Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL34666

Digital Reading as Mobile Attention Allocation: A Weighted Machine Learning Analysis of Reading-App Usage

Jingxuan Zhou

The High School Affiliated to Renmin University of China, Beijing, China
yinggang030702@gmail.com

Abstract. Reading is closely tied to learning, communication, and personal development, but the way people read has changed with the spread of mobile devices. Reading apps now sit on the same phone as short-video platforms, social media, games, and messaging tools. This makes digital reading not only a question of whether people like reading, but also a question of how much of their limited phone time they are willing to allocate to reading. This paper studies reading-app usage using an original questionnaire dataset of 84 respondents. We define the main dependent variable as the ratio of daily reading-app hours to total daily phone-use hours and design a questionnaire that collects demographic background, daily time use, reading platforms, reading habits, motivations, social reading behavior, and content preferences. Methodologically, we compare a full weighted least squares baseline, a weighted ElasticNet feature-selection model followed by post-selection WLS, and a weighted Random Forest. The results show that the unregularized full WLS model severely overfits the high-dimensional survey data. By contrast, the ElasticNet-selected WLS model reduces 110 encoded predictors to 19 selected features and achieves a cross-validated RMSE of 0.156, close to the Random Forest benchmark of 0.154. Across models, WeRead usage, reading routines, social reading behavior, app tenure, and competing app-use patterns are the strongest predictors. These findings suggest that digital reading engagement is shaped more by platform embeddedness and daily behavioral routines than by demographics alone.

Keywords: Digital Reading, Machine Learning, Society Research, User Behavior Analysis

1. Introduction

Reading is one of the most basic human activities. People read to learn, to relax, to follow stories, to improve themselves, and to stay connected with cultural and social conversations. For this reason, reading behavior is not only a personal habit. It also reflects education, lifestyle, media choice, and changes in the surrounding information environment. In recent years, reading has increasingly moved onto mobile devices. Reading apps make books, web fiction, notes, recommendations, and reading communities available through a phone. At the same time, the same device also provides short videos, social media, games, and many other services that compete for attention. This changes the nature of the reading question. In a mobile environment, it is not enough to ask whether a person reads. It is also important to ask how much of that person's phone time is allocated to reading rather than to other digital activities. This paper studies reading-app usage from this perspective. We treat

DOI: 10.54254/2755-2721/2026.GL34666

reading-app engagement as a problem of mobile attention allocation. Instead of using only raw reading time, we focus on the ratio of reading-app time to total phone-use time. This ratio is useful because two respondents can report the same reading-app time but have very different behaviors. A person who reads for one hour out of two total phone hours gives a much larger share of attention to reading than a person who reads for one hour out of ten total phone hours. Several existing studies have examined mobile reading behavior. Wang studies mobile reading among university students in Henan Province using questionnaires and interviews, analyzing platform usage, content preferences, and demographic differences [1]. Yi and Wu focus on WeChat Reading and discuss how platform design and mobile Internet use shape university students' reading habits [2]. Jiang examines college students' socialized reading behavior from the perspective of self-efficacy theory, emphasizing psychological and social drivers behind reading participation [3]. These studies provide valuable background for understanding mobile reading, platform use, and social reading behavior. However, there are two limitations in the existing literature. First, many studies focus on whether students use mobile reading platforms or how they evaluate a specific app, but they do not directly model reading-app time as a share of total phone time. Second, most studies remain descriptive. They summarize questionnaire responses or compare user groups, but they rarely use regularized machine learning methods to select important predictors from a high-dimensional questionnaire. This matters because reading behavior can be associated with many factors at the same time: demographics, phone use, competing apps, reading habits, motivations, social behavior, and content preferences. Simple descriptive comparisons may miss this joint structure.

Our study addresses this gap by building a data-driven framework that links reading-app engagement to daily behavior and personal characteristics. We designed and distributed a questionnaire, collected 84 valid responses, cleaned and transformed the data, constructed the main target variable, and estimated weighted predictive models. The empirical analysis uses three models: a full weighted least squares model as a diagnostic baseline, an ElasticNet-selected WLS model as the main interpretable model, and a weighted Random Forest as a nonlinear benchmark. The main contribution is twofold. Substantively, we identify which types of variables are most related to reading-app engagement. Methodologically, we show why regularization is necessary in small-sample survey research with many predictors. A full WLS model can look excellent in sample, but cross-validation reveals severe overfitting. The regularized model is less impressive in sample, but it is far more stable and interpretable.

