Analysis of Social Media Data to Find Involvement of Users: A Machine Learning Approach

Submitted by

Sumaiya Islam Mim

ID: 181-35-2322

Department of Software Engineering

Daffodil International University

Supervised by

Ms. Nusrat Jahan

Assistant Professor

Department of Software Engineering

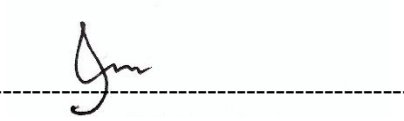
Daffodil International University

The Thesis report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APPROVAL

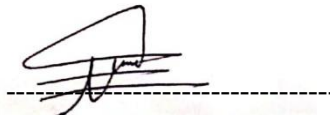
This thesis is being titled as " Analysis of Social Media Data to Find Involvement of Users: A Machine Learning Approach " submitted by Sumaiya Islam Mim, 181-35-2322 to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and for approval as per it's inner styles and contents inside.

BOARD OF EXAMINERS



---------- Chairman

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

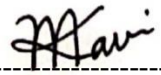


---------- Internal Examiner 1

Nusrat Jahan
Assistant Professor
Department of Software Engineering
Daffodil International University

---------- Internal Examiner 2

Khalid Been Badruzzaman Biplob
Senior Lecturer
Department of Software Engineering
Daffodil International University

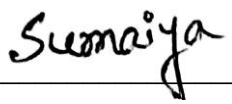


---------- External Examiner

Professor Dr M Shamim Kaiser,
Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

It's been declared that this thesis including all the research-based experimental works has been completed by me under the supervision of Ms. Nusrat Jahan (Assistant Professor), Department of Software Engineering of Daffodil International University. I also declare that neither this thesis nor any part of this whole research-based experiment has been submitted elsewhere for award of any degree.



Sumaiya Islam Mim
181-35-2322
Department Of Software Engineering
Daffodil International University

Nusrat Jahan
Assistant Professor
Department of Software Engineering
Daffodil International University

ACKNOWLEDGMENT

I am grateful to the Almighty Allah (SWT) for allowing me to complete this Bachelor of Science study with this research. I am thankful to my Parents who have always been a support for me throughout my life. They always let me have faith and belief that I can achieve something and have been supporting me. I want to thank my well-wisher who always inspires me for doing something creative.

I would also love to show my respect and gratitude to my Supervisor, **Ms. Nusrat Jahan**, supporting me in completing this research and encouraging me in research and implementation of this work.

Besides, I would like to thank **Dr. Imran Mahmud**, Associate Professor & Head InCharge of the Department of Software Engineering, Daffodil International University for motivating us for quality research. I also want to thank my teacher who has been continuously supporting us throughout this undergrad. Last but not the least, I want to thank all the staff of the Department of Software Engineering and Daffodil International University for continuously working hard to provide us with the best facilities in classrooms and laboratories.

Table of Contents

APPROVAL ..... **i**
DECLARATION ..... **ii**
ACKNOWLEDGMENT ..... **iii**
Table of Contents ..... **iv**
LIST OF FIGURES ..... **vi**
LIST OF TABLES ..... **viii**
ABSTRACT ..... **ix**
CHAPTER ONE **1**
INTRODUCTION **1**
1.1 Background 1
1.2 Motivation of the Research 1
1.6 Research Scope 3
1.7 Thesis Organization 4
CHAPTER TWO **5**
LITERATURE REVIEW **5**
2.1 Studies on Social Media User 5
2.2 Studies on Social Media Impacts 6
2.3 Studies on Other Research 7
CHAPTER THREE **9**
RESEARCH METHODOLOGY **9**
3.1 Data Requirement 10
3.2 Data Set Information 11
3.3 Data Collection 11
3.4 Data Preparation 12
3.5 Exploratory Data Analysis 13
3.6 Data Visualization 14
3.7 Algorithms Used 20
3.7.1 Linear Regression 20

3.7.2 Simple Linear Regression 21
3.7.3 Time Series Forecasting Model 23
3.8 Re-Sampling.....24
3.8.1 Re-Sampling by Train-Test-Split.....25
3.9 Evaluation.....25
3.9.1 Mean Absolute Error.....25
3.9.2 Root Mean Squared Error 26
CHAPTER FOUR.....27
RESULTS AND DISCUSSION.....27
4.1 Simple Linear Regression 27
4.2 Time series Algorithm Using Arima Model 28
4.3 Decision Making 37
4.4 Social Media Impacts.....38
4.4.1. Use of social media for academic and nonacademic purposes 38
4.4.2 Social Media Effects 39
4.4.3 Effects of social media during (COVID-19) pandemic 40
CONCLUSION AND RECOMMENDATION.....42
5.1 Findings.....42
5.2 Contributions.....42
5.3 Recommendation on future works 43
REFERENCES.....44
APPENDIX – A.....47
List of Abbreviation.....47

LIST OF FIGURES

Figure 3.1: Proposed Research and Methodology 9

Figure 3.2: Visualization of Monthly active users and Platform 14

Figure 3.3: Visualization of Platform based on Year 15

Figure 3.4: Visualization of Facebook Monthly active users based on Year 16

Figure 3.5: Visualization of Instagram Monthly active users based on Year 17

Figure 3.6: Visualization of YouTube Monthly active users based on Year 18

Figure 3.6: Visualization of Netflix Monthly active users based on Year 19

Figure 3.7: Visualization of Youth Users based on Entity 20

Figure 3.8: Train Test Split Re-sampling 25

Figure 4.1: For Facebook monthly active user and youth user 28

Figure 4.2: For YouTube monthly active user and youth user 28

Figure 4.3: For Netflix monthly active user and youth user 29

Figure 4.4: For Instagram monthly active user and youth user 29

Figure 4.5: For Twitter monthly active user and youth user 30

Figure 4.4: For LinkedIn monthly active user and youth user 30

Figure 4.6: For WhatsApp monthly active user and youth user 31

Figure 4.7: For TikTok monthly active user and youth user 31

Figure 4.8: For Snapchat monthly active user and youth user 32

Figure 4.9: For Facebook Youth User Forecast 32

Figure 4.10: For YouTube Youth User Forecast 33

Figure 4.11: For Netflix Youth User Forecast 33

Figure 4.12: For Instagram Youth User Forecast 34

Figure 4.13: For Twitter Youth User Forecast.....34

Figure 4.14: For LinkedIn Youth User Forecast.....35

Figure 4.15: For WhatsApp Youth User Forecast 35

Figure 4.16: For TikTok Youth User Forecast 36

Figure 4.17: For Snapchat Youth User Forecast.....36

LIST OF TABLES

Table 3.1: Data Attributes and Types 12

Table 3.2: Mean of features 13

Table 4.1: Social Media Users Prediction With Regression.....27

Table 4.2: Social Media Users Prediction With Time Series 37

ABSTRACT

Analysis of Social Media Platforms provides helpful information for users on social media. Recent papers about user interaction on social media explore methods for predicting user interaction. These analyses of Social Media Platforms have included Active Users and Youth Users analysis based on year. Yet, the studies have not incorporated text data. This research explores the usefulness of incorporating text data to predict user interaction. The study incorporates two types of machine learning models: Linear Regression, Time Series Model(Arima). The models are unique in their use of the data. The research collects 208 Users based on social media platforms such as Facebook, Instagram, YouTube, LinkedIn, Twitter, Netflix, Whatsapp, Snapchat, Google+, TikTok, Pinterest etc per year. The models learn and test on Youth users in order to predict user interaction. The study found that text data produced the best models. The research further demonstrates that Time series models perform best.

Keyword: Social Media, Machine learning, Time series, The use of social media analysis, Social Platforms, Youth Users

CHAPTER ONE

INTRODUCTION

1.1 Background

We have entered a digital age and every day it's all about technology. Social media has exploded as a type of online discourse in which people create, share, bookmark, and network at an alarming rate. Since the pandemic started people are getting more involved with social media. In this time Social media plays an important role in human communication and enhances business applications. Introducing popular social media apps that have received more attention from users in the past few decades. Our younger generation is getting more robotic every day. Users have no choice but to rely on social networks or the computer or virtual device to participate in any type of entertainment. They are dependent on social media for communication. But that has a huge negative impact on our youth.

1.2 Motivation of the Research

There are numerous algorithms that interact with social media. But there is little work about social media users, but couldn’t find out the amount of users based on year. These studies oversimplify the data. The vast majority of users on social media consist of platforms, year, country, user’s of youth data. It’s not a surprise that the users are also young. In recent years, more users have been created on social media platforms. The increase in machine learning libraries includes implementations for prediction based algorithms. Social media provides an optimal environment for such research. A great deal of its data is publically available. This data can be downloaded, transformed, and organized for large-scale machine learning analysis. These new tools, techniques, and data availability offer a fantastic opportunity to improve on current social media analysis.

These also allow for the development of new techniques and the discovery of important features for social media analysis. This paper aims to contribute to a largely unexplored field within academia.

1.3 Problem Statement

In this dynamic world, Social media is the biggest way of communication and there are many things to do. There are Billions of users on social media. So it is very difficult to find out the amount of users. Why people are more addicted to social media than other things. Some people are active on social media for work, daily news. Many people used social media only for entertainment. This thesis aims to explore to find out average monthly users by predicting social media active users and youth users based on platform and year. This helps companies better understand how many users will perform on social media. It also gives social media impact feedback. The research performs computations on the data and predicts users. The aim of the thesis is to predict users. The thesis will explore this prediction with different machine learning models.The research builds many types of machine learning models. The research will detect which can best predict users. The study will measure how well models can predict users based on year.

1.4 Research Question

1. Can Machine Learning algorithms help to predict user interaction in social media?

1.5 Research Objectives

1. We have continuous data so we need to apply a regression algorithm. Also we have data that is time basis which is year. So we also apply a time series algorithm for better results.
2. Want to determine the factors behind the users and social media usage impacts.

Provide some favorable suggestions regarding this issue. Hopefully, this research will be helpful.

1.6 Research Scope

This document provides a comprehensive overview of various applications that use social media analytics through machine learning approaches. We start with the algorithms and machine learning approaches used for social media analysis. We discuss the consequences of social media analysis, which uses machine learning in addition to statistical analysis. The study involves understanding user interaction on social media. The study also underlines which models best predict user interaction. The research will expose how well models can measure user interaction. The paper will also talk about how models can predict with data. On a practical point, the models could be implemented for platforms. This study predicts user interaction with social media. The users include all social media users from child to old. This research mainly focused on our young generation This study explores predicting user interaction with different types of machine learning models. The models are unique in their incorporation of data. The goal is to understand which model-types best predict users. The study also distributes as an overview for approaching user interaction prediction. People's addiction to social networks in the world is growing, which

promotes new fields of research. As a consequence, we end up with aggressive future directions for researchers as open topics in this paper.

1.7 Thesis Organization

This thesis consists of five sections. The first section is just the introduction part. In this section a short introduction about the topic of the research, it's social media user interaction viewpoints and overall summary is shown. Next sections include Literature Review in Chapter two based on different paper studies, Research Methodology including feature selection, exploratory data analysis, model development and visualization in Chapter three. After that we discussed the Result in Chapter four. And we finally conclude that future works can be done from here in Chapter five.

