DOI:10.19181/kongress.2020.571

Александрова Марина Юрьевна
Национальный исследовательский университет
«Высшая школа экономики»,
Москва, Российская Федерация
m.42laksandrova@gmail.com

Предсказание частичного неответа на примере данных european social survey с использованием логистической регрессии

Аннотация. Пропуски в данных представляют собой актуальную проблему в социологических исследованиях. Одним из источников пропусков в данных являются частичные неответы, связанные с нежеланием респондента отвечать на вопрос, затруднением с ответом или другими причинами. Причину возникновения неответов видят как в способе проведения опроса или характеристиках респондентов, так и в характеристиках самой анкеты. В данной работе будет показано, как может прогнозироваться возникновение частичного неответа с помощью логистической регрессии с использованием данных опросов Европейского социального исследования [European Social Survey, ESS]. Были обучены модели предсказания отказов от ответа, отсутствия ответа и затруднений с ответом на основе текстовых характеристик вопросов с использованием частот слов и метрики важности слов TF-IDF. Все полученные модели сравнивались между собой с точки зрения качества получаемых с их помощью предсказаний, кроме того, наиболее важные слова из формулировок вопросов были разделены относительно того, повышают или понижают они вероятность появления неответа. В частности, было выявлено, что слова сензитивных тематик ведут к увеличению доли частичного неответа, а также некоторые слова-инструкции к самим вопросам.
Ключевые слова: частичный неответ; отказ от ответа; отсутствие ответа, «затрудняюсь ответить»; логистическая регрессия; текст-майнинг; европейское социальное исследование; машинное обучение; качество измерения

Aleksandrova Marina Yurievna
National Research University «Higher School of Economics»,
Moscow, Russian Federation
m.42laksandrova@gmail.com

Prediction of an item nonresponse error using a logistic regression based on the european social survey data

Abstract. Missing data represent an urgent problem in sociological research. One of the sources of the missing data is an item nonresponse, which can be related to the respondent's reluctance to answer the question, difficulties that occur during the answering process, or other reasons. The reason for the nonresponse is seen in the method of conducting the survey or in the characteristics of the respondents, and also in the characteristics of the questionnaire itself. This research will show how item nonresponse can be predicted by logistic regression model using European Social Survey data (ESS). Models for predicting rejection answer, no answer, and "don't know" option were trained based on the textual characteristics of the questions using word frequencies and the word importance metric TF-IDF. All the models

obtained were compared with each other in terms of the quality of the predictions can be made with them, in addition, the most important words from questions were divided as to whether they increase or decrease the likelihood of an item nonresponse. In particular, it was revealed that words connected to the sensitive topics lead to an increase in the proportion of an item nonresponse, as well as some words connected to the instruction on how to answer particular question.
Keywords: item nonresponse; refusal to answer; no answer; “neither agree nor disagree”; “Don't know” option, logistic regression; text-mining; European Social Survey; ESS, machine learning; measurement quality

Пропуски в данных представляют собой актуальную проблему в социологических исследованиях. Одним из важнейших источников пропущенных данных являются отказы от ответа респондентов. С одной стороны, отказ от ответа может быть связан с внешними, не связанными напрямую с исследователем, причинами: социальными нормами в обществе, текущей социальной, культурной, экономической или политической ситуацией [Докторов, 1979: 58]. С другой стороны, пропуски в данных, связанные с нежеланием респондента ответить на анкетный вопрос, могут быть свидетельством низкого качества подготовленной анкеты в целом или отдельных вопросов [Клюшина, 1990: 101]. Вне зависимости от того, какая проблема скрывается за неответом респондента, все равно это заставляет исследователя столкнуться со множеством сложностей.

Методы работы с пропущенными данными довольно разнообразны: наблюдения с пропусками могут удаляться или заполняться по определенным правилам, имеющиеся данные могут взвешиваться для достижения необходимого объема выборки и соотношения в ней определенных групп. Тем не менее, нельзя упускать из внимания вероятность того, что работа с пропусками требует предельной осторожности со стороны исследователя – без обоснованной уверенности в случайном характере пропусков существует опасность прийти к сильно смещенным результатам.

Поэтому представляет особый интерес в работе не с уже свершившимися пропусками, а в предотвращении и минимизации их появления на основе четкого знания, что в созданном опросе может повысить вероятность появления неответа.

Выделяют два основных типа отказа от ответа: полный и частичный [Groves, Couper, 1998: 25]. Исследователи различают несколько типов полного неответа в опросах: это отказ из-за отсутствия контакта (невозможность связаться с респондентом предполагаемым способом), отказ от сотрудничества (нежелание респондента участвовать в опросе), отказ из-за невозможности участвовать (например, неспособность пройти опрос из-за языковых ограничений или различий) [Groves, Couper, 1998: 84]. Частичный отказ от ответа возникает, если участники исследования не отвечают на некоторые вопросы, в то время как на другие вопросы дают какой-то ответ, причем содержащий информацию о мнении, отношении, поведении респондента [Spitzmüller C. et al.,2006: 29]. За отсутствием ответа скрывается, как правило,

незнание, отсутствие какого-то выраженного мнения, или нежелание отвечать на вопрос. Выбор варианта ответа «Затрудняюсь ответить» исследователи считают также своего рода отказом, который может скрывать за собой как незнание, так и нежелание отвечать, скрытое за подобной мягкой формой отказа [Colsher, Wallace, 1989: 47]. В то же время, явный отказ от ответа может скрывать нежелание признаваться в незнании какой-то темы, отсутствии выраженного мнения [Sicinski, 1970: 129].

