Modeling User Adherence for MyYouthspan

Karunya Iyappan
M.S. in Data Science Candidate, University of Virginia
Fairfax, VA
ki5rb@virginia.edu

Sam Knisely
M.S. in Data Science Candidate, University of Virginia
Arlington, VA
sck4jh@virginia.edu

Alec Pixton
M.S. in Data Science Candidate, University of Virginia
Jonesville, VA
etk3pu@virginia.edu

Trenton Ribbens
M.S. in Data Science Candidate, University of Virginia
Charlottesville, VA
kvu2et@virginia.edu

Abstract—**This project supports MyYouthspan, a personalized wellness and longevity application. By building a data pipeline and interpretable AI models to evaluate adherence patterns of MyYouthspan users, the team developed user personas that aim to suggest more achievable plans and improve user adherence to the platform. By analyzing MyYouthspan's user plan and log data, the team modeled user adherence likelihood and identified clusters of behavioral personas to enhance recommendation personalization and user engagement.**

I. Introduction

MyYouthspan is a software platform rooted in its commitment to improving the health of its users by providing an interface for daily logging of behaviors and habits. Using this data, MyYouthspan analyzes longitudinal trends and recommends realistic, user-specific goals for life longevity.

This project supports MyYouthspan's objective of identifying user adherence patterns in order to refine and personalize lifestyle recommendations. By understanding which plans users are likely to follow, the platform can better tailor guidance to individual needs, which would ultimately increase the likelihood that users remain engaged over time. The goal is to prevent drop-off by ensuring that recommendations feel achievable and aligned with each user's unique context.

This project investigates user adherence to personalized lifestyle plans by analyzing the relationship between planned activities and logged behaviors. The primary objective is to identify the factors that influence adherence. By developing a method to quantify adherence, the team aims to uncover behavioral trends that can be used to cluster users into distinct personas.

In essence, this project's key research questions are: How can we define and measure adherence? What are the key predictors of adherence? Can we segment users by behavioral patterns to tailor interventions?

II. Literature Review

First, understanding the factors that influence user adherence is important to building effective behavior change platforms. Prior research has demonstrated the value of using predictive models to identify these factors. For example, one study developed an intelligent system using ANNs and a genetic algorithm to predict adherence to diet plans. The model incorporated lifestyle variables such as age, weight, BMI, education level, marital status, smoking status, and physical activity [1]. By identifying which variables most significantly impacted adherence, the study demonstrated how machine learning techniques can uncover patterns that inform personalized health recommendations.

In a related study, researchers used cluster analysis to explore changes in physical activity and sedentary behavior among patients undergoing cardiac rehabilitation. Participants were grouped based on baseline characteristics and behavioral shifts during and after the intervention [2]. The analysis revealed distinct user clusters, each with different support needs, underscoring the importance of tailoring interventions to specific user profiles.

Together, these studies highlight the potential of predictive modeling and clustering approaches to support adaptive, individualized health guidance—an approach aligned with MyYouthspan's mission to keep users engaged through personalized, achievable recommendations.

III. Description of the Data

The MyYouthspan dataset includes 187 rows of self-reported user data across a wide array of wellness categories, including nutrition (e.g., meat-processed, fruits-and-veggies, refined-sugar), movement (e.g., cardio-low, strength-training, balance, stretching), supplementation (e.g., vitamin-d, omega-3, zinc), and lifestyle factors (e.g., alcohol, cigarettes, sleep, sauna-frequency). Each row represents one user's log for a given day and target values from their plan, with additional metadata such as user-id, submission-date, and completed status. Because users select and track their own goals, not every behavior is logged each day, resulting in a dataset that is both personalized and sparse. Additionally, many variables contain zero or null values, reflecting either non-participation or non-response. This sparsity, combined with the small sample size, limits the robustness and generalizability of our models. As a result, the clustering and recommendations we present should be interpreted with caution.

To address these limitations, we suggest future modeling approaches that go beyond unsupervised clustering to include more predictive models of adherence. These models can be used by the MyYouthspan team once more robust data collection becomes available, and could enable proactive engagement strategies by anticipating which users are most likely to drop off before it's too late.