CHAPTER TWO

LITERATURE REVIEW

Throughout this research we have performed some of the condemnatory, constructive and comparative reviews on some of the applicable studies. In the next few sections we will connect some of our reviews over studies being performed on studies. Many works were undertaken to cover the problems and to create a research report. This article focuses on the year based Users dependent on datasets.

2.1 Studies on Social Media User

The penetration of social media into the lives of internet users is increasing. According to the most recent data, there will be 3.78 billion social media users worldwide in 2021, a 5% increase from the previous year. It is also 920 million more than the number of social media users in 2017, representing a 32.2% increase in just five years. The average annual growth rate during this time period has been 7.2%. While the number of social media users is expected to increase in the future, growth is expected to slow. The average annual growth rate from 2022 to 2025 is expected to be 3.9%. [Statista 2021]

In this fast-paced world, text is the most popular mode of communication, with over 18.2 million text messages sent every minute. Among the jpg, png, tiff, and videos, the posting of gif images has increased by 22% year on year. Users are being forced to use high-speed networks by YouTube, Amazon, Netflix, and Facebook. Netflix users worldwide watch 69,444 hours of video per minute.[1]

Almost 70% of adults use Facebook (Gramlich 2018). Running such a large site for so many users is costly. Users, on the other hand, do not pay a fee to use Facebook. Google, like Facebook, provides free services to its users. Because platforms generate revenue from advertising, the services are provided for free.[2]

2.2 Studies on Social Media Impacts

Find out the association of positive and negative experiences using social media in a national survey of young adults. With the increased use of social media, more research is needed to determine the extent to which users have negative and positive experiences on social media, as well as whether these experiences are associated with sleep disturbance over time, and whether this is mediated by rumination about the negative events. Furthermore, 40% of adults report various types of negative online experiences, such as name calling, purposeful embarrassment, and exposure to unwanted graphic or sexual images. However, these experiences are related to sleep disruption.[3]

Social media has more negative than positive effects (Woods and Scott, 2016). Because students spend more time on social media for reasons other than education, this causes distraction from the learning environment, affecting their academic progress (Bekalu et al., 2019; Hettiarachchi, 2014). Furthermore, spending a lot of time on social networking sites can lead to a sedentary lifestyle and a decrease in daily physical activity levels, making them susceptible to noncommunicable diseases like obesity, diabetes, and hypertension (Melkevik et al., 2015; Zou et al., 2019; Hu et al., 2001). Furthermore, social media use has a negative impact on mental health and can lead to depression and anxiety.[4]

2.3 Studies on Other Research

In[1] the author wanted to apply Ann, Naive Bayes, K-means, SVM, Apriori algorithm, Linear regression, Logistic regression, Decision tree, Random forest, KNN but the only collected dataset and no algorithm applied. Later, they discuss the consequences of social media analysis using machine learning along with the statistical analysis.

In [2] the author evaluated the text-based NN and image-based CNN. These predictions were compared with the actual user engagement metrics. The random model was correct 50% of the time. The combined was correct 57%, 55%, and 53% of the time for comment sentiment, comment count, and share count. The model performance serves as a benchmark for future research.

In [4] Authors says, A questionnaire was used to collect data. Chi-squared (Fisher's exact test) test was used to analyze the data. The results showed that 97% of the students used social media applications. Only 1% of them used social media for academic purposes. Whereas 35% of them used these platforms to chat with others, 43% of them browsed these sites to pass time. Moreover, 57% of them were addicted to social media. Additionally, 52% of them reported that social media use had affected their learning activities, 66% of them felt more drawn toward social media than toward academic activities, and 74% of them spent their free time on social media platforms. The most popular applications (i.e., based on usage) were Snapchat(45%), Instagram (22%), Twitter (18%), and WhatsApp (7%). Further, 46% and 39% of them reported going to bed between 11 pm and 12 am and between 1 am and 2 am, respectively. Finally, 68% of them attributed their delayed bedtime to social media use, and 59% of them reported that social media had affected their social interactions.

In[3] the author applied logistic regression, [5] Ann, Naive bayes, K-means, SVM, generalized Linear model, First large margin, Decision tree, Random forest applied and here linear regression

achieved best performance. [6] applied Social network analysis (SNA), Latent Dirichlet Allocation (LDA) and Results showed that political, economic and legal posts had dense clusters around the technology policy of EV, the institutional discourse of electrify-cation of the federal vehicle fleet, and tax and credit framework politics. The environmental and social di-mansions had a higher discourse for social justice, clean air, and better health and well-being.

In[7] the author says, A computational social science methodology was adopted using a mixed-method application of social network analysis and machine learning-based topic modelling through Latent Dirichlet Allocation algorithm on a 600,000-text corpus extracted from the Facebook posts.

In[10] The author uses a simple taxonomy, this paper provides a review of leading software tools and how to use them to scrape, cleanse and analyze the spectrum of social media.

In[11] the author applied a linear regression model for predicting box-office revenues of movies in advance of their release. In this paper, they demonstrate how social media content can be used to predict real-world outcomes and how sentiments extracted from Twitter can be utilized to improve the forecasting power of social media.

In[13] the author worked on a cross-sectional survey of parents of children 0 to 18 years seen in clinics and an inpatient medical unit. A total of 258 parents completed the survey. The mean age was 39.8 years, 83% were female, 59% were white. The most common topics parents read about online were: sleep, mental health, and car safety. Nearly all parents (96%) used social media, with 68% using social media for health information.

CHAPTER THREE

RESEARCH METHODOLOGY

Our proposed research methodology starts with the social media active users per year by analyzing the social media applications such as: Facebook, Instagram, Twitter, LinkedIn, YouTube, Snapchat, WhatsApp, TikTok. Following the analysis, we determined the necessary data for our research work to be carried out. Afterwards, we have collected our required data by Statista and Napoleon cat.

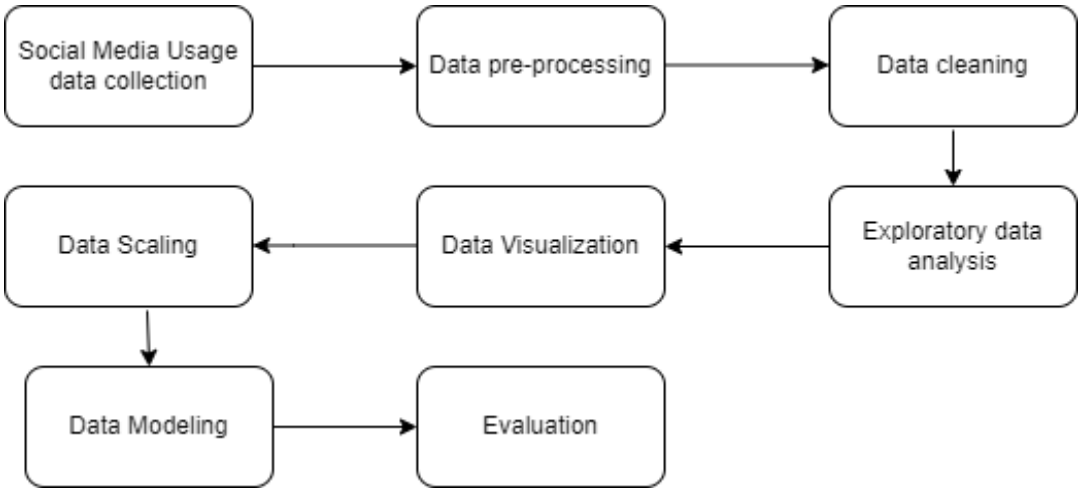


Figure 3.1: Proposed Research and Methodology

Then we have driven some Exploratory Data Analysis (EDA) on our collected social media usage data to understand what we can do further with our experiments. To avoid abundance we have driven our Min Max Scaler/Standard Scaler. This step has been the most critical process throughout our research work. Afterwards, we have evaluated the training results from our feature combinations and elected the best model for our further works. In this term, our research work is ready to be started for implementation.

3.1 Data Requirement

Our research work starts with making decisions on what data we need to drive further research tasks. In this step, we have to use the insights found from our related work reviews. In our study on social media usage, we have got to acknowledge the using issue and know different points about social media. So based on our previous studies we are deciding to collect these information:

1. For proper analysis, firstly, we need to make a list of websites that people visit most based on trending topics.
2. We need Active user's features because we want to make predictions based on user performance.
3. Based on our analysis, then we will collect the Active Users and Youth Users of our listed sites based on Year and Users are going to be our target variable. Also we have collected some Countries and countries Code for our research purpose. So, we will try to find the best suited model that fits best for predicting the user's.
4. However, our primary goal is to find the best model for our features.

3.2 Data Set Information

The data context consists of some Platforms and their monthly active users and Youth users. Also there is data which shows some countries and their youth users based on those countries. So, we have Datasets with 207 rows and 6 columns.

In this dataset, we have six Columns. The features are: Platform(An Online Platform like Facebook, Instagram, WhatsApp, YouTube, LinkedIn, Twitter, Snapchat, Google+, WeChat etc), Year(2008-2021), Monthly active users(Number), Entity(some countries like Bangladesh, India, Malaysia, Japan), Code(Country code like BDG, IND, JPN), Year(2008-2021), Youth User[Users of young generation(age:18-35)].

Source:

1. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
2. https://napoleoncat.com/stats/

We collected data from these two sites.

3.3 Data Collection

From our data requirement understanding step we got to know about the data we require to collect. So, first we used Google's Search Engine to find the sites .Then we collected data from a website(Napoleoncat.com). Also we have collected some data from a website(Statista). We have gathered each collection in separate Comma Separated Value (CSV) files as it is a well accepted format for Data Science experiments.

After the data's attributes have been visualized in a development platform (in this case, a Jupyter notebook), the data's attributes are shown in a table format below:

Table 3.1: Data Attributes and Types

Attribute	Types
Entity	Categorical
Year	Numerical
Monthly Active Users	Numerical
Country	Categorical
Code	Categorical
Youth User(%)	Numerical

3.4 Data Preparation

Before starting our Data Analysis, we prepared our data to prevent noise while exploring social media users' experience data. It is not a new procedure, but it is crucial prior to doing EDA. Our data preparation tasks are as follows:

1. At first, in our data preparation step we collected data from two websites that are Napoleoncat.com and Statista. And kept data in different CSV files.
2. Then we merged CSV files.
3. Sort data.
4. Category transformation
5. Type transformation.

3.5 Exploratory Data Analysis

Before our scaling and modeling, we target learning from data. To achieve this goal, we first conducted Qualitative Exploratory Data Analysis (EDA) on our performance report. At first, we tried to survey our issue data distribution by using the Mean formula. Mean of each web performance issue is calculated as follows where n is the total number of entries we have in our dataset and i is addressed for iteration:

$$\textbf{MEAN (Peformance Issue)= 1 n}\sum \textbf{ i=1 n issue} \tag{3.1}$$

By doing summation of each performance issue entries addressed as i-th iteration in the equation 3.1 and then dividing by total occurrence we got mean for each of our features as follows:

Table 3.2: Mean of features

Attribute	Mean
Monthly Active Users	6379927e+08
Youth User(%)	88.065575

Our observation from mean performance issues indicates there are possibilities of existence of outliers. We check Null values. There are one null values. Then we have categorical data so we encoded them. Then our next iterative process of numerical EDA is Scaling. Here we used MinMaxScaler. This is important for performance issues.

