Survey Design and Optimization: Machine Learning for Question Selection and Weighting Adjustment

Keywords: data mining, longitudinal study, classification, missing data, categorical information

Extended Abstract

Using survey data from a relatively small group of people to infer broad public opinion is an important and practical method in social science. This classic strategy has been widely employed, but some aspects can be optimized, especially for longitudinal study.[1]

Two main aspects of survey optimization are explored in our work: accuracy and cost. For example, some questions could be pruned to reduce costs, owing to correlation and redundancy of information. But how should we design the survey and choose these questions? We need to be more quantitative in making trade-offs between efficiency and accuracy, rather than relying on common sense. For example, it's possible to build a model based on background information that could predict the respondent's answers. And we all know that some machine learning models are likely to be able to meet our requirements.

A critical challenge is how to make the sample in the survey more representative. In general, it's almost impossible to know what kind of sampling would be perfectly representative of the entire group and almost impossible to predict who will complete the survey. In the most ideal case, we can directly apply the proportion of responses from the survey data to the overall target group, which actually always results in large errors. Thus, there are different approaches of calculating the opinions of a group based on samples consisting of survey data. A natural idea is to think of the survey data as a stratified sample, but on which variables should we stratify? For different choices of stratification, almost different results are produced. Many theoretical approaches have been proposed in the literature, such as post-stratification and propensity weighting, but always under some additional assumptions.[2]

At the same time, predicting respondents' answers to certain questions by using some background information or simple questions, such as gender, age, education background, can not only reduce expenses, but also improve the accuracy of the survey. In other words, the connection between the survey sample and the whole group could be better established, which is useful for weighting adjustments.

From a more practical point of view, a high-quality prediction model will tell us which background variables of the respondents have a greater impact on the responses. This kind of models can also be used to stratify the survey data and and finally can improve the accuracy of the survey results.

Our work is based on a private data set on Italian people, as well as the Dutch public data set LISS (Longitudinal Internet studies for the Social Sciences) panel[3], where the first challenges we faced were categorical information and missing data. Using some encoding approaches[4] and imputation methods[2], with the help of feature selections[5], we have initially solved these problems and improved the prediction results, not only in terms of average accuracy, but also in terms of stability.

As we all know, neural networks have been widely used in an army of fields. Meanwhile, the survey data we are currently working with is mainly tabular data, for which we could still believe in tree-based models now.[6][7] Therefore, the tree-based models, such as random forests, are also the focus of our current research. Using the same prediction methods and parameters, we can also compare the efficiency of previous methods dealing with categorical information and missing data.

In the end, we could use some tree-based models to predict the respondents' opinions on certain issues based on their background information or some simple questions, which is useful for the survey design. And with a few easily monitored variables, we could keep abreast of shifts in public thinking and also make appropriate predictions about certain potential changes: increasing educational attainment, aging population trends, etc.

In such a framework, more applications could be foreseen in the future, such as how to identify the swing voters and better predict their choices, how to predict the evolution of opinions in longitudinal study, and so on.

References

[1] Naresh K Malhotra. Questionnaire design. *The handbook of marketing research: Uses, misuses, and future advances*, page 83, 2006.

[2] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[3] Annette C Scherpenzeel. "true" longitudinal and probability-based internet panels: Evidence from the netherlands. In *Social and behavioral research and the Internet*, pages 77–104. Routledge, 2018.

[4] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[5] Vipin Kumar and Sonajharia Minz. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014.

[6] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*, 2022.

[7] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.