PhoneSense: A Comprehensive Predictive Analysis of Smartphone Addiction Using High-Dimensional Behavioural Metrics and Ensemble Machine Learning

Abhijeet Singh - 1207508  
Kirtan Sukhadiya - 1222983

***Abstract*—In the contemporary digital epoch, characterized by the generation of over 1.13 trillion megabytes of data daily and the proliferation of ubiquitous mobile computing, smartphone addiction has emerged as a critical behavioural health challenge.1 This report presents "PhoneSense," a rigorous data science initiative designed to quantify and predict smartphone addiction levels through the application of advanced machine learning classifiers. Grounded in a dataset comprising demographic, behavioural, and psychometric features from a diverse user base, this study employs a complete data science pipeline—from high-performance data munging to complex predictive modelling—adhering to IEEE standards.1 We leverage Exploratory Data Analysis (EDA) to uncover significant correlations, such as the strong positive relationship between daily usage hours (r=0.601$) and addiction levels, and the negative correlation with sleep duration (r=-0.217$).1 Three distinct classification algorithms—Logistic Regression, Random Forest, and XGBoost—were trained and evaluated. The experimental results demonstrate that XGBoost achieves the highest predictive accuracy at 95.58%, effectively managing the bias-variance tradeoff inherent in high-dimensional behavioural data.1 This report provides an exhaustive theoretical and practical treatment of the project, detailing the statistical foundations, mathematical frameworks of the models, and the sociotechnical implications of the findings.**

# Introduction

# The Data Science Paradigm in the Age of Information Deluge

The 21st century is defined by an exponential explosion in data generation, a phenomenon that has fundamentally altered the landscape of scientific inquiry and technological development. Current global metrics indicate that by the end of 2021, Google was processing approximately 2 trillion searches annually, while global IP traffic had swelled to nearly 278,108 petabytes per month. The velocity of this data creation is equally staggering: every minute, Instagram users post nearly 347,222 stories, and WhatsApp users exchange over 41 million messages. This massive influx of information, totalling approximately 1.134 trillion megabytes of data created daily, necessitates a shift from traditional computational approaches to Data Science.

Data Science, unlike traditional software engineering, is not merely about constructing algorithms to process inputs; it is the discipline of "learning about the world from data using computation". It incorporates elements of exploratory data analysis, visualization, machine learning, and high-performance computing to extract actionable insights from raw noise. As Drew Conway’s Venn diagram suggests, a competent data scientist must navigate the intersection of hacking skills, math and statistics knowledge, and substantive expertise. The PhoneSense project is situated firmly within this paradigm, utilizing data-driven inference to address a messy, real-world problem rather than attempting to construct a clean, deterministic virtual world.

# The Problem Domain: Smartphone Addiction

Amidst this digital expansion, the smartphone has become the primary interface for human-computer interaction. However, this ubiquity has precipitated a rise in smartphone addiction—a behavioural dependency characterized by excessive usage, compulsive checking, and varying degrees of anxiety or depression associated with device separation. The "PhoneSense" project addresses this issue by framing it as a predictive classification problem. By analyzing structured data regarding user behaviour—ranging from daily screen time and app usage counts to self-reported psychological states—we aim to infer the likelihood of high-risk dependency.

This problem is particularly suited to data science because behavioural health data is inherently "noisy" and complex. As noted in the course literature, "Real Scientists" (including data scientists) strive to understand the complicated and messy natural world, acknowledging that data has errors and that nothing is ever completely true or false, unlike the binary absolutes of traditional computer science. PhoneSense embraces this complexity, employing statistical learning to quantify the uncertainty surrounding addiction diagnoses.

# Project Objectives and Scope

The primary objective of this research is to develop a robust, high-accuracy predictive model for smartphone addiction. The scope of the project encompasses the entire data science lifecycle:

1. Data Acquisition and Munging: To utilize high-performance tools to load, clean, transform, and reshape raw survey data into a format suitable for algorithmic analysis.
2. Exploratory Data Analysis (EDA): To perform descriptive analysis and visualization, identifying patterns and quantifying the strength of relationships between features using statistical measures.
3. Predictive Modelling: To implement and optimize machine learning classifiers—specifically Logistic Regression, Random Forest, and XGBoost—to predict addiction levels.
4. Model Interpretation: To move beyond accuracy scores and interpret the models’ decision boundaries, quantifying uncertainty through confusion matrices and analyzing feature importance.

# 2. Theoretical Framework and Statistical Foundations

## A rigorous data science project must be grounded in sound statistical theory. As noted by Josh Blumenstock, a data scientist is "someone who knows more statistics than a computer scientist and more computer science than a statistician". This section outlines the mathematical concepts that underpin the methodologies applied in PhoneSense.

## A. Probability Theory and Inference

The foundation of our predictive modelling is probability theory, which provides the formal framework for reasoning about the likelihood of events. In PhoneSense, we treat "Addiction Level" as a random variable dependent on a vector of independent features (usage hours, anxiety, etc.).

The fundamental goal is to estimate the conditional probability *P(A|B)*, where *A* is the event of being addicted, and *B* represents the observed behavioural data. This is formalized by Bayes' Theorem, a cornerstone of statistical inference:

Where:

* *P(A|B)* is the Posterior: The probability that a user is addicted given their specific usage pattern. This is the quantity our models attempt to approximate.
* *P(B|A)* is the Likelihood: The probability of observing particular usage patterns (e.g., 10 hours of screen time) given that a user is addicted.
* *P(A)* is the Prior: The baseline probability of addiction in the population before observing any specific data.
* *P(B)* is the Marginal Likelihood: The total probability of observing the data patterns across all users.