3.6 Data Visualization

To observe more details in between relations between features and target, we have prepared some statistical plots demonstrated below:

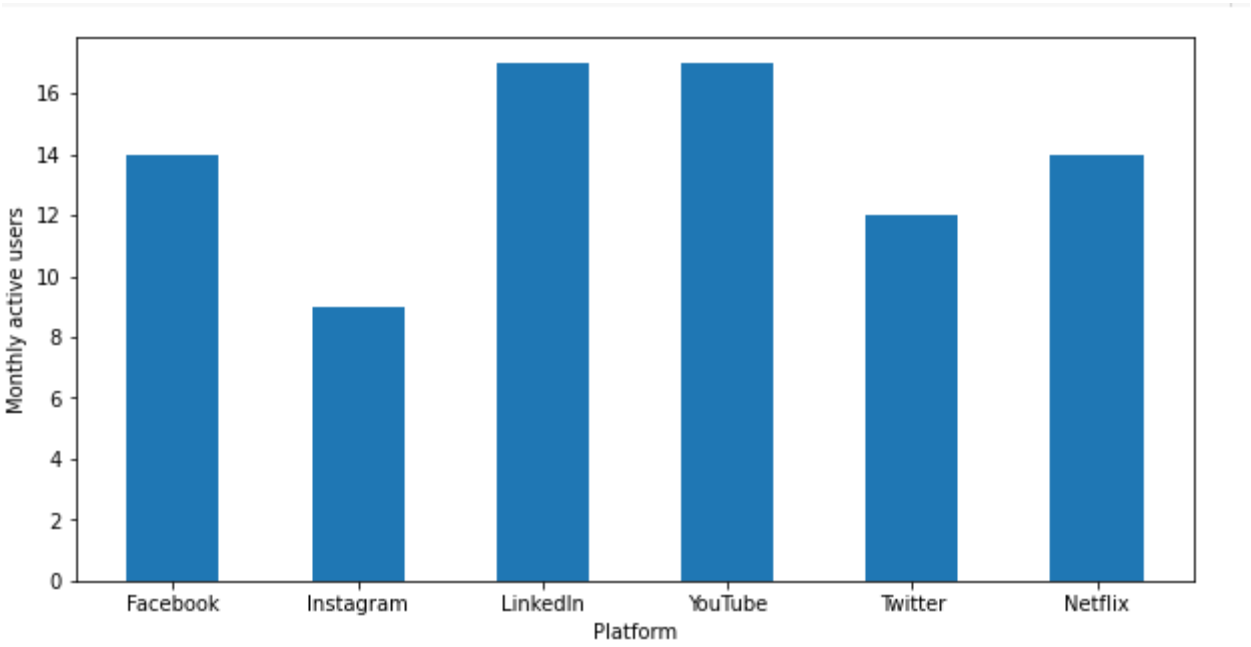


Figure 3.2: Visualization of Monthly active users and Platform

In the plot above, we observed the distribution of relation between Platform and Monthly active users. Here, we see that monthly active users depend on the platform. Here we see people are more active on LinkedIn and YouTube. They are also active on Facebook , Instagram, Netflix and Twitter.

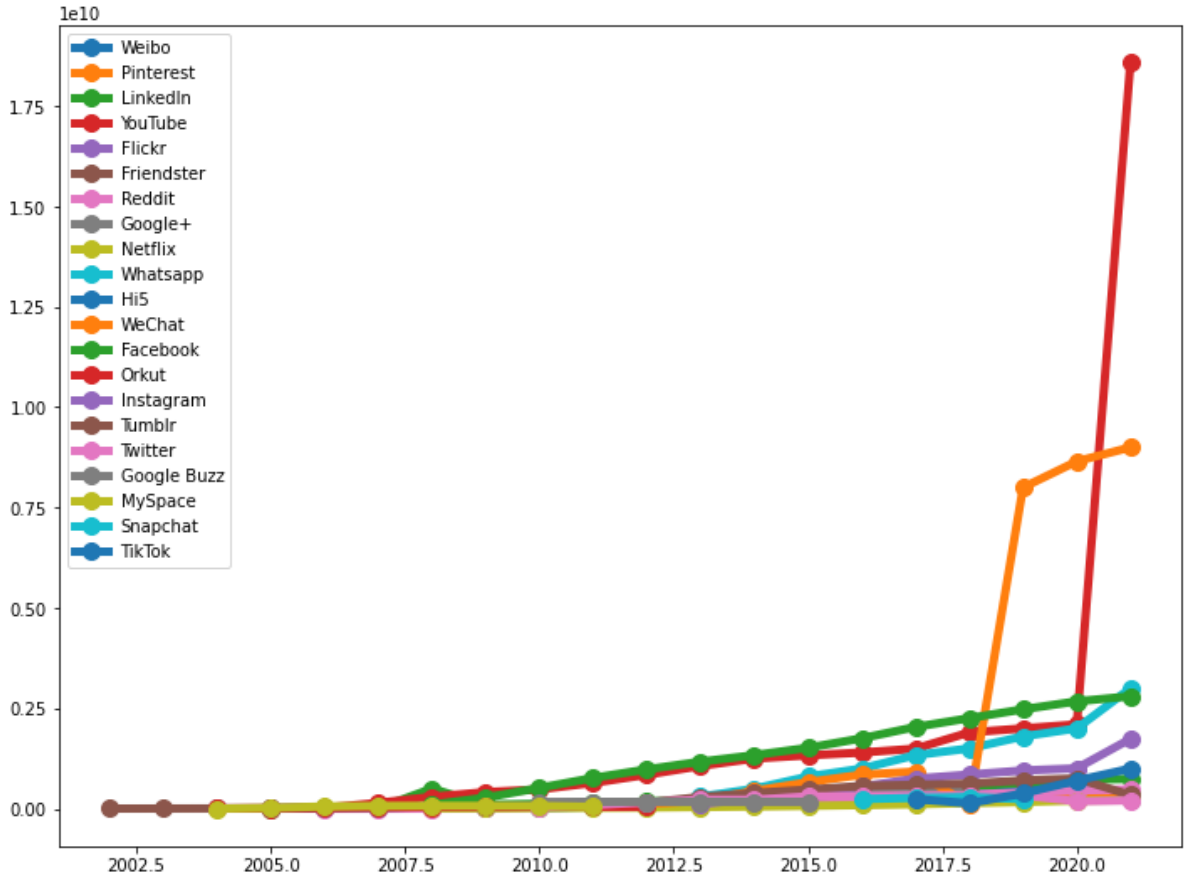


Figure 3.3: Visualization of Platform based on Year

In the plot above, we observed the distribution of relation between Year and Platform based on Users. Here, we see that Platform depending on Year and we see the most active platform is Youtube.

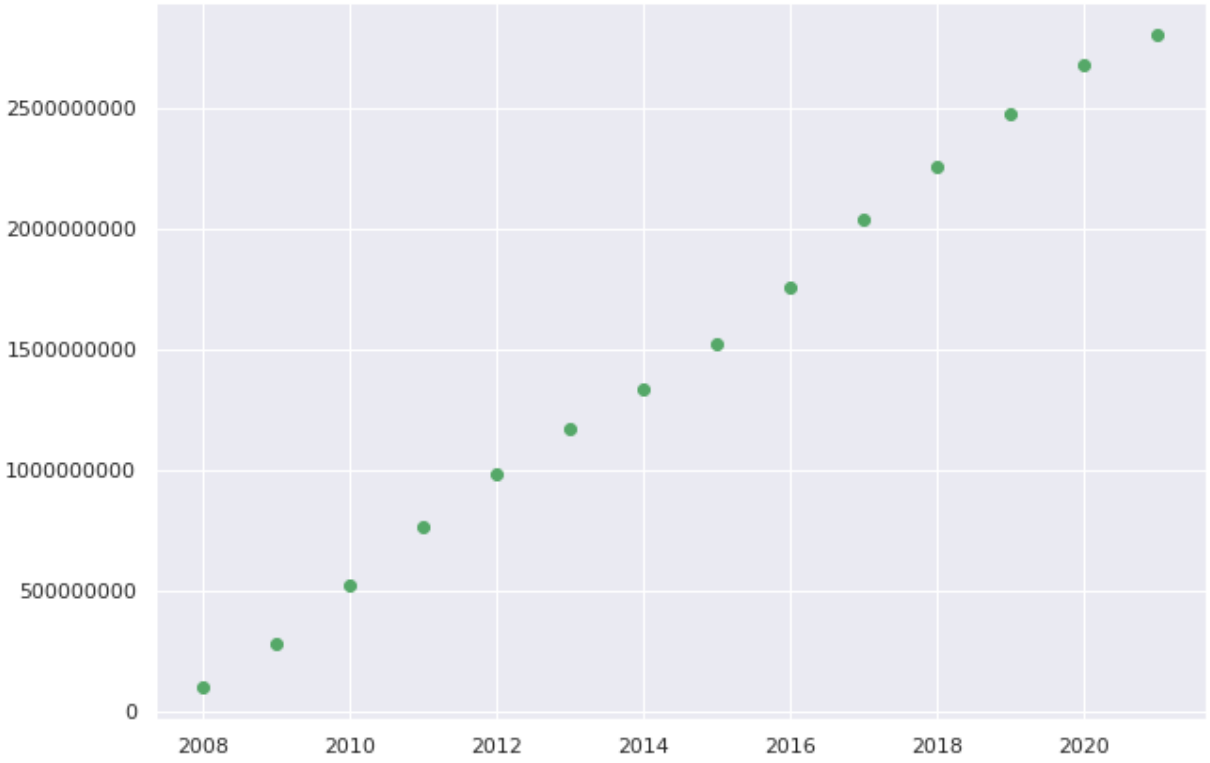


Figure 3.4: Visualization of Facebook Monthly active users based on Year

In the plot above, we observed the Facebook active users per year.

