Data-Driven Support for Substance Addiction Recovery Communities

Benjamin Fischman

University of Minnesota
Minneapolis, MN 55455, USA
fisch885@umn.edu

Abstract

This project aims to assist those suffering from substance addiction so that they might better understand their mental health. I propose a user level in-app relapse report classifier that has an overall accuracy of 0.882 in a dataset where 18% of users report a substance relapse. Interactions between engagement and views of inspirational messages of the day are complex, yet significant. Predictive analytics inspire design implications for sociotechnical systems, which aim to facilitate the addiction recovery process.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada

© 2018 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-5621-3/18/04.

<https://doi.org/10.1145/3170427.3180288>

Author Keywords

Quantitative Methods; Health – Wellbeing; Personal Data/Tracking

ACM Classification Keywords

H.5.2. [Information Interfaces and Presentation]: User Interfaces: User-Centered Design.

Introduction

How might we inform the addiction recovery community on how to improve the likelihood of long-term substance abstinence? Substance addiction is a public health epidemic in the United States, which costs the government an estimated \$740 billion per year and afflicts 11.9% of Americans [1]. Recovering from substance addiction can be an arduous endeavor when attempted alone, but research suggests clinical intervention paired with long-term maintenance programs is an effective treatment strategy [15]. Twelve-step fellowships (e.g., Alcoholics Anonymous (AA) or Narcotics Anonymous (NA)) are some of the most popular and effective maintenance programs. Honesty, commitment, and fostering strong interpersonal relationships are the hallmarks of both NA and AA. The app economy is experiencing a boom in personal health informatics applications [4,12], but to the best of my knowledge, the health informatics community has yet to apply data mining techniques to analyze a large addiction recovery dataset in the

	Engagement	Daily Message
N	381	455
R^2	0.510	0.389
Z	-12.198	-10.242

Table 1: Missing data pre-processing imputation for random forest decision tree analysis. N represents the number of users that were imputed via nearest neighbor mean matching. The distributions of engagement and daily message views are not normal, so this step helps preserve their natural distributions ($p < .000001$).

*Note for confusion matrix interpretation, left hand columns are predicted values, and top row are actual values.

C&RT Test Fold*	Never Report	Relapse Report
Never Report	237	12
Relapse Report	28	50

Table 2: C&RT Confusion Matrix

RF Test Fold*	Never Report	Relapse Report
Never Report	256	24
Relapse Report	16	43

Table 3: RF Confusion Matrix

context of a mobile application. This work aims to classify in-app relapse reporters from in-app relapse never reporters, and proposes design implications for future systems oriented toward the support of substance addiction recovery communities.

Related Work

Recent work in HCI and health informatics journals have begun to reflect on technology's role in fostering salubrious habits. This work is at the intersection of two domains: mobile health app informatics and technology's role in the addiction recovery process.

Mobile App Health Informatics

Many studies have quantitatively and qualitatively evaluated and analyzed a myriad of mobile health applications and reflected on best practices for statistical interpretation of personal data. Qualitative interviews with people who collect and reflect on their personal information brought insight into user motivations and practices [8]. Others have performed systematic surveys of mobile health tools in app marketplaces (i.e., Apple and Google) [4,12], while some have examined the characterization of user engagement metrics for health analytics [13]. A methodological overview of classification and regression tree analysis in the context of public health exists [7]. Previous literature suggests the best practice for mobile health app analysis is a classification tree or a decision tree.

In this work, I leverage data mining and machine learning techniques to help provide design implications for future mobile apps oriented toward substance addiction recovery communities.

Role of Technology in Recovery

Several studies provide design implications for systems that aim to support substance addiction recovery communities. Prediction models for relapse already exist in the context of an online forum [10]. Another body of work has proposed and prototyped substance-specific (i.e., ketamine) mobile health apps. For instance, *KeDiary* incorporates Bluetooth-enabled saliva-screening in a smartphone app to facilitate abstinence from ketamine in the user's daily life [16]. Summative evaluations in prototyping studies usually leverage pragmatic mixed-method approaches with relatively few participants in comparison to this studies dataset and methods.

Dataset

The initial dataset consisted of three years' worth of action history for 3346 distinct users and was provided by the *Field Guide to Life* mobile application [17]. In total, the original dataset had 407,806 observations that contained a unique user identifier, action code, and timestamp. Users could perform 32 unique actions while using the *Field Guide to Life* application (e.g., view a message of the day, self-report a relapse, view a sober day counter, etc.). Formative informal interviews with recovering addicts suggest that bucketization of actions into seven categories is a suitable heuristic in the domain of a 12-step recovery program. The seven categories include meetings, relapse report, referencing any of the 12 steps, tracking progress, seeking support, opening the application, and interacting with the message of the day. I pre-process the dataset to exclude superfluous noise. Any observations that occur after 2016-10-26 were considered an error (e.g., the date 2557-08-18

has not happened yet so that observation is an error). Additionally, users that used the app for less than thirty days were excluded, and individuals that only open the application with no other actions were considered out of scope. In total 87 users and 5527 observations were removed (i.e., $N=402,648$ and user count = 3259).

Methods

In this section, I outline the statistical methods applied to the self-reported *Field Guide to Life* dataset. The nature of a self-reported observational dataset implies the distribution amongst users may vary wildly. For example, some users report relapse daily with little other in-app actions, while others check the daily message, report few relapses, and consistently track progress. Event Flow helped simplify temporal event sequence analysis during initial data exploration [11]. Predictor variable creation was an iterative ad-hoc process, and each factor is summarized in Table 5. Classification & Regression Tree (C&RT) analysis and Random Forests (RF) provide two different statistical approaches to robust model validation and evaluation [6]. The intuition behind mixed model validation stems from the complexity of the dataset, and intertextuality concerning the biostatistics, data mining, and machine learning communities. Both models were originally introduced by Leo Breiman [2,3]. The R packages `RandomForest` and `Rpart` were used for model training and classification evaluation [9,14]. C&RT analyses are non-parametric, and handle potentially skewed distributions with recursive node splits. RF is an ensemble learning approach to decision tree construction, which creates collections of decision trees

with controlled variance, and determines the variables of importance via permutation. RF requires all missing values in the dataset be imputed (filled in) before model training, so I used the R package `Hmisc` to bootstrap nearest neighbor mean matching and took the mean of ten permutations to fill the missing values [5] (Table 2). `Rpart` handles missing values implicitly, so this step only applies to the RF analysis. Normalization was applied such that all variables range between 0 and 1. Additionally, I minimize error for both models (e.g., overfitting and tree pruning at optimal depths). All statistical analyses were conducted in R.

Model Training

I perform a ten-fold cross-validation for the binary classification task of user-level in-app relapse reports (i.e., predict users that never report a relapse from those that report one or more relapses). I ensure each sample fold has an approximately equal proportion of never report users to relapse reporting users (i.e., $596 / 3258 \approx 18\%$ report relapse). I train both models on 90% of the dataset, and reserve 10% for testing and validation. I used 100 tree permutations for the Random Forest approach. `Rpart`'s minimum split and minimum node size criterion were set to 30 ($30 \approx 1\%$ population), and the complexity parameter were set to 0.01667.

Results

The models' performance is surprisingly accurate considering predicting who might self-report a relapse *should* be difficult.

Predictor Variable	Description
Engagement [13]	Actions / Usage Sessions [7]
Daily Message	Daily Message Views / Total Actions
Action Diversity	Count of Unique Features Interactions / All Potential Feature Interactions
Meetings	Meeting Actions / Total Actions
Seek Support	Support Seeking Actions / Total Actions
Track Progress	Track Progress Actions / Total Actions
12 Step View	12 Step Actions / Total Actions
Platform	Mobile Device Operating System

Table 5: Predictor variables for the *Field Guide to Life* Dataset ranked by the rpart model's importance measure.

Figure 1: Gini Index represents the mean decrease in node impurity (and not the mean decreases in accuracy). Engagement and Daily Messages seem to be salient predictor variables.

The RF classifier had a precision of 0.72 while maintaining a recall of 0.64 in a test fold of approximately 18% relapse reporting users. Rpart's C&RT model had a precision of 0.64 and a recall of 0.81.

	Accuracy	Error
C&RT	0.877	0.135
Random Forest (RF)	0.882	0.120

Table 4: 10-fold cross-validation of summary statistics results from test fold ($p < .0001$). Error represents the mean squared error. C&RT stands for Classification & Regression Tree analysis.

These results make intuitive sense since a random forest ensemble algorithm has a greater tendency to overfit outliers in a dataset than a non-parametric recursive approach, yet the recursive nature of C&RT analyses imply these models have trouble identifying and adapting to nuanced signals. One of the benefits of using RFs in health informatics is robust variable selection via tree permutations (Figure 1). C&RT methods can also provide an importance measure for

prediction features, but rpart's summary output is a bit esoteric, so rather than copy and pasting the output from a statistical tool I have ranked Table 5 by rpart's variable importance. Indeed, both models agree that Engagement and Daily Message Views are salient prediction variables for the classification task of identifying in-app relapse reporters from never reporters. Rpart did not make use of the imputed data points for Engagement and Daily Message Views, which increases the certainty of these results.

Figure 2: Scatterplot representation of interactions between Engagement and Daily Message Views. Maroon points represents user that report one or more relapses, and gold points represent users that never report a relapse.

Discussion

Visualization and inspection have guided much of this analysis. Interactions between the most salient variables seem complex and nuanced (Figure 2), but that should not hinder efforts to support addiction recovery communities. For example, imagine two potential *Field Guide to Life* users Adam and Joey. Adam uses the app to perform a diverse set of actions and reports a relapse once a week. Adam exhibits behavior that suggests he engages with the application and does not abuse the relapse report as a mechanism to record mainly bad behavior. Joey uses the application to report relapses, and let's assume that every time he opens the application, he reports at least one relapse. Initial exploratory analysis using Event Flow's temporal event sequence visualizations confirm such users exist, which indicates abnormal behavior [11]. Forty-one users (7% of users that report at least one relapse) reported their initial relapse within, approximately, ninety minutes of their first interaction with the application.

System Design Implications

In this section, I highlight implications and insights meant to inspire the design of future systems in the substance addiction recovery space. Systems should aim to expose users to a diverse set of features on a regular basis. For instance, one might imagine leveraging push notification to prompt the user to read the daily message of inspiration, and include a button or a splash screen that would direct the user to an unexplored feature. Subtle prompts that encourage the user to interact with unexplored features should respect the user's privacy while encouraging healthy habits.

Limitations

Proving causal inference from a descriptive self-reported dataset is a non-trivial matter. The analysis performed in this study is only correlational. Descriptive data and controlled experimental studies can help increase the degree certainty concerning the implications and conclusions of this work.

Future Work

Avenues for future work lie at the intersection of mobile health informatics, inferential statistics, and support for substance addiction recovery communities. Substance addiction and mental health in the United States is a poignant matter that requires additional research efforts. I plan to perform propensity matching to better elucidate potential mechanisms for successful substance addiction interventions.

Conclusion

In this study, I perform a descriptive statistical analysis on a self-reported dataset at the intersection of mobile health applications and substance addiction recovery communities. Tenfold-cross validation shows the proposed model has a precision of 0.72 when trained via ensemble decision tree learning in a user classification task. I describe design implications for sociotechnical systems, which aim to facilitate the addiction recovery process.

Acknowledgements

Thank you to the Hazelden Betty Ford Foundation for providing the *Field Guide to Life* dataset. I would like to acknowledge my advisor, Svetlana (Lana) Yarosh, for her support and feedback. Also, thanks to Sabirat Rubya for her help and feedback. Lastly, thanks to all the GroupLensers for your continued support – I would

not be who I am without the guidance of my colleagues. This work was funded by the NSF grant (1464376).

References

1. National Institute on Drug Abuse. 2017. Trends & Statistics. Retrieved December 18, 2017 from <https://www.drugabuse.gov/related-topics/trends-statistics>
2. Leo Breiman. 1984. *Classification and regression trees*. Belmont, Calif.: Wadsworth International Group.
3. Leo Breiman. 2001. Random forests. *Machine learning* 45, 1: 5–32.
4. Juan M. García-Gómez, Isabel de la Torre-Díez, Javier Vicente, Montserrat Robles, Miguel López-Coronado, and Joel J. Rodrigues. 2014. Analysis of mobile health applications for a broad spectrum of consumers: a user experience approach. *Health informatics journal* 20, 1: 74–84.
5. Frank E Harrell Jr and Maintainer Frank E Harrell Jr. 2017. Package 'Hmisc.'
6. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer New York, New York, NY. <https://doi.org/10.1007/978-1-4614-7138-7>
7. Stephenie C. Lemon, Jason Roy, Melissa A. Clark, Peter D. Friedmann, and William Rakowski. 2003. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine* 26, 3: 172–181.
8. Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 557–566.
9. Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomForest. *R news* 2, 3: 18–22.
10. Diana MacLean, Sonal Gupta, Anna Lembke, Christopher Manning, and Jeffrey Heer. 2015. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. 1511–1526. <https://doi.org/10.1145/2675133.2675146>
11. Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. 2013. Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics* 19, 12: 2227–2236.
12. Preethi R Sama, Zubin J Eapen, Kevin P Weinfurt, Bimal R Shah, and Kevin A Schulman. 2014. An Evaluation of Mobile Health Application Tools. *JMIR mHealth and uHealth* 2, 2. <https://doi.org/10.2196/mhealth.3088>
13. Katrina J. Serrano, Kisha I. Coa, Mandi Yu, Dana L. Wolff-Hughes, and Audie A. Atienza. 2017. Characterizing user engagement with health app data: a data mining approach. *Translational Behavioral Medicine* 7, 2: 277–285. <https://doi.org/10.1007/s13142-017-0508-y>
14. Terry Therneau, Beth Atkinson, Brian Ripley, and Maintainer Brian Ripley. 2017. Package 'rpart.'
15. Svetlana Yarosh. 2013. Shifting dynamics or breaking sacred traditions?: the role of technology

in twelve-step fellowships. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3413–3422.

16. Chuang-Wen You, Ya-Fang Lin, Cheng-Yuan Li, Yu-Lun Tsai, Ming-Chyi Huang, Chao-Hui Lee, Hao-Chuan Wang, and Hao-Hua Chu. 2016. KeDiary:

Using Mobile Phones to Assist Patients in Recovering from Drug Addiction. 5704–5709. <https://doi.org/10.1145/2858036.2858185>

17. Mobile More Field Guide to Life App -- Hazelden. <https://www.hazelden.org/web/go/fieldguide>