IV. Methodology and Results

A. Defining Adherence

Defining adherence was challenging due to the large number of zero and null values. Different users had varying amounts of defined features in the plan data, making comparing adherence even more complicated. For each user, we decided to only include the log values for columns that were also defined in the plan data for that user. For a goal to be considered defined in the plan data, the variable had to be non-null and non-zero (unless zero was a reasonable goal like with cigarettes or alcohol). We then defined adherence as the number of log values that met or exceeded the goal divided by the number of defined values in the plan. For example, if 4 goals were defined and 2 were met, adherence would be 50%. Meeting or exceeding the goal was feature-dependent. For sugar and alcohol intake, being below the goal is considered exceeding the goal. For fruits and vegetables or exercise, being above the goal is considered exceeding the goal.

B. Identifying Key Features with PCA

To explore the structure of the dataset and reduce dimensionality, Principal Component Analysis (PCA) was applied to the cleaned data. Given the high number of features and sparsity in the dataset, PCA helped identify the most influential patterns while minimizing noise. The first three principal components explained a meaningful proportion of the variance (approximately 63%) and offered interpretable insights into user behavior. The first component had strong positive loadings on fitness, nutrition, and social goals, capturing overall engagement with health-related plans. The second component reflected variation in how consistently users logged their activities, while the third component appeared to separate users based on their tendency to meet versus exceed goals.

C. Clustering Analysis of Users with K-Means

We conducted an unsupervised clustering analysis using K-Means to identify users with similar patterns of adherence and ways of using the software product. The first step was to group variables into four categories – diet, lifestyle, exercise, and supplements. We suspected based on the PCA that some users may only be interested in one or two of the categories. Defining adherence by category allowed us to dive deeper into investigating if this held true. Adherence was defined as described earlier except broken out by category. For the overall adherence variable, it was the average of the four categorical adherence values. After adherence was defined for each category, we took the mean value for each unique user and ran the clustering analysis using the K-Means machine learning model. We got the best results (balance of most distinct clusters and interpretability) by only including adherence for each category and using 3 clusters (k=3). Including the variables for total number of log entries and overall adherence led to less distinct clusters. The silhouette score of the resulting clusters, which ranges from -1 (poor fit) to +1 (great fit), was 0.314. The following two tables and figure summarize the variable values for each cluster.

TABLE I: Adherence Metrics by Cluster

Cluster	**Diet**	**Lifestyle**	**Exercise**	**Supplement**	**Overall**
Light	0.136	0.180	0.084	0.011	0.103
Core	0.352	0.627	0.052	0.009	0.260
Power	0.485	0.670	0.249	0.628	0.508

TABLE II: Entry and User Counts by Cluster

Cluster	**Log Entries**	**User Counts**
Light	13.20	20
Core	28.14	14
Power	29.50	10

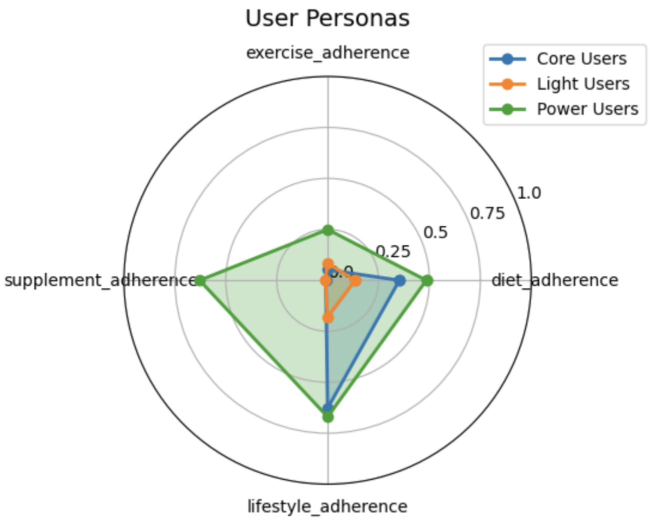

Fig. 1: User Cluster Comparison

The Light Users have low adherence in every category and low overall adherence. They also have far fewer log entries than the other two personas. The Core Users have moderate adherence in diet and lifestyle but still low adherence overall (although much higher than the Light Users). These users seem to be primarily interested in tracking diet and key lifestyle variables such as sleep and water intake. The final group, Power Users, uses every category. They have low adherence in exercise but moderate adherence in both diet and supplements. They have high adherence in lifestyle and moderate adherence overall (almost double that of the Core Users). Both Core Users and Power Users have more than double the log entries of the Light Users. To gain some insight into adherence over time, we selected a few users with a larger number of log entries and graphed the overall adherence over time.