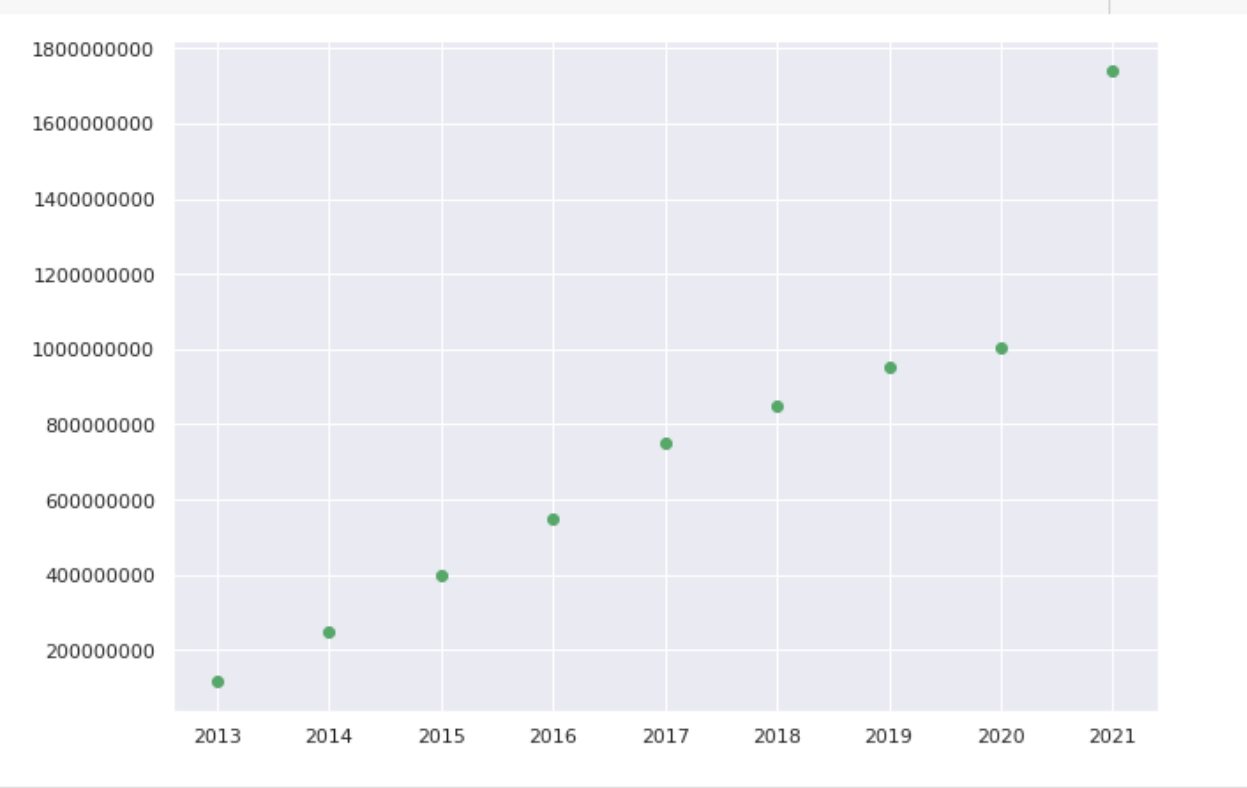


Figure 3.5: Visualization of Instagram Monthly active users based on Year

In the plot above, we observed the Instagram active users per year.

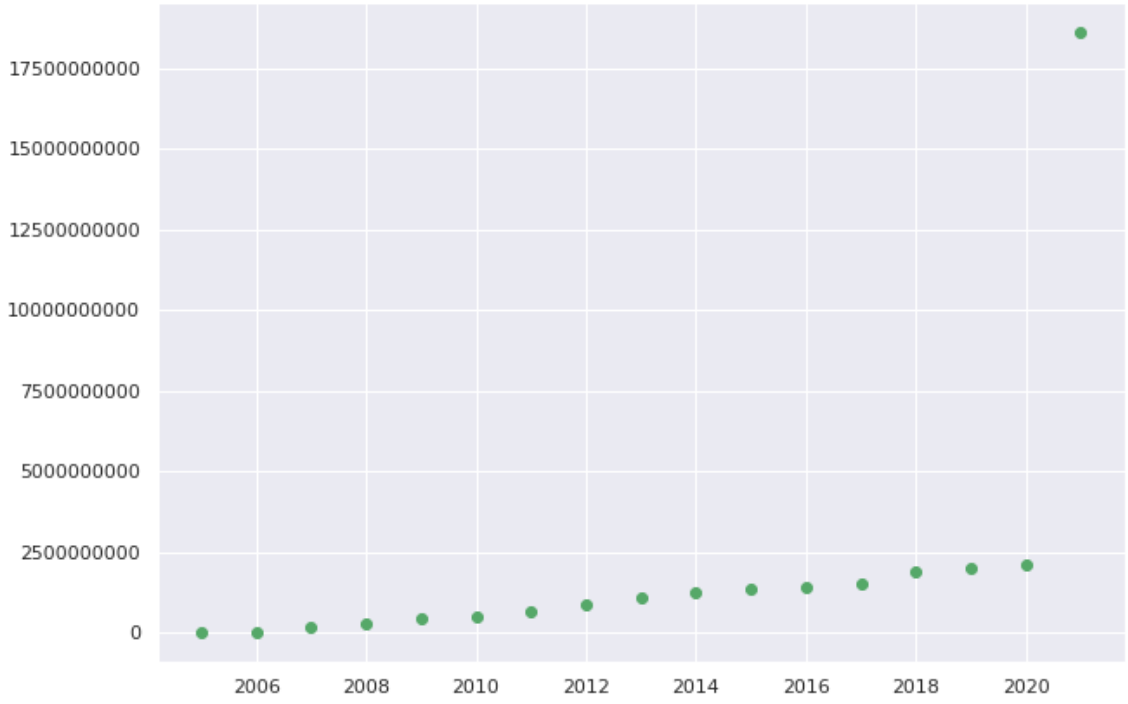


Figure 3.6: Visualization of YouTube Monthly active users based on Year

In the plot above, we observed the YouTube active users per year.

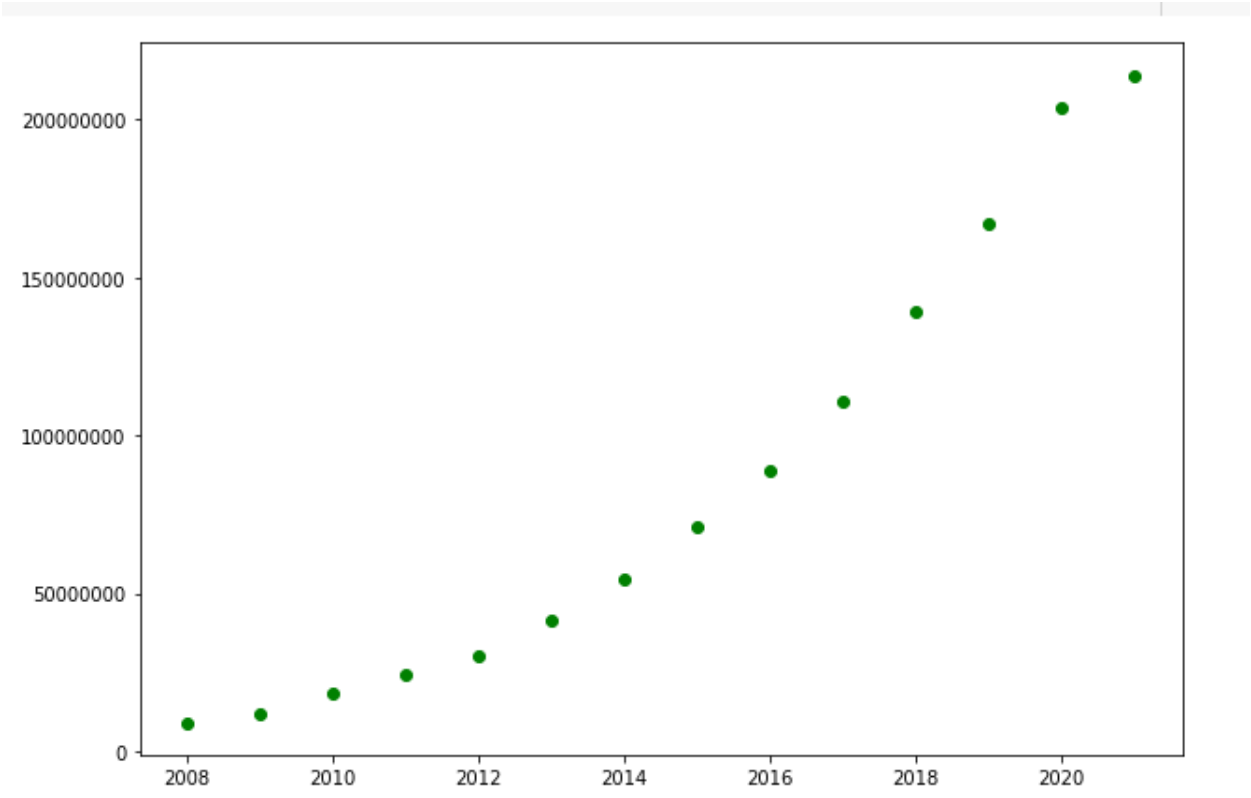


Figure 3.6: Visualization of Netflix Monthly active users based on Year

In the plot above, we observed the Netflix active users per year. So, In these three visualizations we see that there is the connection between Platform, Year and Monthly active users.

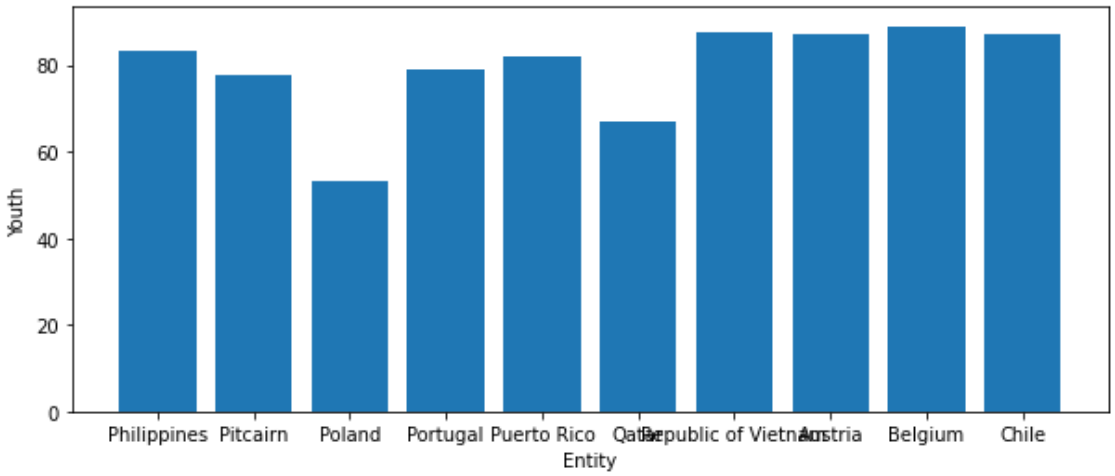


Figure 3.7: Visualization of Youth Users based on Entity

In the plot above, we observed the distribution of relation between Entity and Youth Users. Here, we see that Youth Users depend on the Entity.

3.7 Algorithms Used

In this section we will discuss each of the algorithms we have used throughout our model training. We will have some short overview on how performance issues and experience fits in these algorithms to solve regression problems. And After Data visualization we see that this dataset is based on year. Our target value depended on Year. So we applied a time series algorithm.

3.7.1 Linear Regression

The linear regression (LiR) technique demonstrates how a change in one characteristic affects the other, and it particularly defines how these two qualities are understood. It explains how the dependent attribute and the independent attribute communicate. The dependent property determines why a prediction is made, whereas the independent attribute provides the prediction meaning. For predictive analysis, linear regression is commonly utilized.Linear regression is good

when the relationship between a response variable and covariances are identified to be linear. This method is beneficial because it shifts the focus from statistical modeling to pre-processing and data analysis. Linear regression allows us to learn to experiment with data without worrying about the model's detailed features. In conclusion, it is appropriate for the data analysis learning process. However, because it generalizes real-world problems, it will not be recommended for a variety of practical applications.