Причины возникновения неответов могут быть связаны со способом сбора данных, характеристиками респондента или характеристиками опросного инструментария – анкеты. Характеристики респондентов определенно могут влиять на желание отвечать, однако прогнозирование неответов на их основе может быть проблематичным. Анализ литературы показывает, что одни и те же характеристики респондентов могут давать различные результаты. Так, Гровс связывает неответы отчасти с умственными или когнитивными способностями респондентов, анализируя влияние возраста и образования на вероятность появления отказа от ответа, и приходя к выводу, что пожилой возраст и прохождение меньшего числа ступеней в образовании может приводить к росту неответов [Groves, 1979: 199]. Данная точка зрения противоречит результатам других исследователей, – Герцог и Дильман не выявили сколь-либо статистически значимых различий в тенденции к неответу у разных возрастных групп [Herzog, Dielman, 1985: 353]. Аналогично, уровень образования также не доказывает свое однозначное влияние на долю неответов в различных исследованиях: эта связь или не обнаруживается вовсе [Messmer, Seymour, 1982: 274], или наблюдается обратная связь (тенденция к выбору варианта «затрудняюсь ответить» с повышением уровня образования респондента) [Schuman, Presser, 1980: 1221].

В то же время, исследования влияния характеристик анкет и вопросов показывают меньшее разнообразие, что позволяет предполагать возможность прогнозирования их влияния на неответы.

В данном исследовании использовалась логистическая регрессия для прогнозирования частичного неответа на данных Европейского социального исследования (European Social Survey, ESS). Логистическая регрессия строилась по отдельности – на основе частот слов и TF-IDF (term frequency – inverse document frequency, статистическая мера важности отдельного слова в тексте [Hirschberg, Manning 2015: 262].

Для всех моделей рассчитывался коэффициент точности предсказаний [accuracy score] – доля тестовой выборки, предсказания для которой оказались верными и матрица ошибок [confusion matrix] – двумерная матрица, показывающая распределение правильных и ошибочных предсказаний, сделанных с помощью обученной модели. Все построенные модели предсказания неответов показывают примерно одинаковые показатели качества, кроме того, были получены списки слов, которые с большей или меньшей вероятностью будут вести к возникновению одного

из частичных неответов – затруднению с ответом, отсутствию ответа или отказу от ответа.

Анализ данных показал, что ведут к частичному неответу слова – маркеры сензитивных тематик, так как респондентам может быть некомфортно, неловко отвечать на какие-то вопросы анкеты. С одной стороны, это достаточно очевидный и не новаторский результат, но его полезность заключается в том, что он служит своего рода подтверждением достоверности полученных результатов данного исследования. Кроме того, частичный неответ могут провоцировать те слова, которые направлены на пояснение процедуры ответа на поставленный вопрос, что может быть свидетельством сложности подобных вопросов. Существенных различий в списках слов, которые ведут или не ведут к частичному неответу, в зависимости от предсказания типа частичного неответа, не было выявлено.

Библиографический список

Докторов Б. З. О надежности измерения в социологическом исследовании Л.: Наука, 1979. 127 с.

Клюшина Н. А. Причины, вызывающие отказ от ответа// Социологические исследования. 1990. № . 1. С. 98–105.

Colsher P. L., Wallace R. B. Data quality and age: Health and psychobehavioral correlates of item nonresponse and inconsistent responses //Journal of Gerontology, 44 (2). 1989. P.45–52.

Groves R. M. Actors and questions in telephone and personal interview surveys. Public Opinion Quarterly, 43 (2). 1979. P. 190–205.

Groves R. M., Couper M.P. Nonresponse in Household Surveys. New York: Wiley. 1998.

Herzog A., Dielman L. Age differences in response accuracy for factual survey questions //Journal of Gerontology. 40 (3). 1985. P. 350–357.

Hirschberg J., Manning C. D. Advances in natural language processing //Science. Т. 349. № 6245. 2015. P. 261–266.

Messmer D. J., Seymour D. T. The effects of branching on item nonresponse //Public Opinion Quarterly. 46 (2). 1982. P. 270–277.

Schuman H., Presser S. Public opinion and public ignorance: The fine line between attitudes and nonattitudes //American Journal of Sociology. 85 (5). 1980. P. 1214- 1225.

Sicinski A. "Don't know" answers in cross-national surveys //Public Opinion Quarterly, 34 (1). 1970. P. 126–129.

Spitzmüller C. et al. “If you treat me right, I reciprocate”: Examining the role of exchange in organizational survey response //Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior. Т. 27. № . 1. 2006. P. 19–35.