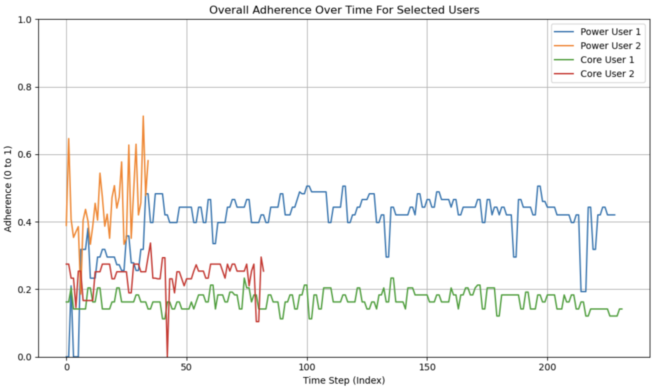

Fig. 2: Adherence Over Time for Selected Users

From the above figure it can be seen that adherence was fairly stable over time. The users did not start strong and then fall off over time. As additional data becomes available, adherence over time should be further investigated to see if these findings hold true or if there are differences between personas.

The clustering analysis gave us three distinct groups of user personas. These personas could be used to help inform additional features and marketing efforts. They could also be used to create persona-level strategies, such as reminders or different goal recommendations, to help more users reach their health goals. These personas should be further refined with additional data and future analysis should include evaluating how user plans compare to initial values when users start using the software.

D. *Predicting Adherence with XGBoost*

We then ran an XGBoost model as a more quantitative way of modeling adherence. In this context, adherence is calculated as the proportion of goals achieved in the log data over goals set in the plan data. For each user, only variables that have non-null/non-zero values in the plan data are used for this calculation of adherence. For example, if a user has filled out 10 goals in the plan data and achieved 5 of those goals in the log, they have 50 percent adherence. The definition of success for each variable is determined based on if exceeding or going below the goal is beneficial to health.

The XGBoost model uses the machine learning gradient boosting method to build an ensemble of decision trees for predicting the adherence proportion. It is well-suited for regression tasks and has built-in mechanisms to handle sparse data. For missing values, XGBoost learns the optimal direction to take at each split. It selects the features and splits based on their contribution to reducing prediction error.

We tested 3 models: 1) using all of the plan and log variables, 2) a pruned model dropping all variables below the median importance, 3) and a SHAP-pruned model. The reason for testing pruned models was to reduce dimensionality of the data and prevent overfitting. However, we actually find that the first model with all predictors yields the lowest RMSE, the highest R-squared, and does not show significant signs of overfitting when evaluating using K-fold cross-validation with 5 folds. In fact, the model with all of the predictors only saw a standard deviation of 2 percent for R-squared and a standard deviation of 0.0038 for RMSE when comparing between the folds. The model diagnostics on the test dataset for each model are shown below.

TABLE III: XGBoost Model Diagnostics

Model	RMSE	R^2
All Predictors	0.0322	92.95%
Median Importance Pruning	0.0401	89.09%
SHAP Pruning	0.0694	67.32%

The learning curve below shows that the XGBoost model with all of the parameters begins to generalize better as more data is added. The study did contain limited data, and ideally as more data becomes available the Validation and Training RMSE lines would continue to move towards convergence.

