Investigating the Impact of Social Media Usage on Attention Span and Daily Routines

Nithya Konduru
Raashi Maheshwari
Tanvi Yamarthy

Professor Chaturvedi
Department of Computer Science, Rutgers University

Introduction

This project aims to explore whether high levels of social media usage are associated with reduced attention spans and disruptions in the lives of daily users. The specific question that we aim to answer is: Does increased engagement with social media, measured with variables like number of platforms used, screen time, and daily behavioral patterns, correlate with self-reported distraction and attention span challenges? In an attempt to answer this question, we combined a variety of techniques, including clustering methods, exploratory data analysis (EDA), and simple regression techniques. These approaches are taken from the following topics learned in class: data preprocessing and visualizations (Weeks 2 - 4), regression modeling (Week 5), and unsupervised learning and clustering (Week 6). Using these techniques will allow us to derive meaningful patterns from the data and assess whether variables from the dataset ("scrolling first thing in the morning", "number of platforms used", etc.) contribute to digital distraction and a lower attention span. This project will combine both practical and theoretical applications of what we have learned through homework labs and lectures.

Motivation

The motivation behind this project comes from a concern about how social media usage impacts users' focus, routines, and overall well-being. As students, we often get distracted from scrolling on social media platforms, which impacts our daily life behaviors. We noticed many students at the library pick up their phones and mindlessly scroll rather than finishing up their work or enjoying the beautiful weather outside. This study is important because it centers the user's perspective, which is often overlooked by other studies. We are excited to explore this issue by taking into account how user-reported habits and distractions reveal behavioral trends. We made sure to account for independent predictors of user distraction as well as multiple predictors together. There are existing studies on this topic that highlight the addictive nature of

social media, but rare studies use real user self-reported data to explore this behaviorally. By doing this project, we are attempting to understand if users are aware of social media's impact, directly or indirectly, on their daily lives.

Methods

Our dataset, "Social Media Usage and User Behavior," was taken from Kaggle and contains responses from 310 users. It is structured as a tabular CSV file consisting of 19 columns that have both demographic data and behavioral patterns. The dataset includes perceptual awareness measures that span categorical and numerical data types. We began with data-cleaning and pre-processing steps, such as standardizing formats and encoding ordinal features. For unsupervised analysis, we conducted exploratory data analysis through bar plots, box plots, and platform engagement. Next, we applied clustering algorithms like K-Means to a random forest model to group users based on social media engagement. Using K-Means clustering enhanced our random forest model, lowering our MSE value. Random Forest Regression works by creating many decision trees from the dataset and then combining those trees into a singular model, which handles non-linear patterns better than linear patterns. We observed non-linear patterns in our dataset, so this random forest model is ideal. K-means clustering groups similar data points together, allowing us to find hidden patterns. To figure out the ideal number of clusters, we used the elbow method, which essentially plots how much. Within-Cluster Sum of Squares (WCSS) decreases as the number of clusters increases. These methods help us evaluate how social media usage patterns align with user-reported impacts on focus and daily life. Further, to refine our analysis on an even more detailed level, we applied regression methods.

Since we previously cleaned out our data in the EDA and pre-processing steps, applying regression methods such as simple regression and multi-regression, allowed us to draw valuable insights from our data. For simple regression, we chose one predictor to see what the MSE and r^2 value would look like. The predictor that we chose was "What is

the total number of social media platforms you use?". The reason that we ultimately chose this as our predictor variable was because it is an easily quantifiable and fairly reasonable predictor to gauge social media engagement, and thus relevant to our hypothesis regarding distraction levels. We will discuss the results of this predictor in the next section. In addition to using simple regression, we also employed multiple regression. With multi-regression, we were allowed to use several predictor values and see how these combined could impact the distraction levels of users. The predictor values that we used for multiple regression include the number of hours spent, whether social media scrolling is the first thing a user does in the morning, if they have a premium subscription, as well as a crucial demographic: their age group. For the age group, we used an important technique that we learned in class: One Hot Encoding. We used the in-built Linear Regression method and .predict method, using our training and testing data to receive the best possible, unbiased outcome. Lastly, we used a coefficient table to model all the values that we received via our linear regression methods. In addition to this, for each of our predictors and a depiction of our simple vs multiple regression r^2 values, we created visualizations to better help interpret our results. We will discuss these outcomes in the next section.

