EXAM BUSINESS INTELLIGENCE SEPTEMBER 2024

MARIA MARINELA CHEAPTANARIU

September 23, 2024

1 Abstract

In this project, a dataset was selected containing information on social media usage, the platforms used, and the emotions experienced by users while interacting with these platforms. The project can be divided into two main areas of analysis: the first part focuses primarily on preparing the dataset for proper analysis, followed by an Exploratory Data Analysis (EDA). The second part of the project attempts to apply predictive models to the data in order to determine whether it is possible to predict which platform users prefer based on their characteristics, or which emotions are likely to be triggered by the use of social media in general. Finally, clustering techniques are applied, which could be particularly useful for developing specific marketing strategies.

2 Introduction

In recent years, social media has become a vital tool for companies to target, segment, and influence consumers. The largest marketing campaigns are now carried out through social platforms, where companies analyze online behavior to optimize their business strategies. Beyond being a powerful tool for businesses, social media has also become a significant part of individuals' daily lives, impacting them to the extent that it can shape mood and overall well-being. In this analysis, the aim is to replicate the type of analysis a company might conduct to better target its consumers, and explore to what extent the time spent on social media affects mood, whether positively or negatively.

3 Dataset Description

The dataset contains 10 columns and 1,001 rows, with some missing values in certain columns (e.g., ”Gender”, ”Platform”, and several daily usage statistics). The dataset includes:

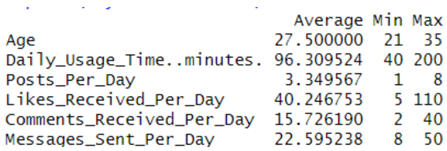
- **User_ID:** User identifier (object type).
- **Age:** Age of the user (object type, possibly some issues with data type).
- **Gender:** Gender of the user (object type, one missing value).
- **Platform:** Social media platform used (object type, one missing value).
- **Daily_Usage_Time (minutes):** Time spent on social media daily (numeric).
- **Posts_Per_Day:** Number of posts per day (numeric).
- **Likes_Received_Per_Day:** Likes received per day (numeric).
- **Comments_Received_Per_Day:** Comments received per day (numeric).

- **Messages_Sent_Per_Day:** Messages sent per day (numeric).
- **Dominant_Emotion:** Predominant emotion reported by the user (object type).

The first step is cleaning the dataset by removing all missing values. After performing a summary analysis, certain variables are converted from 'categorical' to 'numeric'. At this stage, the dataset, named 'clean_data', consists of 984 observations. Next, we analyze the outliers using the IQR method (based on quantiles). It is observed that many features do not have outliers, while others do (such as a high number of posts published or likes), but these are consistent across the same users. In my opinion, these should not be considered outliers, as they likely reflect the user's behavior. For example, they might have a public Instagram profile, engage more frequently on LinkedIn while actively job hunting, or be pursuing a career as a content creator or influencer. Since these numbers are not particularly unusual, it is deemed appropriate to retain them.

4 Exploratory Data Analysis (EDA)

The visualizations presented below illustrate the distributions of key variables in our dataset:



```
                           Average Min Max
Age                        27.500000  21  35
Daily_Usage_Time..minutes. 96.309524  40 200
Posts_Per_Day               3.349567   1   8
Likes_Received_Per_Day     40.246753   5 110
Comments_Received_Per_Day  15.726190   2  40
Messages_Sent_Per_Day      22.595238   8  50
```

Figure 1: Distribution of Key Variables

The analysis reveals that females are the most active users of social media, spending significantly more time on these platforms compared to males and non-binary individuals. They not only spend more time online but also engage more frequently by posting more often and receiving a higher number of likes. Interestingly, while males also report experiencing ”happiness” as a dominant emotion, their engagement levels appear to be lower than those of females. In contrast, non-binary individuals predominantly report a ”neutral” emotional state when using social media.

When examining platform usage, Instagram leads with the highest average time spent at 153 minutes per day. This is followed by Snapchat at 91 minutes and WhatsApp at 88.7 minutes. Conversely, LinkedIn shows the lowest mean usage time at 55.8 minutes, accompanied by a relatively low standard deviation of 6.46. This suggests that LinkedIn usage patterns are more consistent among its users.

