Prediction of Social Media Consumer Behaviour Using Machine Learning

Team-2

Muhammad Sajid - MA23M013
Prabhat Kumar - MA23M017
Pritam Ray - MA23M018
Sourav Majhi - MA23M022
Suman Das - MA23M023
Vinod Kumar - MA23M026

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS

Project Report submitted for

MA5770-Modelling Workshop

Under the Supervision of
Dr. Priyanka Shukla
Associate Professor
Department of Mathematics
Indian Institute of Technology Madras, India

Abstract

Social media platforms have become an essential part of our everyday lives in the digital age, impacting the way we connect, communicate, and consume information. It is imperative for organizations to comprehend consumer behavior on these platforms if they aim to improve brand perception, customize marketing tactics, and stimulate customer interaction. The following major questions are the focus of our research: How do consumers view and engage with brands across various social media channels? What elements affect user attitude, purchasing decisions, and engagement? Is it possible to forecast customer behavior with the data at hand? We gather information from well-known social media sites like Facebook, Twitter, LinkedIn, YouTube, Instagram, and Pinterest in order to answer these questions. A wide variety of user-generated content, including posts, comments, likes, shares, and hashtags, are included in our collection. Through the amalgamation of data from various platforms, we establish an all-encompassing perspective on consumer behavior. In order to guarantee the integrity and quality of the data, we employ pragmatic preprocessing methods such as outlier identification, noise reduction, error correction, and duplicate record elimination. Instead of depending exclusively on data analytics, we investigate a range of machine learning algorithms, encompassing linear regression, polynomial regression, decision trees, random forest, and XGBoost. Our comparative analysis produces a number of findings, including platform-specific behavior and feature significance. The comprehension of consumer behavior enables marketers to customize content, enhance advertising campaigns, and forecast trends. Potential areas of inquiry in the future include real-time monitoring, user segmentation, and the influencer marketing phenomenon.

Keywords: Data Analytics, Predictive Model, Consumer behaviour, Linear Regression, Decision Tree, Random Forest, Xg-Boost.

Acknowledgement

We would like to express our profound gratitude to **Dr. Priyanka Shukla,** Associate Professor , from the Department of Mathematics at IIT Madras, India.

Her invaluable guidance and deep expertise this project possible and also incredibly interesting. The project was very interesting through out the semester and we always looked forward to work on it and with her.

We would also like to extend our sincere thanks to **Professor Mansaf Alam** from the Department of Computer Science at Jamia Millia Islamia, India. His generous provision of the data set was a critical component of our study.

Lastly, we would like to acknowledge our classmates for their consistent attendance and engagement during all presentations throughout the semester. Their participation enriched our learning experience and contributed to the project's success.

We are deeply grateful for everyone's contributions and look forward to future collaborations.

Team 2
M.Tech. IMSC, 2023-2025
Department of Mathematics
IIT Madras

Contents

1 Introduction

Now-a-days social media has become the hub for all kind of activities, including shopping. Data from social media provides significant insights into individual behavior and is extensively analyzed by researchers[2]. Data analytics and machine learning algorithms are instrumental in monitoring social media to refine brand management strategies for luxury hotels[3]. Data from thousands of startups on Twitter has been used to develop machine learning models that predict social media engagement, with deep learning yielding the highest accuracy[4]. The importance of a company's social media activity, such as tweets, retweets, and likes, is crucial for assessing the effectiveness of social media marketing. This paper also examines how big data analytics can assess consumer perception and assertiveness towards social media, influencing overall consumer behavior on these platforms.

We can't even imagine how people make decisions while scrolling through their feeds. But through this project that's exactly what we will try to explore. We gathered data from various platforms like Facebook, Twitter, LinkedIn, You Tube, Instagram etc. to analyze and predict consumer behaviour. We will make sure our data is clean and does not contain any kind of information that may mislead us. Also we want to ensure that the quality of the result to be good, so for that we pre-processed the data using various techniques like Pandas and Numpy which are famous libraries of python to detect outliers, noises, errors and duplicate records.

Basically, in this project we will utilize method of data analytics to process and analyze the data, aiming to forecast consumer behaviour on social media platforms. We will examine consumer behaviour on the selected social media platforms based on certain parameters and criteria including consumer perception and attitude towards the platform i.e. we have focused more on number of likes/downloads/visits on social media posts. Here we will demonstrate our work that how knowledge from user-generated data will help in understanding and improvement in the supply chain.

But to meet the demand of this project we will be needed mathematical and machine learning algorithmic support. So for that we will use various kind of supervised machine learning models or algorithms and will also check the result for each algorithms and will try to find which model will be best for this problem and up to which extent we can predict the behaviour of the consumer.

We will use linear regression, polynomial regression, decision tree regression, random forest regression and xg-Boost regression. These algorithms are solely based on the mathematical concepts. So we will also explain mathematical concepts for all the algorithms listed above in details. We will also write the codes for these algorithms to obtain the requires results. We will write code both from scratch and also by the use of scikit learn which is gain a very popular library of python. We will also use the another library of python i.e. matplotlib to visualise the results of models in plots.

Basically we will train these models using the data we have. But we can't just train the model, so we will not use the whole dataset to train the model, in spite we will split the dataset into two parts i.e. training dataset and testing dataset in the ratio of 80:20. Where training dataset will be for training the model and testing dataset will be for testing the model. By testing the model we can get the information about the working of model on unseen data.

So we can say that we will not just build the models, also will test these models rigorously. We will use different metrics to evaluate our model's performance, including error metrics like mean squared error(MSE), accuracy metrics like R-squared and root mean squared error(RMSE). We will use scikit learn to calculate these metrics and also will write code directly using mathematical formula. These accuracy scores and error analysis will help us in choosing the best model out of various models.

1.1 The Power of Social Media Data

We cannot underestimate the influence of social media on decision-making. As users scroll through their feeds, they encounter a wealth of information, opinions, and brand interactions. To explore this phenomenon, we gathered data from various platforms, including Facebook, Twitter, LinkedIn, YouTube, and Instagram. Our focus lies in analyzing and predicting consumer behavior based on specific parameters, such as the number of likes, downloads, and visits on social media posts.

1.1.1 Insightful Visualization

As we deep dive into our findings, we'll visualize the results of our findings using Matplotlib. But before that let's set the stage with two impact-full visuals:

- **Year-wise User Growth**: The below bar plot(Figure-1) illustrating the exponential rise in social media users over the years. This growth shows the platform's significance in shaping consumer behavior.