3.7.2 Simple Linear Regression

Regression models discuss the relationship between features by fitting a line to the observed data. Linear regression models conduct a straight line, while logistic and nonlinear regression models conduct a curved line. Regression allows us to list how a dependent variable_changes as the independent variable(s) change.

Simple linear regression is conducted to list the relationship between two quantitative variables.You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (Example: The relationship between Year and Youth users).
2. The value of the dependent variable at a certain value of the independent variable (e.g. the amount of Youth users based on Year).

Simple linear regression formula:

$$Y = B0 = B1X + e$$

If We have more than one independent variable, use multiple linear regression instead.

Multiple linear regression is conducted to list the relationship between two or more independent variables and one dependent variable. We can use multiple linear regression when we want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable (Example: how Platform and Year added Monthly active users).
2. The value of the dependent variable at a certain value of the independent variables (Example: the expected Monthly active users depend on Year and Platform).

In this research paper, we do simple linear regression. Our target variable is Monthly active users. Our dependent variable is Year. We try to find how well we can predict Monthly active users with that performance issue. For all performance data points we tried to find the mean of all those residual errors which shows how poorly the performance issue fits with the whole Users. In simple linear regression we consider all our performance issues which are being selected as the best feature set from Feature Selection to fit enough to predict Monthly active users. Our goal of doing Linear Regression here is to identify the strength of the impact of performance issues on the user's visual progression. For Monthly active users, we can summarize the whole LR equation. Here we have expressed the visual progression based user experience, respectively representing performance issues. In linear regression we consider all our performance issues which are being selected as the best feature set from Feature Selection to fit enough to predict Monthly active users. Our goal of doing Linear Regression here is to identify the strength of the impact of performance issues on the user's visual progression. Here we have expressed the visual progression based user experience.

3.7.3 Time Series Forecasting Model

Time series is a sequence of observations noted at regular time intervals. Depending on the frequency of observations, a time series may generally be hourly, daily, weekly, monthly, quarterly and annual. Forecasting is the next step where you want to predict the future values the series is going to take. Here, we use Univariate Time Series Forecasting. We use the Arima Model for forecasting. Using the ARIMA model, we can forecast a time series using the series past values.In this research, we build an optimal ARIMA model to eliminate and extend it to Seasonal ARIMA (SARIMA) and SARIMAX models. We will also build auto arima models in python. Here we use Univariate Time Series Forecasting.

Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q

Where,

p is the order of the AR term,

q is the order of the MA term,

d is the number of differencing required to make the time series stationary.

If a time series has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'. More on that once we finish ARIMA.

A pure **Auto Regressive (AR only) model** is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt'.

Similarly a pure **Moving Average (MA only) model** is one where Yt depends only on the lagged forecast errors.

Arima model In words,

Predicted Yt = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags)

Auto-Arima

auto_arima() uses a stepwise approach to search multiple combinations of p,d,q parameters and chooses the best model that has the least AIC. In Our dataset we try to build the Auto Arima model.

Null Hypothesis

A null hypothesis is one that there is no statistical significance between the two variables in the hypothesis. The researcher is attempting to disprove the hypothesis. Here the P-value is 0.36 which is greater than 0.05, which means data is accepting the null hypothesis, which means data is non-stationary.

3.8 Re-Sampling

We have discussed earlier that our approach to experiment was test driven model development. Generally, computation of accuracy of the model and reducing error are the main concerns while enhancing training in data science experiments. For this purpose, we have applied one specific Re-Sampling technique in our Linear Regression algorithm. This technique is discussed below in terms of enhancement of the model.

3.8.1 Re-Sampling by Train-Test-Split

In the Train-Test-Split (TTS) re-sampling technique, we divided the dataset into two parts. The larger part (70-80%) is called the training dataset and the smallest part (20- 30%) is the test dataset. This is demonstrated in the figure below:

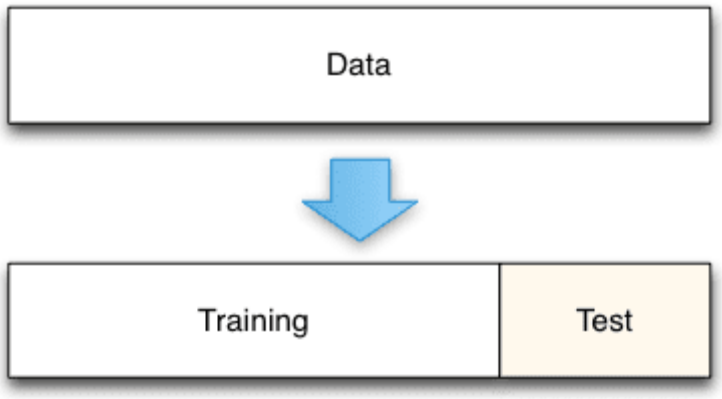


Figure 3.8: Train Test Split Re-sampling

Our goal of doing this re-sampling is to enhance user experience prediction and evaluation on out of sample performance-experience dataset.

3.9 Evaluation

Our final step before evaluating and comparing the models generated in the steps by using the re-sampling techniques used mentioned above. To do this, we have worked with several evaluation techniques that measure errors raised by prediction results.

3.9.1 Mean Absolute Error

In statistical and our research relative terms, Mean Absolute Error (MAE) is the error between coupled entries addressing the same occurrence. It is measured by using predicted and actual.

3.9.2 Root Mean Squared Error

Earlier when we discussed Residual Errors in the SLR section, we measured the distance of performance issue from the fitted regression line.

To calculate Root Mean Squared Error (RMSE) we calculate as follows:

$$RMSE=\sqrt{MSE}$$

Goal of doing RMSE is that we want to normalize the unit we got from MSE which is (millisecond)2 . After RMSE, we get the value back into its original unit millisecond.

CHAPTER FOUR

RESULTS AND DISCUSSION

In this section, we will discuss the results of each experiment we have performed. This section demonstrates all the results starting from Data Visualization. This research includes two types of machine learning models. The two models are the Linear Regression and the Time series model(Arima) . The models train to predict user interaction. The user interaction metrics include Platform like Facebook, Instagram, WhatsApp, Twitter, YouTube, LinkedIn and many more.

First of all, our target variable is Monthly active user based on Platform and selection Year. After visualization we selected some platforms and we will see what the users of these platforms might look like in the future. And for this we apply simple linear regression.

4.1 Simple Linear Regression

Table 4.1: Social Media Users Prediction With Regression

Platform	Predicted Users
Facebook	2022(3069667894.044) 2023(3282252897.574) 2024(3494837901.103)
Instagram	2022(1581773611.111) 2023(1751120111.111) 2024(1920466611.111)
YouTube	2022(6220146074.904) 2023(6689186874.181) 2024(7158227673.458)

4.2 Time series Algorithm Using Arima Model

Here we visualized Monthly active users and Youth users some selected Platforms like Facebook, Instagram, YouTube, Netflix, Twitter, LinkedIn, Whatsapp, Snapchat and TikTok.