2. Research framework and questionnaire design

2.1. Research framework

The goal of this study is to connect reading-app usage with respondents' daily activities and personal characteristics. In the framework, reading apps are not treated as isolated platforms. They are part of a wider mobile environment where users divide time among reading, video, social media, games, work, study, and communication. The full workflow is as follows. First, we designed a questionnaire around factors that may affect reading behavior. Second, we distributed the questionnaire through classmates, friends, teachers, family members, parents' friends, and online platforms. Third, we cleaned the data, renamed variables into script-readable formats, converted survey answers into numerical or categorical variables, and constructed the target variable. Fourth, we created a modeling dataset using the cleaned questionnaire variables. Finally, we evaluated the models using cross-validation and feature-importance analysis. The key idea is that reading-app engagement should be measured relative to total phone use. If reading is one of many activities on a phone, then

DOI: 10.54254/2755-2721/2026.GL34666

its importance is better captured by its share of phone time than by raw time alone. This is why the main dependent variable is defined as:

$$\text{ReadApp\_to\_Phone}_i = \frac{\text{Reading app hours}_i}{\text{Total phone hours}_i}. \quad (1)$$

In the empirical analysis, values above 1 are capped at 1 because the ratio is conceptually bounded between 0 and 1. Values above 1 likely reflect reporting inconsistencies or cases where respondents included paper reading or other non-phone reading time.

2.2. Questionnaire design

The questionnaire was designed to cover six groups of variables. The full questionnaire can be included in the appendix. Table 1 summarizes the main variable groups and why each group is relevant.

Table 1. Questionnaire variable groups

Variable group	Examples	Why included	Expected relationship
Background variables	Gender, age, education, major, identity, annual income, study or work time	Describe respondents' characteristics and time constraints	Demographic and socioeconomic variables may affect reading habits, although they may not be the strongest predictors
Core time-use variables	Phone time, reading-app time, listening-app time, paper or Kindle reading time, video-app time, social-media time, game time	Measure the mobile time budget and competing activities	Reading-app share may decrease when phone time is absorbed by other activities
Reading habits and experience	Common platforms, continuous reading habit, age started reading, fixed reading time, reading multiple books, stopping disliked books	Measure whether reading is routine-based and stable	Stronger reading routines are expected to increase reading-app engagement
Reading motivations	Relaxation, self-improvement, habit, achievement, escape, connection, recommendation, rereading	Capture why people read	Motivation may shape whether users protect time for reading
Social and output behavior	Reading notes, making notes, recording reading, recommendations, fixed partners, bookstores, online or offline activities	Capture social and expressive dimensions of reading	Social reading behavior may increase engagement by turning reading into a community or identity activity
Content preferences	Literature, web fiction, science fiction, mystery, social science, history, Chinese, Japanese, Western, Chinese language, English language	Capture what people like to read	Content preferences may affect app usage if some content types are more convenient through mobile platforms

This design lets us ask a broader question than "who reads more?" We can examine whether reading-app engagement is associated with platform use, competing app time, reading routines, motivations, and social behavior.

DOI: 10.54254/2755-2721/2026.GL34666

3. Data description

3.1. Data collection

We first set the goal of finding factors related to reading-app duration and reading's share of phone time. We then considered variables that may be linked to reading behavior, such as reading habits, phone-use patterns, video-app usage, social-media usage, game-app usage, reading motivations, social reading activities, and preferred content. These variables were included in the questionnaire. The questionnaire was promoted among schoolmates, friends, teachers, parents, parents' friends, and online platforms. In total, we collected 84 valid responses. The answers were recorded in Excel and then processed in Python for analysis. Because the sample was collected through personal and social networks, it should be understood as a convenience sample rather than a random sample. This is important for interpretation, because people who read more or who use reading apps more often may be more willing to answer a reading-related survey.

3.2. Data cleaning and variable construction

The raw survey data were cleaned before analysis. Unnecessary columns were removed. Column names were changed into script-readable variable names. Text responses were converted into numerical or categorical variables. For respondents without reading habits, reading-related behavior variables were set to 0 where appropriate. Missing income values were handled by creating an income missingness indicator and using an imputed income variable for modeling. The main target variable is ReadApp_to_Phone, defined as reading-app hours divided by total phone-use hours. Since this variable is a ratio, values greater than 1 are not conceptually meaningful. In the descriptive analysis, we report the raw ratio to show the data issue. In the modeling analysis, we use ReadApp_to_Phone_capped, which caps values above 1 at 1.