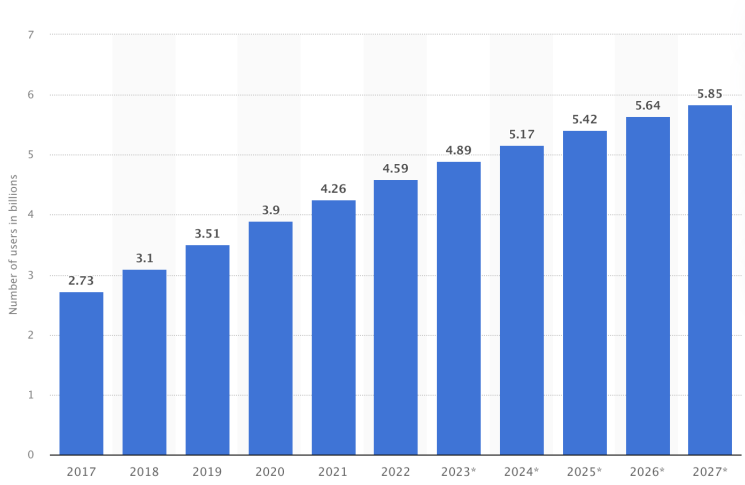

Figure 1 Increase in the Number users Year-wise [5]

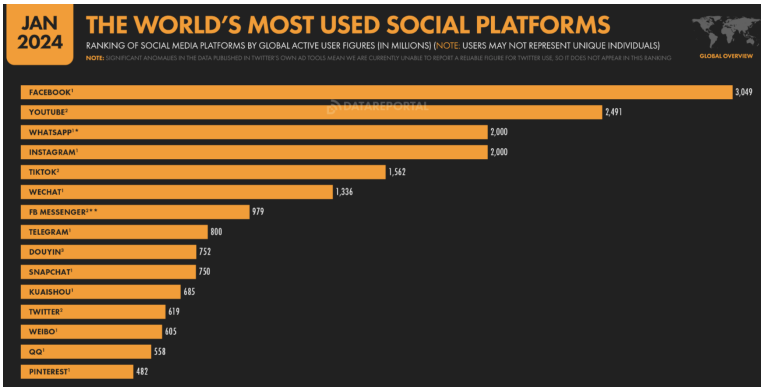

Figure 2 Most Popular Social Media Platforms [6]

- **Popular Social Media Snapshot**: The above figure-2 showcasing the most influential platforms like Facebook, YouTube, Twitter, Instagram, and more. These platforms wield immense power in shaping our opinions, market trends, and purchasing decisions of people.

By combining data analytics, machine learning, mathematical modeling and compelling visuals, Our goal is to understand the complex working of consumer activity on social media .

1.2 Machine Learning

Machine learning (ML) is a branch of computer science and artificial intelligence that focuses on using data, identify the patterns in the data using sophisticated mathematical algorithms to assist AI replicate human learning processes and get gradually more accurate and decisions .

How does machine learning works:

1. **A Decision Process:** Machine learning algorithms are generally used for making predictions and classifications. The algorithm will generate an approximation of a pattern in the data based on particular input data that might or might not be labeled.
2. **An Error Function:** An error function or loss function is a tool used to calculate a model's prediction. An loss function compare the known examples with the predicted values in order to evaluate the model's correctness. it helps to find the balance between bias(under-fitting) and variance(over-fitting).
3. **A Model Optimization Process:** If the model fits the training set's data points more accurately, the weights are modified to reduce the distance between the model estimate and the known example. Model optimization is the name given to this approach. The algorithm will repeat this "evaluate and optimize" process iteratively, updating weights (Gradient Descent is a popular one) on its own until the desired level of accuracy is attained.

1.3 Supervised Machine Learning

The machine learns under supervision of human i.e., explicit labeled data when we use supervised learning. It has a predictive model that uses a labeled dataset to make predictions or decisions. A labeled dataset is one in which the desired outcome is known beforehand. Our goal here is to predict the outcomes for new unseen data.

Pictorial Representation: The Figure-3 shows the different types of machine learning algorithms based on their workings.

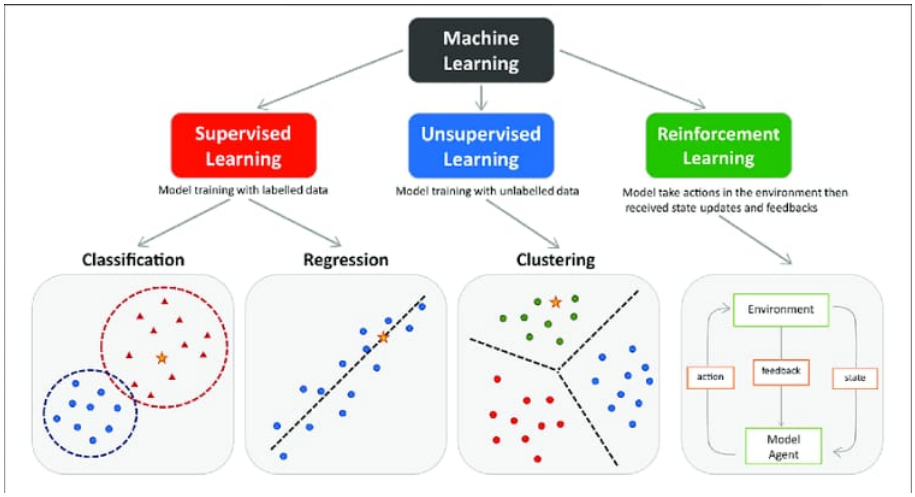

Figure 3 Machine Learning Divisions

Supervised learning can be further divided into two types:

1. Classification
2. Regression

Classification: When an output variable has two or more classes, it is categorical and uses classification. For instance, true or false, male or female, etc.

Regression: When the wanted output variable has a real or continuous value, or we want the output to be numerical value regression algorithms is used. This situation involves a relationship between two or more variables (input features), meaning that changing one will inevitably affect changing another. For instance, pay determined by job experience, weight determined by height, etc.

2 Motivation

The application of machine learning techniques to social media data provides marketers with a wealth of opportunities to enhance their strategies. By analyzing this data, they can gain enhanced insights into consumer behavior, preferences, and sentiments, allowing for more effective tailoring of marketing strategies.

Enhanced Customer Insights: By analyzing social media data using machine learning techniques, marketers gain deeper insights into consumer behavior, preferences, and sentiments. This understanding allows them to tailor marketing strategies more effectively.

Personalized Marketing: Machine learning models can predict individual user preferences based on their social media interactions. This enables personalized content recommendations, targeted ads, and customized promotions, leading to better engagement and conversion rates.

Trend Prediction: Social media platforms generate vast amounts of data. Machine learning algorithms can identify emerging trends, hashtags, and popular topics. Marketers can leverage this information to stay ahead of the curve and align their campaigns with current interests.

Optimized Ad Campaigns: Predictive models help optimize ad placement, timing, and content. By analyzing historical data, machine learning can determine the most effective channels, ad formats, and posting schedules for maximum impact.

Sentiment Analysis: Machine learning algorithms can analyze user sentiments expressed in social media posts. Marketers can gauge public opinion about their brand, products, or services and respond accordingly.

We are motivated by the paper "Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics"[1], our research adopts similar techniques to advance the field. The original work presented an integrated approach for implementing and analyzing social media marketing efficacy using machine learning, In our study, we have chosen Python as the analytical tool to conduct our data analysis, reflecting its robust capabilities in handling complex datasets within the domain of social media marketing.

3 Literature Review

A numerous number of research paper have been published in the area of consumer behaviour using machine learning approach. The paper [some number] which is our main reference paper includes the predictor methods of online consumer reviews using the machine learning and data analytics approach. (note: some more points need to be added for this paper) The main focus of paper [7] is on understanding the behaviour of consumer by taking the different types of social media into consideration. Basically, this paper's work is on predicting the psychological aspect of consumer's needs and not just on understanding it. In this paper they also have worked on the method in improving target advertising, and they demonstrated this thing by studying the consumer's behaviour and predicting their behaviour and also they suggested that by knowing what consumer will buy just after purchasing some particular goods helps too understand perception of the consumers towards brand.

In research paper [11] they investigate the both kind of sentiments i.e. positive and negative of people and also they investigate the reason behind their sentiments whatever it is. For analysis of brand authenticity and polarity of the sentiment, they used the 2204 code tweet's database. The examines the tweets in a qualitative way that helps them in knowing the sentiments of the consumers related to authenticity of the brand and then this can create a framework in a quantitative way where authors forecast both the things i.e. brand dimension's authenticity and their sentimental polarity. Based on numerous categories like heritage uniqueness, quality, commitment and symbolism authors have classifies the tweets. For extracting common words in each category LSA(Latent Semantic Analysis) is used and for brand authenticity predictions of dimensions and consumers polarity on sentiment the result shows higher accuracy.