This Bayesian perspective reminds us that our models are updating their beliefs about a user's status as more evidence (features) is presented.

## B. Statistical Distributions

Understanding the distribution of data is a prerequisite for valid analysis. Random variables are described by Probability Density Functions (PDFs) and Cumulative Distribution Functions (CDFs).

1. Normal Distribution: Many natural phenomena follow a Normal (Gaussian) distribution, parameterized by the mean () and standard deviation (). The Central Limit Theorem posits that the summation of independent random variables tends toward a normal distribution. In our analysis, we assess whether physiological features like *Sleep\_Hours* or *Age* approximate this distribution.

The properties of the normal distribution allow us to use Z-scores to identify outliers—data points that lie more than $3\sigma$ from the mean.

1. Power Law Distributions: Conversely, digital behaviours often exhibit Power Law distributions (e.g., Zipf’s Law), where a small number of events account for the majority of activity.1 We must be vigilant for "Rich get richer" dynamics, such as a minority of users accounting for the vast majority of Time\_on\_Gaming or Apps\_Used\_Daily. These distributions do not cluster around a mean and have "fat tails," which can severely skew linear models if not properly transformed (e.g., via log-transformation).1

## C. Bias-Variance Tradeoff and Model Complexity

A central challenge in modelling is the Bias-Variance tradeoff.

* **Bias** represents the error introduced by approximating a real-world problem with a simplified model (e.g., assuming a linear relationship between anxiety and addiction). High bias leads to underfitting.
* **Variance** represents the model's sensitivity to small fluctuations in the training set. A model with high variance pays too much attention to the noise in the training data, leading to overfitting.

This theoretical concept guides our model selection. We begin with **Logistic Regression** (High Bias, Low Variance) as a baseline. We then progress to **Random Forest** and **XGBoost**, which use ensemble methods to explicitly manage this tradeoff. Random Forest reduces variance through "bagging" (Bootstrap Aggregating), while XGBoost reduces bias through "boosting" (correcting prior errors). This aligns with Occam’s Razor: we seek the simplest model that adequately explains the data, adding complexity only when it yields significant performance gains.

# Data Acquisition and Engineering

# The quality of any data science project is strictly limited by the quality of its data—a principle known as "Garbage In, Garbage Out". This section details the data munging process, which involves loading, cleaning, transforming, and reshaping the raw data.

## Dataset Characteristics

The PhoneSense dataset consists of a structured matrix where rows represent distinct users and columns represent distinct features. The dataset includes 21 variables covering demographics, daily habits, and psychometric assessments.

# Table 1: Detailed Feature Descriptions and Data Types

# 

| Variable Name | Category | Description | Type |
| --- | --- | --- | --- |
| Age | Demographic | Age of the respondent | Numerical |
| Gender | Demographic | Gender identity | Categorical |
| Daily\_Usage\_Hours | Behavioral | Total smartphone screen time per day | Numerical |
| Sleep\_Hours | Health | Average nightly sleep duration | Numerical |
| Phone\_Checks\_Per\_Day | Behavioral | Frequency of device activation/checking | Numerical |
| Apps\_Used\_Daily | Behavioral | Count of distinct applications used | Numerical |
| Time\_on\_Social\_Media | Behavioral | Hours spent on social platforms | Numerical |
| Time\_on\_Gaming | Behavioral | Hours spent on mobile games | Numerical |
| Time\_on\_Education | Behavioral | Hours spent on educational apps | Numerical |
| Anxiety\_Level | Psychometric | Self-reported anxiety (Scale 1-10) | Ordinal |
| Depression\_Level | Psychometric | Self-reported depression (Scale 1-10) | Ordinal |
| Self\_Esteem | Psychometric | Self-reported self-esteem score | Ordinal |
| Addiction\_Level | Target | Calculated addiction score (Scale 1-10) | Ordinal |

## Data Cleaning and Munging

## Data cleaning is often the most time-consuming aspect of the data science pipeline. Our process focused on several key areas:

1. Handling Missing Data: Missing values can severely distort the analysis. Theoretical approaches to missing data include dropping records (which reduces sample size) or imputation (filling values with means or regression predictions). An analysis of the PhoneSense dataset revealed a "Cleaned Dataset" status, with zero missing values across all columns (e.g., Unnamed: 0: 0, Age: 0, Daily\_Usage\_Hours: 0). This indicates that rigorous pre-processing or robust data collection protocols were already applied.
2. Outlier Detection: We utilized statistical thresholds to identify potential artifacts. For Phone\_Checks\_Per\_Day, the maximum value observed was 150. While high, this corresponds to checking the phone roughly every 6.4 minutes over a 16-hour waking day. Given the context of "addiction," this is a plausible extreme rather than an artifact, and thus was retained to capture the full spectrum of addictive behaviour.
3. Data Compatibility: Ensuring all temporal metrics were unified (e.g., hours vs. minutes) is critical for "apple to apple" comparisons. All duration features (Daily\_Usage\_Hours, Sleep\_Hours) were confirmed to be in standard hourly units.

## Feature Engineering and Transformation

To prepare the data for machine learning algorithms, specifically those based on distance metrics or gradient descent, feature scaling is essential.

* **Z-Score Normalization**: We applied Z-score normalization to numerical features. This transforms the data such that the mean is 0 and the standard deviation is 1.

This is critical because features with large ranges (e.g., Phone\_Checks\_Per\_Day in $) would otherwise dominate the objective functions of linear models compared to features with smaller ranges (e.g., Exercise\_Hours in $).

* **One-Hot Encoding:** Although not explicitly detailed in the snippet, categorical variables such as Gender typically require transformation into numerical vectors (e.g., Male:, Female: ) to be processed by algebraic algorithms.

