Markup Language) is the code that is used to structure a web page and its content. For example, content could be structured within a set of paragraphs, a list of bulleted points, or using images and data tables.
- CSS-CSS is the acronym of "Cascading Style Sheets". CSS is a computer language for laying out and structuring web pages

Back-End
- PhpMyadmin - phpMyAdmin is a free software tool written in PHP that is intended to handle the administration of a MySQL or MariaDB database server.

Algorithm used for testing:
- Linear Regression- Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable.
- Xg boost- XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

- Random Forest - Random forest is a supervised learning algorithm Random forest algorithm can be used for both classifications and regression task. It provides higher accuracy through cross validation.
- **Python** is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python is designed to be highly readable.
- **Scikit-learn** is a free software machine learning library for the Python programming language.Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.
- **NumPy** stands for Numerical Python.NumPy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices.
- **Matplotlib** is a low level graph plotting library in python that serves as a visualization utility. Matplotlib is open source and we can use it freely.
- **Numeric**, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities. In 2005, Travis Oliphant created the NumPy package by incorporating the features of Numarray into the Numeric package

2.7   Benefits for Environment and Society

- You can develop a more insightful, data-based marketing strategy. Nothing beats data-based strategy. ...
- Understand your customers. ...
- Measure your marketing campaign. ...
- Take a look at brand perception. ...
- Find industry leaders and influencers. ...
- Give extra boost to your customer service.

