**Understanding Engagement for Addiction Treatment**

Alex Brulliea1, Tyler Staffin1, Labib Nazer1, Alvin Mathew1, James Grissel1, Satish Srinivasan1, Jesse Gardner 1, Dusan Ramljak[[\*]](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=en-US&wopisrc=https%3A%2F%2Fpennstateoffice365-my.sharepoint.com%2Fpersonal%2Fdzr5484_psu_edu%2F_vti_bin%2Fwopi.ashx%2Ffiles%2F4defb11ccf7e45528ee4b77de654bd68&wdenableroaming=1&mscc=1&wdodb=1&hid=9CBC33A0-D0FC-1000-A6BE-D59939586886&wdorigin=Other&jsapi=1&jsapiver=v1&newsession=1&corrid=452aa150-b940-41f9-b8ba-751028158a56&usid=452aa150-b940-41f9-b8ba-751028158a56&sftc=1&mtf=1&sfp=1&instantedit=1&wopicomplete=1&wdredirectionreason=Unified_SingleFlush&rct=Medium&ctp=LeastProtected" \l "_ftn1)1, Pierce Baugh2, and Adi Jaffe2

1School of Graduate Professional Studies  
 The Pennsylvania State University  
 Malvern, PA, 19355, USA

2igntd.com   
University or Company  
 City, State, Zip Code, Country

***Abstract***

There are approximately 20 million individuals in the US with a substance use disorder. Smart Personalized Adaptive Recovery System SPARx is the first recovery fully online platform that helps treating addiction. It is customized, interactive, discrete, and shame-free, all while being easy to use and affordable experience. We used its database of 432 patients whose anonymized data shows engagement in the period of 1 year to try to understand how to improve user engagement that could help in addiction treatment.

**Keywords:** Engagement, Addiction Treatment, Machine Learning, Data-Driven Decision-Making

**1. Introduction**

The global opioid epidemic is challenging our current understanding of managing and treating addiction, posing a significant financial burden and strain on our already over-utilized medical systems. It is difficult to identify the appropriate intervention for a person struggling with mental health or addictions. As is the case with other modern challenges, the introduction of data-driven technology could address this issue and improve the likelihood of long-term patient success. Data-driven decision making technology could aggregate data across many individuals to identify both intrapersonal (data within a specific person on preferences and likes/dislikes) and interpersonal (data across people on preferences based on grouping variables and more) features that would lead to appropriate intervention.

SPARx, a Smart and Adaptive Recovery System developed by IGNTD, utilizes machine learning and customized programming to help individuals throughout the addiction recovery process. SPARx aims “to transform the current addiction industry [by promoting] empowerment, solution-oriented and compassionate recovery.” (IGNTD Announced the launch of SPARx, A Smart Personalized Adaptive Recovery System [1] . In this work, we decided to pursue engagement and liking first, because those are more proximal outcomes and that, without engagement, the probability of improved long-term outcomes are low (given the disengagement). Therefore, we can look at improving engagement and subjective assessment of the material/intervention to improve the length and frequency of engagement. Long-term platform usage is critical to the success of all addiction patients; thus, it is essential that IGNTD continues to identify and implement methods for improving user engagement. This research will serve to address the viability of the SPARx platform through the scope of the following question: Can data-driven techniques be leveraged to increase patient engagement on the SPARx platform?

We tried to answer the following questions to help support the overall research goal more effectively. What content should we be offering users of the platform? What the platform usage patterns are? How do different types of users engage? Given each user's history, how are they more likely to engage with the platform next? What are the best types of content to present to ensure patients are engaging with the platform and progressing in their road to recovery?

[[\*]](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=en-US&wopisrc=https%3A%2F%2Fpennstateoffice365-my.sharepoint.com%2Fpersonal%2Fdzr5484_psu_edu%2F_vti_bin%2Fwopi.ashx%2Ffiles%2F4defb11ccf7e45528ee4b77de654bd68&wdenableroaming=1&mscc=1&wdodb=1&hid=9CBC33A0-D0FC-1000-A6BE-D59939586886&wdorigin=Other&jsapi=1&jsapiver=v1&newsession=1&corrid=452aa150-b940-41f9-b8ba-751028158a56&usid=452aa150-b940-41f9-b8ba-751028158a56&sftc=1&mtf=1&sfp=1&instantedit=1&wopicomplete=1&wdredirectionreason=Unified_SingleFlush&rct=Medium&ctp=LeastProtected" \l "_ftnref1) Corresponding author: Tel.: (610) 648-3299; E-mail: [dusan@psu.edu](mailto:author@company.com)

The rest of the manuscript is organized in the following way. Section 2. contains dataset description and detailed descriptive analysis. Section 3 addressed the methodology used to answer the above questions. Section 4 provides the prescriptive solutions and Section 5. concludes the paper.

**2. Dataset Exploration and Analysis**

***2.1. Data Exploration***

The Data Dictionary for the database and underlying columns can be found in [Table 1.1](#_Appendix) in the Appendix.

In this worksheet, there are 2 tabs. The first is all the column-level names, definitions, measurement scales, data types, null counts, relevance flag (for the analysis), and any associated notes with the column. The second is all the table-level names, definitions, and row counts.

We loaded all data files sent over by the IGNTD team into an Azure MS SQL relational database to provide accessibility to all team members and establish an environment for query execution. We have also set up a [private GitHub repository](https://github.com/labibnazer/DAAN-881-Project) to manage version control for all code employed for the remainder of this project.

The data provided by IGNTD consists of 10 delimited text files. These files fit the profile for data extracted via IGNTD’s web application relational database in that they are normalized with primary and foreign key references. There are also numerous metadata fields including record creation and update date/time stamps.

The tables contain logs of the users’ assessments and engagements while using the SPARx platform. Of the available data fields, 29 were initially have been identified as key fields, which are integer values used to uniquely define records and make references to one or more records in other tables. Although they are essential to the database model, these fields will not be used in analysis.

The table with volume of empty strings of each dataset has been included in Table 1 below (note that all empty strings and NULL values were converted to NA). The statistical summary of all the fields by table in the dataset are shown in the Figure 6 in the Appendix.

*Table* 1*: Selected Predictor Variables*

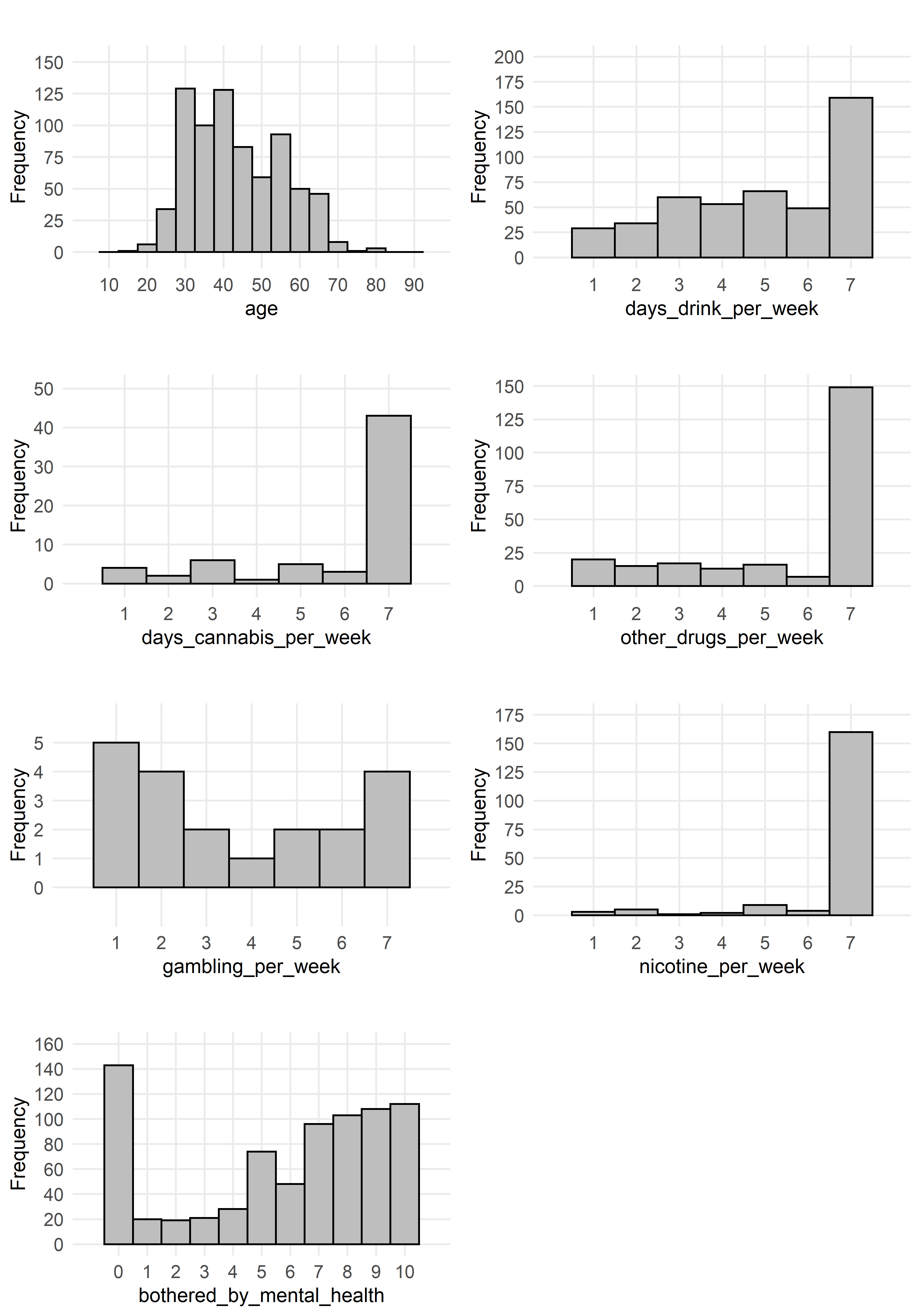
|  |  |
| --- | --- |
| **Table** | **Number of empty strings** |
| completed\_lessons | 2280 |
| daily\_assessment | 2214 |
| intake\_assessment | 6208 |
| lesson\_ratings | 506 |
| sds\_intake\_question\_options | 304 |
| sds\_user\_activations | 3078 |
| site\_visits | 9066 |
| user\_points | 0 |
| user\_rewards | 0 |
| wol\_items | 0 |

***2.2. Exploratory Analysis***

When performing any statistical analysis or developing models, a crucial preliminary step involves thorough quantitative and qualitative analysis of all available data. The Cross-Industry Standard Process for Data Mining (CRISP-DM) advocates for the importance of proper data understanding prior to pursuing any data preparation or model development steps. The initial dataset consists of 13 tables, each of which contains a ranging number of variables and records. After further assessment of each variable and its relevance to the research question, the team decided to focus on the following 18 features as presented in Table 2:

*Table* 2*: Selected Predictor Variables*

|  |  |  |
| --- | --- | --- |
| **Table** | **Variable** | |
| **Name** | **Description** |
| intake\_assessment | age | User age |
| sds\_user\_activations | gender | User gender |
| intake\_assessment | days\_drink\_per\_week | Days per week user drinks |
| intake\_assessment | days\_cannabis\_per\_week | Days per week user uses cannabis |
| intake\_assessment | other\_drugs\_per\_week | Days per week user uses other drugs |
| intake\_assessment | gambling\_per\_week | Days per week user gambles |
| intake\_assessment | nicotine\_per\_week | Days per week user uses nicotine |
| intake\_assessment | bothered\_by\_mental\_health | Scale of how bothered user is by mental health habits from 0 to 10 |
| sds\_user\_activations | welcome\_video | If user watched welcome video |
| sds\_user\_activations | bought\_hero | If user bought Hero package |
| sds\_user\_activations | intake | If user has finished an intake assessment |
| sds\_user\_activations | daily\_assessment | If user has completed a daily assessment |
| sds\_user\_activations | started\_lesson | If user has started a lesson |
| sds\_user\_activations | upload\_worksheet | If user has uploaded their worksheet |
| sds\_user\_activations | rated\_lesson\_1 | If user has rated the first lesson |
| sds\_user\_activations | finished\_module\_1 | If user has finished the first module |
| sds\_user\_activations | created\_community\_acct | If user has created a community account |
| sds\_user\_activations | completed\_lesson | If user has completed a lesson |

As described in the prior section, the team developed an Azure MS SQL relational database to properly store and maintain all provided data. The R statistical computing language was utilized in this phase and for the remaining steps of the CRISP-DM process. The team prepared histograms for each continuous predictor using ggplot2 to derive an improved quantitative perspective of their underlying distributions, as displayed in Error: Reference source not found. Since most of the instances for variables capturing “number of days” were zeros, the distributions were significantly skewed. To address this (and for visual clarity), all zero values were removed from the histogram plots, but not removed from the underlying data.

*Figure* 1*: Histograms of Continuous Predictors*

As mentioned above, most of the variables expressed frequency values of 0. This is likely due to a combination of both true and inaccurate responses, as individuals may be reluctant to share truthful data about their drug, alcohol, and gambling habits. Although the data for these self-reported variables is contingent upon the validity of the user responses, it is critical that all subsequent analysis and model evaluation processes operate under the assumption that the information recorded by users of the SPARx platform is accurate.

Evidently, the age of SPARx platform users varies widely, with the lowest and highest recorded ages (excluding zeros) are 16 and 79, respectively. The average and median ages are currently 41.91 and 41 years, respectively, with a standard deviation of approximately 14.1 years. It appears that most platform users fall between 30 and 40 years of age, but there is a large volume of users with ages on the higher and lower ends of the distribution. Through the model development and evaluation process, the team will certainly explore the impact of user age on their continued engagement with the SPARx platform. The summary statistics for the remaining continuous variables were computed and displayed as follows in Table 3:

*Table* 3*: Summary Statistics for Continuous Predictors*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Min.** | **Q1** | **Med.** | **Mean** | **Q3** | **Max.** | **SD** |
| age | 0 | 32 | 41 | 41.91 | 53 | 79 | 25.89 |
| days\_drink\_per\_week | 0 | 0 | 2 | 2.84 | 6 | 7 | 2.97 |
| days\_cannabis\_per\_week | 0 | 0 | 0 | 0.49 | 0 | 7 | 2.82 |
| other\_drugs\_per\_week | 0 | 0 | 0 | 1.73 | 3 | 7 | 2.76 |
| gambling\_per\_week | 0 | 0 | 0 | 0.1 | 0 | 7 | 2.85 |
| nicotine\_per\_week | 0 | 0 | 0 | 1.58 | 0 | 7 | 2.8 |
| bothered\_by\_mental\_health | 0 | 3 | 7 | 5.8 | 9 | 10 | 3.76 |

Additionally, there are several Boolean variables the team identified as being potentially relevant for answering the business question. Although the original dataset did not characterize these variables as Boolean, the team modified their variable classes to accurately reflect their types. While the first factor is gender (descriptive of the user), the remaining variables relate to the engagement with and usage of the SPARx platform. The bar plot in Error: Reference source not found captures these factors in terms of their relative percentage of the total number of instances:

Timeline

Description automatically generated

*Figure* 2*: Bar Plot of Boolean Variables and % of "True" Values*

The above bar plot provides some interesting initial observations into the refined data subset. It appears that roughly two-thirds of the platform users are female, with a population consisting of 292 females and 140 males. 68 instances of gender were removed due to NULL or “other” values (specifically, 67 NULL and 1 “other”). Most users completed an intake assessment (~88%), and just over two-thirds of users reported having completed at least one daily assessment. While 65% of users started the first lesson, a much smaller number of participants completed lesson in its entirety. Just about 15% of platform users successfully completed of the first module. Throughout this research, the team will look to identify which factors contributed to longer platform engagement to ensure that users continue to monitor their treatment process with the SPARx platform.

For correlations between values, we decided to focus on the numeric values of the intake assessment. Based on the values, there were no outstanding unexpected correlations negatively or positively. 2 correlations of interest, however, were nicotine\_per\_week and cigarettes\_per\_day both positively correlated with sex\_issues\_per\_week and other\_drugs\_per\_week. Based on that correlation, we could make the assumption that those who have sex issues and other drug use tend to also have a reliance on nicotine. The full correlation matrix can be seen in figure 4 below, however, a more easily readable document called CorrelationMatrix.xlsx has been attached for reference.

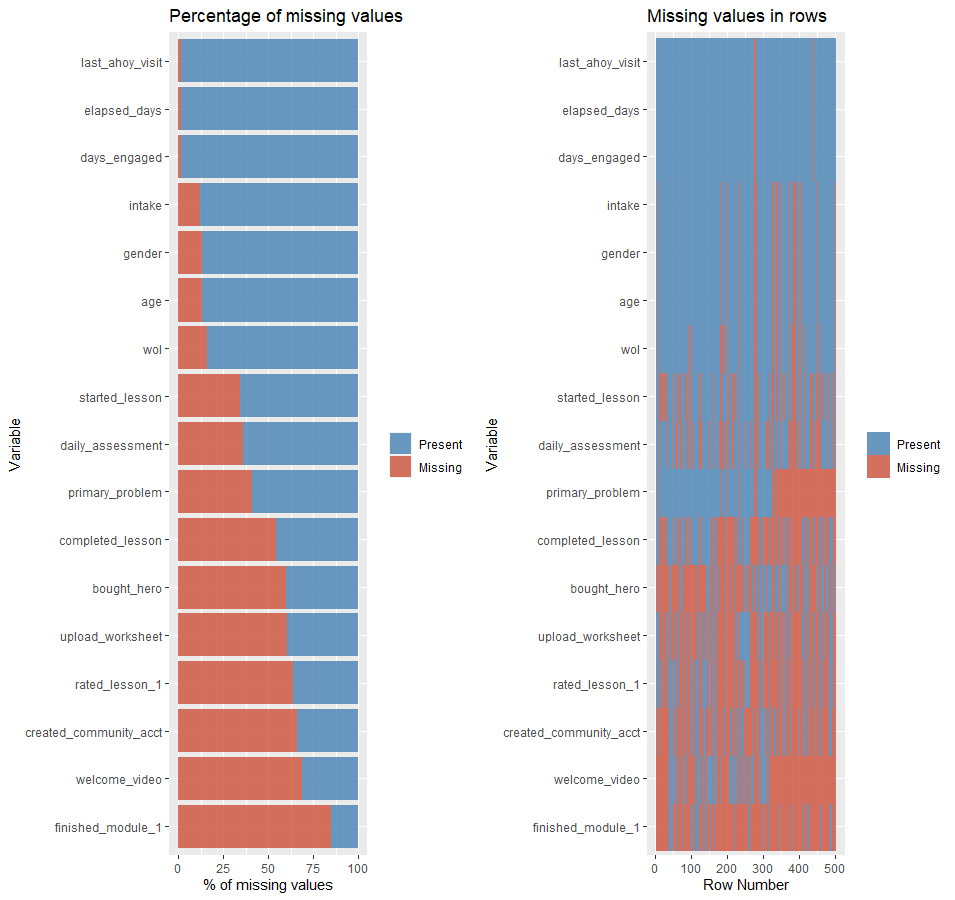
Graphical user interface

Description automatically generated

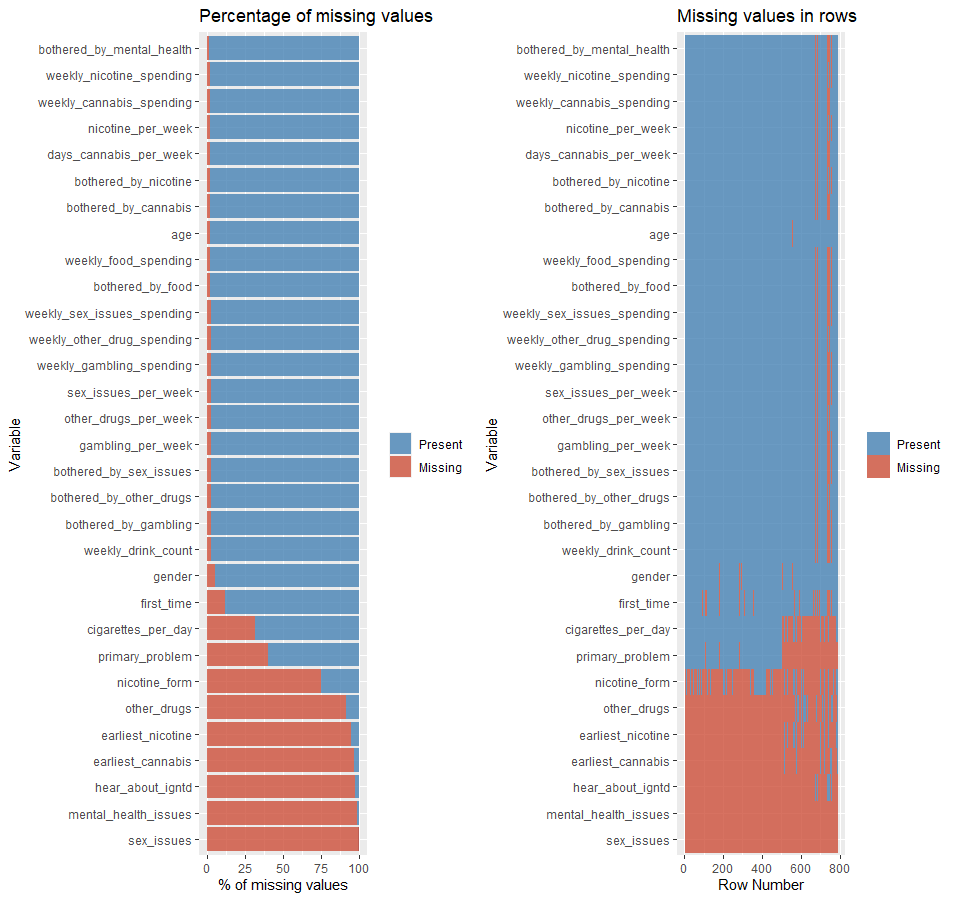
*Figure* 3*: Correlation Matrix of intake\_assessment*

***2.3. Missing Values***

To visualize missing values in the data, any NULL instances were converted to NA values and plotted. The tables sds\_user\_activations and intake\_assessment contain the variables selected for analysis, and the number of present and missing values for all variables in these tables is captured in Error: Reference source not found below. We can see that many of these fields contain missing data and that some fields are missing data in the majority of instances.



*Figure* 4*: Bar Plot of Missing Values in sds\_user\_activations by Variable and Row*



*Figure* 5*: Bar Plot of Missing Values in intake\_assessment by Variable and Row*

***2.4. Data Issues***

The following list captures the data issues identified throughout the initial phases of this research. Additionally, potential solutions are listed below each issue and will be modified accordingly as solutions are implemented in future deliverables:

1. Boolean variables listed as “NA” instead of 0
   * To address this issue, the team with utilize R’s is.na() function on all boolean variables and set all instances where is.na() is TRUE to 0. This should resolve this data issue.
2. Boolean data is inconsistently recorded
   * For example, the first\_time field has “TRUE” and “FALSE” values in the intake\_assessment table, whereas the bought\_hero attribute in the sds\_user\_activations tables contains either a “1” or “NULL”. The team will need to standardize these values in the data.
3. NULL values prevalent in data
   * NULL values are represented in inconsistent ways in the character variables. For example, the mental\_health\_issues field has four different representations of NULL, including “NULL”, “{NULL}”, “{}”, and “”. The team will need to implement a consistent pattern for representing NULL values across all variables.
   * NULL values are represented as “NULL” character values in columns that should only contain integers. For example, the fields bought\_hero and upload\_worksheet demonstrate this issue, while others including wol and daily\_assessment do not. The team will need to implement a consistent solution for representing NULLs across all integer fields.
4. Outlier values in data (i.e., age has values of 0, etc.)
   * Outliers will be addressed on a case-by-case basis, with potential systematic solutions resulting from outlier detection methods.
5. Some of the tables have entries for user\_id (1629, 1630, and 1631) that do not exist in sds\_user\_activations, which should contain all users. The team will need to decide whether to omit this data entirely or create dummy user\_id values.
6. The intake\_assessment table contains more than one response in some cases for specific users. After clarification from our collaborators, we now understand that this is intended within the data, as intake assessments are supposed to occur monthly for each user.
7. sds\_intake\_question\_options was provided as a table with enum descriptions that give information about what the data in the tables refer to. These will need to be mapped to the corresponding data fields in the intake\_assessment table.

***2.5. Methodology and Model Building***

To address the previously mentioned data issues in Deliverable 3, the team has taken the following actions to ensure the issues are mitigated using data cleansing techniques in both SQL and R.

1. NA values in Boolean variables were set to zeroes.
2. TRUE/FALSE values were converted to dummy/binary values for consistency with Boolean variables within the IGTND platform.
3. NULL values across the tables were inconsistent based on the IGTND platform’s front-end usage, therefore the team normalized the NULL values irrespective of the data type.

|  |  |  |
| --- | --- | --- |
| **SDS\_USER\_ACTIVATION**  **(Pre- Boolean conversion)** | | |
| Column | Column Value | Count |
| bought\_hero | 1 | 201 |
| bought\_hero | NULL | 299 |
| completed\_lesson | NULL | 273 |
| completed\_lesson | 1 | 227 |
| created\_community\_acct | 1 | 170 |
| created\_community\_acct | NULL | 330 |
| daily\_assessment | NULL | 179 |
| daily\_assessment | 1 | 321 |
| finished\_module\_1 | 1 | 75 |
| finished\_module\_1 | NULL | 425 |
| intake | NULL | 61 |
| intake | 1 | 439 |
| rated\_lesson\_1 | 1 | 182 |
| rated\_lesson\_1 | NULL | 318 |
| started\_lesson | NULL | 173 |
| started\_lesson | 1 | 327 |
| upload\_worksheet | 1 | 197 |
| upload\_worksheet | NULL | 303 |
| welcome\_video | 1 | 157 |
| welcome\_video | NULL | 343 |

|  |  |  |
| --- | --- | --- |
| **INTAKE\_ASSESSMENT**  **(Pre-Boolean conversion)** | | |
| Column | Column Value | Count |
| First\_Time | TRUE | 551 |
| First\_Time | NULL | 93 |
| First\_Time | FALSE | 141 |

|  |  |  |
| --- | --- | --- |
| **SDS\_USER\_ACTIVATION**  **(Post Boolean conversion)** | | |
| Column | Column Value | Count |
| bought\_hero | 0 | 299 |
| bought\_hero | 1 | 201 |
| completed\_lesson | 0 | 273 |
| completed\_lesson | 1 | 227 |
| created\_community\_acct | 0 | 330 |
| created\_community\_acct | 1 | 170 |
| daily\_assessment | 0 | 179 |
| daily\_assessment | 1 | 321 |
| finished\_module\_1 | 0 | 425 |
| finished\_module\_1 | 1 | 75 |
| intake | 0 | 61 |
| intake | 1 | 439 |
| rated\_lesson\_1 | 0 | 318 |
| rated\_lesson\_1 | 1 | 182 |
| started\_lesson | 0 | 173 |
| started\_lesson | 1 | 327 |
| upload\_worksheet | 0 | 303 |
| upload\_worksheet | 1 | 197 |
| welcome\_video | 0 | 343 |
| welcome\_video | 1 | 157 |

|  |  |  |
| --- | --- | --- |
| **INTAKE\_ASSESSMENT**  **(Post-Boolean conversion)** | | |
| Column | Column Value | Count |
| First\_Time | 1 | 551 |
| First\_Time | 0 | 234 |

The table sds\_intake\_question\_options contains text enum mappings that describe the data in the table intake\_assessment for gender, primary\_problem, and nicotine\_form. However, there are many observations in intake\_assessment with NULL values. To handle this problem, we created a SQL view titled intake\_assessment\_question\_texts and will leverage it as a replacement for the intake\_assessment raw data table. The SQL statement for this view is in Appendix B.

The team developed a script in R (Appendix C) to handle the remaining issues regarding NULL and missing (NA) values. The first step was to identify the columns in the intake\_assessment\_question\_texts view and sds\_user\_activations that contained NULLs, and then treat them as missing values. A function titled get\_null\_cols() was developed to search for NULL values (i.e., NULL, {NULL}, and {}) in these tables to determine where to resolve this problem. NULL values were only present in character variables, so the first step was to convert all NULL values to NAs. The presence of NULLs in numeric variables caused the values to be imported as characters, thus the classes of these variables were converted back to numeric. The class of the first\_time variable in the intake\_assessment\_question\_texts view was converted to logical to enable the proper conversion to the numeric variable type.

The next step involved the conversion of missing values in Boolean variables with zeroes, as the unique values for these variables only consisted of 1s and NAs. Lastly, zeroes were regarded as outliers in the age column of the intake\_assessment\_question\_texts view and replaced with NAs. All steps were necessary to ensure adequate preparation of the data prior to performing any model development-related steps.

**3. Methodology and Model Building**

In order to proceed with building our model,

***3.1. Feature Selection***

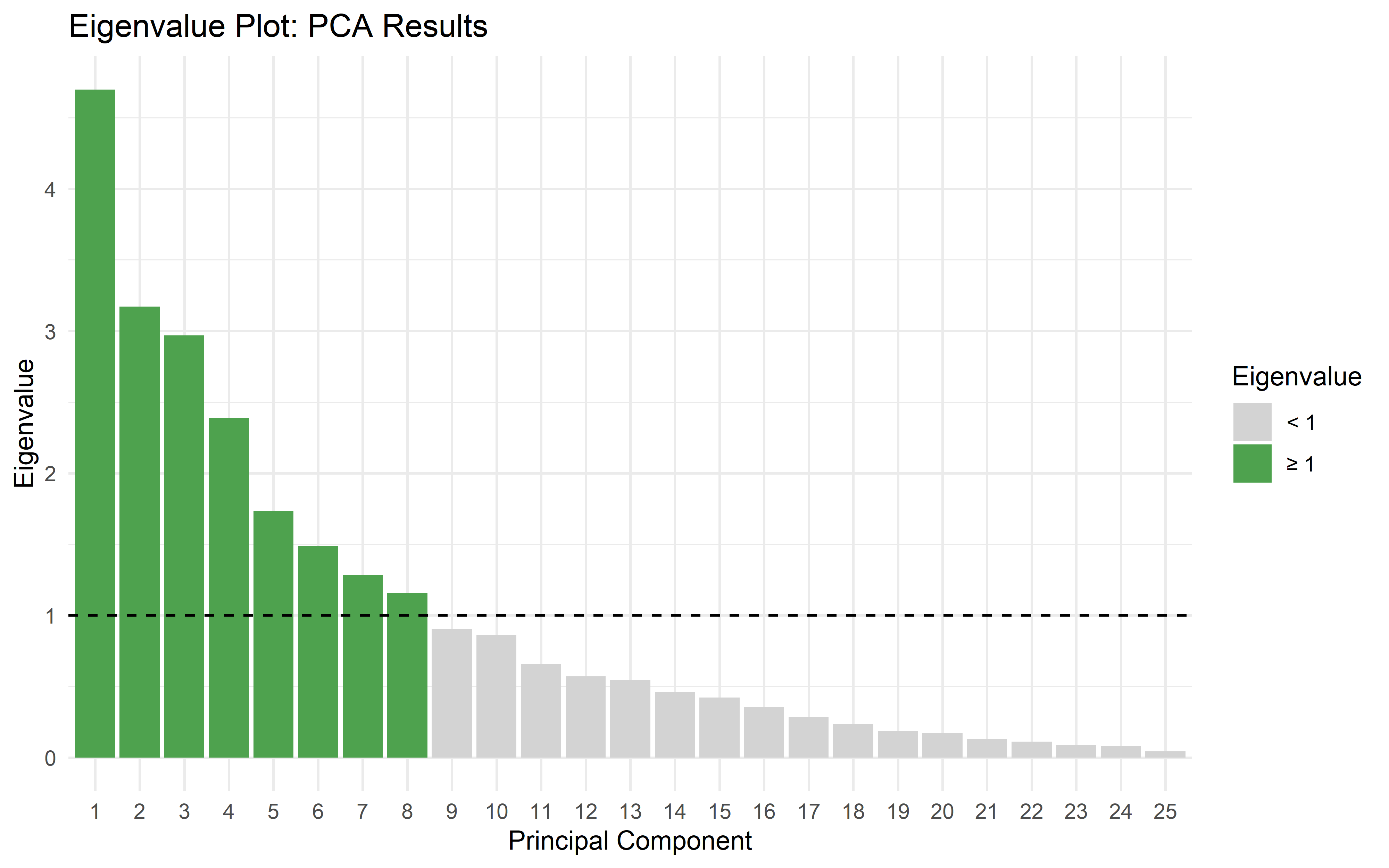
One of the major steps in preparation for model development is the identification and selection of variables. Not only does reducing the number of variables increase the speed of the modelling process, it can also reduce noisiness while improving both model performance and accuracy. The team employed various feature selection and engineering approaches in this step, and determined that the Principal Component Analysis (PCA) results would be used as predictor variables for fitting the proposed models in the next phase.

Prior to performing PCA, the data was split into training and test data subsets, with unique partitions of 80% and 20%, respectfully. Additionally, the team surveyed and evaluated a variety of other feature selection techniques, including Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), and the Random Forest Model (via the Boruta R package). The following sections outline these feature selection procedures in further detail.

**Principal Component Analysis (PCA)**

As discussed in Lecture 5, the use of PCA helps to reduce the amount of correlation in the underlying dataset through the elimination of redundancy. The team decided to leverage the Eigenvalue-One Criterion, in which principal components with eigenvalues greater than one are maintained, and the other principal components are discarded. The Proportion of Variance approach was also implemented, but required that a larger number of principal components were maintained (using a 90% variance threshold).

These computed principal components enable significant dimension reduction for this data subset. These results will be used to decrease the number of variables in this subset from 23 to 8 without drastically impacting the amount of explained variance. These resulting principal components will be combined with the remaining variables in the prepared data set and then used in the subsequent model development phase. Figure 6 graphically depicts the eigenvalues of each principal component, and distinguishes signficiant eigenvalues by color. Additionally, the team compiled the resulting percentage contribution of the predictor variables to each principal component, with these results presented in Table 4. The top 5 contributing variables for each principal component are highlighted in green to provide insight into the most influential variables within the training data. Subsequently, these PCA results of the will be applied to the test dataset to ensure the principal component structure is preserved when evaluating model accuracy on out-of-sample observations.



*Figure* 6*: PCA Eigenvalue Plot*



*Table* 4*: Variable Contribution to Principal Components*

**Random Forest Model (Boruta Algorithm)**

Random Forest models were also surveyed in the variable reduction process. The model identifies the importance of each variable in the decision-making process of the algorithm, outputting a numerical value of importance to how each variable interacted with the completed\_lesson output variable. For this, we utilized the Boruta algorithm, which is a wrapper around Random Forest. We had identified factors of potential relevance and chose to use these across all our explored variable reduction methods. When plugged into the Boruta training model and searching for predictors of completed\_lesson, the following was found:

Chart

Description automatically generated

*Figure* 7*: Boruta Feature Importance Boxplots*

The most important factor was weekly\_drink\_count, followed by bothered\_by\_mental\_health, weekly\_alcohol\_spending, other\_drugs\_per\_week, and bothered\_by\_gambling. Days\_drink\_per\_week and weekly\_other\_drug spending were just on the threshold and were left out of the final feature selection by Boruta.

Table

Description automatically generated

*Figure* 8*: Boruta Feature Importance and decision*

**Factor Analysis**

Factor Analysis (FA) was also evaluated to determine the underlying factors from which the observed variables were generated. Both Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were used for the variable assessments.

**Exploratory Factor Analysis**

Exploratory Factor Analysis (EFA) can be used to identify constructs that help explain covariation amongst a set of observed variables. In our case, we identified common factors derived from variables in the intake assessment survey. The team determined that the optimal number of factors was 8 after further analysis:

***Varimax Rotation***

Under a varimax rotation, the factors nearly perfectly lined up with a number of central themes. The output generated in R is presented in Figure 9. These factors and themes included:

**Factor 1:** Days\_drink\_per\_week, weekly\_drink\_count and bothered\_by\_drinking; aligned with “drinking” theme

**Factor 2:** Nicotine\_per\_week, cigarettes\_per\_day, bothered\_by\_nicotines; aligned with “smoking” theme

**Factor 3:** Gambling\_per\_week, weekly\_gambling\_spending, bothered\_by\_gambling; aligned with “gambling” theme

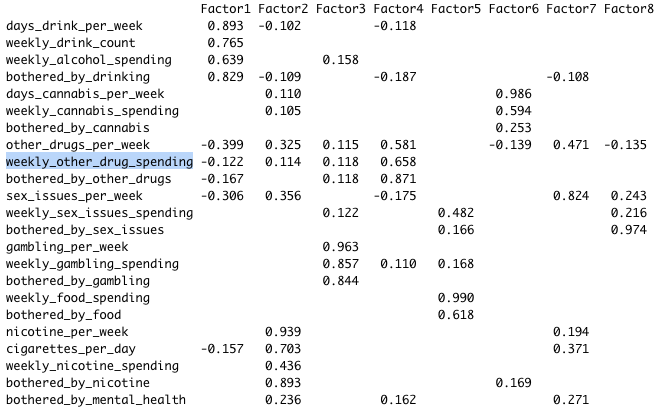
**Factor 4:** other\_drugs\_per\_week, weekly\_other\_drug\_spending, bothered\_by\_other\_drugs; aligned with “other drugs” theme

**Factor 5:** weekly\_food\_spending, bothered\_by\_food; aligned with the “food” theme

**Factor 6:** days\_canabis\_per\_week, weekly\_canabis\_spending, bothered\_by\_canabis; aligned with the “cannabis” theme

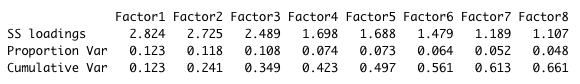
**Factor 7:** no apparent central theme for this factor, but consisted of variables pertaining to sex, drugs and smoking

**Factor 8:** weekly\_sex\_issues, sex\_issues\_per\_week, bothered\_by\_sex\_issues; aligned with the “sex” theme



*Figure* 9*: EFA Results Under Varimax Rotation*

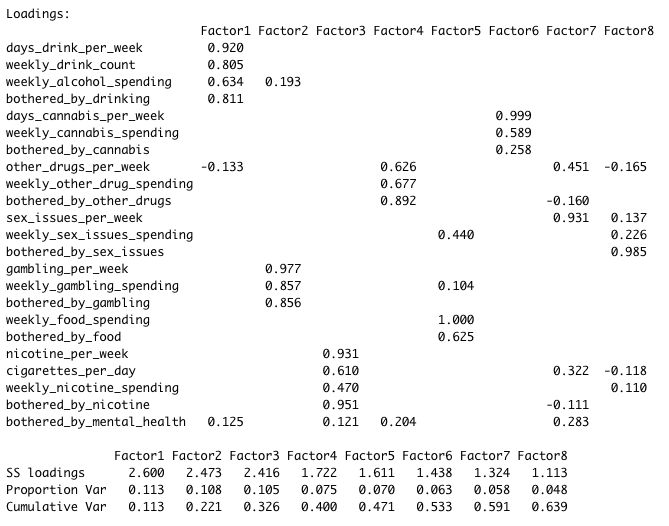
As captured in Figure 10, 66% of the variance could be explained under these 8 factors under the varimax rotation approach:



*Figure* 10*: Variance Explained by Factors Under Varimax Rotation*

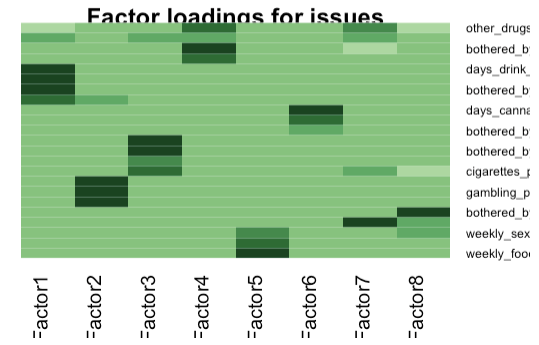
***Oblique Rotation***

The factor alignment under the oblique rotation approach was similar to that of the varimax rotation approach, suggesting that 63% of the variance can be explained under the 8 factors. This information is presented in the following output:



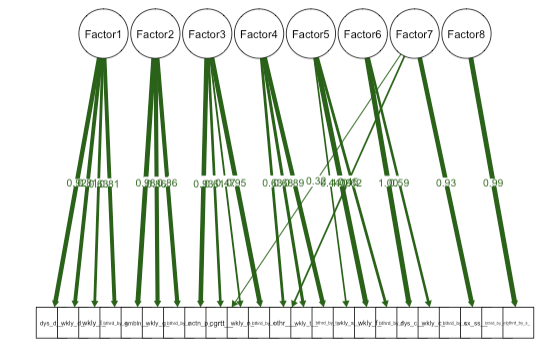
*Figure* 11*: EFA Results Under Oblique Rotation*

Clear boundaries are visible for each of the factor loadings, as demonstrated in Figure 12:



*Figure* 12*: EFA Factor Loadings*

The following path diagram shows the clear distinctions for each factor and their respective loadings:



*Figure* 13*: EFA Path Diagram*

**Confirmatory Factor Analysis**

Confirmatory Factor Analysis (CFA) can be used to evaluate factor significance using a hypothesized model. In this analysis, we construct a model that groups factors as:

* “Alcohol” – any factors that have to do with alcohol or drinking
* “Substance” – any factors that are substance-related (excluding alcohol), including smoking, cannabis, or other drugs
* “NonSubstance” – any factors related to addiction that are not alcohol, smoking, or drugs, such as eating, gambling, sex, or mental health

Our initial model included all factors in our intake assessment data as grouped into the above categories. The data was also scaled for normalization purposes. As presented in Figure 14, the model did not fit well, producing a CFI of 0.635 and RMSEA of 0.143.

A screenshot of a computer

Description automatically generated with medium confidence

*Figure* 14*: First CFA Model Summary*

A second attempt was constructed by removing several factors that had weak estimates or high *p*-values. [Some analysts](https://benwhalley.github.io/just-enough-r/gof.html) suggest that simplifying a CFA model, especially in the presence of a low number of observations, can be helpful for improving the model fitness.

Evidently, this second model exhibited a greatly improved fit, producing a CFI of 0.933. However, the RMSEA was still relatively high at value of 0.121. Additionally, the estimates of the modeled latent variables were not strong, and deemed to be insignificant with *p*-values greater than 0.05 in each case. This information is presented in Figure 15 and Figure 16.

A screenshot of a computer

Description automatically generated with low confidence

*Figure* 15*: Second CFA Model Summary*

A picture containing text, receipt

Description automatically generated

*Figure* 16*: Second CFA Model Latent Variable Estimates and Significance*

This updated CFA model produced the path diagram illustrated in Figure 17.

Diagram

Description automatically generated

*Figure* 17*: CFA Path Diagram*

Although interesting to perform, the hypothesized CFA models are not strongly supported by the available data. This could possibly be improved with larger data samples or by developing different hypotheses to analyze.

***3.3. Data Modeling Plan***

Once the data reduction procedure is completed and resulting preparation steps are employed, consideration of the data model to be built should begin. There are several factors that must be taken into consideration when selecting model candidates:

* How large is our data set?
* How many variables will be used as predictor variables?
* What is our response variable? What data type is it (i.e., continuous, binary, categorical)?
* How much time does the team have to run the model?

Taking these factors into consideration, the team will initially aim to build a logistic regression model. This model was selected due to its reliability, the small underlying data size, and its inherent ability to perform efficient binary classification. Additionally, models including Artificial Neural Networks, Stochastic Gradient Boosting Models, and *k*-Nearest Neighbors (KNN) will also be fitted for comparison of performance and prediction accuracy.

***3.4. Methodology and Model Building***

The next step of the process is centered around model development. In the prior section, the team leveraged the PCA results for dimensionality reduction, significantly reducing the number of predictor variables without compromising the explained level of variance. The training dataset used to generate the PCA model was already devoid of potential outliers in this phase. One of the benefits of using PCA is that it combines highly correlated variables into a set of uncorrelated variables, thereby eliminating existing multi-collinearity between the features. Additionally, the PCA process produces a set of uncorrelated and scaled predictor variables. The PCA model results were then applied to the test data in a similar fashion prior to evaluating each model. The following model was used to fit the various predictive models, where completed\_lesson is the target variable and the other variables are the eight predictors engineered via PCA:

completed\_lesson ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8

**Data Modeling**

As mentioned in the prior section, the models selected for prediction of the binary target variable include Logistic Regression, Stochastic Gradient Boosting, Artificial Neural Network, and KNN. These models were fitted and evaluated using a 80% to 20% training to test data split based on our 257 available complete observations (excluding missing values). The following section outlines the details of these models and presents an evaluation of their prediction accuracy:

**Logistic Regression**

Logistic Regression is a statistical method used for the prediction of binary outcomes by analyzing the relationship between one or more independent variables.

First, we must consider the assumptions of logistic regression:

1. The outcome is binary: completed\_lesson, our outcome variable is binary.
2. No influential outliers: during our outlier analysis, we found that because our data was predominately categorical, and that we did not have an issue with outliers.
3. Absence of multicollinearity: from the Variance Inflation Factor analysis, we can see that there is no multicollinearity between our component variables.

Logo

Description automatically generated with medium confidence

*Figure* 18*: Logistic Model Variance Inflation Factor*

1. Independence of observations: each observation was collected from a unique user, so we can assume that there is no dependency in our observations.