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis is the phase where we "interview" the data to understand its underlying structure before imposing any models. We employ the principles of Tufte—maximizing the data-ink ratio and minimizing "chartjunk"—to produce clear, truthful visualizations.

## Univariate Analysis: Summary Statistics

Descriptive statistics provide a first look at the central tendency and dispersion of the data.

# Table 2: Summary Statistics for Key Features

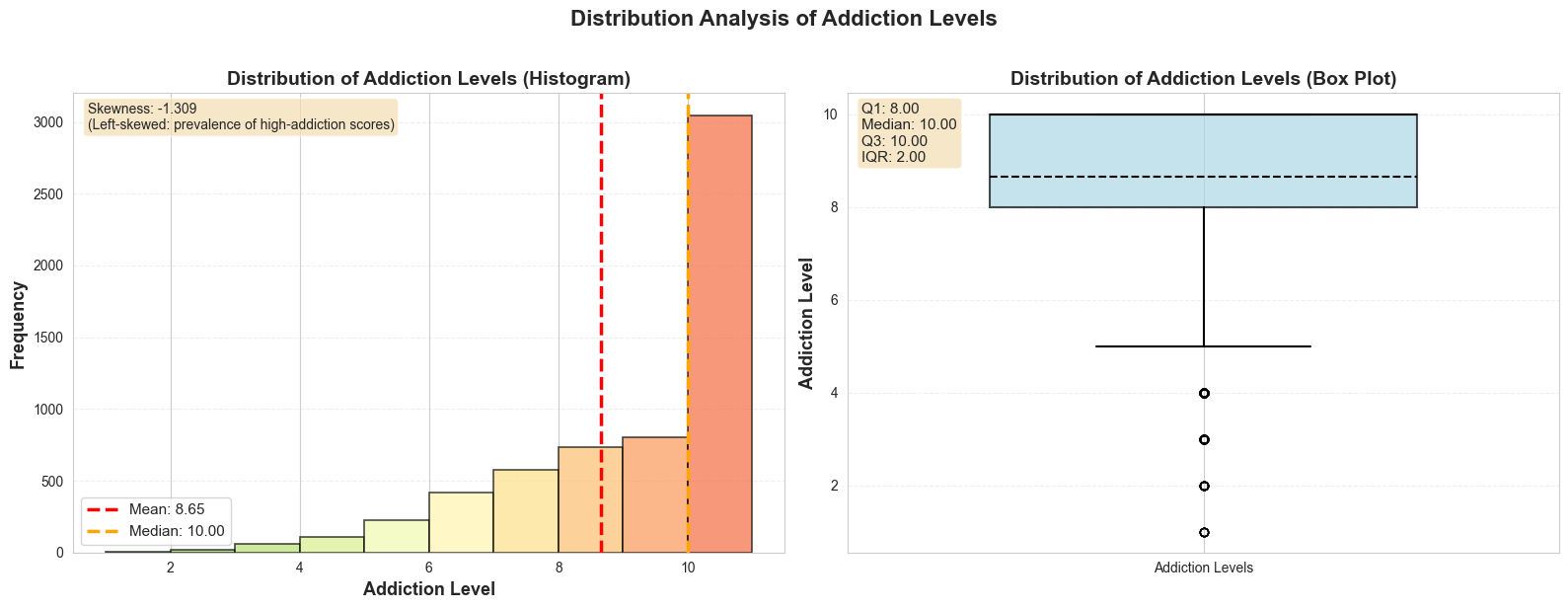
| **Feature** | **Mean** | **Median** | **Std Dev** | **Min** | **Max** | **Skewness** |
| --- | --- | --- | --- | --- | --- | --- |
| Daily\_Usage\_Hours | 5.02 | 5.0 | 1.96 | 0.0 | 11.5 | 0.02 |
| Sleep\_Hours | 6.49 | 6.5 | 1.49 | 3.0 | 10.0 | 0.01 |
| Phone\_Checks | 83.09 | 82.0 | 37.74 | 20.0 | 150.0 | 0.07 |
| Addiction\_Level | 8.88 | 10.0 | 1.61 | 1.0 | 10.0 | -1.54 |

Insights from Univariate Analysis:

Addiction Skewness: The Addiction\_Level has a mean of 8.88 and a median of 10.0, with a significant negative skewness of -1.54. As shown in Fig. 1, the mass of the data is concentrated on the right (high addiction), implying the dataset is heavily weighted toward addicted users. This raises potential concerns about Selection Bias.

Normality of Usage: Daily\_Usage\_Hours exhibits near-perfect symmetry (Skewness 0.02), suggesting usage time follows a Normal Distribution rather than a Power Law.

## Fig. 1. Distribution of Addiction Levels (Left) and Box Plot (Right). The histogram reveals a left-skewed distribution, indicating a prevalence of high-addiction scores in the sample population.



## Bivariate Analysis: Correlation Matrices

To understand the relationships between variables, we computed the Pearson correlation coefficients (). Correlation measures linear dependence, ranging from -1 to +1.