Results

The first step we took was to run exploratory data analysis on the dataset, which yielded compelling evidence supporting our hypothesis that increased social media usage correlates with reduced attention spans and disrupted daily routines. Bar plots comparing occupation with the number of social media platforms used revealed that students, particularly aged 18 - 24, tend to use more platforms, aligning with the age group most likely to report using social media first thing in the morning. This behavior was further confirmed through a stacked bar chart showing that younger users had the highest percentage of morning scrolling habits. This indicates early routine disruption. A boxplot examining the relationship between hours spent online and the number of platforms used demonstrated that individuals engaging with more platforms tend to

spend more time online daily, which may limit time for other focused tasks. Crucially, graphs correlating distraction frequency with platform count revealed that users reporting frequent distraction while working or studying were more likely to be high-platform users. The correlation heatmap also showed a modest but clear positive relationship between platform usage, distraction frequency, and morning usage behavior. These findings collectively support our research question by illustrating how social media behaviors, especially high engagement across multiple platforms, are associated with attention-related difficulty and altered daily routines. No evidence from the data directly contradicted our hypothesis, though further modeling and clustering will be valuable for identifying more nuanced user segments.

```
 Simple Regression
R² Value: -0.02252746754974866
MSE Value: 0.68336968369805

 Multiple Regression
R² Value: 0.0890879611481582
MSE Value: 0.6087755015115456

Coefficients Table:
                          Feature  Coefficient
0          num_platforms_numeric     0.101415
1                    hours_score     0.132178
2              first_thing_score     0.306697
3                  premium_score    -0.105862
4  what is your age group?_18-24     0.048046
5  what is your age group?_25-34    -0.244604
6  what is your age group?_35-44    -1.116678
7  what is your age group?_45-54    -0.364957
8    what is your age group?_55+    -0.336517
```

One crucial step of our process was applying linear regression (both single and multiple regression) on our cleaned data, and seeing what the regression told us about different predictors and their relation to distraction level. These predictors, such as number of platforms, hours spent, age group, and morning scrolling, served as independent variables for a dependent variable (testing variable) that we coined called “distraction score”. To the right is our resulting coefficient table that our regression methods yielded. Now, let us go through each one of these predictors.

Straight off the bat, our r^2 value for simple regression was a weaker indicator than our r^2 value for multiple regression. What this tells us is that our single value predictor, “number of platforms used,” is not solely enough to predict how distracted a user is. The MSE Value is 0.68 for simple regression and 0.60 for multiple regression, indicating that our model has low errors. Now let us understand the r^2 value for single regression,

because it's in the negative values, (-0.0225), which indicates to use that the number of platforms does not explain the variance in distraction. It is not a strong enough indicator to communicate any significant information. If we look at the multiple regression, where we took in 4 different predictor values, we got an r^2 value of 0.089. This is a significant improvement from our simple regression's r^2 value. It shows us that our model explains 8.9% of the variance in distraction, which is fairly responsible for behavioral human data. Let us look through each of the predictor values in the coefficient table. We see that the number of platforms has a positive value of 0.101, indicating that more platforms are positively/ slightly correlated with a greater distraction tendency. The same appears for the hours spent. Because of the positive +0.132 value, we can conclude that more screen time is also associated with greater distraction levels. Another great insight from our data is seen by the +0.307 in the first_thing row. This shows that users who use social media first thing in the morning report greater distraction levels in their behavior. This is by far our strongest predictor. And our last positive value in our coefficient table is the +0.048 seen by the age group of 18-24 years. This shows us that individuals in this age range are slightly more susceptible to being distracted due to social media compared to the baseline, but are fairly more distracted compared to other age groups. The rest of the predictors yield negative values, showing us that with the data we have right now, there is not much we can say about their correlation to distraction levels. There is no apparent, significant relationship. However, one thing we can seem to notice is that older age groups appear to have less significant distraction levels compared to younger age groups.

