Analyzing the Impact of User Metrics on Social Media Addiction

AMS 595: Foundations of Computing
Professor Chenyu You
Group Members: Valentina Tillmann, Anthony Degennaro , Ankita Mohanty , Chaeeun Shin

Introduction

Over the past two decades, social media platforms have become deeply embedded in people's daily lives, with their influence growing at an unprecedented rate. Recent studies estimate that approximately 10% of Americans are addicted to social media. This ever-increasing dependency has sparked concern among experts, particularly regarding its potential impact on the developing brains of young people. For instance, in response to such problems, Australia recently implemented a ban on social media use for children under 16.

The objective of this investigation is twofold: first, to conduct exploratory data analysis (EDA) to uncover insights about the relationship between user characteristics and levels of social media addiction; and second, to evaluate the effectiveness of two predictive modeling approaches to determine which is better suited for this dataset.

The data for this project was sourced from Kaggle.com and was generated using advanced synthetic techniques. While artificially created, the dataset was designed to closely mimic real-world social media usage trends. We selected this dataset for its rich and realistic representation of user behavior and its suitability for analysis.

For further understanding, here is the link to the data:
https://www.kaggle.com/datasets/zeesolver/dark-web

Exploratory Data Analysis (Explain data cleaning/preparation, various graphs, trends in the data, etc.)

The data contains 1000 observations with 34 variables. No missing value is observed. A new variable, Generation, is created based on the variable 'age'. The participants under 28 are considered as Generation Z, between 28 and 44 as Millennials, between 45 and 60 as Generation X, and above as Boomers. Daily watching time is converted into an integer and a new variable, Watch Time of Day, is generated based on the converted data. Watching before noon is considered as morning, before 6 p.m. is afternoon, before midnight as evening, and after midnight as night. Also, the typo in the location column is fixed where the value 'Barzil' is corrected to 'Brazil' to ensure consistency and accuracy in location-related analysis.

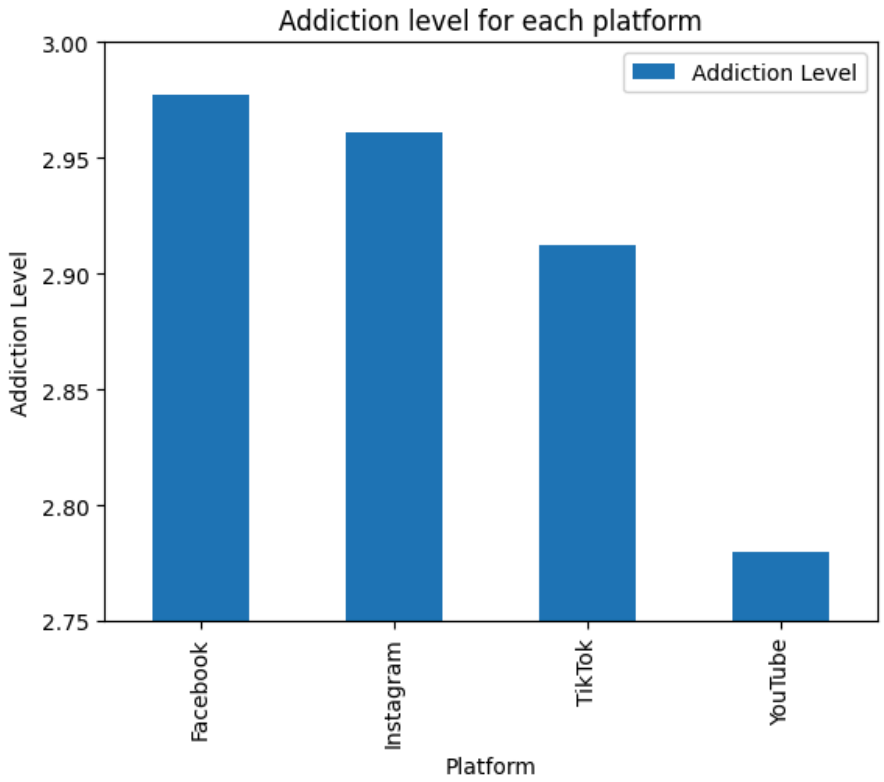

Figure 1 Addiction level vs different platforms

The plot displayed above in figure 1 describes the mean addiction level by each platform. Addiction level does not vary greatly by platform. However, it is observed that Facebook users have the highest mean addiction level.