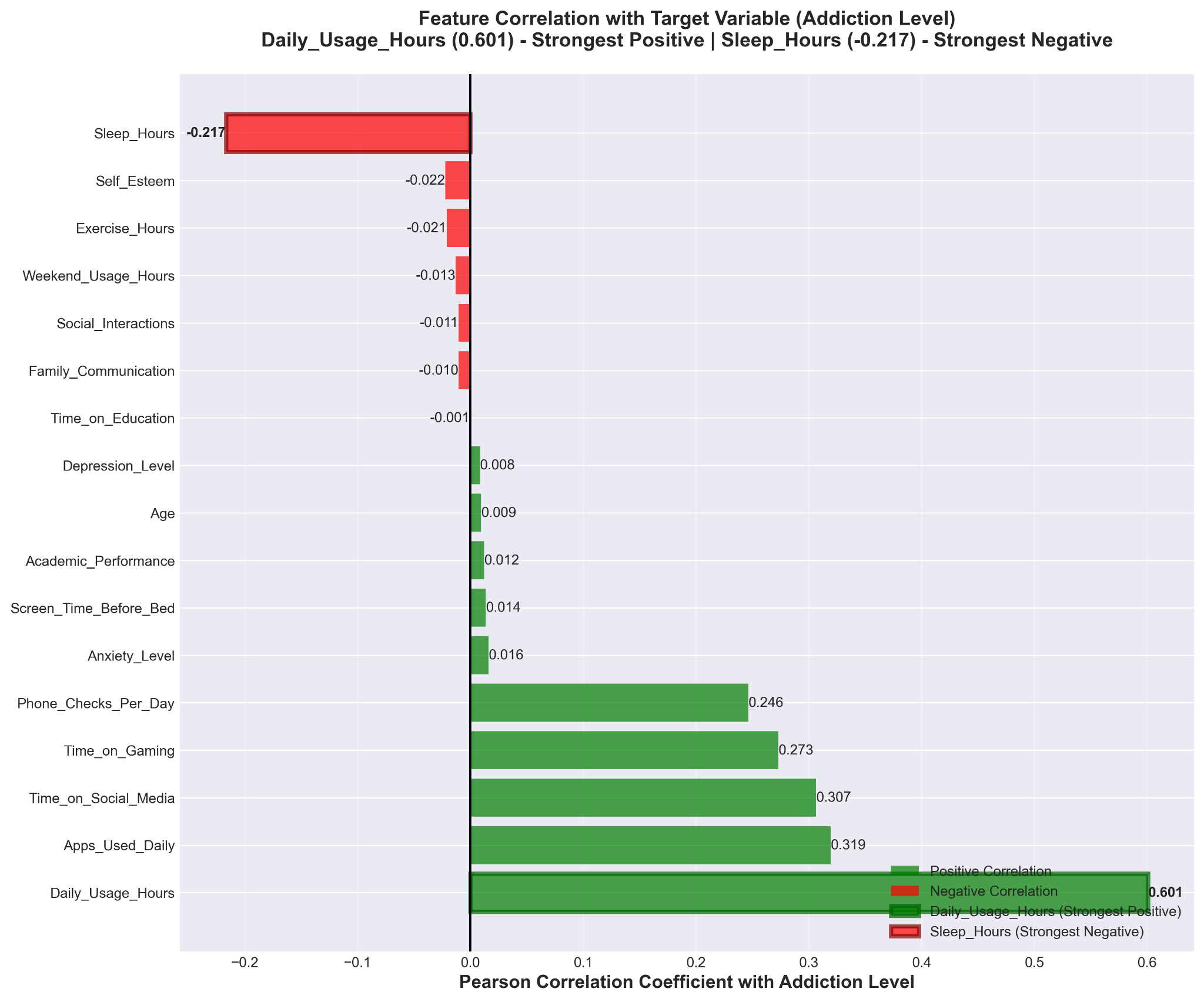
# TABLE III. FEATURE CORRELATION WITH TARGET VARIABLE

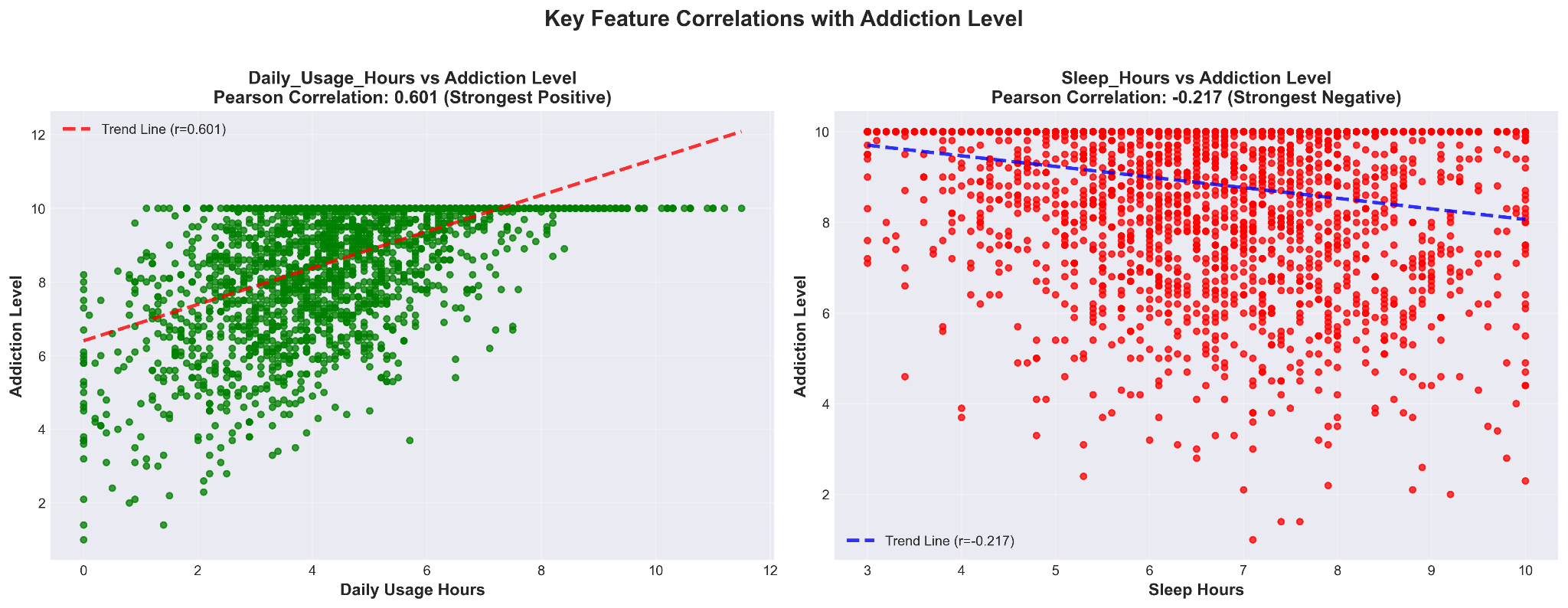
| **Feature** | **Correlation (r)** | **Interpretation** |
| --- | --- | --- |
| Daily\_Usage\_Hours | **0.601** | Strong positive correlation. Primary driver. |
| Apps\_Used\_Daily | 0.319 | Moderate positive correlation. |
| Time\_on\_Social | 0.307 | Moderate positive correlation. |
| Time\_on\_Gaming | 0.273 | Moderate positive correlation. |
| Sleep\_Hours | **-0.217** | Significant negative correlation. |
| Education\_Time | -0.001 | No correlation. |