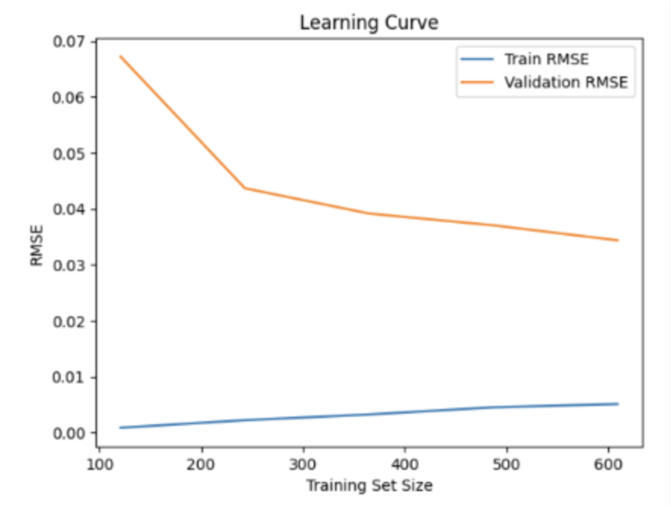

Fig. 3: Learning Curve of the XGBoost Model

The following figure shows that the XGBoost adherence model fits the actual adherence data well.

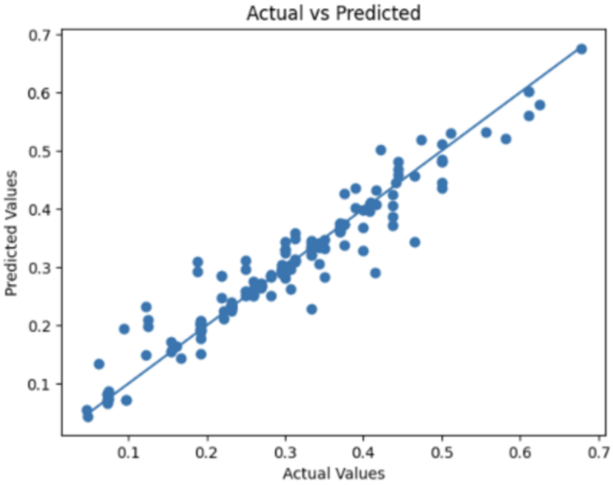

Fig. 4: XGBoost Model Fit

E. Future Clustering using Recurrent Neural Networks

Our goal using Recurrent Neural Networks is to gain a deeper, more granular understanding of user adherence patterns, including an adherence variable corresponding to each individual attribute. Currently, with the available data, directly applying Recurrent Neural Network (RNN) models does not yield usable results. However, as additional data is collected, these models can be run to provide significant insights.

To use these models, we created a categorical variable corresponding to each logged variable. Failure to adhere to the planned value is represented as -1, successful adherence is represented as 1, and untracked variables are represented as 0. We defined four groups of variables with different rules on what counts as adhering. The first group is ratio_range. If the logged value is within a certain percentage (25%) of the planned value, the user is considered to be adhering(dairy, sleep). The second group is adhering if the logged values are greater than the planned value(water, cardio). The third group is adhering if the logged values are less than the planned value(alcohol, cigarettes). The fourth group are binary variables so the value must match exactly(calorie_restriction, fasting). Each rule also considers whether a 0 value in the plan should be considered as an actual goal or simply an untracked variable for that user.

Recurrent neural networks require a sequence as an input. In our case, we are providing each user's logs as a time series. Since each user has submitted a different number of logs, we must add extra rows filled with the value 99 until each user's sequence has the same number of rows as the user with the maximum. This padding allows the model to process users with shorter histories while retaining as many entries as possible. Currently, the data only allows for the model to be run using a sequence length of 5. Most users have fewer than 5 real entries, with many having only, one which does not enable the model to evaluate time-dependent adherence. If the recurrent neural network receives too many padded entries, it will not be able to effectively train. When these models are trained again in the future, the sequence length should be adjusted to include as many entries as possible while excluding users without sufficient entries.

The basis for each of our RNN models is the Long Short-Term Memory layer. This layer contains gates that control what information gets stored in the model over a time series. To create embeddings for users, we build an encoder by training an autoencoder. The autoencoder consists of an encoder that turns the time series into an embedding. A decoder then tries to reconstruct the original time series from the encoding. As the loss, measured using mean-square error, decreases, the encoders ability to capture relevant information about users improves. The encoder is then used to create a unique embedding for each user. These embeddings are analyzed using clustering analysis techniques such as the k-means used previously.