Next, we performed a machine learning pipeline that combines KMeans clustering and a random forest model to predict user distraction based on social media behavior. Then, we analyzed which features proved the be the most important predictors. We start by normalizing the data on a scale from 0 to 1 because KMeans is sensitive to feature scale. Normalization ensures that each feature is proportionally evaluated. Next, we used the Elbow Method to find the optimal number of clusters to split our data into. Our code takes an iterative approach to doing this for 1-10 clusters and then collects the WCSS

(within-cluster sum of squares), which measures the compactness of the clusters formed by a clustering algorithm like KMeans. It quantifies how close data points in a cluster are to the center of that cluster, so the lower the WCSS, the more compact the clusters are, which is better. The elbow point of the graph is the point at which adding more clusters does not drastically improve the model, indicating the best number of clusters. We found the optimal number of clusters to be 3 because any number up to 3 significantly reduces the within-cluster variation, and anything above 3 presents no change. Finally, we performed random forest regression with clustering. This allows us to predict user distraction using the original features and cluster labels. Combining cluster patterns allows us to capture complex user behaviors more accurately.

```
Random Forest Regression with KMeans Cluster Feature
R² Score: 0.06404445450639429
MSE: 0.6255124340200501

Feature Importances:
                      Feature  Importance
                  hours_score    0.289856
        num_platforms_numeric    0.195523
            first_thing_score    0.113403
                      cluster    0.111753
what is your age group?_18-24    0.108990
what is your age group?_25-34    0.053001
                premium_score    0.047795
  what is your age group?_55+    0.038326
what is your age group?_35-44    0.028312
what is your age group?_45-54    0.013043
```

The random forest model output demonstrates an R^2 score of 0.064, which is 6.4%, representing 6.4% of the variance in distraction scores. We did not get a very high R^2 value, meaning the model explains only a small portion of the variance in user distraction. However, we were still able to deduce some valuable insights, especially based on the feature importance values. The MSE is 0.625, which indicates the average squared error between the predicted and actual distraction scores. When this is paired with the r^2 value, it reinforces that the model has a limited predictive accuracy, but it is not entirely uninformative. The model identifies well the relative influence of behavioral features like time spent online and the number of platforms used. Compared to simpler methods like linear regression, random forest regression has non-linear modeling

capabilities, and can capture the relationship between different features as well us provide us with Feature importance insights.

The feature importance is the relative importance of each predictor in the Random Forest model. Here, we can see the individual impact of each feature/ predictor on user distraction. This bar plot shows that hours_score is the most important predictor, having the highest importance score of 28.9%. This tells us that the more time someone spends online daily, the more likely they are to report higher distraction scores. Num_platforms_numeric has the next highest importance score, indicating that users who are active on more platforms are also more prone to distraction, which confirms our findings from earlier plots. Next, we see that those who begin their day with social media tend to have higher distraction levels. The K-Means cluster assignment is almost as predictive as individual behaviors, confirming that behavioral groups significantly affect distraction. It also confirms that adding the K-Means clusters successfully enhanced our model, meaning there are multidimensional usage patterns, just as our multi regression model supported. Lastly, we see that, based on the age_group variable, younger users are more strongly associated with higher distraction scores.

Final Discussion

Based on all the data that we yielded from our regression analysis, some main takeaways that we can make are that the user's social media habits (like morning scrolling) appear to have a greater impact on distraction levels compared to quantitative measures like the number of platforms a user is on. In addition to this, age is an important measure in determining distraction levels, as we see varying coefficient values for the different age groups. After running our code, you can see the different visualizations that further depict the regression analysis above.

Github Project Link: http://github.com/sk2539/FinalDSProject
Demo Video Link: https://youtu.be/jA82Y6Y-oXc