4 Objective

1. **Developing a Mathematical Model**: The study aims to create a mathematical model that leverages machine learning algorithms. This model is designed to predict how consumers behave on social media platforms, which is a complex task due to the dynamic nature of user interactions and the vast amount of data generated.
2. **Processing and Analyzing Data:** Given the high volume and velocity of data from social media, the study uses data analytics to process and analyze this information. This involves handling data from multiple platforms like Facebook, Twitter, LinkedIn, YouTube, Instagram, and Pinterest to extract meaningful patterns related to consumer behavior.
3. **Data Pre-processing:** Before analysis, data must be cleaned and pre-processed. This step is crucial to ensure the accuracy of the predictive model. The paper discusses various data pre-processing techniques to detect and correct outliers, noise, errors, and duplicate records, which can significantly impact the model's performance.
4. **Analyzing Consumer Perception and Attitude:** The model also aims to analyze consumer perceptions and attitudes towards social media platforms. Understanding these aspects can help businesses to decide their marketing strategies to better align with consumer preferences and behaviors.
5. **Choosing the Best Predictive Model:** We will compare the effectiveness of the models which are linear regression, polynomial regression, decision tree, random forest and xg-boost.Then on the basis of errors and accuracy we will finalise the best model.

5 Research Gap

Here first of all we are going to discuss the work done by the authors of the paper we are taking as a reference.

They have mentioned various supervised machine learning algorithms but have discussed only one algorithm up to some extent. To be specific they have mentioned that they have used algorithms like linear regression, decision tree, random forest, xg-boost. But they have explained only linear regression in details and not explained the working of the other algorithms. Also, they did not instructed that how to take the data of their interest as input for the different different models i.e. it is not clear in their work that how data is being used. It was also necessary to add some snippet of the code to give the better understanding of the working of the model, but it is missing in their work. Also they have directly written the result and concluded that decision tree algorithm gives the best result but we are not fully satisfied with the thing that how they are getting this result.

Now we are going to give some glimpse of our work in this project.

We will explain all the algorithms in detail and also the working of these algorithms mathematically. We all also try to show the compatibility of the data with all these algorithms. We will add some snippets of the codes in between to give more insight on the working of the algorithms. We will also try to elaborate the calculation of the errors in better way i.e. we will also support these calculation with mathematical calculation and do not directly jump into the result. Also, we will try to do the same thing while comparing the models i.e. we will try to give the proper explanation that why we are selecting any specific algorithm as best one.

6 Methodology

6.1 Dataset Collection

First of all we tried to collect data from kaggle and other sites but we were not getting the desirable data of our interest. So we took data from the author of the paper that we are taking as main reference, we can see data from here(link of google drive for data). This dataset includes total of 5280 data and each data is the number of likes/follows/downloads on the posts from different platforms along with agencies and date.

6.2 Data Preprocessing

Although we had total 5280 data but all were not of our work, some data were incomplete which were not of our use at all, so we pre-processed our data by identifying useless data and removing them. For performing these required operation we have used a very famous tool of python called pandas. This preprocessing of the dataset was must for well working of the model with our dataset. After preprocessing the data and by removing the unnecessary data from the dataset the total number of data in the dataset left was around 3700, and we continued our whole work on the basis of this dataset itself.

6.3 Model Selection

We have compared the performance of different supervised machine learning algorithms such as linear regression, decision tree etc for prediction of social media consumer behaviour. These algorithms are well suited for the problem we tried to solve but the performance of different algorithms may differ due to the difference in their architecture.

Model Architecture: The Figure-4 shows us the step by step process we have followed throughout the project work.

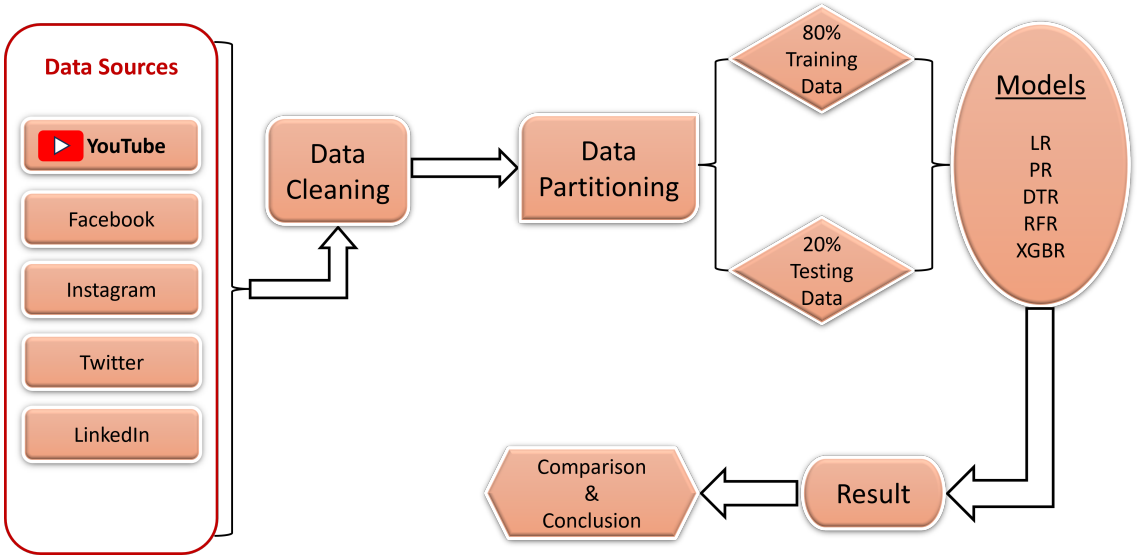

Figure 4 Consumer Behaviour Model

6.3.1 Linear Regression

It is a kind of supervised machine learning algorithm that allows us to fit a linear equation to observed data in order to calculate the linear relationship between the independent feature, which can be one or more than one, and the dependent variable, or label.

Mathematical Representation:

$$y = \beta_0 + \beta_1 X \tag{6.1}$$

Where, y is label, X is features, β_0 is intercept and β_1 are slopes.

This algorithm basically finds the equation of the best fit line that can predict the values based on the independent variables. Here we are using linear regression in our problem, but the format of model is somewhat different from usual linear regression models. We have used same data i.e. likes/downloads/visits for features and labels. Likes/downloads/visits of one social media platform is used as features and another is used as label. We have tried to find relation between different-different social media platforms using the mentioned model.

Thus to find the best fit line one should have the values of parameters β_i for $i = 0, 1$. For that we calculated cost function and tried to minimize it. After minimizing the cost value we will get our desirable value which is given by

Given the cost function $J(w_1, w_2)$, which is defined as:

$$J(w_1, w_2) = \frac{1}{2m} \sum_{j=1}^{n} (y_j - (w_1 + w_2 \cdot x_j))^2$$

We want to minimize this cost function with respect to parameters w_1 and w_2. To obtain the minimum of J, we have to take the partial derivatives of $J(w_1, w_2)$ with respect to both w_1 and w_2 and set them equal to zero.