3.3. Descriptive statistics

Table 2 reports descriptive statistics for the main dependent variables. The distribution of reading-app engagement is highly skewed. The median respondent spends 1 hour per day on reading apps, while the upper tail contains a small number of heavy users. For the main ratio variable, the median of ReadApp_to_Phone is 0.146 and the third quartile is 0.250. This means that for most respondents, reading apps account for a small share of total phone time. The raw ratio has a maximum of 1.459, which motivates using the capped version in the regression and machine learning models. The left panel of Figure 1 presents the distribution of the capped reading-app attention share. Most observations are concentrated near the bottom of the distribution, while a small number of observations remain much higher. The scatter plots provide additional descriptive patterns. The right panel of Figure 1 shows the relationship between reading-app attention share and total phone-use hours. The relationship is weakly negative: the fitted slope is -0.0068, with Pearson correlation -0.114 and Spearman correlation -0.106. This is consistent with the idea that as total phone use increases, additional phone time may be absorbed by other activities rather than reading.

Table 2. Summary statistics of reading outcomes

Variable	N	Mean	Std	Min	P25	Median	P75	Max	Outliers
ReadApp_to_Phone	84	0.199	0.252	0	0.022	0.146	0.25	1.459	5
ReadApp_to_Phone_capped	84	0.191	0.217	0	0.022	0.146	0.25	1	5

DOI: 10.54254/2755-2721/2026.GL34666

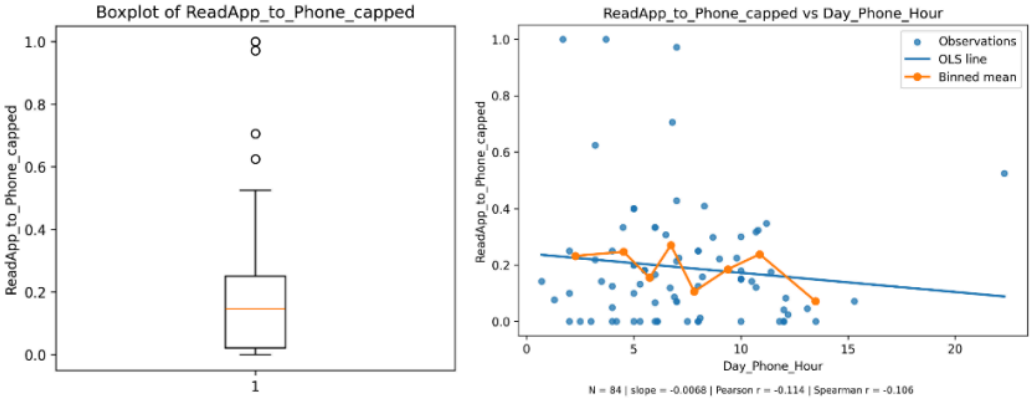

Figure 1. Distribution of reading-app attention share

This figure reports the boxplot of ReadApp_to_Phone_capped (Left) and Reading-app attention share and total phone-use time (Right).

