



**BERLIN SCHOOL OF
BUSINESS & INNOVATION**

Essay / Assignment Title: User Behavior Analysis for Optimizing Engagement on Social Media Platforms

Programme title: MSc Data Analytics

Name: VICTORY CHIEMEKA OKEZIE

Year: 2024

CONTENTS

INTRODUCTION	4
CHAPTER 1: DATA COLLECTION	5
CHAPTER 2: ANALYZING THE DATASET	7
CHAPTER 3: DATASET PREPROCESSING	10
CHAPTER 4: ALGORITHM APPLICATION.....	13
CHAPTER 5: SPECIFYING THE PROBLEM AND MODEL EVALUATION	15
CONCLUSION.....	17
BIBLIOGRAPHY	18

Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program).

Name and Surname (Capital letters): VICTORY CHIEMEKA OKEZIE

.....

Date:21...../....12...../....2024....

INTRODUCTION

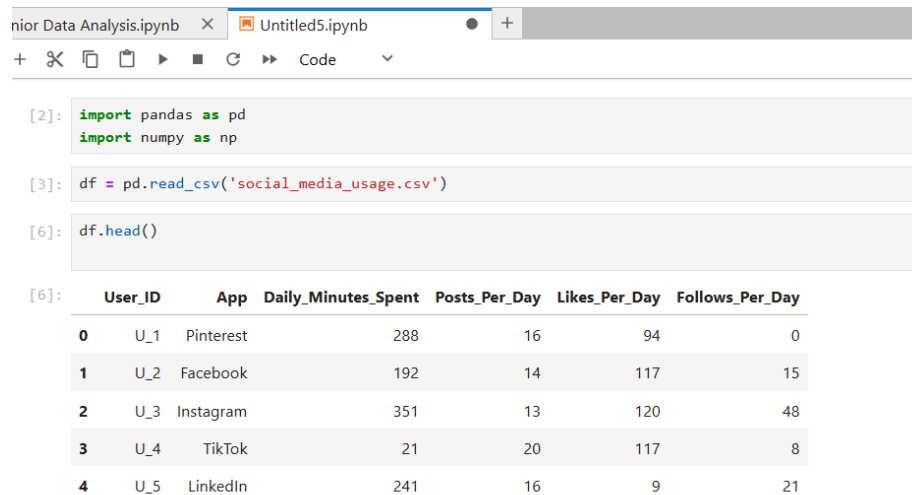
In the dynamic world of social media, platforms continuously strive to enhance user engagement and deliver personalized experiences to maintain a competitive edge. As an operator of a social media platform offering diverse features such as posts, messaging and content recommendations. Understanding user behavior is critical to achieving these goals. By analyzing and categorizing user activities, platforms can optimize content delivery, foster meaningful interactions and improve overall user satisfaction.

This essay aims to analyze and categorize user behavior pattern based on their activity on social media platforms. We will be using a real-world dataset from Kaggle.com, this dataset addresses user engagement in seven popular platforms: Facebook, Instagram, Twitter, Snapchat, TikTok, LinkedIn and Pinterest. The primary objectives that are needed to analyze and categorize user behavior patterns will be to collect specific data points, clean and data preprocessing, statistical algorithms for mathematical basic solutions, and model evaluation and interpretation.

CHAPTER 1: DATA COLLECTION

Throughout this essay, we will analyze a dataset from Kaggle.com. (Kaggle.com is an online platform where users contest, teach and share real world datasets). Kaggle. (n.d.). Discover and share datasets. Available at: <https://www.kaggle.com> (Accessed: 13 December 2024). We will use the dataset that was uploaded by Bhadra Mohit. (October 2024). *Social Media Usage Dataset (Applications)*. Available at: <https://www.kaggle.com/datasets/bhadramohit/social-media-usage-datasetapplications>. (Accessed: 13 December 2024).