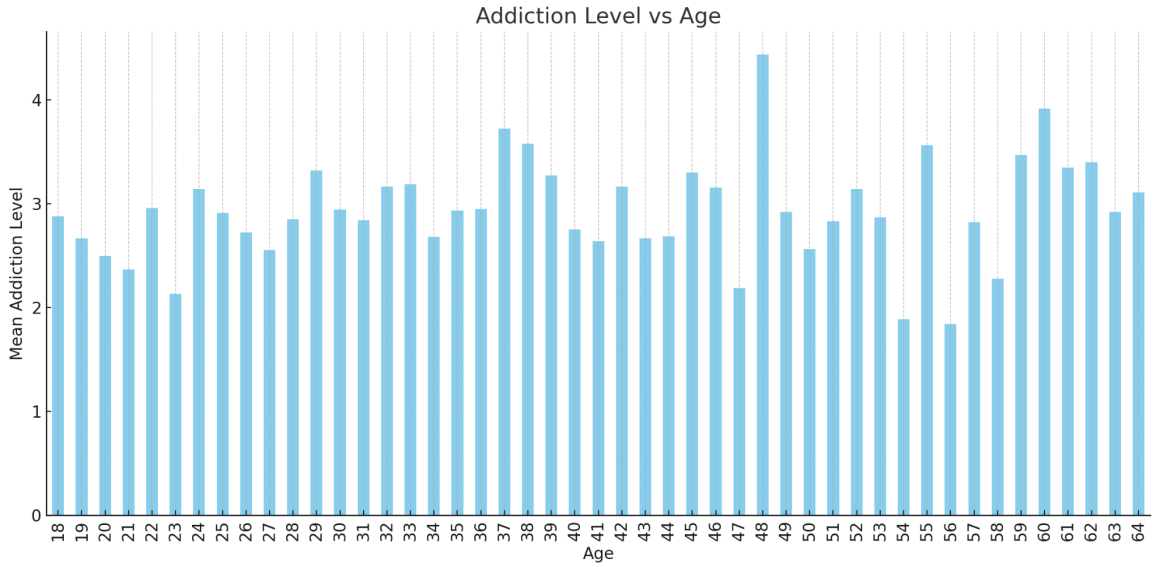

Figure 2 Addiction level vs different ages

The bar plot in figure 2 depicts the mean addiction level across different ages, showcasing a fluctuating trend. While the mean addiction level generally remains between 2 and 4, certain age groups, such as individuals in their late 40s and mid-50s, exhibit higher addiction levels, peaking above 4. Conversely, younger age groups, particularly in their early 20s, tend to have relatively lower addiction levels. These variations highlight the potential influence of age-specific behaviors or external factors on social media addiction. This analysis underscores the need for further investigation into demographic and contextual elements to better understand these trends.