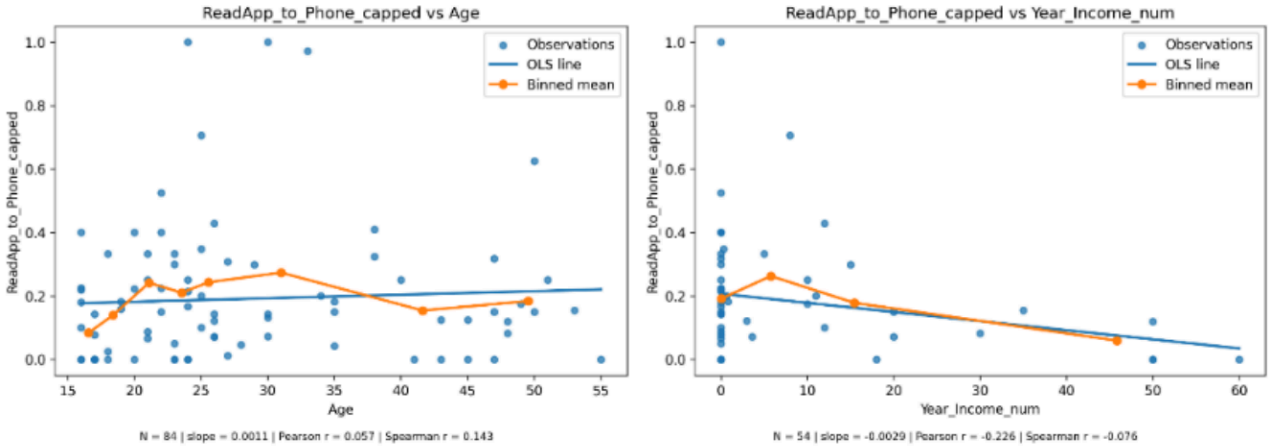

Figure 2. Reading-app attention share and respondent characteristics–Age (left) and Annual Income (right)

Figure 2 compares reading-app attention share with age and annsual income. The relationship with age is weak, with a fitted slope of 0.0011, Pearson correlation 0.057, and Spearman correlation 0.143. The income relationship is also weakly negative: the slope is -0.0029, with Pearson correlation -0.226 and Spearman correlation -0.076. The income result should be interpreted cautiously because income is missing for part of the sample and many respondents are students with zero or low reported income.

4. Methodology

This section provides a detailed explanation of the regression analysis used to examine the relationship between reading time allocation and corresponding independent variables.

4.1. Weighted linear regression

We begin with a weighted linear regression model:

DOI: 10.54254/2755-2721/2026.GL34666

$$Y_i = X_i^{'}\beta + \varepsilon_i \tag{2}$$

where Y_i is $\mathrm{ReadApp\_to\_Phone_capped}$ for respondent i, X_i is a vector of cleaned and encoded questionnaire variables, β is the coefficient vector, and ε_i is the error term. The weighted least squares estimator solves:

$$\widehat{\beta_{WLS}} = \mathrm{argmin}_\beta \sum_{i=1}^{n} w_i \left(Y_i - X_i^{'}\beta\right)^2 . \tag{3}$$

The weights are used because the survey likely overrepresents reading-app users. We collected three external WeRead-to-WeChat contact ratios and used them as a calibration benchmark. The target positive-user share is 19.38%, while 76.19% of the survey sample reports positive reading-app usage. Therefore, positive reading-app users receive lower weights, while zero-use respondents receive higher weights.

Table 3. Calibration weight construction

Group	Sample share	Target share	Raw weight
Positive reading-app users	0.7619	0.1938	0.2543
Zero reading-app users	0.2381	0.8062	3.3862

This weighting procedure should not be interpreted as fully correcting sample-selection bias. It only partially adjusts the sample toward an external benchmark for reading-app participation.

4.2. ElasticNet feature selection

After one-hot encoding categorical variables, the model contains 110 encoded predictors. With only 84 observations, a full WLS regression is too flexible and likely to overfit. To address this issue, we use ElasticNet regularization [4]:

$$\widehat{\beta_{EN}} = \arg\min_\beta \left[\sum_{i=1}^{n} w_i \left(Y_i - X_i^{'}\beta\right)^2 + \lambda \left(\alpha \|\beta\|_1 + (1-\alpha)\|\beta\|_2^2\right)\right]. \tag{4}$$

The L_1 part of the penalty can shrink some coefficients exactly to zero, which makes it useful for feature selection. The L_2 part stabilizes estimates when predictors are correlated. ElasticNet combines both penalties, which is appropriate for questionnaire data where many behavioral variables are likely related. After ElasticNet selects nonzero predictors, we refit WLS using only the selected features. This post-selection WLS model gives interpretable coefficients while avoiding the extreme overfitting of the full model. Because feature selection is data-driven, the p-values from post-selection WLS should be treated as exploratory rather than definitive.

4.3. Random forest benchmark

We also estimate a weighted Random Forest model as a nonlinear benchmark [5]. Random Forest builds many decision trees and averages their predictions:

$$\widehat{Y_i^{RF}} = \frac{1}{B} \sum_{b=1}^{B} T_b\left(X_i\right), \tag{5}$$

DOI: 10.54254/2755-2721/2026.GL34666

where $T_b(X_i)$ is the prediction from tree b. Random Forest can capture nonlinear patterns and interactions among variables. It does not produce linear coefficients, so we interpret it using feature importance.