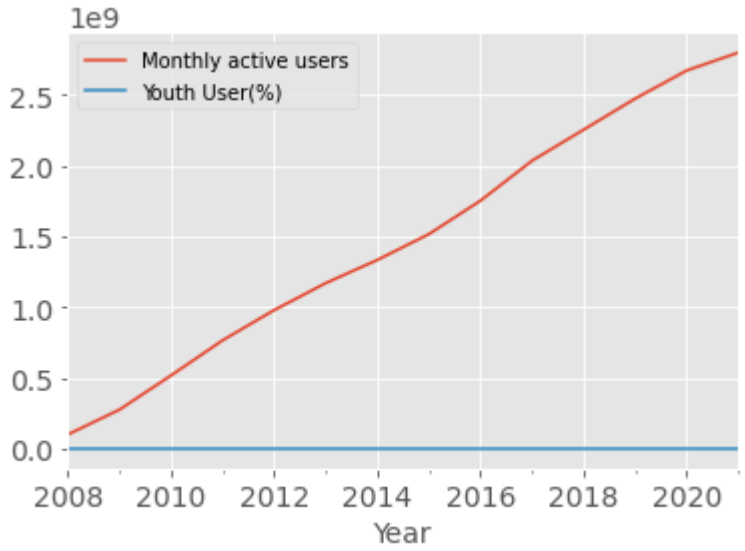


Figure 4.1: For Facebook monthly active user and youth user

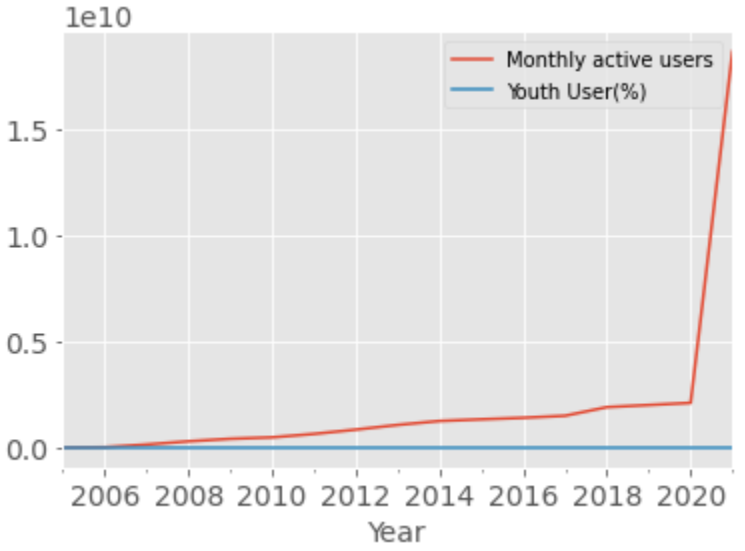


Figure 4.2: For YouTube monthly active user and youth user

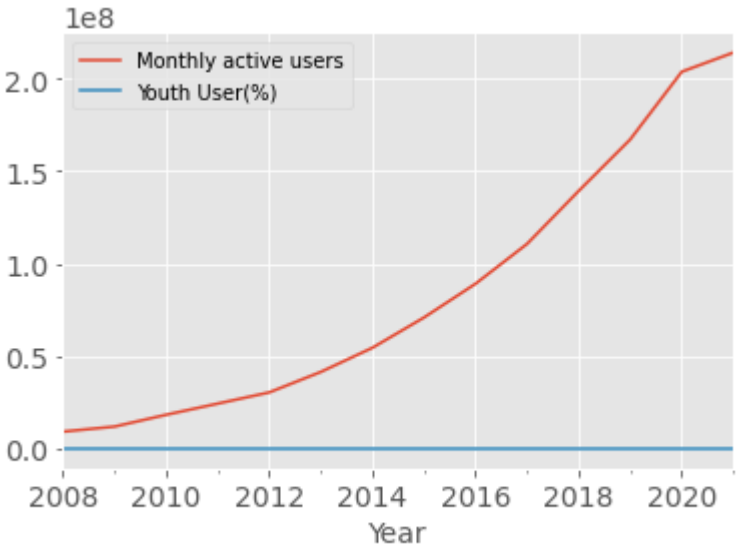


Figure 4.3: For Netflix monthly active user and youth user

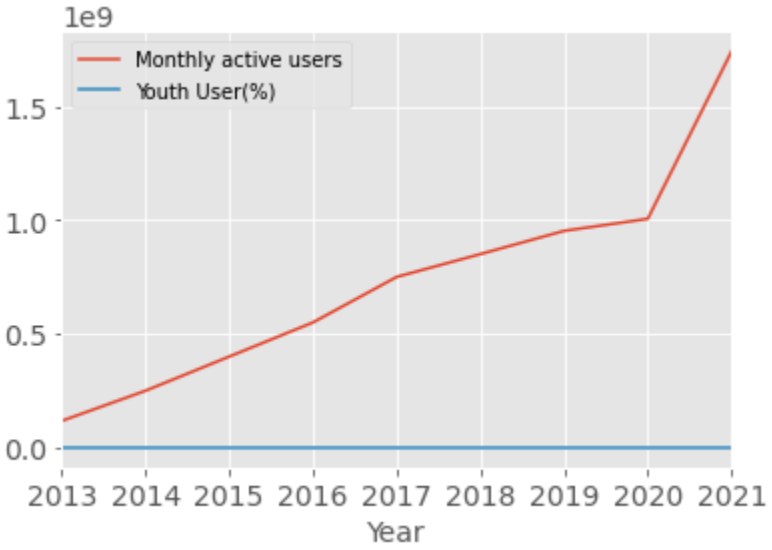


Figure 4.4: For Instagram monthly active user and youth user

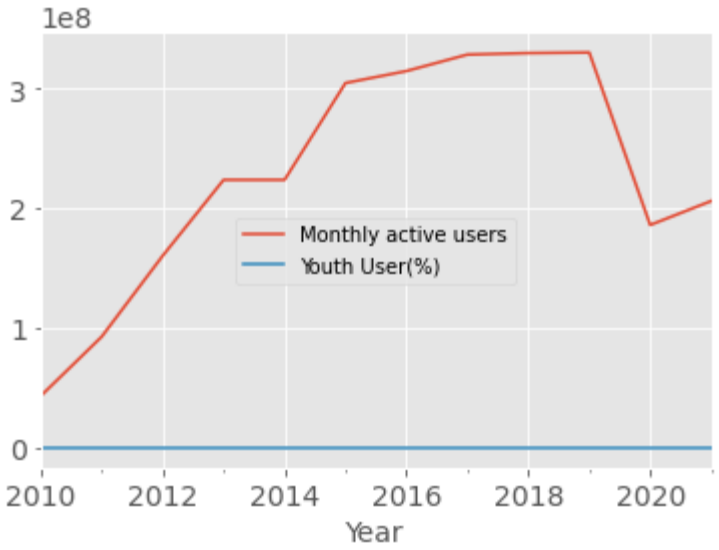


Figure 4.5: For Twitter monthly active user and youth user

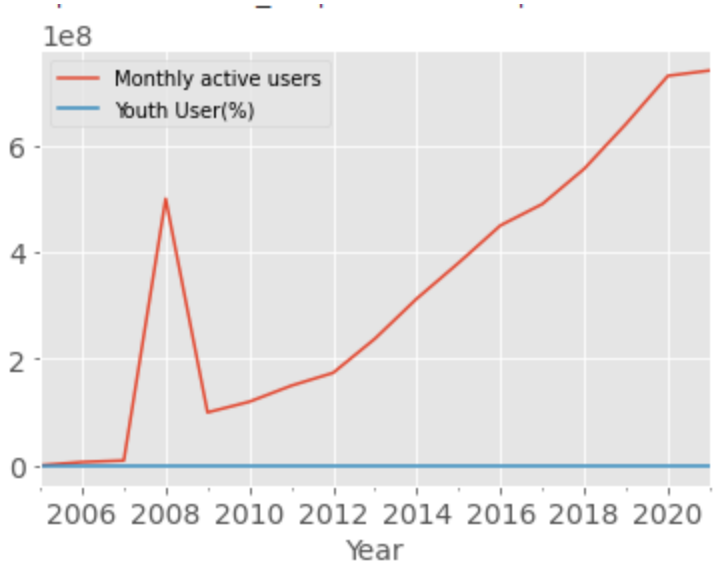


Figure 4.4: For LinkedIn monthly active user and youth user

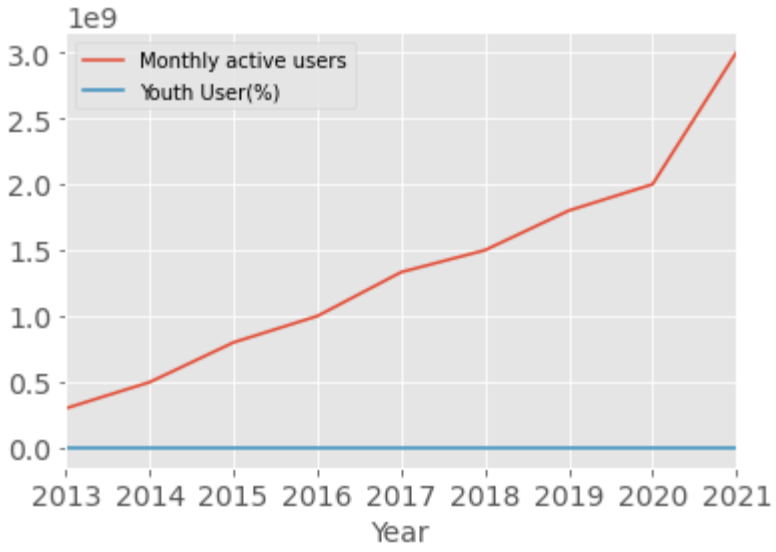


Figure 4.6: For WhatsApp monthly active user and youth user

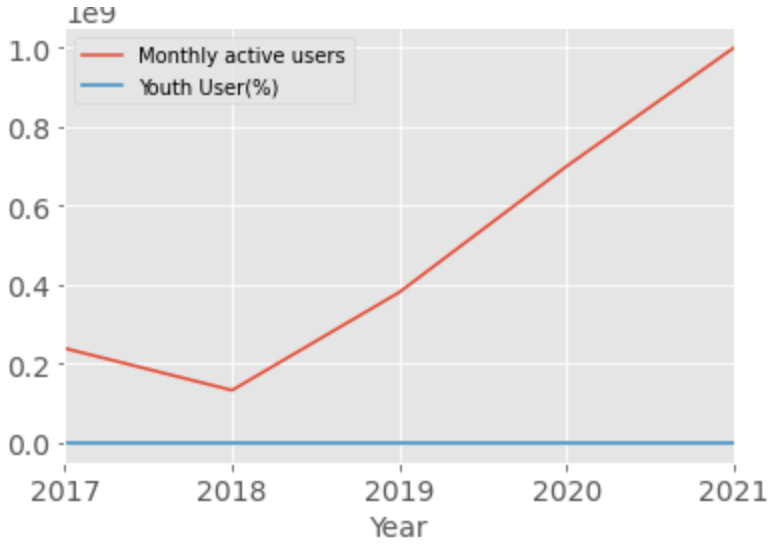


Figure 4.7: For TikTok monthly active user and youth user

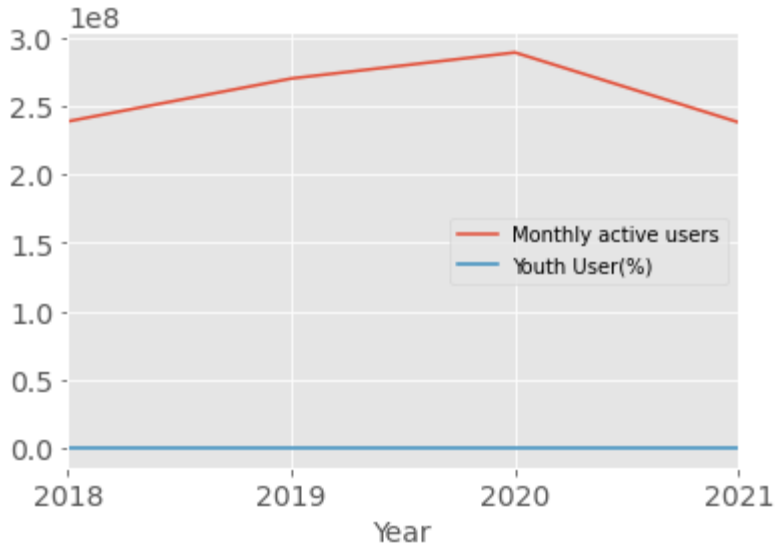


Figure 4.8: For Snapchat monthly active user and youth user

After this visualization we realise that Youth users are more acceptable for the time series algorithm. So, we work on it.

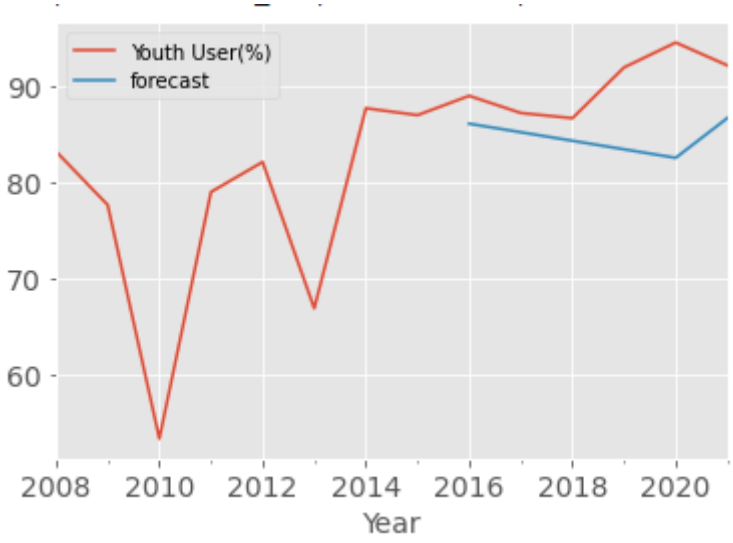


Figure 4.9: For Facebook Youth User Forecast

After this forecast diagram for facebook we evaluated this Platform based on Youth users we found RMSE 6.622340414324915, MAE 5.527537 and Accuracy 93.98%.

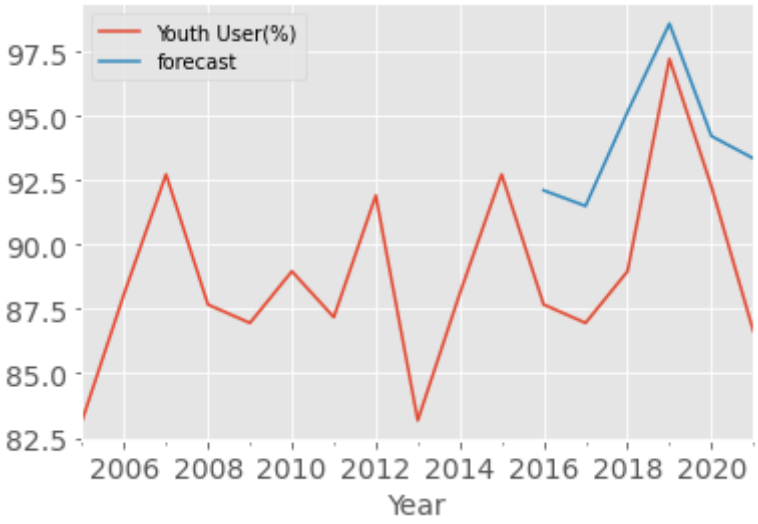


Figure 4.10: For YouTube Youth User Forecast

After this forecast diagram for youtube we evaluated this Platform based on Youth users we found RMSE 4.6423986765682645, MAE 4.200251 and Accuracy 95.25%.