We will proceed to import the data set into our Jupiter notebook for better understanding and to capture the necessary data points needed for this analysis. The code we type to view the first five header is `df.head()`, view figure 1 below



```
[2]: import pandas as pd
import numpy as np

[3]: df = pd.read_csv('social_media_usage.csv')

[6]: df.head()
```

	User_ID	App	Daily_Minutes_Spent	Posts_Per_Day	Likes_Per_Day	Follows_Per_Day
0	U_1	Pinterest	288	16	94	0
1	U_2	Facebook	192	14	117	15
2	U_3	Instagram	351	13	120	48
3	U_4	TikTok	21	20	117	8
4	U_5	LinkedIn	241	16	9	21

Figure 1: The Dataset

We know that there is a total of 1000 users in the dataset and these users are selected randomly across the seven different social media. From the above figure, the best data points to use are those that are related to in app activity such as user likes, post per days, daily minutes spent etc.

The main potential challenges in gathering accurate data on user preferences and interests from social media activity includes data that has lack of context such as posts, likes and minutes spent lack sufficient context to determine a user genuine interest. Also, privacy laws or data

restrictions are ethical laws that limit companies or analyst from extracting user specific data without the user consent. And finally, fake accounts such as bots and counterfeit user account can give wrong or inconsistent data for the analysis.

The author of the dataset we are using (Bhadrat Mohit from Kaggle) gathered data from previous user ratings insights using sites like google and social media review to foster enough data input that will be suited for this analysis. Therefore, the dataset being used for this essay is 95% genuine and free from the potential challenges some analysist may come across when getting data collection.

Understanding linear algebra, statistics and calculus is crucial in structuring the data for more elaborate insights. For example, we will use statistics matrix like mean, median, variance, and standard deviation to summarize user behavior (e.g., average daily active user). Problem solving techniques like mathematics are important to optimize the frequency of our recommendations. And what if we would like to find target customers segment in order to find specific customers that we want to deliver a special kind of personalized content? We can always use k-means clustering for that.

CHAPTER 2: ANALYZING THE DATASET

Now that it is clear where we got the dataset and its authenticity, we can proceed forward to analyze the dataset. Analyzing the dataset is very important because from that we will be able to understand and interpret our dataset better, we will check for missing values and clean the dataset, we also perform some basic descriptive statistics like EDA to find the average and highest user social media in app activity. In the previous chapters, it was stated that the dataset was clean by the author, in order to verify this, we will use the `df.isnull()` and `df.isnull().sum()` to check for the any missing values. Additionally we can type `df.duplicated().sum()` to check for any duplicated value in the dataset. See Figure 2 below.

```
[13]: df.isnull().sum()
```

```
[13]: User_ID          0
      App             0
      Daily_Minutes_Spent 0
      Posts_Per_Day    0
      Likes_Per_Day    0
      Follows_Per_Day  0
      dtype: int64
```

Figure 2: Dataset missing values

We can also draw a visualization that shows a heat map of any missing values across the entire 1000 rows and 6 columns using seaborn heatmap. See Figure 3 below.



Figure 3: Heatmap

We are sure that the dataset is clean and there are no missing values, we can then proceed to check the descriptive statistics of the social media users. The descriptive statistics which are also known as exploratory data analysis (EDA) shows us the social media app that have the highest, lowest and mean in app activity, it is useful for our analysis as we will use it to determine how we can boost user engagement or whom to optimize content delivery. See Figure 4.