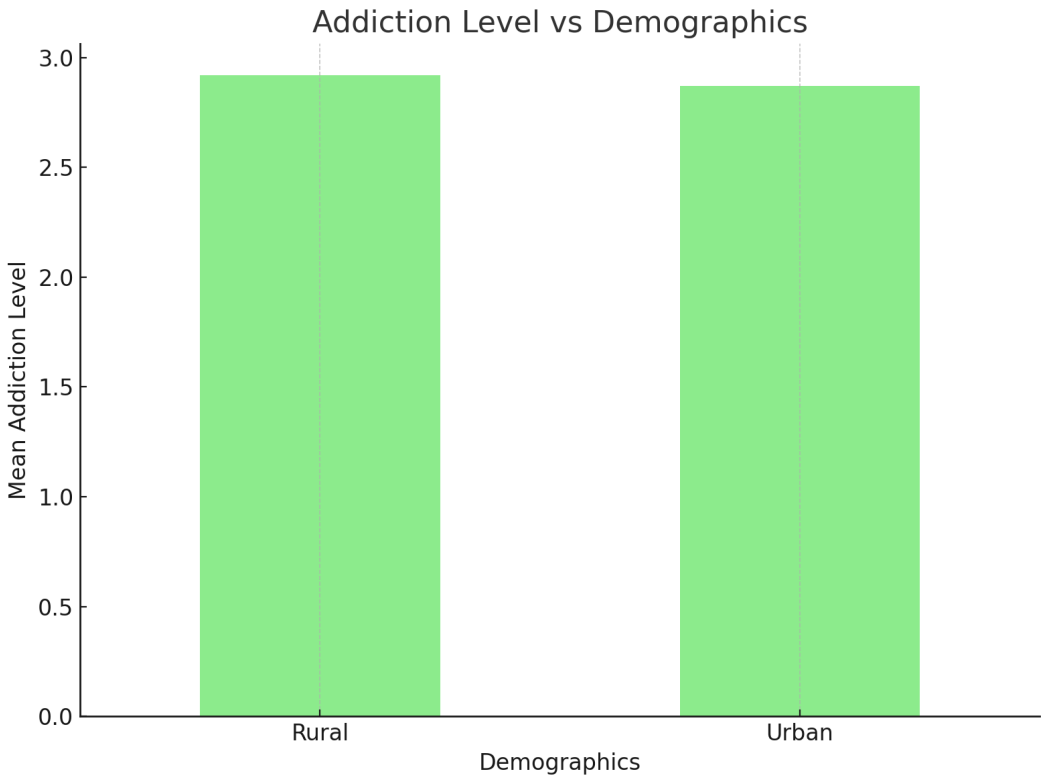

Figure 3 Addiction level vs different demographic regions

The bar plot shown in figure 3 highlights the comparison of mean addiction levels between rural and urban populations. The results indicate that both rural and urban demographics exhibit similar mean addiction levels, with no significant difference between the two groups. This suggests that demographic location, whether rural or urban, may not be a major determinant of addiction levels, indicating that other factors, such as age or platform preferences, could play a more significant role.