4.4. Model evaluation

We evaluate model performance using cross-validation. The main metrics are RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R^2. The definition of RMSE, MAE and weighted R^2 are given by the following, We focus mainly on cross-validated RMSE and MAE rather than in-sample R^2, because in-sample fit can be highly misleading in small, high-dimensional datasets.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2}. \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y_i}\right|. \tag{7}$$

$$R_w^2 = 1 - \frac{\sum_i w_i\left(Y_i - \widehat{Y_i}\right)^2}{\sum_i w_i\left(Y_i - \bar{Y}_w\right)^2}. \tag{8}$$

5. Empirical results

5.1. Full WLS baseline

The full WLS model uses all 110 encoded predictors. It achieves an in-sample weighted R^2 of 0.994 and an in-sample RMSE of 0.010. At first glance, this looks excellent. But the model has only 3 residual degrees of freedom, and the condition number is approximately 4.10×10^{16}. The full WLS regression summary reports that the input rank is higher than the number of observations and warns that the design matrix may be singular, indicating strong multicollinearity or rank deficiency. Cross-validation confirms the problem. The full WLS model has a mean cross-validated RMSE of 1.961 and a mean cross-validated MAE of 1.511. Since the dependent variable is bounded between 0 and 1 after capping, an RMSE greater than 1 indicates severe out-of-sample instability. This model should therefore be viewed only as a diagnostic baseline, not as a valid final specification.

5.2. ElasticNet-selected WLS

ElasticNet selects 19 predictors from the 110 encoded features. These selected variables include platform variables, reading routines, social reading variables, time-use variables, demographic variables, and motivation variables. After refitting WLS on these selected variables, the model obtains an in-sample weighted R^2 of 0.618 and an in-sample RMSE of 0.085. The residual degrees of freedom increase to 64, and the condition number drops to 12.61. The most important coefficient is $\mathrm{Weread\_1}$. Its coefficient is 0.1336 with a p-value of 0.0086. This suggests that, conditional on the selected controls, WeRead users allocate about 13.4 percentage points more of phone time to reading apps. Another statistically significant coefficient is $\mathrm{Reading\_Because\_Habit\_2}$, with a coefficient of -0.1687 and a p-value of 0.0029. The result indicates that habit-related motivation is connected to reading-app engagement.

DOI: 10.54254/2755-2721/2026.GL34666

5.3. Model comparison

Table 4 compares the three main models. The comparison shows why regularization is necessary. The full WLS model looks nearly perfect in sample, but collapses under cross-validation. The ElasticNet-selected WLS model gives up some in-sample fit but performs far better out of sample. The Random Forest has a similar cross-validated RMSE of 0.154, only slightly lower than the regularized WLS model's 0.156. This suggests that the regularized WLS captures much of the predictable structure in the data while remaining easier to interpret.

Table 4. Model comparison

Model	N	Features	In-sample R_w^2	In-sample RMSE	CV RMSE	CV MAE
Full WLS	84	110	0.994	0.01	1.961	1.511
ElasticNet + WLS	84	19	0.618	0.085	0.156	0.11
Weighted Random Forest	84	110	0.546	0.092	0.154	0.114

5.4. Feature contribution and random forest importance

To interpret the regularized model, we compute a drop-column contribution measure. For each variable, we remove it from the model, refit WLS, and measure the decrease in weighted R^2. A larger decrease means that the variable contributes more to the model. Figure 3 reports the two main importance plots. The leading variable in both models is WeRead. In the regularized WLS contribution ranking, removing WeRead produces the largest drop in weighted R^2. In the Random Forest, WeRead is also the most important variable. Other recurring variables include fixed reading partners, reading records, constant reading time, social-media time, game-app time, and app tenure. The overlap between the two rankings strengthens the interpretation that reading-app engagement is related to platform use, routines, and social behavior. Random Forest does not rely on the same linear structure as WLS, so the agreement between the two models is meaningful.