The analysis of standard deviations across platforms indicates variability in user engagement. Snapchat exhibits the highest variability at 25.4, highlighting diverse engagement levels among its users. This variability suggests that while some users may be highly active, others may use the platform less frequently.

5 Principal Component Analysis (PCA)

The principal component analysis (PCA) was conducted on the numeric columns from the dataset, which include various metrics related to social media usage. The goal was to reduce dimensionality while preserving the variance in the data. The numeric columns selected for PCA include Daily Usage Time, Posts Per Day, Likes Received Per Day, Comments Received Per Day, and Messages Sent Per Day. The data was standardized to ensure that each variable contributed equally to the analysis, preventing any one variable from disproportionately influencing the results due to differing scales.

The PCA results indicate that Principal Component 1 (PC1) explains 92.85% of the variance in the dataset, while Principal Component 2 (PC2) explains only 2.81%. This substantial difference implies that PC1 captures the majority of the data's structure and trends. The cumulative proportion of variance explained by the first two components is 95.66%, indicating that these two components together effectively summarize the dataset's variability.

PC1, referred to as the "Social Media Usage Level," synthesizes key characteristics of social media engagement, such as time spent and interactions (likes, comments). This dimension provides a comprehensive view of user engagement. PC2, which could reflect the "Engagement Level," captures less variance, suggesting it may provide only marginally useful information for distinguishing between user behaviors.

6 Linear Regression

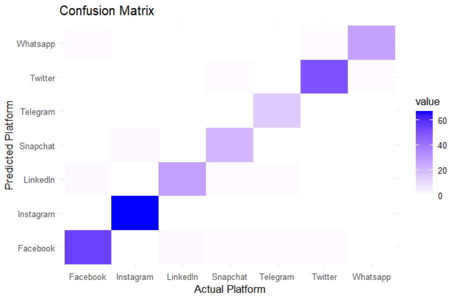
The linear regression model aims to predict daily social media usage time based on user interaction metrics such as posts per day, likes received, comments received, and messages sent per day. The intercept is 23.30 minutes, indicating a baseline usage, while posts per day (1.50 minutes) and messages sent per day (1.50 minutes) significantly increase usage time. Likes received per day (0.80 minutes) also contribute positively, but comments received per day (p = 0.437) do not show a significant effect. With a Multiple R-squared of 0.9096, the model explains approximately 91% of the variability in usage time, reflecting a strong fit, further supported by an Adjusted R-squared of 0.9092 and a low F-statistic p-value (¡ 2.2e-16). However, the Mean Squared Error (MSE) of 138.97 suggests some deviations between predicted and actual values, and a subsequent R-squared of 0.0904 highlights poor predictive power. Visualizations indicate mispredictions, especially at higher usage levels, suggesting that while the model captures substantial variance, it has limitations, particularly with comments. Future improvements could involve refining the model or exploring nonlinear relationships to enhance predictive accuracy.

6.1 Second attempt to do Linear Regression

The multinomial regression model attempts to predict emotional states using PCA1 and PCA2 as predictors. The model converged after 20 iterations, but the results suggest challenges in interpretation. While PCA2 consistently showed a negative influence on all emotions—indicating that increases in PCA2 correlate with decreased probabilities of specific emotions—its overall quality is questionable, undermining the model's validity. The coefficients for different emotions reveal varied effects, but the model fails notably in predicting "Sadness" and "Anxiety," which is evident in the confusion matrix results. The accuracy of the model stands at a low 47.62

7 Random Forest

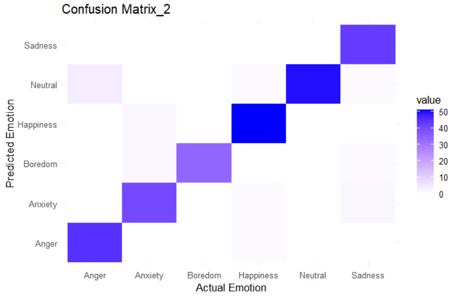
The Random Forest model demonstrates strong performance in predicting social media platforms based on user interaction features, achieving an accuracy of 94.6%. Using a dataset split into 70% training and 30% testing, the model effectively classified most platforms, as shown in the confusion matrix. The model's sensitivity and specificity rates across different classes were notably high, with the Kappa statistic at 0.9347, indicating strong agreement between predicted and actual classifications. However, while the model excels overall, examining the variable importance suggests that certain features play more significant roles in predictions, which could provide insights for further investigation. The visualization of the confusion matrix reinforces these results, illustrating the model's capacity to distinguish between platforms accurately.