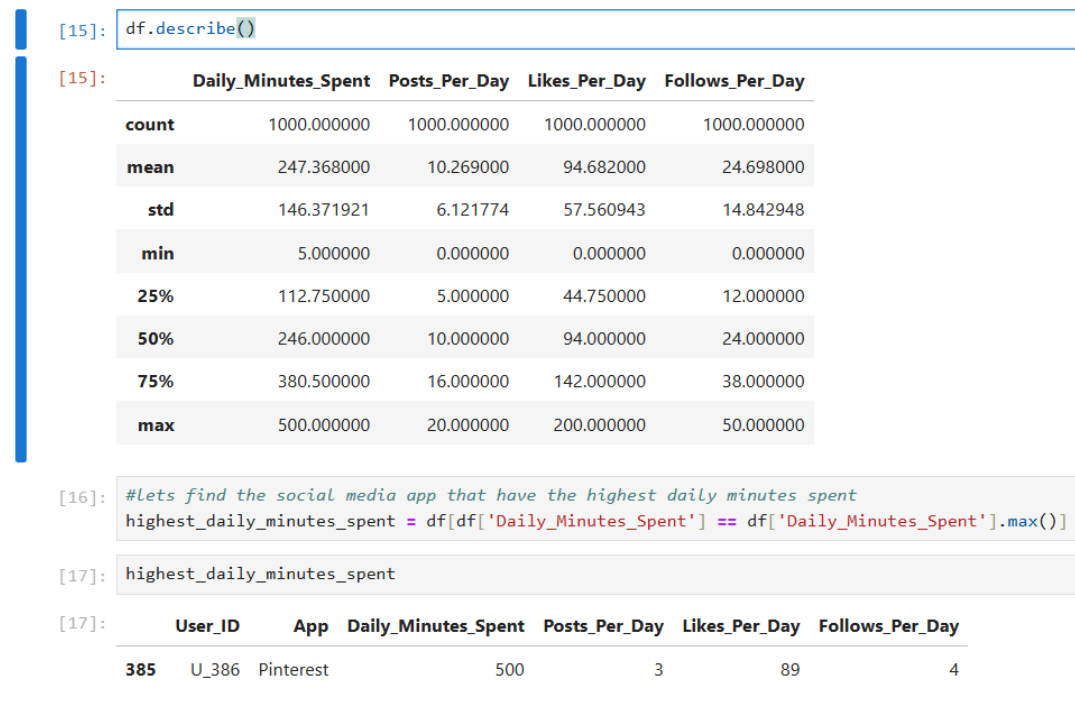


Figure 4: EDA of Dataset

Pinterest has the highest number of daily minutes spent. As a social media platform operator, if I want to improve the user engagement on this platform, I would offer personalized services and special premium qualities to maintain user daily minutes spent. Whereas LinkedIn have the lowest number of daily minutes spent. As a social media platform operator, if I want to improve the user engagement on this platform, I would offer new features and give discounts to premium features.

In the next chapters we will be using k-means clustering to segment users into groups based on their engagement level which is daily minutes spent, in order to create marketing strategies for this user and increase their engagement with it. We will also use logistic regression from the k-

means clustering to predict users that are likely to become inactive or leave the platform, once we find those users we will optimize content delivery to keep them engaged.

Now I am going to add three extra columns which are age, country and demographic to our dataset, this ensures that our models will work effectively and it will tailor to the right users and not just to anyone. We will use NumPy to update the dataset. See Figure 5. Note that this new data was selected and applied at random, it will not disturb our analysis, it will only enhance it.



Figure 5: New Columns

Our dataset contains values that are numbers, these variables like "Likes_Per_Day" and "Daily_Minutes_Spent" are numerical (quantitative) data, so therefore we will have to use specific visualizations that fits best these datasets e.g. Histogram or Box plots. We will also need to use z scores and IQR method to perform a statistical analysis of the collected data to identify patterns or outliers. All this will be covered in the next chapter. For now, we have fully understood what type of dataset we are dealing with and what will be the necessary steps that will improve user engagement and optimize personalized content delivery.