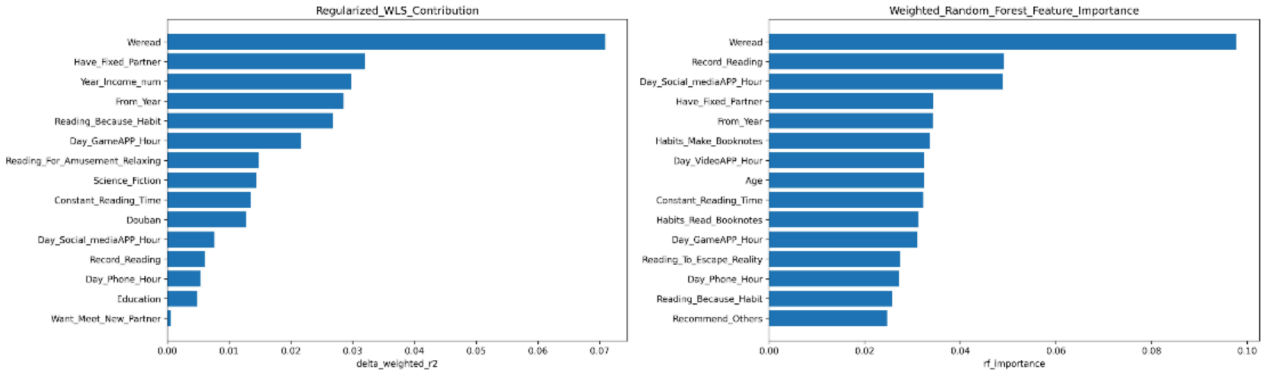

Figure 3. Regularized WLS contribution (left) and Random Forest importance (right)

The results show that reading-app engagement is not mainly explained by demographic variables alone. Age has little direct relationship with reading-app share in the descriptive plots. Education and income appear in the selected model, but they are not the whole story. More important are platform use, habits, social behavior, and daily app-use patterns. The importance of WeRead is particularly notable. It appears as the top variable in both regularized WLS contribution and

DOI: 10.54254/2755-2721/2026.GL34666

Random Forest importance. This may reflect self-selection: people who like digital reading choose WeRead. It may also reflect platform effects: WeRead may encourage reading through convenience, records, social visibility, and recommendation functions. With cross-sectional survey data, we cannot separate these two channels. Still, the result shows that platform embeddedness is strongly associated with reading-app engagement. The weak negative relationship between total phone time and reading-app share also fits the mobile attention framework. People who use phones longer do not necessarily allocate more of that time to reading. Additional phone time may go to videos, social media, games, or messaging. This supports the idea that reading apps compete for attention within the phone environment. The comparison between full WLS and regularized models is also methodologically important. In small survey datasets with many categorical variables, full regression models can look impressive but be unstable. Here, full WLS has an in-sample weighted R^2 of 0.994, but cross-validation reveals that it does not generalize. ElasticNet reduces the feature set and produces a model that is both interpretable and much more stable.

6. Conclusion

This paper studies reading-app usage as a problem of mobile attention allocation. Using an original questionnaire dataset of 84 respondents, we define the main outcome as the ratio of daily reading-app hours to total phone-use hours. Descriptive evidence shows that reading-app engagement is right-skewed: most respondents allocate a relatively small share of phone time to reading apps, while a few users allocate much larger shares. The empirical analysis shows that a full WLS model with 110 encoded predictors severely overfits the small survey sample. Although its in-sample weighted R^2 is 0.994, its cross-validated RMSE is 1.961 and the model has only 3 residual degrees of freedom. By contrast, the ElasticNet-selected WLS model reduces the predictor set to 19 variables and achieves a cross-validated RMSE of 0.156. A weighted Random Forest benchmark obtains a similar cross-validated RMSE of 0.154. Substantively, the results suggest that digital reading engagement is shaped less by demographics alone and more by platform embeddedness, routines, and social behavior. WeRead usage is the strongest predictor across both linear and nonlinear models. Other important predictors include fixed reading partners, reading records, constant reading time, habit-related motivation, social-media time, and game-app time. These findings show that reading apps should be understood not only as reading tools, but also as part of a broader mobile attention ecosystem.

References

[1] Haiyan Wang. Investigation and analysis of mobile reading behavior of university students: evidence from university students in Henan province. Tushuguanxuekan, (7), 2023.
[2] Jingyu Yi, Weide Wu. Research on the influence of the mode of WeChat reading on university students in the era of mobile internet [J]. Library Work in Colleges and Universities, 42(05): 75-80, 2022.
[3] Jingcheng Jiang. Research on the influencing factors of college students' socialized reading behavior. Guangdong University of Finance & Economics, MA thesis.
[4] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67(2): 301–320, 2005.
[5] von Hirsch, A., & Ashworth, A. (2005). Proportionate sentencing: Exploring the principles. Oxford University Press.