The partial derivative of $J(w_1, w_2)$ with respect to w_1 is:

$$\frac{\partial J}{\partial w_1} = -\frac{1}{m} \sum_{j=1}^{m} (y_j - (w_1 + w_2 \cdot x_j))$$

Setting this equal to zero:

$$-\frac{1}{m} \sum_{j=1}^{m} (y_j - (w_1 + w_2 \cdot x_j)) = 0$$

Now, let's solve for w_1:

$$\sum_{j=1}^{m} (y_j - (w_1 + w_2 \cdot x_j)) = 0$$

$$\sum_{j=1}^{m} (y_j - w_1 - w_2 \cdot x_j) = 0$$

$$\sum_{j=1}^{m} y_J - nw_1 - w_2 \cdot \sum_{j=1}^{m} x_j = 0$$

$$\sum_{j=1}^{m} y_j = nw_1 + w_2 \cdot \sum_{j=1}^{m} x_j$$

$$w_1 = \frac{\sum_{j=1}^{m} y_j - w_2 \cdot \sum_{j=1}^{m} x_j}{m}$$

Similarly, let's find the partial derivative of $J(w_1, w_2)$ with respect to w_2:

$$\frac{\partial J}{\partial w_2} = -\frac{1}{m} \sum_{j=1}^{m} x_j (y_j - (w_1 + w_2 \cdot x_j))$$

Setting this equal to zero:

$$-\frac{1}{m} \sum_{j=1}^{m} x_j (y_j - (w_1 + w_2 \cdot x_j)) = 0$$

$$\sum_{j=1}^{m} x_j y_j - w_1 \sum_{j=1}^{m} x_j - w_2 \cdot \sum_{j=1}^{m} x_j^2 = 0$$

Substituting the expression we found for w_1:

$$\sum_{j=1}^{m} x_j y_j - \left(\frac{\sum_{j=1}^{m} y_j - w_2 \cdot \sum_{j=1}^{m} x_j}{m} \right) \sum_{j=1}^{m} x_j - w_2 \cdot \sum_{j=1}^{m} x_j^2 = 0$$

Now, let us solve this for w_2:

$$m \cdot \sum_{j=1}^{m} x_j y_j - \left(\sum_{j=1}^{m} y_j - w_2 \cdot \sum_{j=1}^{m} x_j \right) \sum_{j=1}^{m} x_j - m \cdot w_2 \cdot \sum_{j=1}^{m} x_j^2 = 0$$

$$m \cdot \sum_{j=1}^{m} x_j y_j - \sum_{j=1}^{m} y_j \sum_{j=1}^{m} x_j + w_2 \cdot (\sum_{j=1}^{m} x_j)^2 - m \cdot w_2 \cdot \sum_{j=1}^{m} x_j^2 = 0$$

$$w_2 \cdot (\sum_{j=1}^{m} x_j)^2 - m \cdot w_2 \cdot \sum_{j=1}^{m} x_j^2 = \sum_{j=1}^{m} y_j \sum_{j=1}^{m} x_j - m \cdot \sum_{j=1}^{m} x_j y_j$$

$$w_2 \cdot \left(\sum_{j=1}^{m} x_j^2 - \frac{1}{m} \left(\sum_{j=1}^{m} x_j \right)^2 \right) = \sum_{j=1}^{m} y_j \sum_{j=1}^{m} x_j - m \cdot \sum_{j=1}^{m} x_j y_j$$

$$w_2 = \frac{m \cdot \sum_{j=1}^{m} x_j y_j - \left(\sum_{j=1}^{m} x_j \right) \left(\sum_{j=1}^{m} y_j \right)}{m \cdot \left(\sum_{j=1}^{m} x_j^2 \right) - \left(\sum_{j=1}^{m} x_j \right)^2}$$

These are the expressions for w_1 and w_2. We derived them by setting the partial derivatives of the cost function with respect to w_1 and w_2 equal to zero, and then solving for w_1 and w_2, respectively.

Pictorial Representation:

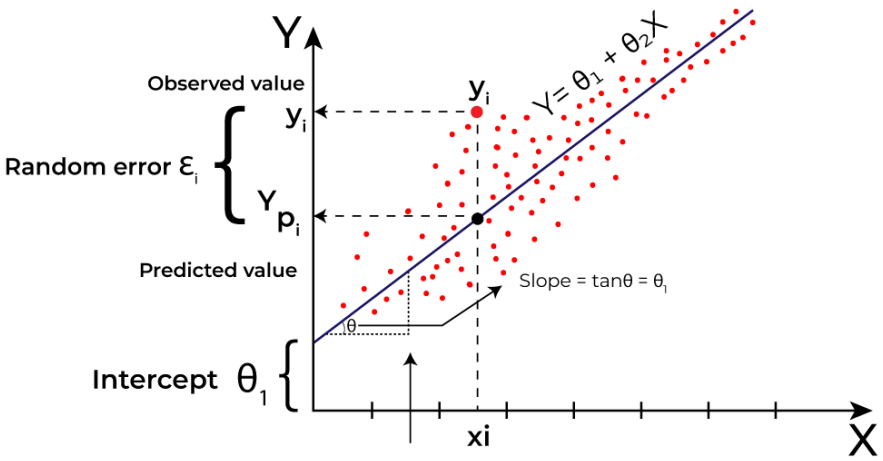

Figure 5 Linear Regression

6.3.2 Polynomial Regression

Polynomial regression is the name given to a special form of linear regression in which the polynomial equation is fitted to the existing data to reveal a curvilinear relationship between the independent variables (features) and dependent variable (label). The n^{th} degree polynomial reflects the relationship between the independent and dependent variables and shows us the nature of their connection .

Mathematical Representation:

Polynomial Regression model is mathematically represented as

$$y = b_0 + b_1 x + b_2 x^2 + ... + b_n x^m \tag{6.2}$$

Where y is label, x is feature, b_0 is bias and $b_i' s$ are weights.

Here also our implementation of polynomial regression is slightly different from the usual one. We have taken the features and labels in the same manner as of linear regression model.

Now, to find the model which fit the data in well manner, one need to find the value of $b' s$ i.e parameters in such a way that the cost function J is minimum. So the value of the vector b is calculated by matrix multiplication and for multiple variable the matrix calculation is done by using the below equation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots x_1^m \\ 1 & x_2 & x_2^2 & \dots x_2^m \\ \vdots & & & \\ 1 & x_n & x_n^2 & \dots x_n^m \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{6.3}$$

Pictorial Representation:

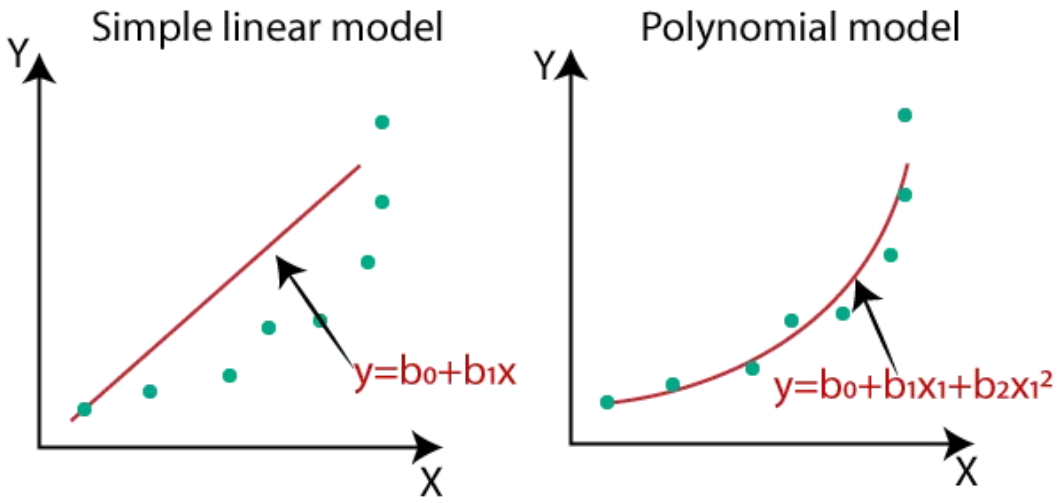

Figure 6 Polynomial Regression

6.3.3 Decision Tree Regression

Decision Tree Regression is a tree structure like algorithm having root node, internal node and leaf nodes. The internal nodes representing features depending on which the tree divides, branches

representing rules, and leaf nodes represents the output(desired result) of the algorithm. Among the supervised machine learning algorithms, it is regarded as one of the most potent. Decision trees use metrics to quantify the degree of impurity randomness in subsets, and during the training process, they choose the most suitable attribute to split the data depending on Information Gain. Our goal is to identify the characteristics that, following splitting, increase information or minimize impurities.