7.1 Second Attempt of random Forest

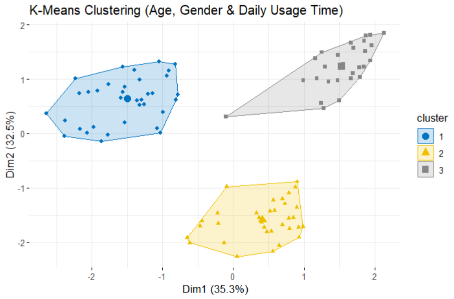
The Random Forest model predicts emotional states using features from principal component analysis (PCA) and participant age. The dataset was split into a training set (70%) and a testing set (30%), achieving an of 94.6%. The confusion matrix indicates robust performance, with 45 instances of Anger predicted correctly and 40 instances of Anxiety identified, but with two misclassifications. The model achieved 100% sensitivity for Boredom and Sadness.

With a Kappa statistic of 0.935, the model demonstrates a strong agreement between predicted and actual values. The balanced accuracy rates exceed 95% for most classes, showcasing reliable emotion differentiation. Visualizing the confusion matrix as a heatmap further clarifies prediction accuracy.

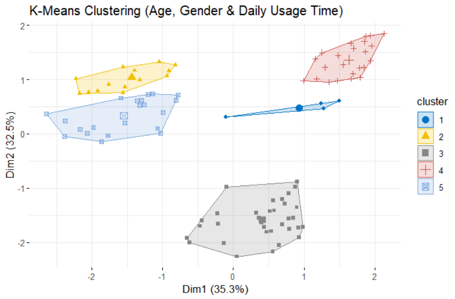


8 Clustering

The k-means clustering analysis identified three distinct clusters based on the following features: Age, Gender and Daily usage time. Cluster 1 (average normalized age 0.453) demonstrates high daily usage time (0.452), posting frequency (4.06 posts/day), and engagement metrics with an average of 51.6 likes and 18.8 comments received per day, suggesting an active and engaged user group. Cluster 2 (average age 0.538) features moderate daily usage (0.343) and lower engagement levels, with 3.39 posts, 40.7 likes, and 16.6 comments, indicating a less active but potentially more reflective user base. Cluster 3 (average age 0.381) shows the youngest demographic with the lowest daily usage time (0.226) and minimal engagement (2.32 posts, 23.9 likes, 10.3 comments), highlighting a group that may not be as invested in social media activity. Overall, these clusters reveal varying levels of engagement and demographic characteristics, which can inform targeted strategies for content delivery and user interaction.



To deepen the cluster analysis, we experimented with increasing the number of clusters from 3 to 5. This adjustment proved feasible, leading to the identification of new clusters with even more specific age ranges and distinct characteristics. This allows for a more detailed segmentation of users.



Cluster 1 features older users who are very active on social media. They demonstrate high posting frequency and receive substantial likes and comments, indicating strong social interaction.
Cluster 2 consists of very young users with the highest activity level. They frequently post and garner the most likes per post, showcasing effective content that resonates well with their peers.
Cluster 3 represents moderately active users across various ages. Their average posting and interaction metrics suggest they may have a smaller following or less impactful content.
Cluster 4 includes low-activity users who are likely casual social media consumers. With minimal posts and engagement metrics, they prefer passive content consumption.
Cluster 5 comprises moderately active users who are slightly older. They have high levels of interaction with good likes and comments, indicating active social media engagement for personal or professional reasons.

9 Conclusions

This report reveals key insights into social media usage and user emotions across different platforms. The analysis identified significant differences in engagement behaviors, with clustering techniques highlighting distinct user segments. Although decision tree methods were attempted, the data complexity led to overfitting and ineffective predictions. Overall, the findings emphasize the diverse interaction patterns among users, providing valuable data-driven insights for optimizing marketing strategies.