Chapter 3

Project Design
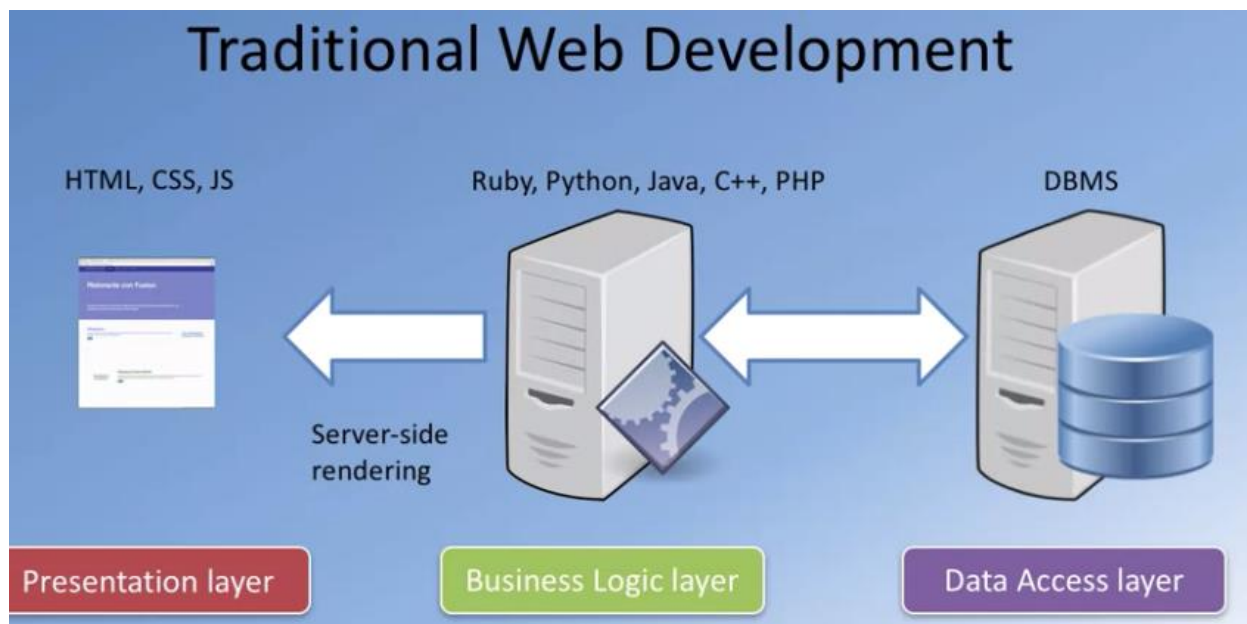
1.1 Proposed System

Fig.1

- The presentation layer which is concerned with delivering the information to the user. So, those are usually the UI related concern that are dealt with at the presentation layer. Here we usually have HTML, CSS, and JavaScript.
- The business logic layer on the other hand is concerned more about the data , the data validation, the dynamic content processing, and generating the content to be delivered to the user. Here we usually have, Node, Python, Java, c++ or PHP.
- This is backed up behind the scenes with the data persistence layer or the data access layer. So, this is concerned with how we store and interact with the data, typically in the form of a database and access this data through an API. SO business logic layer interacts with the database like MongoDB or PostgreSQL.
- **Reference**: https://fbohz.com/posts/2020-08-10-three_tier_full_stack/
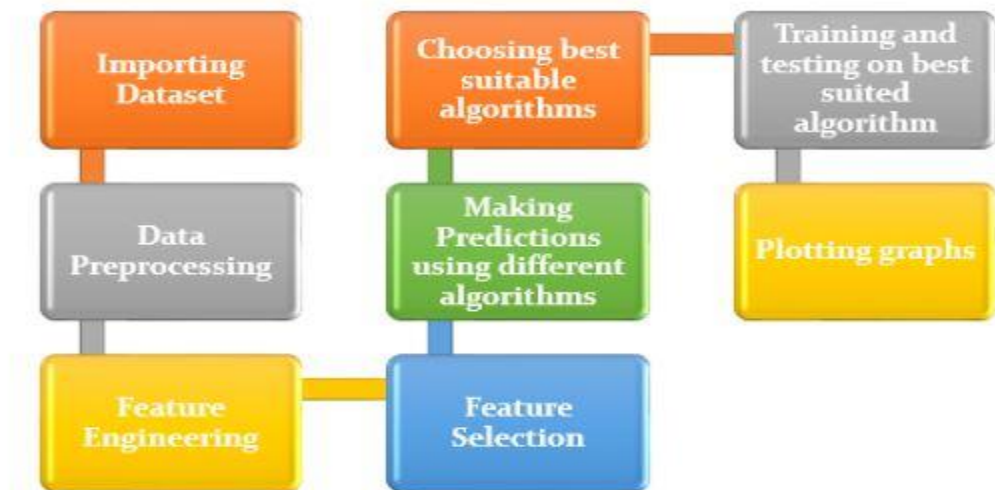
1.2 Design (Flow of Modules)

Fig.2

- Importing Dataset

- Data Preprocessing

- Feature Engineering

- Choosing best suitable algorithm

- Making prediction using different algorithms

- Feature Selection

- Training and testing on best suited algorithm

- Plotting graph

    Above steps are explain in Module 1.4
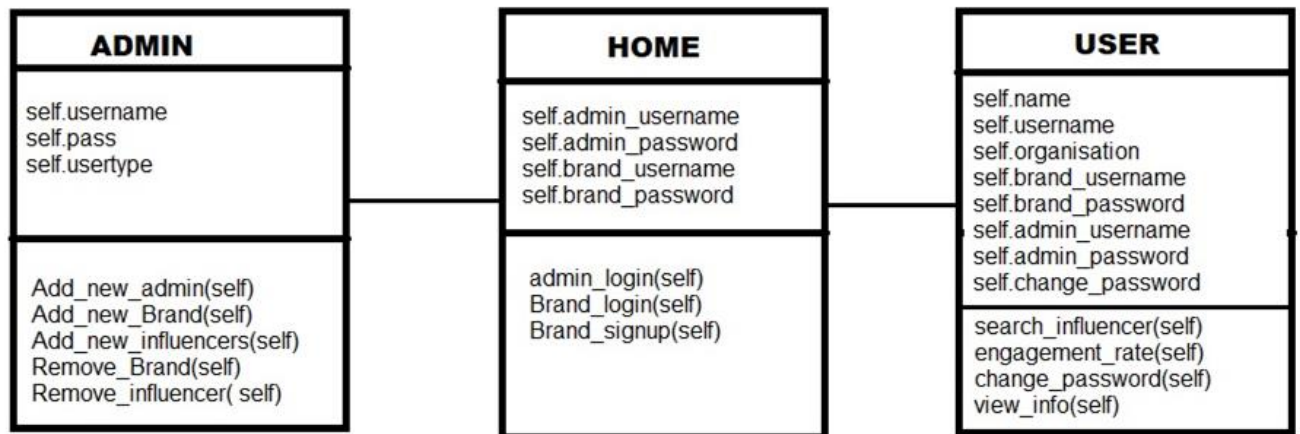
1.3   Class Diagram

Fig.3

- Admin can add new Brands, Influencer and also remove Brand, Influencer. Admin can also add new admin
- In Home admin can login, brand login and brand signup
- User can search the influencer, engagement rate, change password, and view information

1.4   Modules

1.4.1   Module 1

Importing Dataset
- We have taken the dataset from github. It has 11 column and 52687 rows.