CHAPTER 3: DATASET PREPROCESSING

In this chapter, we will compute all of our statistical analysis and we will pave the way to segment customers with k-means clustering in order to prepare the dataset for the two main machine learning model. As mentioned in the previous chapter, our dataset has numerical data therefore it can provide us with the trends and distribution of the social media users in app activity. We have to identify outliers in the dataset because outliers can distort our statistical analysis, it can also make it difficult to impute our two machine learning models. To identify outliers, we will use two methods z-score and interquartile range (IQR). See Figure 6 below.

```
print(minutes_outliers)
print(likes_outliers)
print(post_outliers)
print(follow_outliers)
print(Age_outliers)

Empty DataFrame
Columns: [User_ID, App, Daily_Minutes_Spent, Posts_Per_Day, Likes_Per_Day, Follows_Per_Day, Age, Country, Demographics, z_score, Minutes_Z, Likes_Z, Post_Z, Follows_Z, Age_Z]
Index: []
Empty DataFrame
Columns: [User_ID, App, Daily_Minutes_Spent, Posts_Per_Day, Likes_Per_Day, Follows_Per_Day, Age, Country, Demographics, z_score, Minutes_Z, Likes_Z, Post_Z, Follows_Z, Age_Z]
Index: []
Empty DataFrame
Columns: [User_ID, App, Daily_Minutes_Spent, Posts_Per_Day, Likes_Per_Day, Follows_Per_Day, Age, Country, Demographics, z_score, Minutes_Z, Likes_Z, Post_Z, Follows_Z, Age_Z]
Index: []
Empty DataFrame
Columns: [User_ID, App, Daily_Minutes_Spent, Posts_Per_Day, Likes_Per_Day, Follows_Per_Day, Age, Country, Demographics, z_score, Minutes_Z, Likes_Z, Post_Z, Follows_Z, Age_Z]
Index: []
Empty DataFrame
Columns: [User_ID, App, Daily_Minutes_Spent, Posts_Per_Day, Likes_Per_Day, Follows_Per_Day, Age, Country, Demographics, z_score, Minutes_Z, Likes_Z, Post_Z, Follows_Z, Age_Z]
Index: []
df['Minutes_Z'].hist(bins=30)
plt.show()
```

Figure 6: Z score

Our goal was to find values that are unusual high or low and that are too far apart from the normal data we have. We did this by calculating the z score of every column with numerical values (float or integer) e.g. minutes spent and likes per post, we then added a threshold of 3 for the z score. The above graph shows that there are no outliers or extreme outliers, possible reasons can be that the data may not contain values that are unusually far from the mean, so no z-scores exceed the threshold. We will also use IQR to identify outliers incase the data is skewed and follow a non-normal distribution.

For interquartile range, we will first calculate Q1 and Q3 of each numerical column, then we will add the outlier threshold, doing this for all the numerical column is superfluous because we can easily define a function that calculates the IQR and returns its value. See Figure 7 below.

```
def identify_outliers_iqr(df, column_name):
    Q1 = df[column_name].quantile(0.25)
    Q3 = df[column_name].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = df[(df[column_name] < lower_bound) | (df[column_name] > upper_bound)]
    non_outliers = df[(df[column_name] >= lower_bound) & (df[column_name] <= upper_bound)]
    return outliers, non_outliers
```

Figure 7: Interquartile Range

After inputting this function to all the numerical column, our result is that there are no outliers so therefore our data is good for the machine learning model.

Visualizations such like line charts and bar graphs are powerful tools for understanding trends in numerical datasets, especially for user activity on a social media app over time. Line charts are ideal for showing how user activity (e.g., posts, likes, comments) changes over time. A line chart can reveal rising or falling trends, while bar charts can be useful for comparing user activity across discrete time intervals, such as comparing activity on different days of the week or different months. They help to easily spot which days or time periods saw the highest activity. Next, we will be performing k-means clustering to segment users based on their activities and we will use the clusters to draw a bar chart.

Due to the fact that our dataset has 1000 rows, we will need to determine the number of clusters needed to fit the social media users effectively. To this we will use elbow method to find the optimal clusters. The elbow method helps to identify how many distinct customers segment that exist in the data. See Figure 8.1 below. Once we find the number of clusters we need, we will then label it and group it by their mean. We can now impute the k-means to the clusters. See Figure 8 1.1 below.