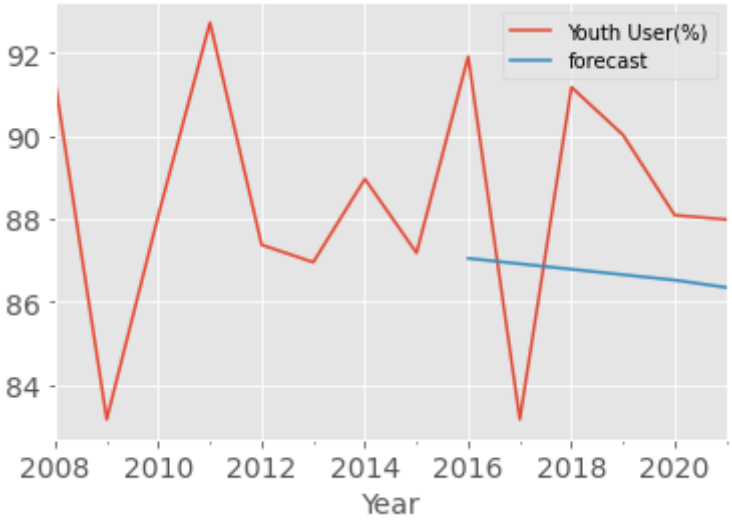


Figure 4.11: For Netflix Youth User Forecast

After this forecast diagram for netflix we evaluated this Platform based on Youth users we found RMSE 3.495796377861954, MAE 3.259307 and Accuracy 96.34%.

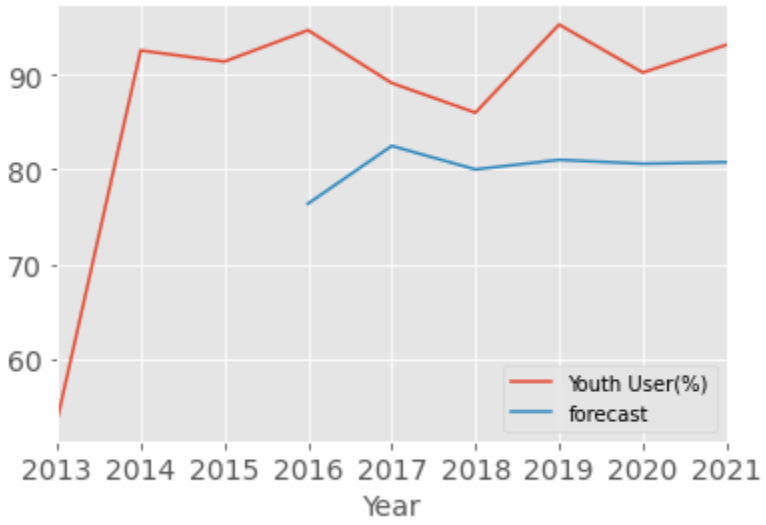


Figure 4.12: For Instagram Youth User Forecast

After this forecast diagram for instagram we evaluated this Platform based on Youth users we find RMSE 12.019671391400825, MAE 4.123056 and Accuracy 87.87%.

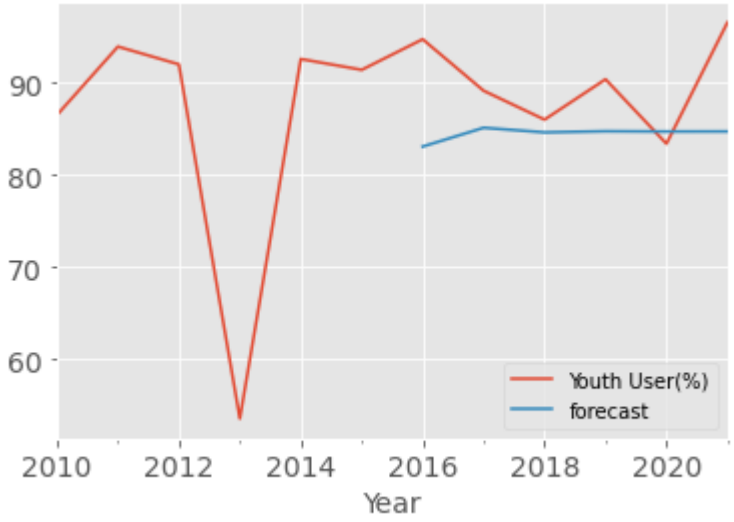


Figure 4.13: For Twitter Youth User Forecast

After this forecast diagram for twitter we evaluated this Platform based on Youth users we find RMSE 7.374665785216909, MAE 5.966157 and Accuracy 93.59%.

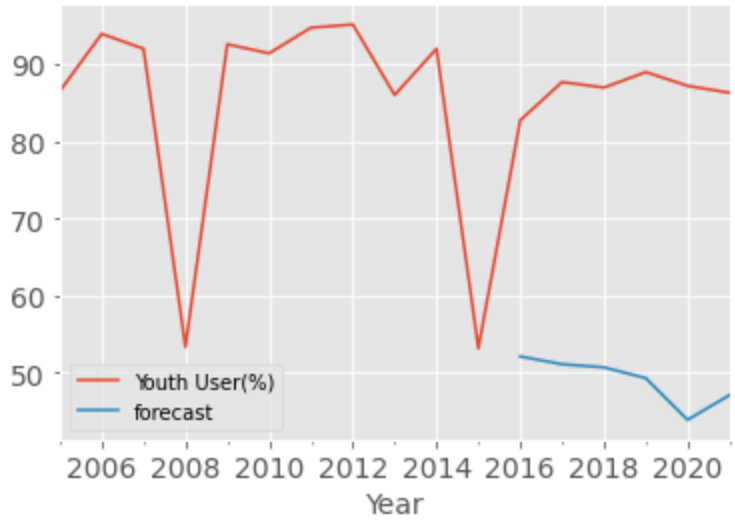


Figure 4.14: For LinkedIn Youth User Forecast

After this forecast diagram for linkedin we evaluated this Platform based on Youth users we find RMSE 37.776607806478374, MAE 37.576181 and Accuracy 56.68%.

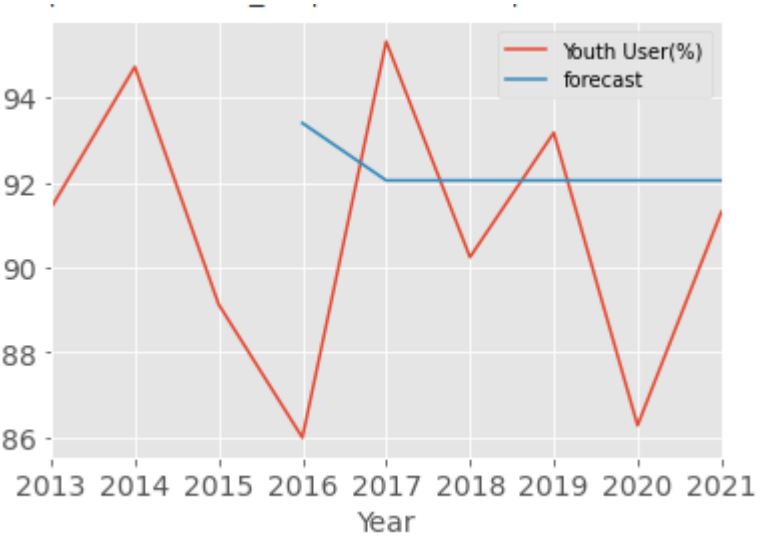


Figure 4.15: For WhatsApp Youth User Forecast

After this forecast diagram for Whatsapp we evaluated this Platform based on Youth users we find RMSE 4.154272903044234, MAE 3.346822 and Accuracy 96.09%.

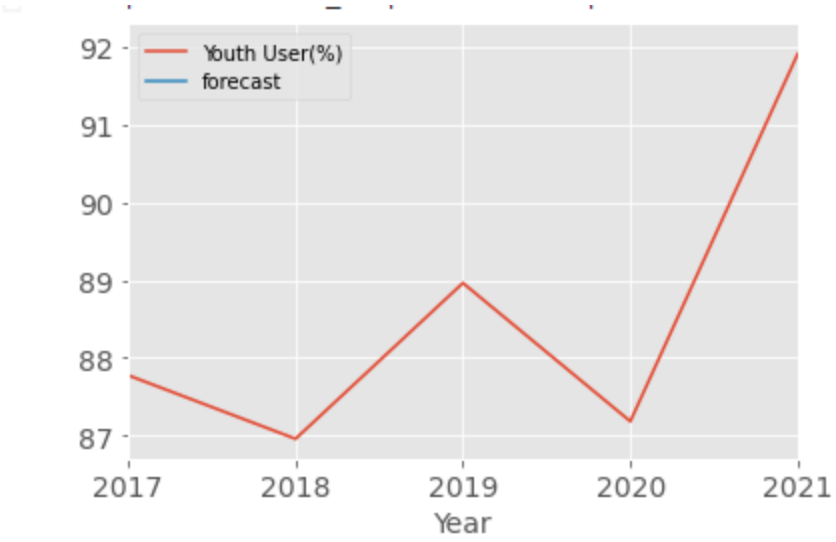


Figure 4.16: For TikTok Youth User Forecast

After this forecast diagram for tiktok we evaluated this Platform based on Youth users we find RMSE 3.2642729030442, MAE 2.329811 and Accuracy 95.14%.

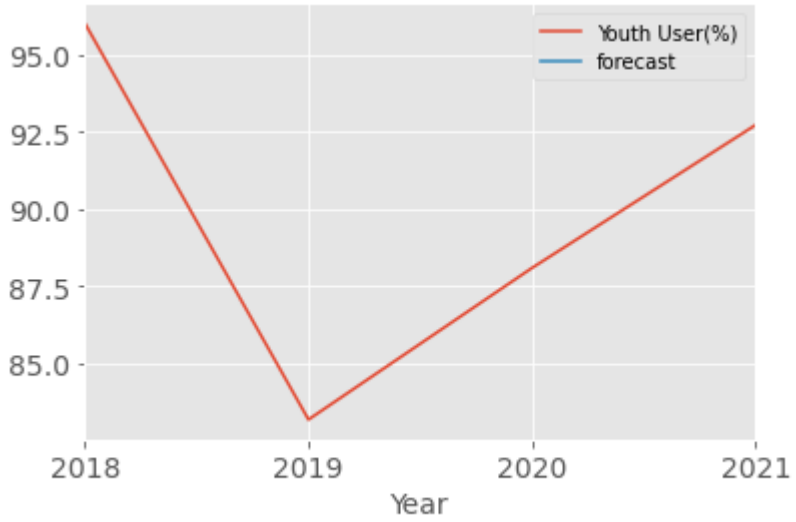


Figure 4.17: For Snapchat Youth User Forecast

After this forecast diagram for snapchat we evaluated this Platform based on Youth users we find RMSE 2.98760981234, MAE 1.78922 and Accuracy 40.09%.