Data Preprocessing
- Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it.
- Preprocessing data starts from data cleansing, feature engineering, feature selection and etc until the data is ready to be consumed by the Machine Learning model.
- Four stages: data cleaning, data integration, data reduction, and data transformation

1.4.2   Module 2

Feature Engineering:-

- Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.

Feature selection:-

- The next step is Feature Selection, The method I use is to look at the Correlation Coefficient between the Predictor and the Target Features.

- The goal of feature selection in machine learning is to find the best set of features that allows one to build useful models of studied phenomena.

- There are supervised and unsupervised techniques.

- From a taxonomic point of view, these techniques are classified as under:

- A. Filter methods

- B. Wrapper methods

- C. Embedded methods

- D. Hybrid methods

Correlation Coefficient

- Correlation is a measure of the linear relationship of 2 or more variables.

- If two variables are correlated, we can predict one from the other. Therefore, if two features are correlated, the model only really needs one of them, as the second one does not add additional information

1.4.3   Module 3

Making predictions using different algorithms

Modeling uses Machine Learning Algorithm (Linear Regression, Random Forest, XGBoost) and also do some Tuning Hyperparameters.

Hyperparameters Tuning

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with   existing data, we are able to fit the model parameters.

GridSearchCV

- In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

In our project we will be considering 4 cases:

- Modelling without Feature Selection and without Tuning Hyperparameters.

- Modelling without Feature Selection and with Tuning Hyperparameters.

- Modelling with Feature Selection and without Tuning Hyperparameters.

- Modelling with Feature Selection and wit Tuning Hyperparameters.

Metrics evaluation that we will be using is Root Mean Squared Error (RMSE)

- Root mean square error or root mean square deviation is one of the most commonly used measures . It shows how far predictions fall from measured true values using Euclidean distance.RMSE is commonly used in supervised learning applications, Lower values of RMSE indicate better fit.

Choosing best suitable algorithms

- The algorithm which will give lowest rmse value will be chosen for final prediction and results will be evaluated

1.4.4   Module 4

Training and testing on best suited algorithm

- Whichever RMSE value is least that will be better fitted so we will choose that algorithm to make our final predictions.Our main target is to predict the engagement rates of different users and based on that the user will be categorized in to "Growing" or "Declining" phase. Brand will be able to select Influencers basis on which phase they are.

Plotting Graphs

- Making graphical representations of the total influencers how many of them are in growing/declining phases.

- Also sorting feature importances in descending order using bar graphs

1.5   References

[1]Kanwal Zahoor,Narmeen Zakaria Bawany,Soomaiya Hamid, "Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning",2020 21st International Arab Conference on Information Technology (ACIT), DOI: 10.1109/ACIT50332.2020.9300098.

[2]Arry Akhmad Arman; Agus Pahrul Sidik, "Measurement of Engagement Rate in Instagram" (Case Study: Instagram Indonesian Government Ministry and Institutions), 2019 International Conference on ICT for Smart Society (ICISS), DOI: 10.1109/ICISS48059.2019.8969826

[3]Badra Al Aufa; Wahyu Sulistiadi; Faizah Abdullah Djawas, "Measuring Instagram Activity and Engagement Rate of Hospital: A Comparison Before and During COVID-19 Pandemic", 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI),DOI: 10.1109/ISRITI51436.2020.9315490.

[4]Yashika Agarwal; Dilip Kumar Sharma; Rahul Katarya, "Sentiment/Opinion Review Analysis: Detecting Spams from the good ones!", 2019 4th International Conference on Information Systems and Computer Networks (ISCON),DOI: 10.1109/ISCON47742.2019.9036249

[5]S M Asiful Huda; Md Mohiuddin Shoikot; Md Anower Hossain; Ishrat Jahan Ila, "An Effective Machine Learning Approach for Sentiment Analysis on Popular Restaurant Reviews in Bangladesh", 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS).

[6]Achmad Fauzi Azmi; Indra Budi, ""Exploring practices and engagement of Instagram by Indonesia Government Ministries", 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE).

[7]Roy Ling Hang Yew; Syamimi Binti Suhaidi; Prishtee Seewoochurn; Venantius Kumar Sivamalai, "Social Network Influencers Engagement Rate Algorithm Using Instagram Data" ,2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)

Chapter 4

Planning for next semester