Four model architectures were deployed, each utilizing a hyperparameter search to optimize the final model values. The first model contains a single LSTM layer for both the encoder and decoder. A second model contains 2 stacked LSTM layers for both the encoder and decoder. In the third model, we introduced a bidirectional LSTM layer for the encoder and a standard LSTM layer for the decoder. For the fourth model, we stacked 2 bidirectional LSTM layers for the encoder and 2 standard LSTM layers for the decoder. Hyperparameter tuning suggests that larger numbers of latent dimensions increase performance. The lowest loss on preliminary data was using the single bidirectional LSTM layer model.

When training is finished, the encoder produces a unique embedding for each user. Once a clustering analysis is complete, each user can be assigned to a group. By looking at mean adherence values across these groups, we can gain more variable-specific insights. We initially performed the clustering analysis with 3 groups to try to match our previous clustering analysis. However, due to the number of features, more than 3 groups are likely to emerge from the final clustering analysis. With the limited current data, increasing the number of groups simply puts individual users into their own group, which is not informative. The elbow and silhouette methods are provided to determine the optimal number of groups.

The following are examples of heatmaps produced by the encoder. Note that these results are preliminary due to the lack of data. However, even with the lack of data, it is clear the model has learned to distinguish somewhat between users who are using the platform and those who are not. Across each category of features, cluster 1 shows only 0 values, indicating a large number of users who are not active on the platform at all. In contrast, cluster 2 shows variables filled out across all categories, including some supplements, as well as significant success in adhering to their plan. Cluster 0 contains people trying to use the platform some, but that are struggling to adhere to their diet and exercise goals.

V. Conclusion

The PCA served as a valuable pre-processing step for clustering analysis. While PCA reduces the interpretability of

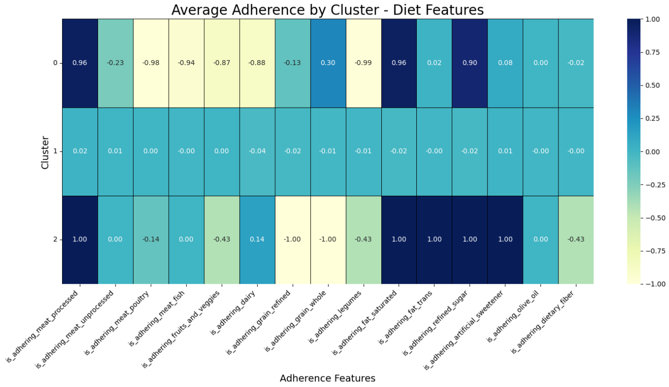

Fig. 5: Diet Features Heatmap

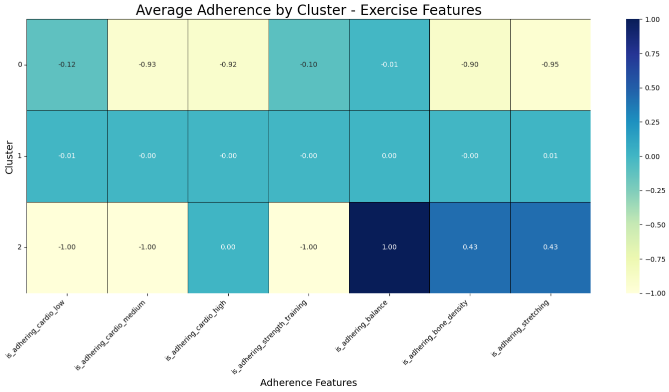

Fig. 6: Exercise Features Heatmap

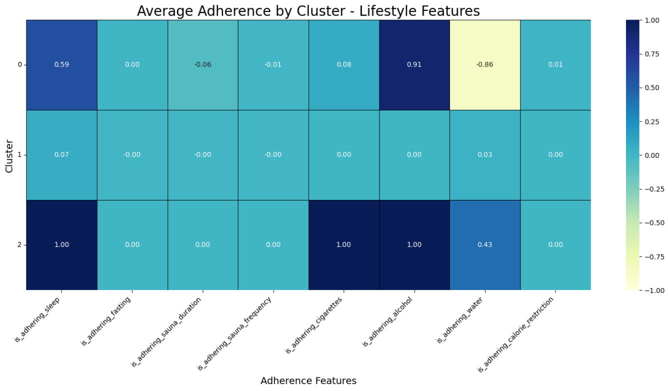

Fig. 7: Lifestyle Features Heatmap

individual features and provided a necessary foundation for uncovering behavioral profiles within the data.