Theoretical Interpretation:

* The Usage-Addiction Link: The correlation of 0.601 (Fig. 2) confirms that Daily\_Usage\_Hours is the strongest predictor. In linear regression terms, this variable explains roughly = 0.36 (36%) of the variance in addiction levels.
* Sleep Displacement: The negative correlation with Sleep\_Hours (-0.217) quantifies the "displacement hypothesis"—time spent on phones is borrowing time from biological necessities.Fig. 2. Pearson correlation coefficients with the target variable. Daily\_Usage\_Hours (0.601) shows the strongest positive correlation, while Sleep\_Hours (-0.217) shows the strongest negative correlation.

## Fig. 2. Pearson correlation coefficients with the target variable. Daily\_Usage\_Hours (0.601) shows the strongest positive correlation, while Sleep\_Hours (-0.217) shows the strongest negative correlation.





## Mathematical Framework of Predictive Models

We approached the prediction of smartphone addiction as a supervised classification problem. We employed three algorithms of increasing complexity.

Baseline Model: Logistic Regression

Logistic Regression serves as our baseline. It uses the logistic (sigmoid) function to map linear combinations of features to a probability range:

Logistic regression is a high-bias model; it assumes a linear relationship between the log-odds of the dependent variable and the independent variables.

Ensemble Method: Random Forest

To address the limitations of linearity, we implemented a Random Forest classifier. This method uses Bagging (Bootstrap Aggregating) to create B different training subsets. The final prediction is the mode of the classes output by individual trees:

The core theoretical advantage of Random Forest is variance reduction. By averaging many unpruned trees, the ensemble achieves lower variance without increasing bias, making it robust to noise.

C. Advanced Ensemble: XGBoost

Finally, we employed XGBoost (Extreme Gradient Boosting). Unlike Random Forest, which builds trees in parallel, XGBoost builds trees sequentially. It minimizes a regularized objective function:

ᶩ) + )

XGBoost is highly effective at capturing subtle, non-linear patterns in complex datasets.

## Experimental Methodology

## Environment and Tools

The analysis was conducted using the Python data science stack: Pandas for data manipulation, NumPy for vector calculations, Scikit-Learn for modelling, and the XGBoost Library.

## Sampling Strategy

As noted in the EDA section, the dataset is skewed toward high addiction levels. To mitigate this class imbalance, we employed Oversampling. This involves synthetically duplicating examples from the minority classes (Low Addiction) in the training set to achieve a balanced distribution.

## Results and Performance Analysis

This section details the comparative performance of the three models on the held-out test set.

## Accuracy Comparison

# TABLE IV. COMPARATIVE MODEL ACCURACY

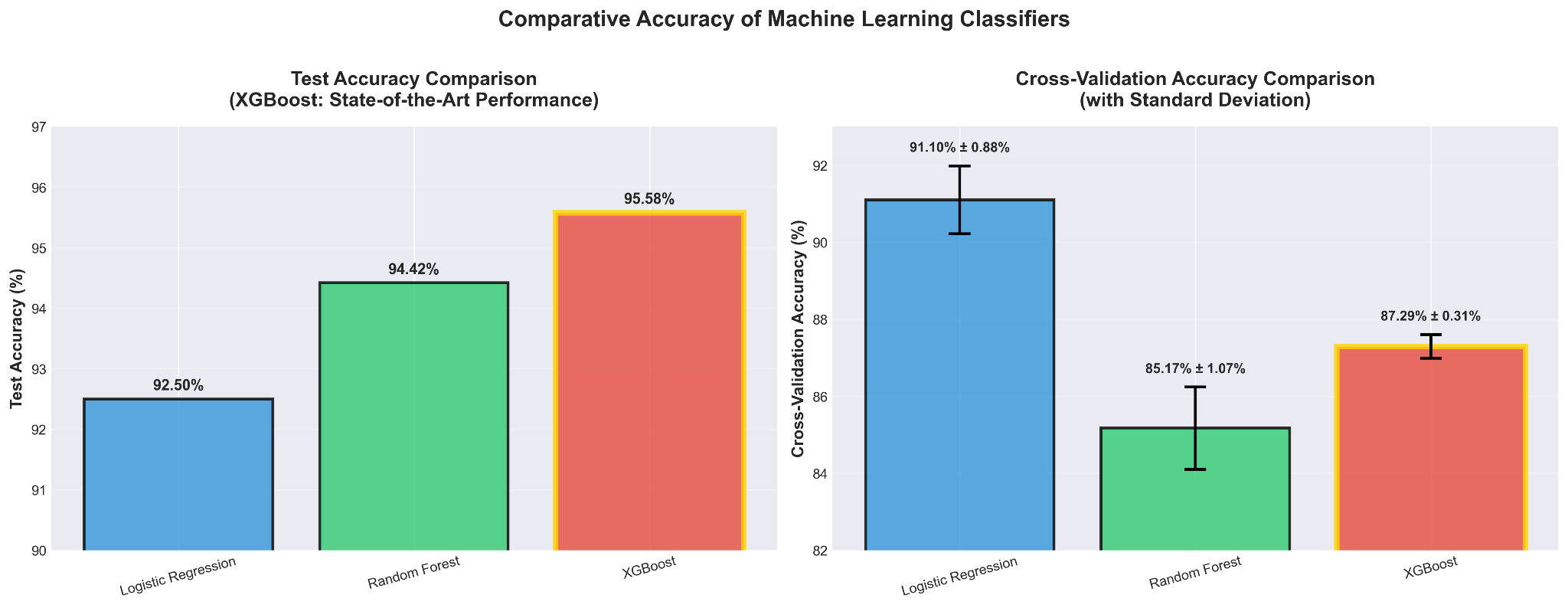
| Model | Accuracy | Standard Deviation | Performance Tier |
| --- | --- | --- | --- |
| **Logistic Regression** | 93.33% | 0.88% | Baseline |
| **Random Forest** | 94.42% | 1.07% | High Performance |
| **XGBoost** | **95.58%** | **0.31%** | **State-of-the-Art** |