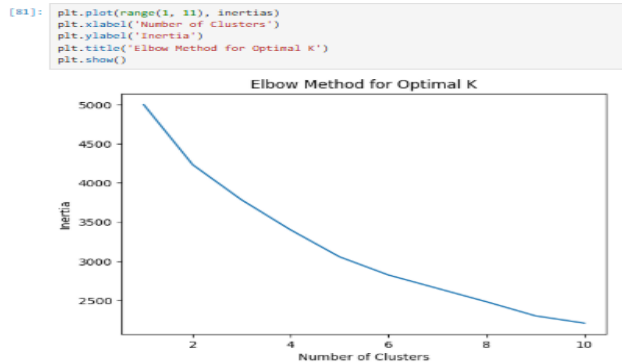


Figure 8.1: Elbow Method for Optimal K

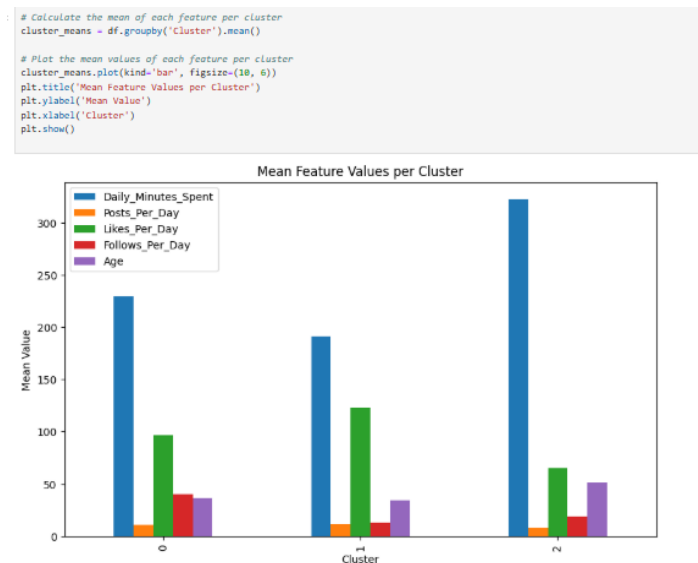


Figure 8.1.1: Bar Chart of mean feature values

From Figure 8.1.1 we can see that cluster 2 has the highest amount of time spent on social media and these are older adults. As a social media platform operator in order to improve user engagement, I will introduce games and targeted activities to cluster 0 and cluster 1 young users, in order to engage them and increase their post per day. I will highlight cluster 1 and 2 post on their feed in order to encourage them to post more and to increase their followers. I will send personalized content related notification to cluster 0 and 2 in order to increase the number of likes click in the app. All of the personalized and improvement of content delivery can be done using machine learning models, in the next chapters we will use Logistic regression to predict user that may want to leave the app.

CHAPTER 4: ALGORITHM APPLICATION

For the logistic regression, we will select two columns from our previous K-Means bar chart which are daily minutes spent and post per day, this column will make our prediction less redundant because they are the exact datapoints that normally determine a user social media interest. Other data points like likes or followers does not directly relate to a user interest in a social media app. See Figure 9 below.

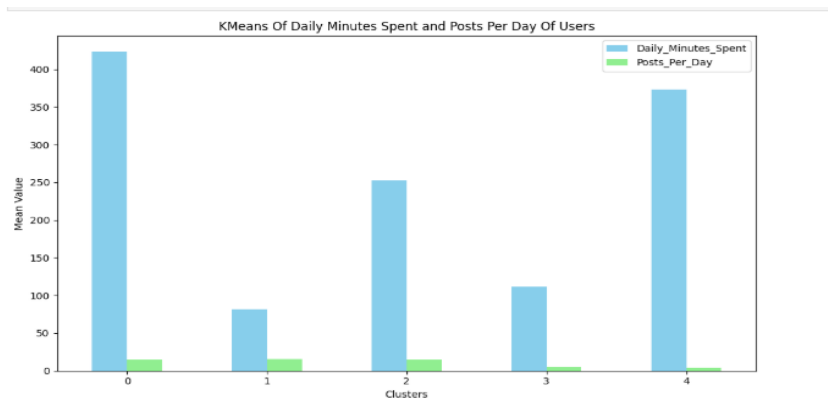


Figure 9: Bar Chart of K-Means of Daily Minutes Spent and Post Per Day

From Figure 9, we can see that on average users spend more time on social media but they post less. With logistic regression, we will be able to determine the exact one thousand users that will leave the app based on Figure 9. For the logistic regression we will use a scatter plot to show the data points, color-coded by their actual Leave status (0 or 1), and overlay the decision boundary of the logistic regression model. 0 will mean to stay while 1 will mean to leave. See Figure 10 below.