Mathematical Representation:

Here we can write the classification and regression tree algorithm for regression as

$$IG(Q_m) = Var(Q_m) - \left(\frac{n_m^{left}}{n_m} Var(Q_m^{left}) + \frac{n_m^{right}}{n_m} Var(Q_m^{right}) \right) \tag{6.4}$$

Where Q_m be the data available at the node m, n_m be it's number of samples, n_m^{left} is the number of nodes came to left sub-tree of node m and n_m^{right} is the number of nodes came to right sub-tree of node m. Also Var is variance of the data points and thus

$$Var = \frac{1}{n_m} \sum_{y \in Q_m} (\bar{y_m} - y)^2 \tag{6.5}$$

Where,

$$\bar{y_m} = \frac{1}{n_m} \sum_{y \in Q_m} y \tag{6.6}$$

Pictorial Representation:

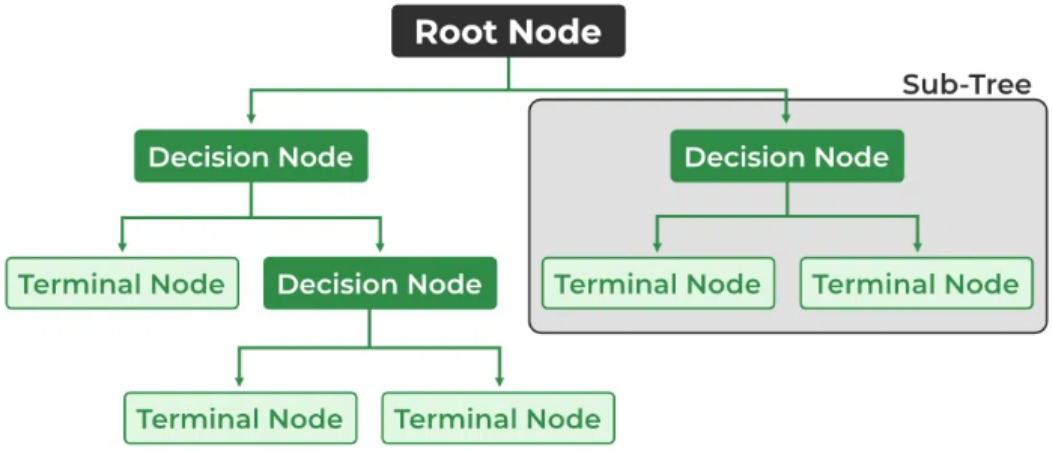

Figure 7 Decision Tree

6.3.4 Random Forest Regression

Random Forest Regression might be most popular bagging algorithm out there. Random Forest combines the output of several decision trees to produce a single output. Its versatility and ease of use, compatibility with large data set and other library combined with its ability to handle both regression and classification problems, have driven its popularity. Usually this algorithm is used to solve the problem of over fitting in decision tree algorithm. In over fitting training accuracy is high i.e. low bias and testing accuracy is low i.e. high variance, so we aim to reduce the variance to solve the problem of over fitting in decision tree algorithm using this i.e. random forest algorithm.

Working:

Let's take a training dataset of size m, then take k new training set from the original training dataset of size $m_i \leq m, i = 1, 2, ...$ Then train a decision tree on each of these dataset and then finally aggregate the output of these k results to get the final result.

So, basically the working of decision tree algorithm is based on bagging technique bagging comprises of two terms i.e. Bootstrapping and Aggregation. The step we performed by taking subsets of training dataset and training of base model(in random forest case decision tree is the base model) comes under bootstrapping and after that we combined the results got by training decision tree on different dataset to get the final result comes under aggregation.

6.3.5 Pictorial Representation:

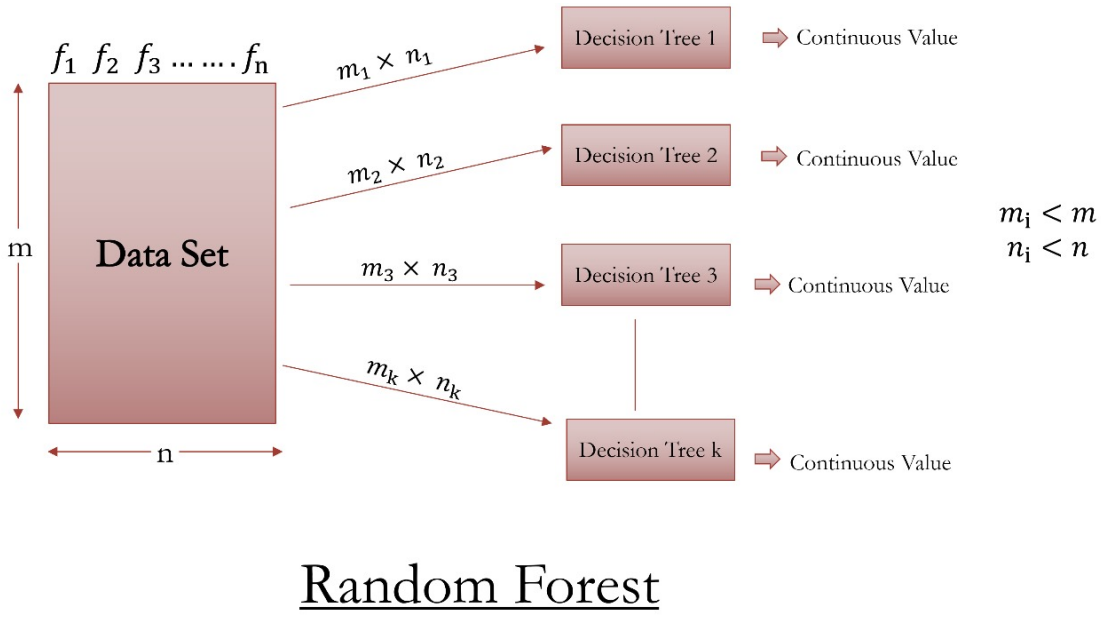

Figure 8 Random Forest

6.3.6 XG-Boost Regression

XGBoost is a generalized gradient boosting toolkit designed to maximize scalability as well as performance in the training of machine learning models. By combining the estimates of multiple weak models, this method of ensemble learning produces a better prediction. The machine learning technique known as Extreme Gradient Boosting, as well as XGBoost, has been well-liked and widely used due to its ability to manage enormous datasets and attain cutting-edge results in numerous machine learning applications, such as classification and regression. One of XGBoost's main benefits is its ability to accept missing values effectively, which allows it to deal with data from the real world containing incomplete data without having a great deal of pre-processing. Additionally, XGBoost has integrated support for processing in parallel, making it possible to swiftly train models on large datasets.