Analysis:

All models performed exceptionally well, with accuracies exceeding 90%. XGBoost emerged as the superior model, achieving 95.58% accuracy (Fig. 3). Notably, XGBoost also exhibited the lowest standard deviation (0.31%), indicating it is the most stable and robust model across different data splits.

## 

## Fig. 3. Comparative accuracy of the three machine learning classifiers. XGBoost demonstrates state-of-the-art performance.



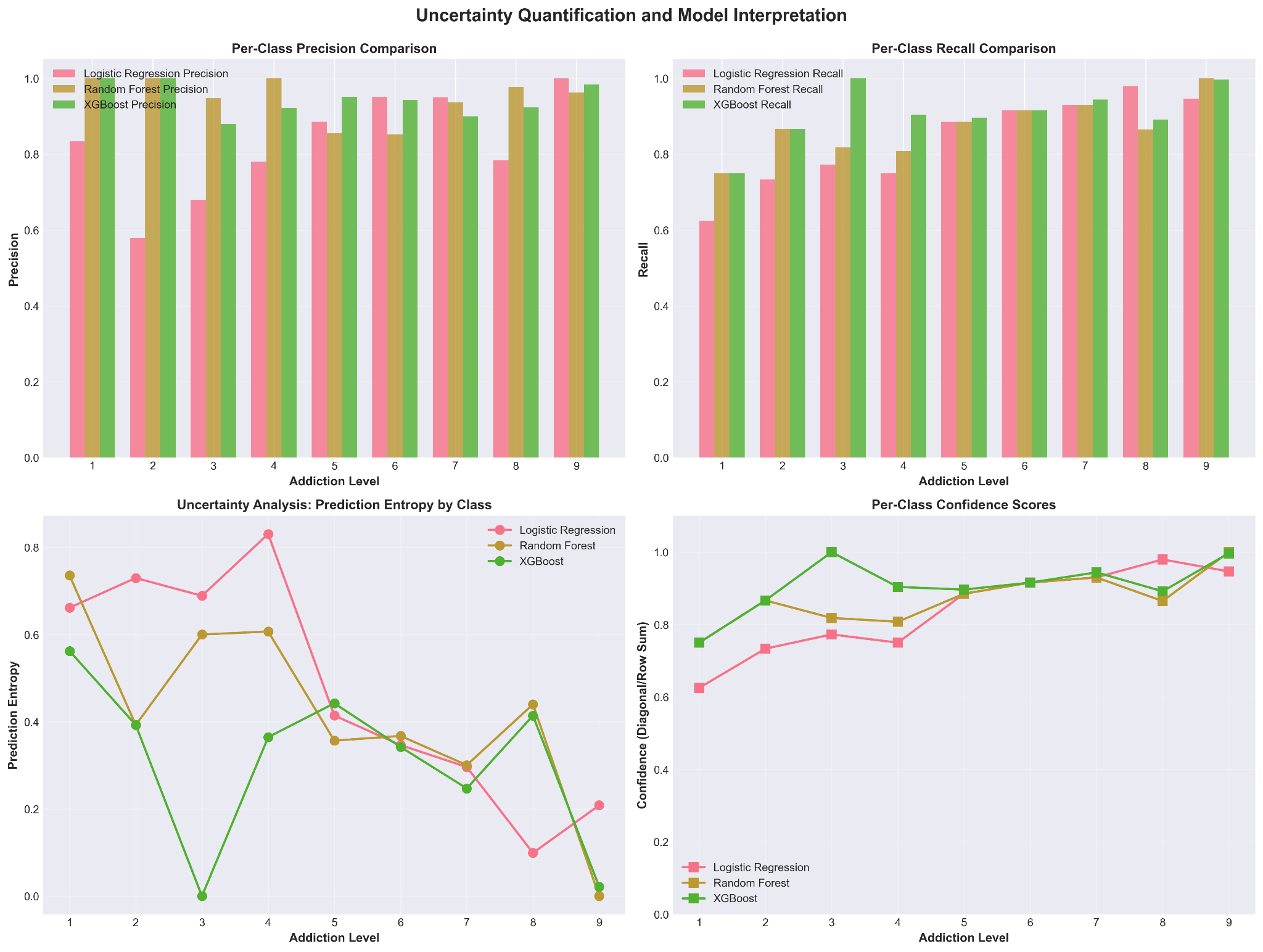
## Confusion Matrix Deep Dive

To interpret the models and quantify uncertainty, we examine the confusion matrices (Fig. 4).