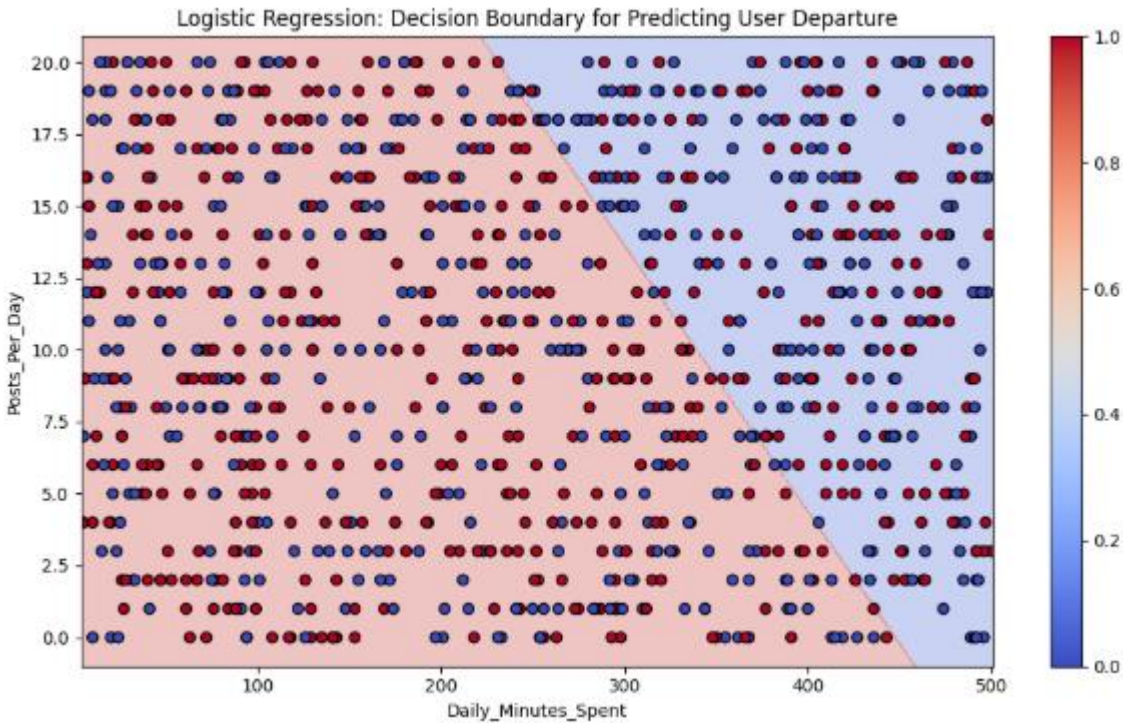


Figure 10: Logistic Regression for Predicting User Departure

From the above Figure 10, they are more users that will leave the app than those that would like to stay. If we apply a threshold of half (which is 0.5) to `predict_proba()` code, we will get 674, which is the exact number of users from Figure 10 that may leave the app based on daily minutes spent and post per day. Analyzing user preferences and in app activities on a social media app is not constant; it varies from user to user, predicting a user minute spent and post per day doesn't 100% correlate to the user interest in the app, the user may be spending less time posting but could be highly interested in the app, user preferences is always subjective to the user.

As a social media operator, I will definitely tailor recommendations and incentives to the 674 users to reduce the probability of them leaving the social media app. I will leverage new features and monitor or improve my K-Means Clustering to keep up with trends.

CHAPTER 5: SPECIFYING THE PROBLEM AND MODEL EVALUATION

The primary problem is to analyze and categorize user behavior patterns based on activity on a social media platform, with the goal of improving user engagement and optimizing personalized content delivery. This problem ‘identifying user engagement patterns’ should be considered as a clustering problem rather than a classification problem because in clustering, there are no predefined categories or labels for the data. We will have to discover natural groupings in the data based on similarities. Moreover, clustering is often used in exploratory data, where the goal is to uncover hidden patterns and structure in the data.

Having well-defined user activity segments on a social media platform can provide significant benefits, such as by segmenting users based on activity, the platform can deliver more relevant tailored content to each user. Brands can create targeted ad campaigns based on user behavior patterns which will increase ad effectiveness. Segments can be used to foster and strengthen communities within the platform. Well-defined user activity segments provide a social media platform with actionable insights that can improve user engagement and overall platform success.