Mathematical Formulation:

This algorithm is based on boosting algorithm which is an ensemble technique that tries to construct a strong prediction using the various weak predictions and this task is performed by making a model by joining weak models in series. Boosting algorithm's main motive is additive model i.e.

$$G(y) = g_0(y) + g_1(y) + g_2(y) + ... \tag{6.7}$$

So as discussed, in our case each $f_i's$ will be decision tree model which is then multiplied with the learning rate.

Objective function:

Here the objective function $L(g)$ is the addition of the cost function $\mathcal{L}$ and the regularization term Ω:

$$L(g) = \sum_{k=1}^{M} \mathcal{L}(y_k, g(x_k)) + \Omega(g)$$

Where:

- M represents the number of training examples.
- y_k represents the ground truth value for example k.
- $g(x_k)$ represents the predicted value for example k.

Regularization Term: The regularization term $\Omega(g)$ penalizes the complexity of the model:

$$\Omega(g) = \gamma J + \frac{1}{2}\lambda \sum_{j=1}^{J} w_j^2$$

Where:

- J is the number of leaves.
- γ and λ are regularization coefficients.
- w_j are the leaf weights.

Loss Computation: At the mth step, the loss function $L_m(F_m)$ includes the current predictions F_m and the regularization term:

$$L_m(F_m) = \sum_{k=1}^{M} \mathcal{L}(y_k, f_{m-1}(x_k) + F_m(x_k)) + \Omega(F_m) + \text{const}$$

Optimization: The optimization involves computing Taylor expansion upto second order of the loss function:

$$L_m(F_m) \approx \sum_{k=1}^{M} [\mathcal{L}(y_k, g_{m-1}(x_k)) + r_{km} F_m(x_k) + \frac{1}{2} h_{km} F_m^2(x_k)] + \Omega(F_m) + \text{constant}$$

Where:

- r_{km} and h_{km} are the 1^{st} and 2^{nd} derivatives (Hessian) of the cost function w.r.t. the predictions at mth step.

Tree Structure: For regression trees, the prediction function is $G(y) = w_l(y)$, where l denotes the leaf node and w be the weights of leaf.

Leaf Weight Optimization: The optimal weights w_l^* for each leaf are given by:

$$w^*l = -\frac{Rlm}{H_{lm} + \lambda}$$

Where:

- R_{lm} and H_{lm} are the sums of first and second derivatives for data points in leaf l.

Tree Structure: For regression trees, the prediction function is $F(x) = w_q(x)$, where q specifies the leaf node and w are the leaf weights.

Leaf Weight Optimization: The optimal weights w_j^* for each leaf are given by:

$$w^*j = -\frac{Gjm}{H_{jm} + \lambda}$$

Where:

- G_{jm} and H_{jm} are the sums of first and second derivatives for data points in leaf j.

Node Splitting: Nodes are split based on the gain in the objective function. A split is considered if the gain is greater than a threshold γ.

Fast Approximation: A fast approximation is used for evaluating the objective function without sorting features, especially for choosing optimal thresholds for splitting.

Pictorial Representation:

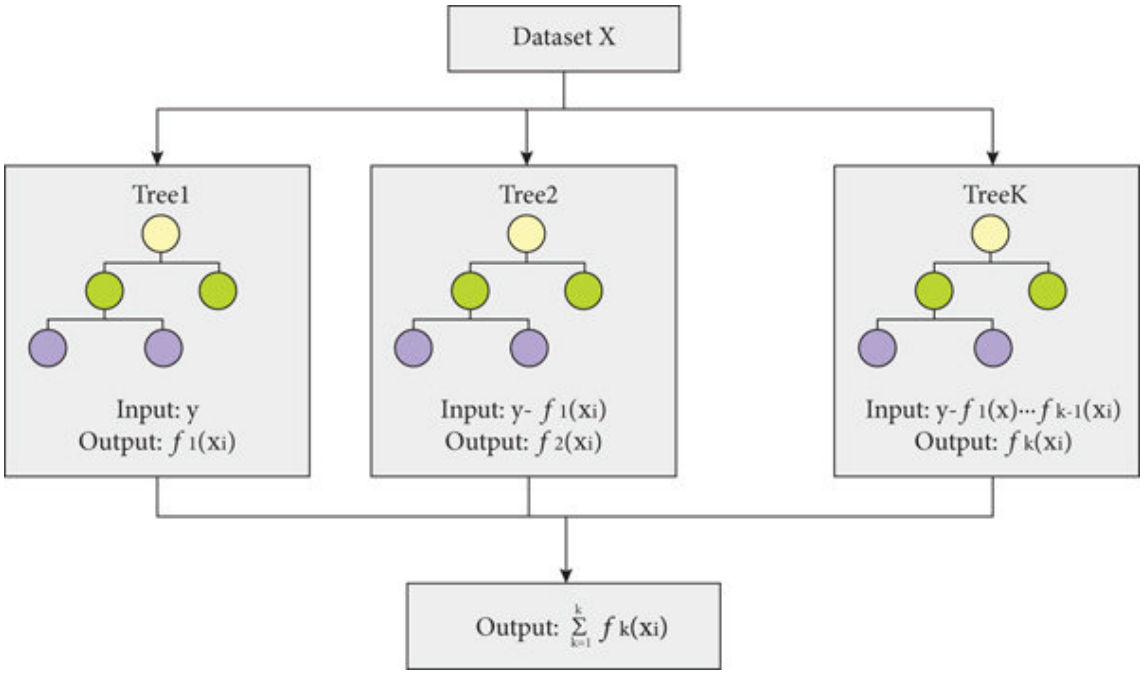

Figure 9 Boosting Algorithm

6.4 Comparison and Interpretation

We have evaluated and compared the performance of above mentioned algorithm by feeding them the data of our interest. The evaluation and comparison is based on two metrics: R-2 score and Root Mean Squared Error. We evaluated these metrics on both training and testing data, by doing this we can easily identify the problem of over-fitting.
In addition to this we analysed accuracy and errors caused by above stated algorithms by just seeing and visualising the figures. Then from these observations we finalised the best algorithm for our project.

Graphs to visualise:

Linear and Polynomial Regression:

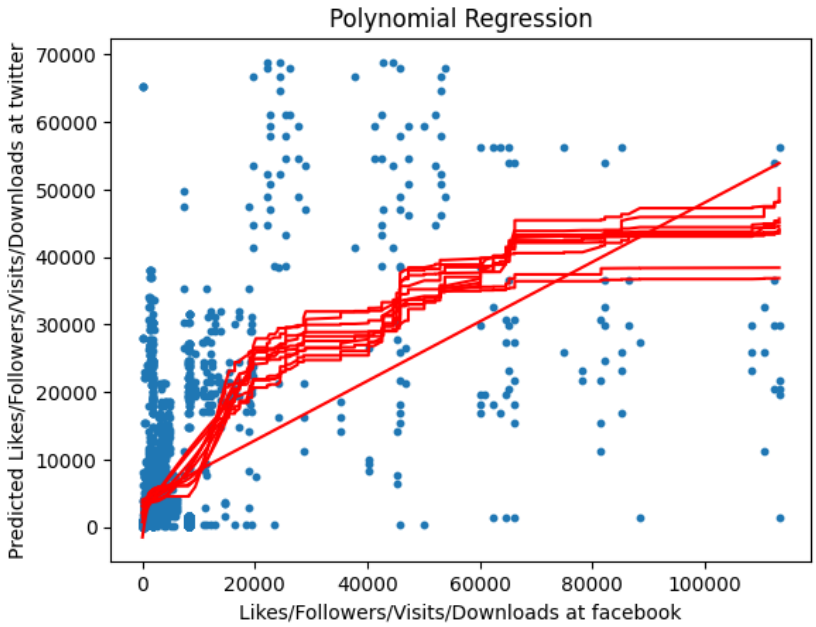

Figure 10 Polynomial Regression

From above figure we can see a straight line, that represents the best fit line for linear regression or we can say that it is polynomial regression with degree 1. Further we see the other curves are the best fitted curves for polynomial regression with degree ranging from 2 to 11. Also, below we have listed the R-Squared accuracy and Root Mean Squared Error.

```
Root mean Squared Error for 1 degree polynominal model : 9364.284480308022
R-squared for 1 degree polynominal model : 0.28254772330050293
Root mean Squared Error for 2 degree polynominal model : 8685.293888367753
R-squared for 2 degree polynominal model : 0.38281855995371017
Root mean Squared Error for 3 degree polynominal model : 8675.02057170554
R-squared for 3 degree polynominal model : 0.38427775075439563
Root mean Squared Error for 4 degree polynominal model : 8496.152655319147
R-squared for 4 degree polynominal model : 0.4094068104481213
Root mean Squared Error for 5 degree polynominal model : 8389.214202092862
R-squared for 5 degree polynominal model : 0.4241804745083818
Root mean Squared Error for 6 degree polynominal model : 8396.57044501124
R-squared for 6 degree polynominal model : 0.4231701950338934
Root mean Squared Error for 7 degree polynominal model : 8364.255835194848
R-squared for 7 degree polynominal model : 0.4276015665488533
Root mean Squared Error for 8 degree polynominal model : 8246.669225884634
R-squared for 8 degree polynominal model : 0.4435822558082604
Root mean Squared Error for 9 degree polynominal model : 8134.642680136141
R-squared for 9 degree polynominal model : 0.4585968446869382
Root mean Squared Error for 10 degree polynominal model : 8048.571870670845
R-squared for 10 degree polynominal model : 0.46999316112770684
Root mean Squared Error for 11 degree polynominal model : 8231.53960566854
R-squared for 11 degree polynominal model : 0.44562202885820834
```

Figure 11 Metrics of Polynomial Regression

From above snippet we can see that, for linear regression we are getting around 28% accuracy and in best case for polynomial regression(10-th degree polynomial) we are getting around 47% accuracy. By seeing these results we can conclude that on any given case it is not a good accuracy.

Decision Tree:

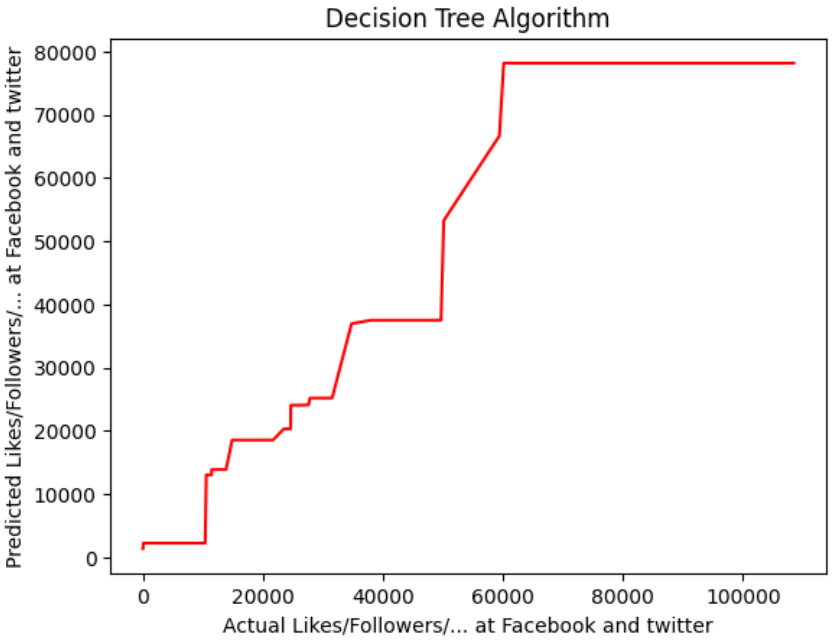

Figure 12 Decision Tree Regression

Above plot is the graph of true values vs predicted value using decision tree regression. The horizontal line at top is due to less number of data in that range. By visualising the graph we can state that the decision tree algorithm giving us good prediction than linear regression or polynomial regression.

```
The root mean squared error for Decision Tree is: 4927.188790331218
The R2 score for Decision Tree is: 0.8089764600439032
```

Figure 13 Metrics of Decision Tree Regression

From the above snippet we can see that it is supporting the claim we just made above. Here what we see is that we are getting around 81% accuracy after using decision tree algorithm, obviously it is giving better accuracy than the above two i.e. linear and polynomial regression. However, on any given day 81% accuracy can be considered as good one.

Random Forest:

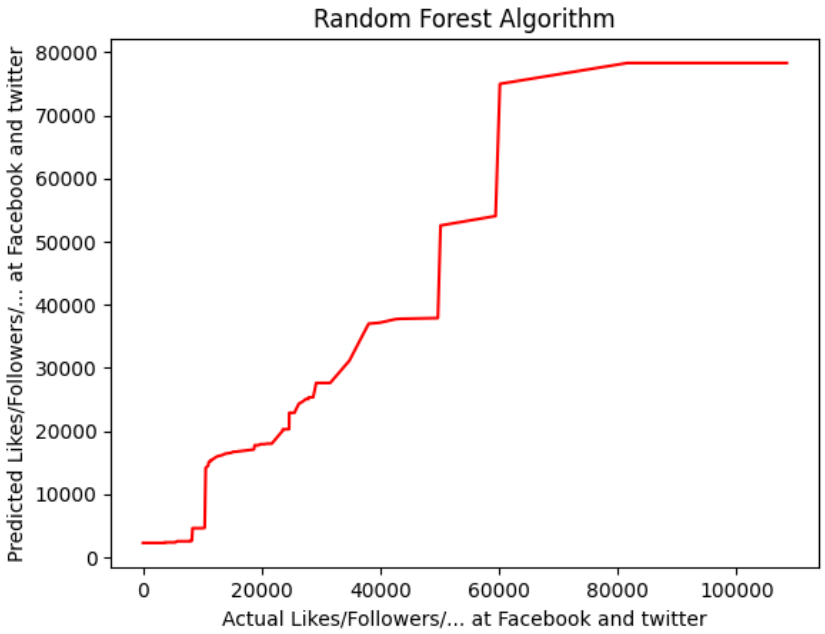

Figure 14 Random Forest Regression

Above we see another figure which is a plot of actual value of likes/followers/downloads vs predicted value of likes/followers/downloads by our model in this case it is random forest. By visualising the graph we can say that it looks similar to graph of decision tree regression algorithm. So, comparing the actual result we can look the metrics of random forest regression which is given below.

```
The root mean squared error for Random Forest is: 4920.012979010444
The R2 score for Random Forest is: 0.8093891906951729
```

Figure 15 Metrics of Random Forest Regression

Here we can see that for random forest regression algorithm also accuracy is around 81% which is almost similar as decision tree regression. By visualising we observed that both graphs are looking similar and we can see from metric that the observation got correct.

XG-Boost:

In the plot of Xg-boost regression, by visualising the graph we can claim that it is almost similar to previous two algorithms i.e. random forest and decision tree. But one thing to notice that, in this graph we are encountering more horizontal line than the previous ones, so there is possibility of getting lower accuracy if we use this algorithm. We must see the metric of this algorithm to reach at any kind of conclusion.