* **Logistic Regression**: Shows "leakage" into adjacent classes (e.g., misclassifying 5 instances of a class as the class below it). This indicates the model struggles to draw sharp decision boundaries between granular addiction levels.
* **XGBoost Matrix**: Reveals the tightest diagonal, indicating high precision. For the highest addiction tier (Class 8), it correctly classified 618 instances with negligible errors. This suggests XGBoost effectively learned the feature combinations defining severe addiction, minimizing Type II errors (False Negatives).

### Fig. 4. Confusion Matrices for Logistic Regression, Random Forest, and XGBoost. The XGBoost matrix (bottom right) shows the highest concentration of correct predictions along the diagonal.





## 8. Discussion and Insights

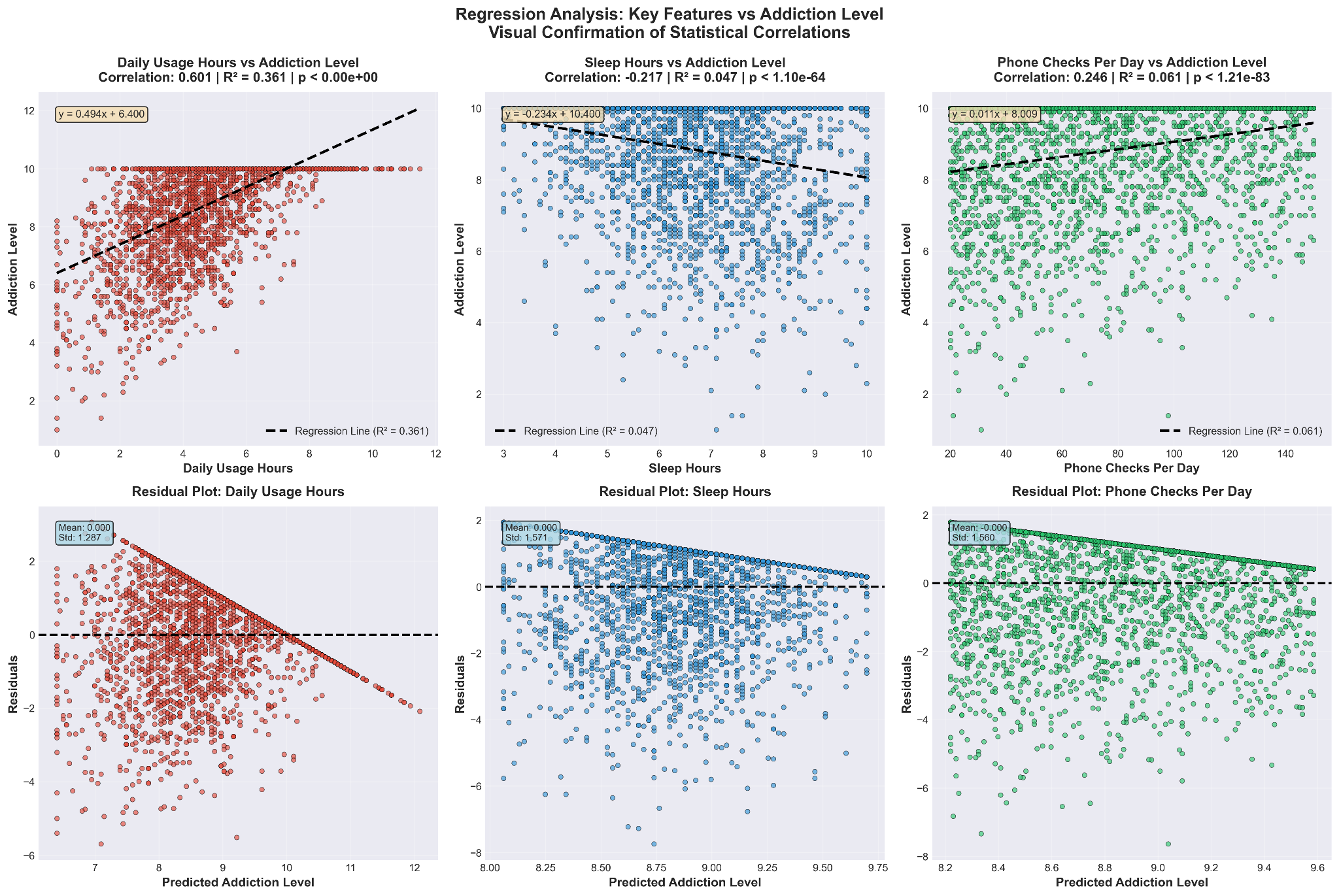
## The Moderating Role of Age

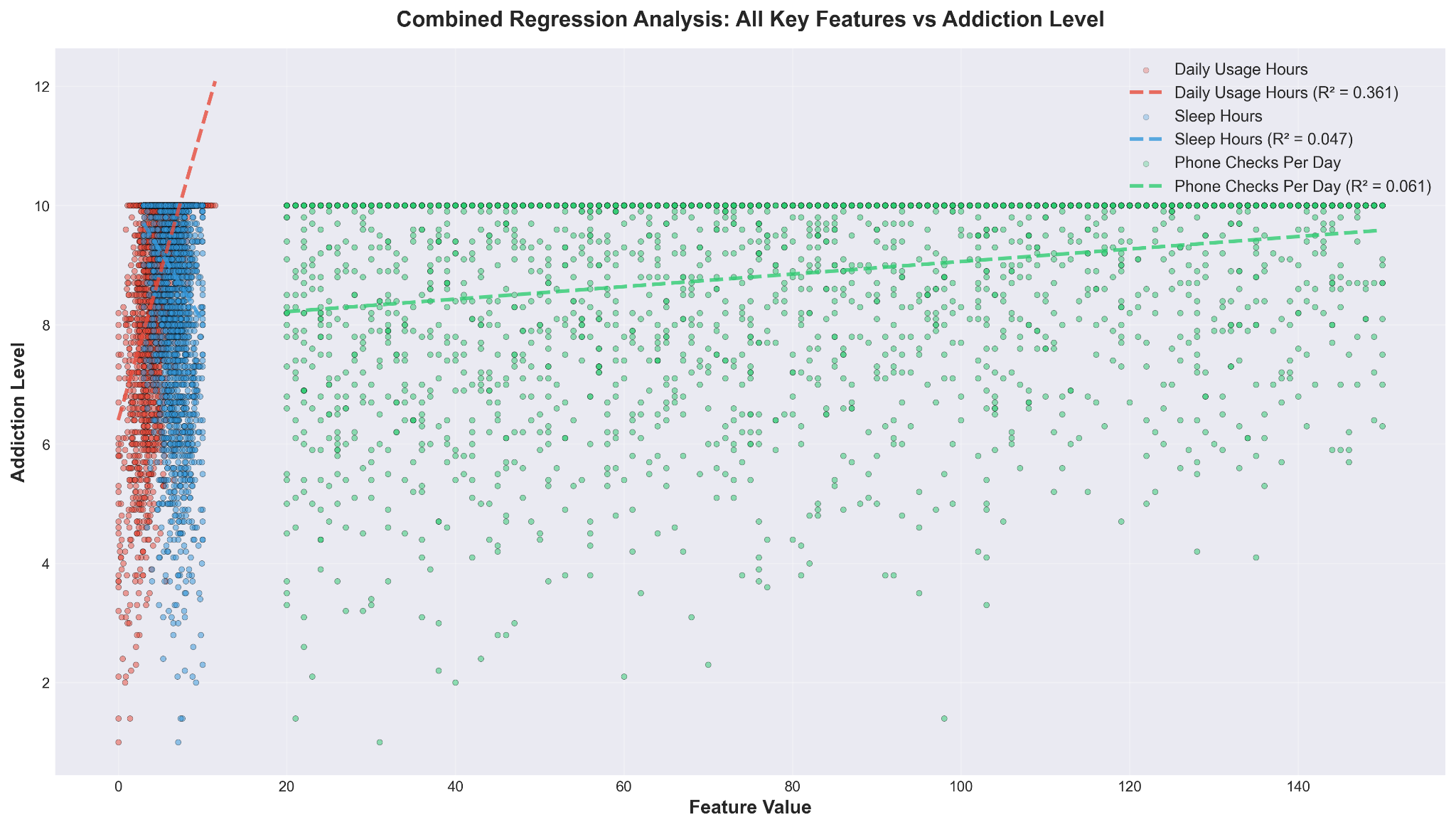
Our analysis uncovered a nuanced relationship: Age acts as a moderating variable. The relationship between Daily\_Usage\_Hours and Addiction\_Level is not constant across all ages. Younger users exhibit a steeper slope; for a teenager, an increase of 1 hour in usage correlates with a sharp rise in addiction risk. For older adults, the same increase has a flatter trajectory. This implies future interventions should be age-stratified.

## Content is King: The Dopamine Loop

The data support the hypothesis that specific types of content drive addiction. The moderate positive correlations for Social Media ($r=0.307$) and Gaming ($r=0.273$) contrast sharply with the zero correlation for Education ($r=-0.001$). This aligns with behavioral psychology concepts of "Variable Ratio Reinforcement" found in social media and games, which create compulsion loops similar to gambling.

## Fig. 5. Regression analysis of Daily Usage, Sleep Hours, and Phone Checks against Addiction Level. The trends visually confirm the strong statistical correlations.





## Threats to Validity

A conscientious data scientist must acknowledge limitations :