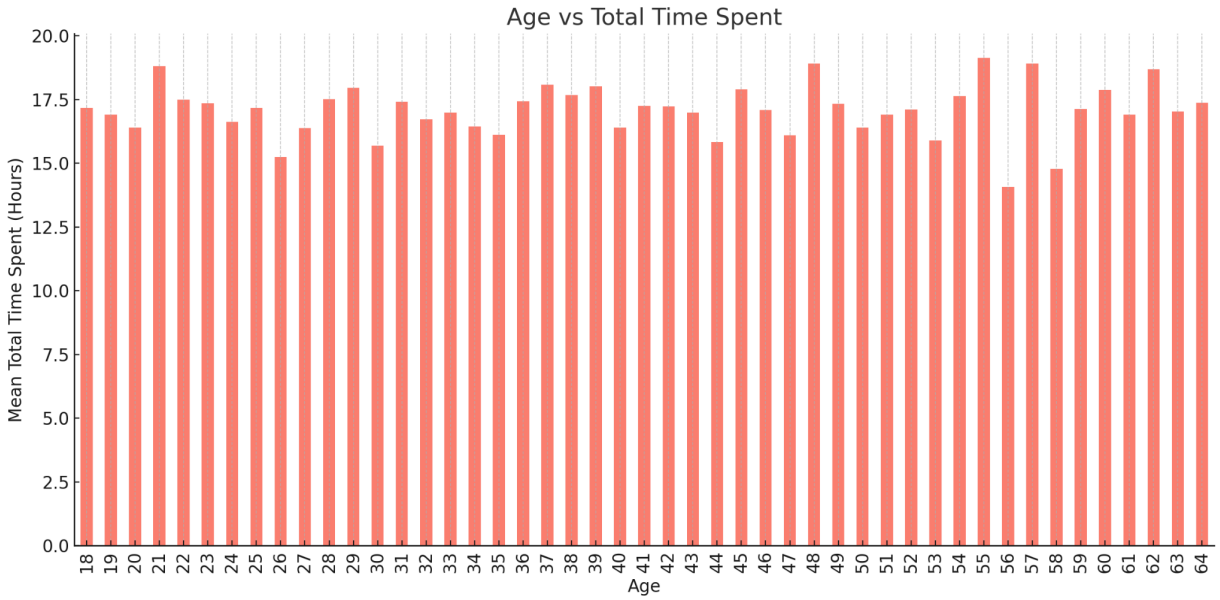

Figure 4 Total time spent vs Ages

The bar plot in figure 4 highlights the average time spent on social media across different age groups. The data reveals a relatively consistent pattern, with most age groups spending between 16 and 20 hours on social media. Notably, age groups in their early 20s and late 50s exhibit slightly higher engagement, peaking near 20 hours. This consistent engagement across all ages underscores the widespread use of social media, suggesting that factors beyond age, such as platform preferences or usage purposes, may influence the total time spent.

Multinomial Logistic Regression (explain how it works and results)

Multinomial logistic regression is a statistical modeling technique used to predict the probabilities of multiple categorical outcomes for a target variable based on several independent variables. In our analysis, the independent variables included 'Location,' 'Gender,' 'Profession,' 'Age,' 'Platform,' 'Operating System,' 'Watch Time,' and 'Reason for Watching.' The target variable was 'Addiction Level.'

An important step in training the model was creating dummy variables for each independent variable using pd.get_dummies(). This process was necessary because the independent variables were categorical, and machine learning models require numerical inputs. By transforming these features, we ensured the model could interpret the data correctly.

Leveraging these user characteristics, our goal was to predict the most likely social media addiction level for each individual. The data was split into 80% training and 20% testing, a standard practice in logistic regression, to evaluate the model's performance on unseen data. Model Structure:

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

We start by defining the softmax function, which transforms raw model outputs into probabilities by ensuring all values are nonnegative and sum to 1. Next, we define the cost function, which calculates the negative log-likelihood of the predicted probabilities, measuring how well the model aligns with the true labels.

To optimize the cost function, we use scipy.optimize.minimize, providing the gradient of the cost function to enhance the convergence speed. Once the model is trained, we use the softmax function to make predictions, allowing us to identify the most likely class for each instance.

The training and test accuracies achieved are 87% and 55%, respectively. These results indicate that the model performs well on the training data but struggles to generalize to unseen data, suggesting potential overfitting or the need for additional feature engineering, regularization, or a larger dataset.

Xgboost

XGBoost (eXtreme Gradient Boosting) is a machine learning algorithm that is meant for speed and accuracy in supervised learning. It is built on gradient boosting, where decision trees are sequentially added to minimize the loss function. Each tree that gets made tries to correct the errors of the previous trees which improves the model incrementally. XGBoost includes both lasso and ridge to control the model complexity and helps prevent overfitting. Learning rate controls the step size in updating the weights, a small learning rate allows for gradual improvements but also requires more iterations which can be computationally expensive. It has a tree pruning algorithm which helps to prevent unnecessary splits in the model. XGBoost handles missing data well which makes fitting a model on incomplete datasets easy.

XGBoost is considered one of the most powerful algorithms for tabular data. It has the ability to produce extremely accurate results with minimal feature engineering and handles complex datasets with ease.

For the XGBoost model we made for our data, we got an accuracy of 99.50%. This tells us that the model is very reliable for predicting addiction levels for this given dataset.