Through K-Means clustering, users were effectively segmented into groups based on activity levels, enabling tailored strategies to target specific user demographics. Logistic regression further allowed predictions of user attrition, highlighting actionable insights to retain at-risk users through personalized recommendations and incentives. Although challenges such as class imbalance and the subjective nature of user preferences persist, these models offer robust foundations for improving social media platform performance.

K-Means doesn’t give a direct accuracy score like supervised models since it’s unsupervised. The performance is measured by how well it clusters the data. My clusters are meaningful (e.g. they group users who share similar in-app activities) therefore my k-means should be successful even without an explicit accuracy score.

My Logistic Regression model has an accuracy score of 0.499999, which means that accuracy alone is not the best way for dealing with the dataset (as it can be misleading). I applied confusion matrix and classification report, the report gave me a model that correctly predicts all instances of class 1. My F1-score is 0.66, this is a decent score for class 1 because it is showing a good balance of precision and recall. The model performed poorly to identify any instances of class 0 meaning all class 0 may be predicted as class 1. To get a better understanding of the model performance, we can consider the use of techniques like SMOTE or adjusting the dataset. The Below table explains both models in terms of their accuracy, robustness, speed, interpretability and scalability.

Table 1: Table Showing both models in terms of their accuracy, robustness, speed, interpretability and scalability

ASPECT	K-Means Clustering	Logistic Regression
Accuracy	It relies on clustering quality	Accuracy is 0.50 due to class 0 imbalance
Robustness	Sensitive to outliers and cluster shape	Sensitive to class imbalance, can underperform with imbalanced datasets
Speed	Relatively Fast for moderate-sized datasets; slow for large k or high-dimensional data.	Fast for smaller datasets, but may need optimization for large datasets
Interpretability	Easy to interpret cluster assignments.	Very interpretable through feature coefficients.
Scalability	Scales well, especially with mini-batch K-Means	Scales well with large datasets and many features

CONCLUSION

In this essay, we were successfully able to analyze and categorize user behavior patterns based on their activity in the platform, using two distinct machine learning models; K-Means clustering and Logistic Regression. The evaluation of this models gave us significant insights on how to improve user engagement and optimize personalized content delivery. Key findings include the number of users that are inactive or may leave the platform, the number of users that will need similar recommendations based on their segment clusters and the what is the average user in app activity.

By leveraging the real-world data we had from Kaggle, this essay demonstrated the power of data collection, preprocessing, statistical analysis, and machine learning techniques like K-Means clustering and logistic regression in understanding user activities. These methods revealed key user engagement patterns and helped identify strategies to improve user retention and personalize content delivery.

Moving forward, incorporating additional demographic data and addressing model limitations, such as through techniques like SMOTE for imbalance adjustment, can enhance predictive accuracy and cluster quality. By continuing to refine these analytical approaches, social media platforms can foster deeper user engagement, optimize content delivery, and maintain a competitive edge in a dynamic digital landscape.

BIBLIOGRAPHY

References

Bhadra, M. (2024, October). *Social Media Usage Dataset (Applications)*. Kaggle. Retrieved December 13, 2024, from <https://www.kaggle.com/datasets/bhadramohit/social-media-usage-datasetapplications>

Kaggle. (n.d.). Discover and share datasets. Retrieved December 13, 2024, from <https://www.kaggle.com>

List of Figures

Figure 1: The Dataset	5
Figure 2: Dataset missing values	7
Figure 3: Heatmap.....	7
Figure 4: EDA of Dataset	8
Figure 5: New Columns	9
Figure 6: Z score	10
Figure 7: Interquartile Range.....	11
Figure 8.1: Elbow Method for Optimal K.....	12
Figure 9: Bar Chart of K-Means of Daily Minutes Spent and Post Per Day.....	13
Figure 10: Logistic Regression for Predicting User Departure	14

List of Tables

Table 1: Table Showing both models in terms of their accuracy, robustness, speed, interpretability and scalability	16
--	----