As we claimed that the accuracy may reduce and here we see that the accuracy given by xg-boost algorithm is around 78% which is less than random forest and decision tree.

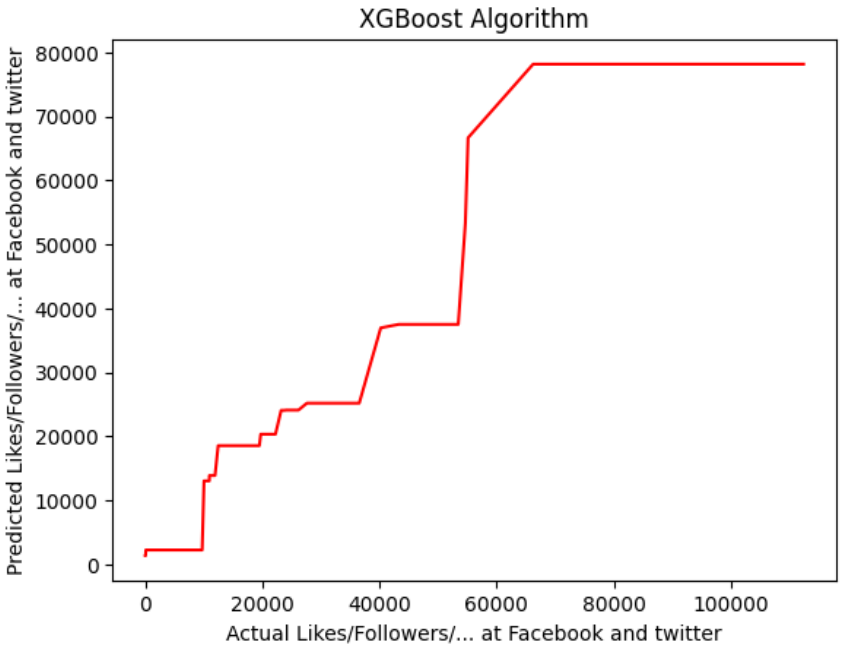

Figure 16 XG-Boost Regression

```
The root mean squared error for XGBoost Algorithm is: 5335.944330454404
R2 Score: 0.7816510041779375
```

Figure 17 XG-Boost metrics

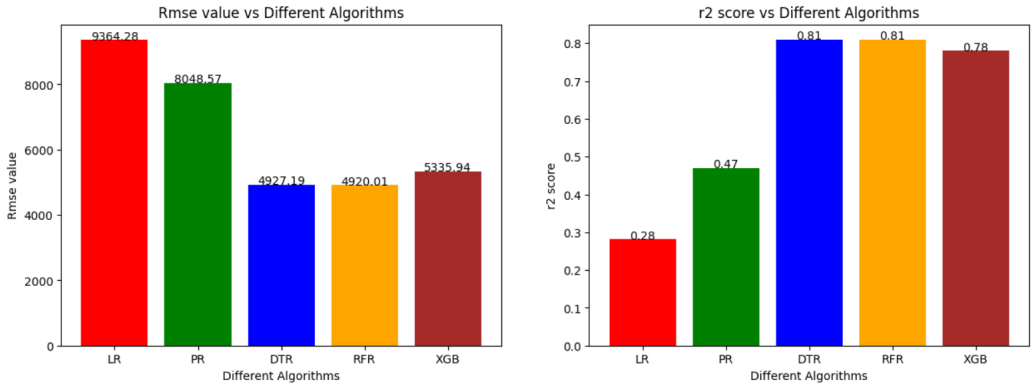

Figure 18 rmse vs. different algorithms & r2 vs. different algorithms

6.4.1 Summary:

Here is the summarized charts for Root mean square error and R squared accuracy for different algorithms. The above bar plot (Figure-18) clearly shows the errors and accuracy for different models.

7 Conclusion

This project provides the detailed analysis of the 5 supervised machine learning models for regression problem. This project also considers the different parameters while predicting the output. As we already seen in result section that for random forest regression and decision tree regression we are getting around 81% accuracy which is highest amongst all the algorithm we are using. But we will consider random forest as best one because it is usually used to overcome with the problem of over fitting in decision tree algorithm. So if we use any other complex dataset then decision tree regression may not perform well as compared tp random forest regression. Also we have seen that linear regression and polynomial regression are performing in horrible way with test dataset, so it is clear sign of over fitting. Other than these two algorithms, all are performing decently well but random forest regression is the best one.

8 Future Works

For the future scope of this project we can expand the dataset by incorporating data from additional social media platforms and diversifying user interactions.This expansion could be complemented by experimenting with a wider range of machine learning and deep learning models such as Neural Networks beyond the ones we've used, such as linear regression, polynomial regression, decision tree regression, random forest regression, and XGBoost regression.

To enhance these models, one may consider delving into more sophisticated feature engineering techniques that could extract richer information from the social media data.
An interesting extension to our project could be the integration of sentiment analysis, which would provide insights into the sentiments behind social media posts, enriching the understanding of consumer behavior.

9 Usefull links

The data set used for this study and all the codes are available here GitHub Repository.

References

1. Kiran Chaudhary, Mansaf Alam , Mabrook S. Al-Rakhami and Abdu Gumaei. Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics.

2. Tufekci Z. Big questions for social media big data: representativeness, validity and other methodological pitfalls.
In: Eighth international AAAI conference on weblogs and social media. 2014.

3. Giglio S, Pantano E, Bilotta E, Melewar TC. Branding luxury hotels: evidence from the analysis of consumers' "big" visual data on TripAdvisor. J Bus Res. 2020;119:495–501.

4. Jung SH, Jeong YJ. Twitter data analytical methodology development for prediction of start-up frms' social media marketing level. Technol Soc. 2020;63:101409.

5. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

6. https://khoros.com/blog/the-2024-social-media-demographics-guide

7. Matz SC, Netzer O. Using big data as a window into consumers' psychology. Curr Opin Behav Sci. 2017;18:7–12

8. XGBoost: A Scalable Tree Boosting System, Tianqi Chen, University of Washington, tqchen@cs.washington.edu

9. Simona Giglioa, Eleonora Pantanob, Eleonora Bilottaa, T.C. Melewar, Branding luxury hotels: Evidence from the analysis of consumers' "big"visual data on TripAdvisor.

10. B. Senthil Arasu, B. Jonath Backia Seelan, N. Thamaraiselvan, A machine learning-based approach to enhancing social media marketing.

11. https://www.geeksforgeeks.org/what-is-machine-learning/

12. Tayebi S, Manesh S, Khalili M, Sadi-Nezhad S. The role of information systems in communication through social media. Int J Data Netw Sci. 2019;3(3):245–68.

13. Stieglitz S, Meske C, Ross B, Mirbabaie M. Going back in time to predict the future-the complex role of the data collection period in social media analytics. Inf Syst Front. 2018;1–15.

14. Jansen BJ, Zhang M, Sobel K, Chowdury A. Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci Technol. 2009;60(11):2169–88.

15. Saif H, He Y, Alani H. Semantic sentiment analysis of twitter. In: International semantic web conference. Springer, Berlin, Heidelberg; 2012. pp. 508–524.

16. Jussila J, Vuori V, Okkonen J, Helander N. Reliability and perceived value of sentiment analysis for Twitter data. In: Strategic innovative marketing. Springer, Cham; 2017. pp. 43–48.

17. Radi SA, Shokouhyar S. Toward consumer perception of cellphones sustainability: a social media analytics. Sustain Prod Consum. 2021;25:217–33.