The classification reports give a breakdown of the model's performance for each class of the target variable (addiction level). Precision tells us how many of the predicted instances were correct, all of them were 1 besides at addiction level of 2 which was 0.98. A high precision tells us that the model does not make false positive predictions. Recall tells us how many of the actual instances were correctly predicted. We had a recall of 1 for everything besides addiction level of 3 which was 0.97. F1-score is the harmonic mean of precision and recall. This combines them into a single number to evaluate the overall performance of the model, all were 1 but addiction levels 2 and 3 were 0.99 and 0.98 respectively. The image below shows how we get all three metrics in terms of formulas and below is the output from the model.

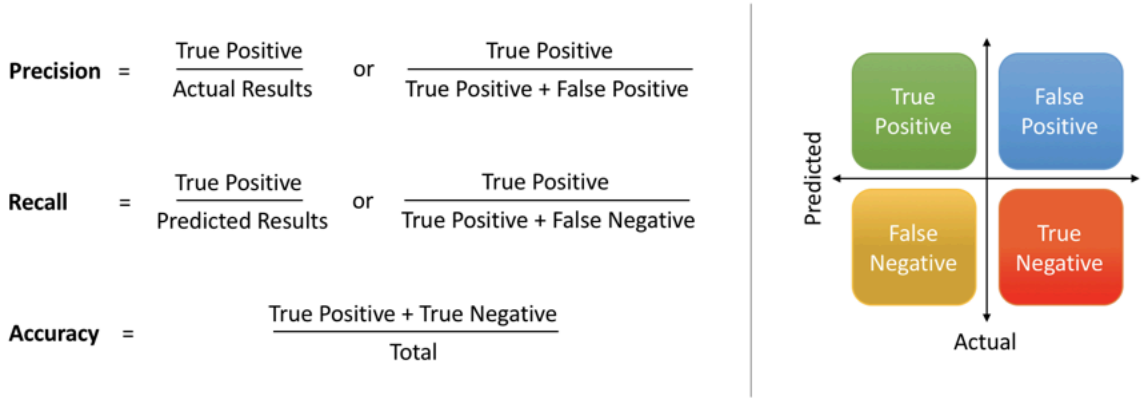

Figure 5 Evaluation metrics for classification models: Precision, recall, and accuracy formulas, along with a confusion matrix representation of prediction outcomes.(Source)

Table 1 Classification report with 99.50% model accuracy: Precision, recall, and F1-scores for each class

Model Accuracy: 99.50%

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	37
1	1.00	1.00	1.00	8
2	0.98	1.00	0.99	41
3	1.00	0.97	0.98	33
4	1.00	1.00	1.00	6
5	1.00	1.00	1.00	47
6	1.00	1.00	1.00	10
7	1.00	1.00	1.00	18

Next we investigated the feature importance of the model. ProductivityLoss (618) and Self Control (186) were the most important features. This shows us that these variables have the strongest relationship with addiction levels. Feature importance helps with interpretability since it can illustrate which predictors matter the most and by how much.

