Секция «Искусственный интеллект и машинное обучение в цифровой экономике»

Редкие продажи: синергия модели Кокса и модели случайного леса

Научный руководитель – Терехов Сергей Александрович

Челошкина Ксения Сергеевна
Студент (магистр)
Национальный исследовательский университет «Высшая школа экономики», Факультет бизнеса и менеджмента, Москва, Россия
E-mail: kscheloshkina@gmail.com

Ритейлеры собирают различные данные о клиентах, что позволяет получить собирательный образ основного клиента, а также выделить сегменты клиентов со схожими ожидаемыми паттернами поведения. Однако помимо пола, количества покупок и доли просмотренных email-рассылок клиент может быть охарактеризован со стороны его готовности к совершению той или иной покупки, что критически важно для персонализации предложений и, в целом, для работы с лояльными покупателями. В данной работе рассматривается задача предсказания вероятности редкой покупки на определенном временном горизонте для каждого клиента компании.

Для решения этой задачи было использовано около 70 признаков клиентов (анкетные данные, rfm-метрики[5], показатели, основанные на истории покупок и динамике бонусного счета).

Задача предсказания вероятности покупки товара заданной группы на каком-либо временном горизонте для конкретного клиента может быть рассмотрена с двух сторон: как задача классификации и как задача анализа выживаемости. Таким образом, было обучено две модели:

1. Модель случайного леса[2].

Имея признаковое описание клиента за день до начала периода прогнозирования и метку совершения покупки товара рассматриваемой категории на заданном временном горизонте, можно обучить искомый классификатор. Для построения случайного леса использовалась библиотека-обёртка для R платформы H2O[7], которая позволяет быстро получать модели машинного обучения на больших данных. Был построен случайный лес из 500 решающих деревьев с использованием балансировки классов, так как на рассматриваемом горизонте (1 месяц) доля клиентов, совершивших покупки, в сравнении с общим числом рассматриваемых клиентов мала (покупок - 0,16%).

2. Модель Кокса

Если провести формальную аналогию между покупкой товара рассматриваемой категории и некоторым «заболеванием», то клиенты компании все время находятся под «риском заболеть». Тогда данная задача может интерпретироваться как задача анализа выживаемости (survival analysis), т.е. задача предсказания времени до наступления события (покупки). При этом стоит заметить, что данные о продажах являются цензурированными. Таким образом, выходными наблюдаемыми переменными служат время до наступления события или момента цензурирования и метка цензурирования (0 - покупка в указанное время, 1 - окончание наблюдения за объектом).

При такой постановке задачи необходимо историю клиентов, совершавших покупки товара рассматриваемой категории, разбить на несколько частей по точкам совершения покупок (см. Рис.1). Такое преобразование данных гарантирует, что для каждого наблюдения событие может наступить только один раз, что отвечает классической постановке задачи моделирования выживаемости. Данные модели позволяют работать с цензурированными данными.

Одной из широко распространенных моделей выживаемости является модель пропорциональных рисков Кокса [1,3,4,8]. Она предполагает, что для каждого объекта помимо наблюдений времени до наступления события доступны также значения признаков. Функция риска представляет собой функцию от этих признаков и неизвестных регрессионных коэффициентов, умноженную на функцию от времени. Данная модель и была обучена на преобразованных данных.

Результаты применения моделей на тестовой выборке представлены в Таблице 1. В контексте данной задачи качество можно сравнивать по динамике показателей в подвыборке в зависимости от ее доли (подвыборка заданного объема из клиентов с максимальной вероятностью покупки). При сравнении результатов было замечено, что пересечение клиентов в выборках этих моделей достигает максимального значения 40% при отборе рассматриваемых долей выборки, что является серьезным основанием для создания ансамбля из данных моделей. Были рассмотрены два варианта ансамблей[6]:

1. Расчет нормализованного в интервал [0;1] среднего ранга вероятности покупки по двум моделям
2. Градиентный бустинг на основе самых важных признаков для каждой из моделей, а также предсказанных ими вероятностей покупки (500 деревьев, повышенный вес для наблюдений с меткой совершенной покупки)

Результаты ансамблей на тестовой выборке представлены в таблице 2.

В рассмотренной задаче каждый подход имел свои недостатки: случайный лес недостаточно четко выделял клиентов, которые совершат первую покупку, а предположения модели Кокса выполняются лишь приближённо. Однако так как эти модели совершенно разные и их предсказания не очень сильно пересекаются, они представляют собой хорошую основу для создания ансамбля. Градиентный бустинг позволил при выделении подвыборки объема 1-5% от всей выборки получить прирост лифта конверсии 9,3-15% от лучшей модели первого уровня (случайный лес). Описанная работа выполнена в АО Связной Логистика под руководством Терехова С.А., к.ф.-м.н .

Источники и литература

1) Родригес Г. Модели выживаемости // Квантиль. 2008. No 5. С. 1-27.
2) Breiman L. Random Forests // Machine Learning. 2001. No 45. С. 5–32.
3) Cox D. R. Regression models and life tables (with discussion) // Journal of the Royal Statistical Society, Series B. 1972. No 74. С. 187–220.
4) Fox J.,Weisberg S. Cox Proportional-Hazards Regression for Survival Data. Appendix to An R Companion to Applied Regression. Second Edition. Thousand oaks; Sage, 2011.
5) Intro Guide to the RFM Model: https://canopylabs.com/resources/an-introduction-to-the-rfm-model
6) Kaggle ensembling guide: https://mlwave.com/kaggle-ensembling-guide/
7) Overview - H2O 3.16.0.4 documentation: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html
8) Residuals and model diagnostics (Patrick Breheny): http://myweb.uiowa.edu/pbreheny/7210/f15/notes/11-10.pdf

Иллюстрации

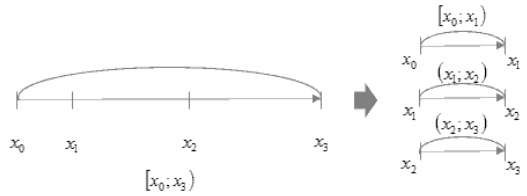

Рис. 1. Рис.1

Модель	Доля выборки	Покрытие	Лифт конверсии
Случайный лес (RF)	0.01	0.15488	15.488
Модель Кокса (Cox)	0.01	0.1309	13.090
Случайный лес (RF)	0.02	0.2398	11.990
Модель Кокса (Cox)	0.02	0.19712	9.856
Случайный лес (RF)	0.03	0.30162	10.054
Модель Кокса (Cox)	0.03	0.24552	8.184
Случайный лес (RF)	0.04	0.35299	8.825
Модель Кокса (Cox)	0.04	0.29029	7.257
Случайный лес (RF)	0.05	0.39435	7.887
Модель Кокса (Cox)	0.05	0.32351	6.470

Рис. 2. Таблица 1

	Доля выборки	0,01	0,02	0,03	0,05
Средний ранг	Покрытие	0,168	0,251	0,313	0,407
	Лифт конверсии	16,751	12,553	10,448	8,144
Бустинг	Покрытие	0,178	0,268	0,334	0,431
	Лифт конверсии	17,808	13,378	11,12	8,618

Рис. 3. Таблица 2