Применение методов Subgroup Discovery для оценки эффекта от воздействия

Холодницкий Сергей Вадимович

Студент

Факультет экономики, менеджмента и бизнес-информатики НИУ ВШЭ, Пермь, Россия

E-mail: serega.holodnitscky@yandex.ru

Научный руководитель *— Бузмаков Алексей Владимирович*

Введение: Задача оценки эффекта от воздействия востребована в случаях, когда необходимо сделать выводы об эффективности, например, нового лекарственного препарата или новой маркетинговой кампании. При этом, распространенного A/B тестирования недостаточно ввиду потребности получения информации о распределении эффекта от воздействия, и его использовании для улучшения характеристик самого воздействия.

Цель исследования: Основной целью этой работы является проверка гипотезы об эффективности использования методов Subgroup Discovery (далее SD) для анализа данных A/B тестов с целью поиска способов улучшения характеристик производимого воздействия.

Методология исследования: В данной работе демонстрируется применение методов SD для решения задачи оценки эффекта от воздействия. SD — это группа методов раздела Data Mining, направленных на поиск человеко-читаемого описания подгруппы наблюдений, которые значительно отличаются от «среднего» наблюдения с точки зрения некоторой целевой переменной, как правило, в терминах задач классификации и регрессии.

В то же время, постановка задачи оценки эффекта от воздействия включает не только целевую переменную, но и переменную воздействия, и поэтому не относится к классическим задачам классификации или регрессии. Для ее решения существуют специальные методы [1,2]. Также эта задача может быть сведена к задаче классификации или регрессии [3]. В частности, бинарная целевая переменная Y и переменная воздействия T, могут быть заменены на бинарную переменную $Z = Y \cdot T + (1-Y)(1-T)$. Именно этот подход используется в этой работе, так как позволяет воспользоваться уже существующими методами SD.

Описание и результаты эксперимента: Эксперимент проводился на данных продуктовой сети, с использованием библиотеки `pysbgroup` для `Python`. В качестве признаков бралась информация о сумме покупок, среднем чеке, времени последней покупки, времени жизни клиента и др. В качестве воздействия выступает информация об отправке или не отправке клиенту смс-сообщения. Целевой является бинарная переменная – пришел ли клиент в течение следующих двух недель после получения рассылки.

Получены следующие результаты: подгруппа клиентов, которые более 14 дней не посещали магазин и чье суммарное время жизни не более 180 дней имеет наибольший шанс прихода после получении смс-сообщения. Общий размер этой подгруппы составил чуть более 10 процентов от общего количества клиентов.

Для сравнения была построена логистическая регрессия и замерены коэффициенты перед признаками. Значимыми признаками в модели оказались: время жизни и время с последней покупки. Здесь логистическая регрессия дала общее представление о значимости переменных, однако методы SD позволили дополнительно описать клиентов с высоким откликом на воздействие.

Выводы: Полученные результаты свидетельствуют о том, что при использовании методов SD для анализа подобных экспериментов можно получить интерпретируемую информацию о клиентах, на которых эффект от воздействия оказался выше среднего. Такую информацию можно учесть при разработке последующих маркетинговых кампаний, для повышения их эффективности и/или снижения затраты на их реализацию.

Благодарность: исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта No 20-31-70047

Литература

1. Buzmakov A. “Machine Learning for Subgroup Discovery under Treatment Effect”, arXiv:1902.10327, 2019.
2. Chekroud A. M. et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach // The Lancet Psychiatry. — 2016. — Vol. 3, no. 3. — Pp. 243–250.
3. Weisberg H. I., Pontes V. P. Post hoc subgroups in clinical trials: Anathema or analytics? // Clinical Trials — 2015 — Vol 12 — no. 4 — Pp. 357–364.