4.3 Decision Making

We compared all the experiment results and visualization we have performed so far in terms of Time Series Algorithm.

Table 4.2: Social Media Users Prediction With Time Series

Platform	RMSE	MAE	Accuracy
Facebook	6.622	5.527537	93.98 %
YouTube	4.642	4.200251	95.25 %
Netflix	3.495	3.259307	96.34 %
Instagram	12.019	4.123056	87.87%
Twitter	7.374	5.966157	93.59%
LinkedIn	37.776	37.576181	56.68%
Whatsapp	4.154	3.346822	96.09%
TikTok	3.264	2.329811	95.14%
Snapchat	2.987	1.78922	40.09%

After evaluation we see that Netflix, Whatsapp and YouTube users are more active on social media platforms.

Finally, after all observations we found Time series with tuning gives the best result in terms of RMSE.

This thesis research concerns machine learning with data. It asks if machine learning models can predict user interaction from this data. Moreover, this thesis explores how well the models can learn the data.

4.4 Social Media Impacts

After all the experiments we found that people are using social media platforms based on trending topics. And after reading various research papers that I already mentioned in Literature Review we found the reason behind user interaction on social media platforms. But the younger generation are more active on social media platforms. Because they found everything there that they wanted. Social media use is ubiquitous; the most recent survey in the US showed that 88% of 18 to 29-year olds used social media. A large proportion of social media users engage with several platforms and they engage with these various platforms often. For example, 50% of Facebook users and Snapchat users visit the site several times a day and around 70% of younger users (18 to 24 year olds) engage with Snapchat multiple times a day (Smith & Anderson, 2018) (https://www.nationalelfservice.net/mental-health/depression/social-media-good-bad-experiences-impact-depression/). In[5] this paper found some effective points that mentioned below:

4.4.1. Use of social media for academic and nonacademic purposes

Social networking sites use is widespread among youth users because of the utility of smartphones and easy access to such sites through home computers. At present, social media platforms can be used to recover necessary information that serves educational purposes. Nevertheless, social media use negatively affected academic improvement and studies have shown a strong positive relationship between academic performance and social media usage. Most participants used social media platforms to talk rather than for academic purposes. Previous research has found that students who spend more time on social media sites perform worse academically. This is because they spend time talking online and making friends on social media sites instead of reading books.

This has a negative effect on their academic performance (Owusu-Acheaw and Larson, 2015; Abbas et al., 2019).Therefore, it is important to condition the duration of time that they spend on social media sites and the ratio of time that is exhausted on social media sites for academic purposes. Social media plays an vital role in education. However, because several social networking sites exist, students spend more time talking, watching movies, shopping, and playing games rather than on educational activities (Abbas et al., 2019). The use of social media has both positive and negative effects. Yet, the negative effects are more pronounced because students use such platforms for entertainment and waste time rather than for academic purposes. This may detract their attention away from learning and academic activities. This study determined the percentage of students who felt more drained to social media than to academic activities, as well as the preference for using social media for fun rather than for academic purposes. The findings emphasize the importance of raising student awareness about the negative effects of such habits on academic performance. This will assist students in excelling academically and gaining adequate knowledge, which will improve their performance in competitive examinations.

4.4.2 Social Media Effects

Social media use has mental health effects and young adults are the most vulnerable one. Studies have shown that social media use is associated with mental disorders, including depression and anxiety (Hu et al.,2001). Social media helps individuals connect with others and develop new relationships. Relying solely on social media (i.e., without physical proximity) to build and maintain relationships can lead to loneliness, alienation, and depression (OwusuAcheaw and Larson, 2015). Smartphones establish a psychological barrier between people by reducing face-to-face interactions between family and friends, which can have a detrimental impact on the quality

of time spent in these relationships. This can have a significant effect on social well-being and satisfaction among friends (Abbas et al., 2019).These changes have important behavioral and social implications. Adults who use social media have less physical activity and spend more time. These modifications have a larger impact on the physiological mechanism. This is associated with impaired lipid profiles and glucose uptake, greater energy intake, higher waist circumferences, and greater mortality risk (Sobaihy, 2017; Healy et al., 2008b, 2007).

4.4.3 Effects of social media during (COVID-19) pandemic

Participants in this study reported prolonged non-academic social media use, social media addiction, distraction from learning, a lack of sleep, and decreased social interactions. These findings are especially concerning given the ongoing COVID-19 pandemic. Colleges and universities have adopted new teaching methods as a result of educational institutions being closed to combat the spread of COVID-19. Traditional methods of instruction have given way to collaborative multimedia distance learning techniques. As a result, universities have implemented distance learning strategies. As a result, they spend less time on social networking sites, sit for shorter periods of time, and engage in some physical activity. Online learning methods, however, have been adopted since the outbreak of COVID-19. This has increased the duration of time spent on mobile devices and computers, resulting in increased sitting time and decreased physical activity levels. These modifications may increase the likelihood of developing metabolic syndrome and noncommunicable diseases. Furthermore, the outbreak of COVID-19 prevented them from socializing with their college friends. This could also have a negative impact on their mental health, leading to feelings of loneliness and depression. As a result, the COVID-19 pandemic has a significant impact on physical activity, face-to-face social interactions, and mental health, causing

significant stress and anxiety. Excessive social media use as a result of the COVID-19 pandemic may have a negative impact on learning. It is recommended to prevent the COVID-19 pandemic from causing negative mental health and cardiovascular consequences as a result of abrupt cessation of physical activity.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Findings

Our study seeks to predict user interaction. This is different, because this study used regression, since the data was continuous. This study is also different because it incorporates the Time series model. In our literature review, we found performance issues and various studies relevant to trending of user experience over time. We have tried different Feature Selection methods and tried the feature sets in various combinations and exchanged them with resampling techniques like Train-Test-Split with different algorithms to best predict the visual progression. After different sets of experiments, we have found the best prediction model for our dataset which is time series algorithm. We want to see the amount of youth users are involved in social media. We found that Youth users are more involved in this area after doing visualization.

5.2 Contributions

In this research our contributions are listed as follows:

1. Analytically collected data from Github and Statista.
2. Understanding Data through Exploratory Data Analysis.
3. Based on Model visual progression, finding the impact of performance issues through feature significance.

5.3 Recommendation on future works

In this research, we have tried to find feature importance based on visual progression based user experience. Obviously, this is not the entire user experience. There are other user experience metrics in our dataset like Country based Youth Users. To expand the research based on a separate Country and Based on Year collected more users, then experiment different Machine learning and Deep learning algorithms, one can measure other importance of performance issues.

REFERENCES

1. Statista, ‘‘Number of social network users worldwide from 2017 to 2025 (in billions).’’ [Online]. Available: http://www.statista.com/statistics/278414/number-ofworldwide-social-network-users/
2. T.K., B., Annavarapu, C. S. R., & Bablani, A. (2021). *Machine learning algorithms for social media analysis: A survey. Computer Science Review, 40, 100395.*
3. Crowe, C. (2018). *Predicting User Interaction on Social Media using Machine Learning* (Doctoral dissertation, University of Nebraska at Omaha).
4. Rzewnicki, Daniel I.; Shensa, Ariel; Levenson, Jessica C.; Primack, Brian A.; Sidani, Jaime E. (2020). Associations between positive and negative social media experiences and sleep disturbance among young adults. Sleep Health, (), S2352721820300796–.
5. Kolhar, M., Kazi, R. N. A., & Alameen, A. (2021). *Effect of social media use on learning, social interactions, and sleep duration among university students. Saudi Journal of Biological Sciences, 28(4), 2216–2222.*
6. Rzewnicki, Daniel I.; Shensa, Ariel; Levenson, Jessica C.; Primack, Brian A.; Sidani, Jaime E. (2020). *Associations between positive and negative social media experiences and sleep disturbance among young adults. Sleep Health, (), S2352721820300796–.*
7. Debnath, R., Bardhan, R., Reiner, D. M., & Miller, J. R. (2021). Political, economic, social, technological, legal and environmental dimensions of electric vehicle adoption in the United States: A social-media interaction analysis. *Renewable and Sustainable Energy Reviews*, *152*, 111707.
8. J. C. Flack and R. M. D'Souza, "The Digital Age and the Future of Social Network Science and Engineering," in Proceedings of the IEEE, vol. 102, no. 12, pp. 1873-1877, Dec. 2014.

9. Xiang, Zheng; Du, Qianzhou; Ma, Yufeng; Fan, Weiguo (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58(), 51–65.
10. Batrinca, B., Treleaven, P.C. Social media analytics: a survey of techniques, tools and platforms. *AI & Soc* 30, 89–116 (2015).
11. S. Asur and B. A. Huberman, "Predicting the Future with Social Media," 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010, pp. 492-499.
12. K. Zhang *et al.*, "SES: Sentiment Elicitation System for Social Media Data," *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 129-136.
13. Bryan, M. A., Evans, Y., Morishita, C., Midamba, N., & Moreno, M. (2020). Parental perceptions of the internet and social media as a source of pediatric health information. *Academic pediatrics*, *20*(1), 31-38.
14. Zhu, Alex Yue Feng; Chan, Alex Lih Shing; Chou, Kee Lee (2019). Creative social media use and political participation in young people: The moderation and mediation role of online political expression. Journal of Adolescence, 77(), 108–117.
15. Fry, John; Binner, Jane M. (2015). Elementary modelling and behavioural analysis for emergency evacuations using social media. European Journal of Operational Research, (), S0377221715004397–.
16. P. Chen, S. Cheng, P. Ting, C. Lien and F. Chu, "When crowdsourcing meets mobile sensing: a social network perspective," in *IEEE Communications Magazine*, vol. 53, no. 10, pp. 157-163, October 2015.

17. Colace, Francesco; Casaburi, Luca; De Santo, Massimo; Greco, Luca (2015). Sentiment detection in social networks and in collaborative learning environments. Computers in Human Behavior, 51(), 1061–1067.
18. Gibson, Kerry; Trnka, Susanna (2020). Young people's priorities for support on social media: “It takes trust to talk about these issues”. Computers in Human Behavior, 102(), 238–247.
19. Ghani, Norjihan Abdul; Hamid, Suraya; Targio Hashem, Ibrahim Abaker; Ahmed, Ejaz (2018). Big Social Media Data Analytics: A Survey. Computers in Human Behavior, (), S074756321830414X–.
20. Whelan, Eoin; Brooks, Stoney; Islam, A.K.M. Najmul (2019). Applying the SOBC paradigm to explain how social media overload affects academic performance. Computers & Education, (), 103692–.

APPENDIX – A

List of Abbreviation

EDA	Exploratory Data Analysis
SLR	Simple Linear Regression
MSE	Mean Squared Error
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
ARIMA	Autoregressive Integrated Moving Average
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Auto-Regressive Integrated Moving Average with exogenous factors
AR	Auto Regressive
MA	Moving Average
AIC	Akaike Information Criterion