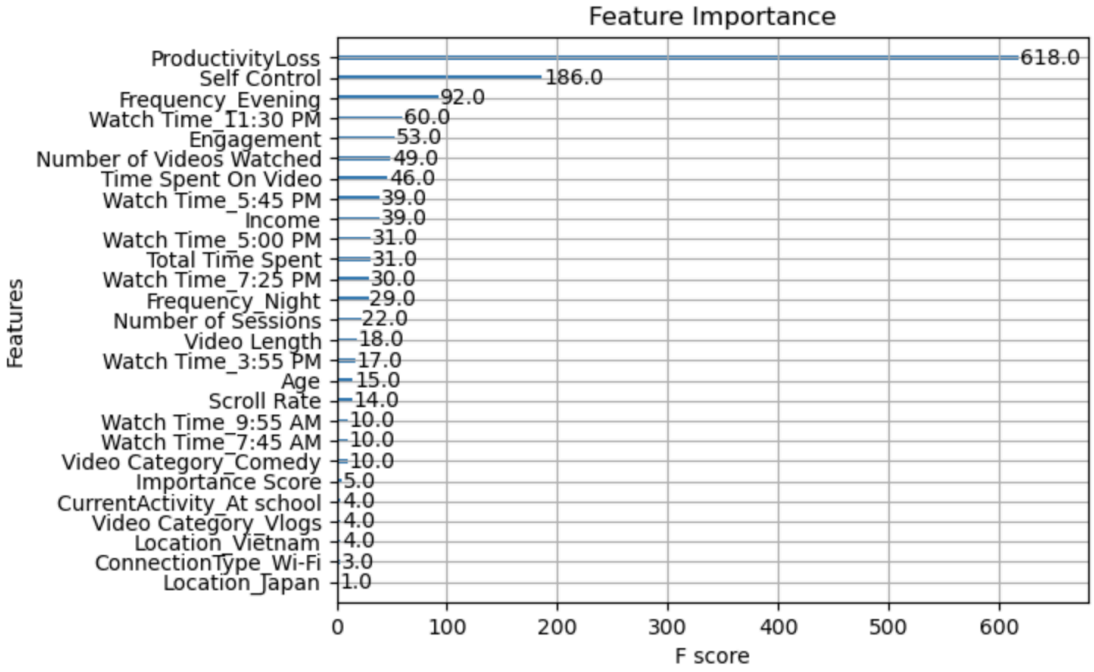

Figure 6 Feature importance plot showcasing the most influential predictors of social media addiction levels

Finally, we implemented a test case prediction. This allows us to import values for the features and the model will estimate what the addiction level would be. We are able to test if the model makes sense and it allows for practical applications in terms of prediction.

Conclusion

This study effectively analyzed social media addiction through a combination of exploratory data analysis (EDA) and predictive modeling, offering valuable insights into user behaviors and characteristics. Using a synthetic dataset that closely mimics real-world patterns, the analysis revealed key trends, such as fluctuating addiction levels across age groups and consistent time spent on social media irrespective of demographic factors like rural or urban settings. Middle-aged individuals exhibited higher addiction levels, while younger and older users showed slightly higher engagement times, underscoring the varied dynamics of social media use.

The predictive modeling efforts provided contrasting outcomes. Multinomial logistic regression achieved reasonable training accuracy (87%) but struggled to generalize with only 55% testing accuracy, highlighting potential overfitting and the need for enhanced feature engineering or regularization techniques. In contrast, the XGBoost model performed exceptionally well, achieving a testing accuracy of 99.5%. This model's robustness and ability to handle complex datasets make it a powerful tool for predicting addiction levels. Feature importance analysis identified "Productivity Loss" and "Self Control" as the most impactful predictors, shedding light on the behavioral attributes most associated with higher addiction levels.

The EDA further emphasized the significance of user-specific attributes. While demographic factors, such as rural versus urban settings, did not show substantial variation in addiction levels, age and platform preferences revealed nuanced patterns. These findings indicate that while certain variables play a minor role, others, particularly behavioral metrics, hold significant predictive value. The integration of categorical and continuous variables in modeling provided a comprehensive view, enhancing the interpretability of results.

In conclusion, this study underscores the multifaceted nature of social media addiction, driven by behavioral tendencies, technological influences, and demographic nuances. The exceptional performance of the XGBoost model highlights the potential for advanced predictive tools in understanding and mitigating social media addiction. Future research can focus on improving logistic regression models, incorporating additional behavioral metrics, and validating these insights through experimental or longitudinal studies. These steps will pave the way for more targeted interventions and a deeper understanding of this critical social phenomenon.

References

1. **https://www.kaggle.com/datasets/zeesolver/dark-web**
2. **https://medium.com/@shrutisaxena0617/precision-vs-recall-386cf9f89488**