Initial K-means clustering identified 3 main groups of users - light users, core users, and power users. Each group uses the software differently in which categories of variables they are tracking and their degree of adherence. This finding is supported by preliminary recurrent neural networks. Further groups may come to light as the recurrent neural networks are retrained on a more full dataset.

The XGBoost model provides an approach to predict a user's adherence to their plan data. The model fit the data well with an R^2 value of 92.95%, which indicates that 92.95% of the variability in adherence can be explained by the model. Overall, this model can help MyYouthspan predict how well a user may adhere to their plan, and intervene accordingly to help raise adherence.

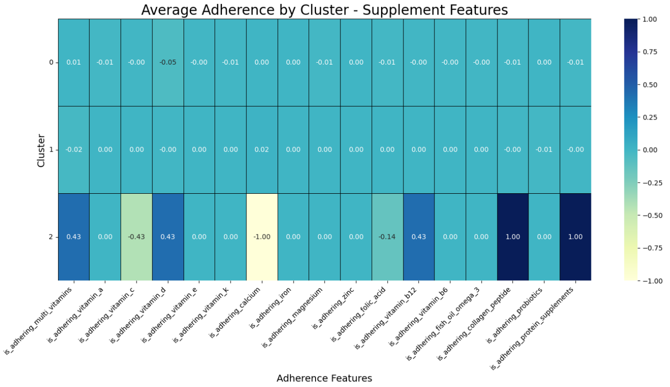

Fig. 8: Supplement Features Heatmap

The best Recurrent Neural Network model was the single bidirectional which enabled assessing user entries as a time series. Most users have less than 5 entries so the models are not very informative. All four models should models should be rerun when more data is available. A basic hyperparameter is implemented but can be tuned as the models progress. Currently the model views adherence at the individual feature level but features can also be evaluated as a group.

Future work should refine the user personas and predictive models as more data is collected. Research should be done on whether different personas would benefit from different types of support. Support could be in the form of reminders to fill out a log entry or different initial goal recommendations.

VI. Student Contributions to Project

Our team as a whole developed definitions of adherence and worked through cleaning the data. Karunya Iyappan developed the PCA exploration, wrote the introduction and data description, and coordinated sponsor communication. Sam Knisely developed the XGBoost model to predict adherence. Alec Pixton developed the Recurrent Neural Network models. Trenton Ribbens conducted the K-Means clustering analysis to develop the personas and performed EDA.

References

[1] Hediye Mousavi, Majid Karandish, Amir Jamshidnezhad, and Ali Mohammad Hadianfard, “Determining the effective factors in predicting diet adherence using an intelligent model,” https://www.nature.com/articles/s41598-022-16680-8

[2] Marlou M. Limpens, Rita J. G. Van den Berg, Iris den Uijl, and Madoka Sunamura, “Physical activity and sedentary behaviour changes during and after cardiac rehabilitation: Can patients be clustered?”, https://www.researchgate.net/publication/372309631_Physical_activity_and_sedentary_behaviour_changes_during_and_after_cardiac_rehabilitation_Can_patients_be_clustered

[3] Ermal Elbasani and Jeong-Dong Kim, “LLAD: Life-Log Anomaly Detection Based on Recurrent Neural Network LSTM,” Journal of Healthcare Engineering, https://pmc.ncbi.nlm.nih.gov/articles/PMC7932773/

[4] Vijendra Pratap Singh, Manish Kumar Pandey, Pangambam Sendash Singh, and Karthikeyan Subbiah, “An LSTM Based Time Series Forecasting Framework for Web Services Recommendation,” Computacion y Sistemas, https://www.researchgate.net/publication/342701624_An_LSTM_Based_Time_Series_Forecasting_Framework_for_Web_Services_Recommendation