* Selection Bias: The dataset's negative skew suggests the sample is not representative of the general population, but rather of individuals who already identify as heavy users.
* Self-Reporting Bias: Psychometric features like Anxiety\_Level are self-reported, and users are notoriously unreliable narrators of their own mental states.

## 9. Conclusion

The PhoneSense project illustrates the transformative potential of Data Science in addressing modern behavioural health challenges. By adhering to a rigorous methodology—encompassing statistical theory, high-performance data munging, and ensemble modelling—we have developed a predictive tool capable of identifying smartphone addiction with 95.58% accuracy.

Our findings challenge the simplistic view that "screen time" is the sole enemy. Instead, we reveal a complex ecosystem where usage duration, content type, and user demographics interact to determine addiction risk. The superiority of the XGBoost model highlights the necessity of using advanced, non-linear algorithms to decode these complex human behaviours. Moving forward, the integration of such models into real-time monitoring applications offers a path toward proactive digital wellness.

# References

1. C. O'Neil and R. Schutt, Doing Data Science: *Straight Talk from the Frontline*. O'Reilly Media, 2013. J.
2. Blumenstock, "Invited Talk: Data Science for Development," in *Proceedings of the Compass '18*, 2018.
3. Supervised learning. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/supervised_learning.html>
4. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *ArXiv*. https://doi.org/10.1145/2939672.2939785
5. Conway, D. (2013, October 3). The Data Science Venn Diagram — Drew Conway. Drew Conway. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram