Survey Data System: a Framework to Enhance Response Rate in Clinical Studies

Giulio De Sabbata[1]

[1] PhD Sudent, ICT Doctorate at DBGroup, University of Modena and Reggio Emilia, Modena, Italy

Abstract

Clinical studies involve typically surveys of patients to gather data. Coaching aims at identifying the least engaged patients to enhance the response rate. To support the coaches we have developed a platform that enables ongoing data pipelines to elicit insights during the study from these patient-generated data. The most challenging aspect derives from the attempt of including the content of the individual responses. Exploiting this information occurs to the detriment of the possibility to generalize the results, preventing any kind of inference across different studies. The framework designs two AI solutions for the analysis of survey data. The first method is a forecast of the response rate. The second consists in an implementation of a descriptive module that generates an user-friendly interface displaying patterns in the response rate behaviour. The coach leverages this visualization tool especially if the predictive module falls short, taking advantage of a tidy view for optimizing the patient engagement strategies during the study. A first approach to the stated problem and an innovative employment of the tree technique are the main contributions of this paper.

Keywords

Data Mining, Questionnaire Processing, Tree Technique, Data Visualization

1. Introduction

For medical research, surveys are the standard to collect data from patients participating to clinical studies such as trials or observational studies. A survey research is essentially a collection of information from a sample of individuals through their responses to questions [1] and we refer, from now on, to them as surveys or studies. One fundamental requirement for the study success is the patient engagement [2], which is typically measured by the survey response rate. To increase the response rate our approach is to exploit questionnaire compiling information (i.e., log information and actual responses to the questions) during the study for defining questionnaire administration (i.e., policy for reminders or scheduling time) [3]. This approach entails processing datasets that pose significant challenges for a predictive task, such as limited observations number and a high number of explanatory variables.

Survey Data Platform for Clinical studies

In this paper we design the implementation of a tool in a digital platform for gathering and processing survey data. This tool is developed to meet the specific needs of the coach, the end-

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy

giulio.desabbata@unimore.it (G. De Sabbata)
0000-0001-7328-441X (G. De Sabbata)

CEUR Workshop Proceedings (CEUR-WS.org)

user of the platform that is an health worker committed to patients monitoring. The coaching activity aims at increasing response rate by contacting the patients to maximize the response rate. The objective of the work is to elicit from survey data relevant insights for the coaching.

To achieve that, we propose a data-driven approach with the implementation of two independent AI solutions. The first functionality enables a forecast of the response rate levels allowing to straightforwardly identify the least engaged patients. Alternatively, the coach could resort to a second module that offers a tidy view of the relevant data affecting mostly the response rate behaviour. The coach, a domain expert with no knowledge about data, identifies supported by an highly interpretable interface which patients to contact.

Contributions

The novelty of the work consists in a new approach to increase the response rate by eliciting insights from survey data while the study is still running and by including the responses. Furthermore, to our knowledge there are no works which exploit the tree technique in such a manner.

The remainder of the paper is organized as follows: Section 2 describes the studies features and the relative implications for the analysis; Section 3 deals with the system managing the data survey platform; Sections 4 and 5 introduce the methods designed to solve the problem resorting to the gradient boosting algorithm for the predictive task and to an innovative implementation of the tree technique for the descriptive interface; Section 6 describes the results of the experiments and the challenges involved in the field-test of the method; in Section 7 are argued the main contributions; lastly, Section 8 draws the conclusions and links them to future improvements.

2. Preliminaries

In the introduction are mentioned two features colliding with predictive task: dataset with a limited sample size and a high number of explanatory variables. The large number of explanatory variables derives from considering each question as an input feature in the framework and, hence, the forecast could suffer from the curse of dimensionality. Whereas, we can identify at least two main reasons that pose the framework in a scenario of narrow data: small surveys are frequent and their processing occurs in the ongoing period. Typically, survey management is expensive and it is common for small and medium research group to aim at keeping their size as small as possible due to limited resources [4]. For the same reason the questionnaires are extremely heterogeneous and characterized by traits that limit analysis possibilities [5].

In the analyzed studies those aspects are actually present *(i)* restricted amount of patients (ranges from 30 to 100 patients) *(ii)* large amount of questions for each study (the count can reach 100 questions) *(iii)* time span of studies differs (lengths of the studies range from one month to one year).

Some may argue that combining studies could overcome the limited size of the dataset. However, the description of the studies suggests that they differ significantly preventing us to join them. Additionally, the lack of consistency in the questions across the studies precludes us from combining the studies and meanwhile exploiting the responses.

3. Survey Data Management System

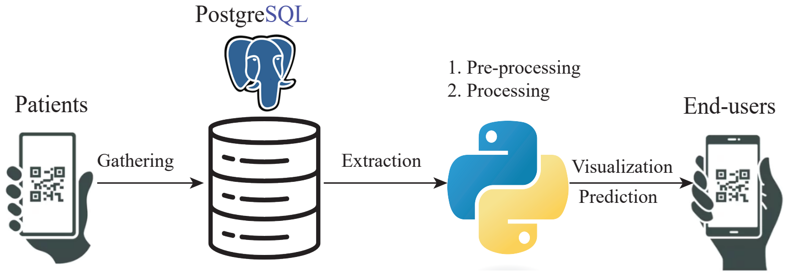

Figure 1: Architecture overview. Once extracted the dataset, its organization enables to effectively perform prediction and visualization tasks.

3.1. Architecture Overview

Figure 1 features the architecture overview and conveys an idea of the data pipeline. The tool should satisfy the following non-functional requirements: *(i)* Systematically executable on different studies, *(ii)* Interactive customization and *(iii)* User-Friendliness. The same script manages different studies just by switching the dataset and the parameter setting. Moreover, the tool should incorporate adjustable setting conveying flexibility to the framework and concurrently maintain highly intuitive usability.

3.2. Survey Data Pre-Processing Module

This module organizes data so that the processing phase runs optimally; the actions performed in the module pertain mostly to the structure or the organization of the dataset. Granularity refers to the number of entities represented by a record in a dataset [6]. The granularity of the data in the raw stage is too fine with each record representing a scheduled question to a patient in a certain date. To identify meaningful patterns, user records are aggregated over a time period, called cycle, such that each row represents the patient behaviour in that cycle. To minimize the loss of information caused by the aggregation on each cycle, it is important to balance the need for shorter cycles and the identification of underlying patterns. The optimal length of the cycle, and the number of cycles, may depend on the length of the study and on the frequency of the questionnaires. Once the time frame is defined, switching from the original indexing to the new one requires to pivot the different questions on the columns, setting them as input features. In each cycle, the value for each question is computed as the average of the corresponding non-missing responses in that specific cycle. In conclusion, it is measured the response rate as the proportion of compiled questionnaires on the total. To get flexible insights we derive a new key indicator, the spread of the response rate, which measures the difference in response rate over consecutive cycles. The coach interactively sets the response rate or the spread as the target variable.

4. Response Rate Forecast

The predictive module gives in output the response rate values of the future cycle, fitted with a Gradient Boosting algorithm. This particular model belongs to a highly effective and extensively used family of machine-learning techniques [7]. Boosting algorithm follows an ensemble learning approach where one of the most performing is the stump-based. The boosting approach generates an additive model that re-weights the data at each iteration to focus on the most difficult examples to predict and, thus, to reduce the risk of overfitting. To optimize the model performance these are the main parameters to tune *(i)* `number of iterations` controls the number of weak models; *(ii)* `maximum tree depth` controls the size of each estimator, if passed one each weak model is a tree stump; *(iii)* `learning rate` controls the amount by which the contribution of each tree is reduced.

The module contains an evaluation modality as well as the predictive one. In the field of predictive modeling, it is crucial to evaluate the dependability of the fitted values. The assessment of the model's performance on the current data provides an indication of its ability to perform well in future predictions. The evaluation modality generates two distinct metrics the `score` and the `progressive recall`. The `score` evaluates the goodness of fit using a re-examined coefficient of determination, which is measured on a scale from -1 to 1. When the score is 0, the model performance is equivalent to the baseline model, and, hence, if the score is negative, the model performance falls shorter than fitting all values with the mean. The `progressive recall` measures the percentage of patients correctly identified by the predictive model as part of the least responding patients.

5. Descriptive Trees: a Method to Discriminate Patient Behaviour

Alternatively to the predictive module, the coach could resort to a tree interface displaying descriptive rules for discriminating patient behaviour based on their response rate. The task consists in clustering patients not only by response rate, but also by the factors that have the greatest impact on it. This visualization may suggest the coach which are the most effective actions for increasing patient engagement. To perform this task a decision tree method is adopted for two main reasons. Firstly, it can handle numerous explanatory variables and perform the task of variable selection [8], overcoming the curse of dimensionality generated from the large number of pivoted questions. Secondly, it is easily displayed and interpretable [9], generating highly intuitive insights. Overall, this technique allows to select relevant variables and displays higly intuitive results.

The work exploits the tree technique in an innovative way since we are not performing any predictive task. In particular, we do not split data in training and test set but we just run the model on the whole dataset. This choice follows the intuition that the method needs just describing/summarising data.

It was possible to process data indifferently either with a regression tree or with a decision tree. A decision tree is adopted since it requires to discretize the dependent variable into a qualitative one, making it more intuitive with the definition of two categories of patients. The

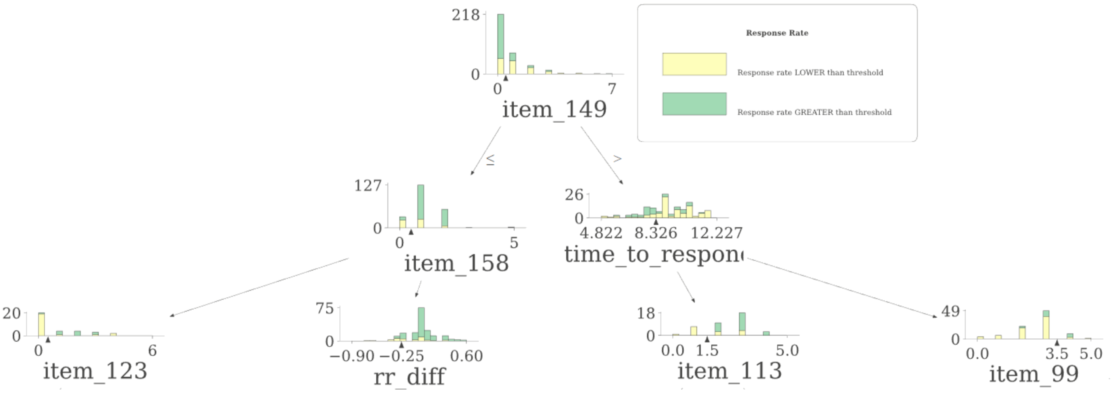

Figure 2: Decision Tree suggests that the patients not responding 0 to the item 149 and employing more time on average to compile the questionnaires tend to be less engaged.

mean of the response rate is the default choice to set the threshold to categorize the response rate. However the coach may customize the tree by setting a different threshold according to her budget. For what concerns (Figure 2) *(i)* the two patient categories are distinguished by different colours (i.e., the most engaged patients are the green) *(ii)* it is reported the text of the input feature that generates the split.

6. Experiments and Open Challenges of Descriptive Trees Field-Test

6.1. Experiments

The framework has been tested on 5 different studies. Initially, different set of explanatory variables are tested to evaluate how the predictor behaves. The results indicate that all available explanatory variables should be included, suggesting that the algorithm can effectively handle the curse of dimensionality. The metrics of each study are generated at an intermediate and a conclusive time point and collected in Table 1.

In general, the performance improves according to the length of the study and to its progress. This result suggests that the size of the study is a critical point and lack of data represents an issue.

6.2. Open Challenges of Descriptive Trees Field-Test

The evaluation of the effectiveness of the descriptive tree method is yet to be tested through a field-test in an ongoing study, and thus, there are no actual validations available. Due to the experimental nature of the method, a well-structured plan is required for the implementation of its field-test. The first challenge is the identification of appropriate metrics to measure the effectiveness of the method. One possible approach could involve soliciting feedback from

Table 1
Evaluation Metrics. Positive scores entails a better performance than the baseline model. Ranking the 10 (or 15) patients with the lowest response rate is a reasonable choice for studies with patients ranging from 30 to 55.

Study	Time	Score	Recall at 10	Recall at 15
Study1	Intermediate	0.49	6/10	10/15
Study1	End	0.68	7/10	13/15
Study2	Intermediate	0.52	6/10	14/15
Study2	End	0.66	7/10	14/15
Study3	Intermediate	0.72	6/10	18/15
Study3	End	0.91	3/10	8/15
Study4	Intermediate	0.03	2/10	8/15
Study4	End	0.10	4/10	8/15
Study5	Intermediate	0.19	7/10	9/15
Study5	End	0.65	6/10	13/15

coaches who have utilized this method, such as through a survey. Secondly, further improvements are necessary to ensure that the resulting clusters are meaningful. The meaningfulness of a cluster depends on the relevance of the insights it provides, where the explanatory variables involved in the splits can suggest the reasons for the cluster behavior and guide the appropriate actions. Thus, the inclusion of domain-specific knowledge or expert input could address the interpretability of the visualization. For instance, the coach could undertake a feature selection task by retaining the most interesting questions related to the treatment. Alternatively, the coach could re-weight the features to give more importance to the most promising ones. The third point, which is closely related to the second, is the balance between a user-friendly tool and its effectiveness. Simplifying the tool usability occurs to the detriment of its capability to provide relevant summaries. For example, achieving the desired level of usability requires careful consideration of the design parameters with particular regard to their number and complexity of the settings. The issue of balancing the size and the accuracy of the displayed tree is an additional example of this challenge. In general, a cost-complexity parameter controls this trade-off, which, in a predictive scenario, trees are pruned according to preventing the tree from overfitting the data. In this framework the parameter regulation depends on other criteria such as the screen size of the coach's smart device or the lack of enough explanatory power of the data itself.

7. Related Works

In the literature there is a plethora of research focusing on the processing phase when the studies have been concluded yet [10]. The novelty of this work lies in the idea of enhancing the response rate directly in the administration phase by processing the questionnaires while the study is still running. Moreover, to our knowledge in the literature they just exploit contextual information and ignore the responses of the questionnaires. To provide for the potential criticisms in the predictive task, a descriptive module supports the end-user decision-making. Despite this module performs a task similar to clustering, the labels, generated by the levels of the response rate, of the data are known and clusters are generated with respect to these labels. An additional feature distinguishing the task from clustering pertains the grouping operation that selects the factors affecting most the response rate behaviour.

8. Conclusion and Future work

This framework aims at eliciting insights from survey data to increase the response rate. In addition to its predictive functionality, the framework includes a descriptive module which offers an alternative solution when the predictive task falls short. The descriptive module employs a highly interpretable tree visualization that groups patient based on their response rate behaviour. However, further improvements and a field-test are necessary to validate the effectiveness of the descriptive trees, given their experimental nature.

In the future, mapping at least the more general questions across different studies and standardizing the size and length of each study could enable the aggregation of survey data from multiple sources. This approach may facilitate the identification of patterns to infer across different studies.

References

[1] J. Check, R. K. Schutt, Research methods in education, Sage Publications, 2011.

[2] L. Duffett, Patient engagement: What partnering with patient in research is all about, Thrombosis Research 150 (2017) 113–120. URL: https://www.sciencedirect.com/science/article/pii/S0049384816306089. doi:https://doi.org/10.1016/j.thromres.2016.10.029.

[3] Z. Dörnyei, T. Taguchi, Questionnaires in second language research: Construction, administration, and processing, Routledge, 2009.

[4] R. W. Beck, Sample size for a clinical trial: why do some trials need only 100 patients and others 1000 patients or more?, Ophthalmology 113 (2006) 721–722.

[5] I. S. Sjetne, O. A. Bjertnaes, R. V. Olsen, H. H. Iversen, G. Bukholm, The generic short patient experiences questionnaire (gs-peq): identification of core items from a survey in norway, BMC Health Services Research 11 (2011) 88. URL: https://doi.org/10.1186/1472-6963-11-88. doi:10.1186/1472-6963-11-88.

[6] T. Rattenbury, J. M. Hellerstein, J. Heer, S. Kandel, C. Carreras, Principles of data wrangling: Practical techniques for data preparation, " O'Reilly Media, Inc.", 2017.

[7] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front Neurorobot 7 (2013) 21.

[8] W. Loh, Variable Selection for Classification and Regression in Large p, Small n Problems, Probability Approximations and Beyond (2012) 135–159.

[9] A. N. Elmachtoub, J. C. N. Liang, R. McNellis, Decision trees for decision-making under the predict-then-optimize framework, in: International Conference on Machine Learning, PMLR, 2020, pp. 2858–2867.

[10] A. Saleh, K. Bista, Examining factors impacting online survey response rates in educational research: Perceptions of graduate students., Online Submission 13 (2017) 63–74.