Сравнение методов выявления индивидуального эффекта от воздействия

Семенова Дарья Владимировна, Темиркаева Мария Рифинатовна

Студент, Студент

НИУ ВШЭ, Пермь, Россия

E-mail: mariya.temirkaeva@mail.ru, dvteterina@gmail.com

Научный руководитель *— Бузмаков Алексей Владимирович*

Мы живем в мире постоянно растущей информации, в связи с чем возникает множество вопросов по хранению, обработке и использованию этих данных. В частности, все чаще задаются вопросы о наиболее релевантных методах решения проблем, связанных с большими объемами данных. В данной работе мы сравниваем несколько подходов решения одной из широко обсуждаемых задач последнего десятилетия – оценки индивидуального эффекта от воздействия. Под воздействием в литературе понимается некое действие, направленное на человека, с целью вызова его ответной реакции. Например, в бизнесе о наличии эффекта от воздействия будет говорить факт отклика покупателя на рекламное сообщение.

Воздействие, оказываемое на людей, может оцениваться как неэффективное для всей выборки, но при этом может быть эффективным для определенной группы людей. Это связано с тем, что популяция индивидов зачастую гетерогенна, поэтому действие эффекта для одних индивидов может быть сильным и положительным, для других – слабым и положительным, а для третьих вовсе отсутствовать. Поэтому методы оценки эффекта от воздействия в среднем по популяции потеряли свою актуальность, им на смену пришли методы оценки индивидуального эффекта от воздействия (PTE). В нашей работе мы сравним разные методы машинного обучения (линейную логистическую регрессию, случайный лес, бустинг и SVM) применительно к нескольким подходам измерения PTE: к методу двух моделей и методу преобразования зависимой переменной, предложенным Jaskowski и Jaroszewicz [3], и Weisberg и Pontes [4]. Также, сравним вышеперечисленные методы измерения PTE с подходом Uplift Random forest [2].

Для сравнения различных комбинаций методов и подходов оценки эффекта от воздействия мы используем набор данных, созданный Diemert и др. [1]. Данные были получены в ходе проведения рекламной кампании, целью которой было увеличение посещений

сайта рекламодателя. Набор данных состоит из 25 миллионов наблюдений. По каждому человеку известны 12 характеристик, его принадлежность к одной из двух групп (экспериментальной – той, которая подвергалась воздействию, или контрольной) и целевая переменная – посетил ли пользователь сайт рекламодателя в течение тестового периода (две недели). Согласно предварительному анализу полученных данных, отклик потребителей в экспериментальной и контрольной группах составил 4,41% и 2,61% соответственно. Разница отклика в контрольной и экспериментальной группах не столь велика. Поэтому крайне важно найти комбинацию подхода и метода, которые позволят наиболее точно оценить PTE. Наш алгоритм, описанный ниже, направлен на достижение этой цели.

1. Случайное деление выборки на обучающую и тестовую: случайным образом распределяем наблюдения либо в обучающую выборку, либо в тестовую. 2. Оценивание модели: на обучающей выборке происходит обучение модели. 3. Предсказание: предсказываем PTE на тестовой выборке. Далее рассчитывается средний эффект от воздействия (AVE) и общий эффект от воздействия (TTE) на подвыборке (30% наблюдений с самым высоким значением PTE) для сравнения методов.

Стоит отметить, что использовалась кросс-валидацию для уменьшения разброса получаемых результатов. Каждая итерация кросс-валидации включала в себя шаги 1-3, описанные выше.

Согласно полученным результатам, метод двух моделей, оцененный логистической регрессией, оказался лучшим среди всех. Далее с небольшой, но статистически значимой разницей следует Uplift random forest. Одной из причин высокого качества метода двух моделей подхода может быть специфика набора данных. В частности, если модель зависимости целевой переменной от характеристик человека и факта, было ли оказано на него воздействие, хорошо вписывается в данные, то данный подход может обеспечить хорошие результаты. Таким образом, для формирования окончательных выводов, необходимо произвести дополнительное сравнение методов на других наборах данных.

Литература

1. Diemert E. Betlei A. Renaudin C. Massih-Reza A. A large scale benchmark for uplift modeling. In: Proceedings of the AdKDD and TargetAd Workshop. 2018.
2. Guelman L. Guillen M. Perez-Marın A. M. Uplift random forests. Cybernetics and Systems. 2015. No 46(3-4). P. 230–248.

3. Jaskowski M. Jaroszewicz S. Uplift modeling for clinical trial data. In: ICML Workshop on Clinical Data Analysis. 2012.
4. Weisberg H. I. Pontes V. P. Post hoc subgroups in clinical trials: Anathema or analytics? Clinical trials. 2015. No 12(4). P. 357–364.