The model trained at an accuracy of 56.3% and performed slightly better on the test data set, producing an accuracy of 60%. The Logistic Regression model also produced an F1 score of 0.4444, which was the highest score observed among the evaluated models:

Table

Description automatically generated

*Figure* 19*: Logistic Regression Test Model Results*

*Figure* 20*: Logistic Regression Test Model F1 Score*

**Stochastic Gradient Boosting**

Stochastic Gradient Boosting is a machine learning technique that employs a variation of boosting, whereby on each iteration, a subsample is gathered at random from the data and used to fit the base learner. The Stochastic Gradient Boosting model resulted in accuracies of ~55% and 54% on the training and test data, respectively. The test case produced an F1 score 0.3429:

Table

Description automatically generated

*Figure* 21*: Stochastic Gradient Boosting Test Model Results*

*Figure* 22*: Stochastic Gradient Boosting Test Model F1 Score*

**K-Nearest Neighbors (KNN)**

KNN is a supervised learning technique in which the algorithm stores all available training data based on similarity into a class, and then calculates the probability of the test data belonging to classes of the training data. The class with the highest probability is selected. For our training data set, the KNN model produced a ~54% accuracy. On the test data subset, KNN demonstrated an accuracy of 52% and a resulting F1 score of 0.3684:

Table

Description automatically generated

*Figure* 23*: KNN Test Model Results*

*Figure* 24*: KNN Test Model F1 Score*

**Artificial Neural Network**

Artificial Neural Networks operate by utilizing node layers containing an input layer, one or more hidden layers, and an output layer. Each node is connected to another node with a known associated weight and threshold, and is activated based on exceeding its threshold. Then, data is sent to the next layer of the model. Using Neural Network on the training dataset, the model achieved an accuracy level of ~59%, whereas with the test dataset, the model achieved an accuracy of 54%. The Neural Network also produced an F1 score of 0.4102, which was the second highest observed F1 score amongst the various models.

Table

Description automatically generated

*Figure* 25*: Artificial Neural Network Test Results*

*Figure* 26*: Artificial Neural Network Test Model F1 Score*

The global opioid epidemic is challenging our current understanding of managing and treating addiction, posing a significant financial burden and strain on our already over-utilized medical systems. As is the case with other

**4. Results and Analysis**

Out of the four fitted models, the Logistic Regression model demonstrated the best in terms of both accuracy and F1 score. The Stochastic Gradient Boosting and Neural Network models both achieved an accuracy of 54%, but the Neural Network carried a higher F-1 Score. The model with the lowest level accuracy was KNN, reflecting an accuracy of 52%. The team also identified the top three principal components by order of importance for each model. PC2 & PC4 were recurring principal components of high importance in these models. The contribution of variables in the original dataset to these principal components is presented in Table 4. Based on the comparability of accuracy scores for the models on training and test datasets, it is not evident that these models suffered from overfitting.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Type** | **Accuracy** | **F1 Score** | **Top 3 Principal Components (Ascending Order)** |
| Logistic Regression | 0.6 | 0.4444 | PC6, PC7, PC2 |
| Stochastic Gradient Boosting | 0.54 | 0.3429 | PC2, PC5, PC4 |
| K-Nearest Neighbor | 0.52 | 0.3684 | PC1, PC4, PC2 |
| Neural Network | 0.54 | 0.4102 | PC3, PC1, PC4 |

*Table* 5*: Overall Model Performance Comparison*

**Results and Analysis**

The global opioid epidemic is challenging our current understanding of managing and treating addiction, posing a significant financial burden and strain on our already over-utilized medical systems. As is the case with other

**5. Conclusion and Review**

Going back to our original business questions, we can evaluate our model findings and look to see how they help to solve the problems presented.

Based on discussions with our collaborators, we decided to create a predictive model to determine whether users of the platform would complete a lesson. The collaborators indicated that saw completing lessons as a key indicator of user engagement, the outcome that they want to maximize.

We used PCA to reduce variables and possibly extract new variables. In our case, we ended up with eight significant principal components. From those eight, PC1, PC3, and PC4 were the most important variables in our strongest linear regression model according to variable importance.

Because the PCA process combines several variables into components, it is difficult to determine which of the original variables have the most influence, but we can see from the predictor loading on principal components table that some variables have a relative high weight where others are significantly lower.

By looking at the highest loading in the components, we see that data related to drugs and gambling were strong, while mental health and alcohol related variables weighed lower. From this we could suggest that making content to speak to those issues could help engage the users that are currently being drawn into the platform. While we expected *gender* and *bought­\_hero* (purchase the platform) to be powerful, they did not turn out to be as important as we expected.

We believe that the performance of our modeling could be improved by collecting more diverse independent variables that describe users or their usage patterns. For example, most of the variables available were variations on the same theme: *botherd\_by*, *usage*, and *spend*. There were relatively few variables that describe the user other than *gender* and *age*. In other data sets it is common to see information such as education level, employment status, marital status, income, etc.   
To this end, a good approach could be to iterate in the web platform development to make sure that data collected is better describes the usage patterns that the team is wanting to influence. It might be ideal for data analysts to partner with business analysts and IT developers to ensure that the platform can be optimized and improved over time.

Another problem that we observed what a relatively short usage history among the user data, with many of them not appearing to be users that did more than sign up - only to never return. Although this could be useful in modelling, the data for these users was typically all NULL making it difficult to glean anything meaningful.

It would also have been helpful if the data observations available were more complete. There was a strong presence of NULL responses, resulting in only having 257 complete observations that we could use.

**Acknowledgements**

Penn State World Campus, DAAN 881 classmates.

**References**

*[1]* *IGNTD Announces The Launch of SPARx, A Smart Personalized Adaptive Recovery System That Finally Fixes The Failing One-Size-Fits-All System That is Killing More People Every Year. 24-7pressreleases.com. (2021, February 8). Retrieved January 23, 2022, from* [*https://www.24-7pressrelease.com/press-release/479198/igntd-announces-the-launch-of-sparx-a-smart-personalized-adaptive-recovery-system-that-finally-fixes-the-failing-one-size-fits-all-system-that-is-killing-more-people-every-year#*](https://www.24-7pressrelease.com/press-release/479198/igntd-announces-the-launch-of-sparx-a-smart-personalized-adaptive-recovery-system-that-finally-fixes-the-failing-one-size-fits-all-system-that-is-killing-more-people-every-year)

*[2]* *Vaughan, Daniel. “3. Learning to Ask Good Business Questions.” Analytical Skills for AI and Data Science: Building Skills for an AI-Driven Enterprise, O'reilly, Beijing, 2020.*