



ФОРМИРОВАНИЕ ПЕРСОНАЛЬНЫХ ПРЕДЛОЖЕНИЙ PERSONAL OFFERS MODELING

DOI: 10.24411/2658-4964-2020-10021

Михайлова Диана Владиславовна, Магистрант, Санкт-Петербургский государственный экономический университет, г. Санкт-Петербург
Mikhaylova D.V. dimihailova@mail.ru

Аннотация

В данной статье представлена модель машинного обучения, используемая в розничной сети магазинов формата дрогери для формирования персональных предложений. Описаны входные данные и ключевые шаги их обработки. Модель определяет вероятность покупки товаров в конкретной категории в течение фиксированного периода времени. Важными особенностями приведённого подхода являются калибровка прогнозируемых значений, подход «клиент товару, но при этом выбор лучшего товара из списка для клиента» и использование вспомогательных вероятностных моделей для повышения экономической эффективности персональных коммуникаций.

Summary

This paper presents a machine-learning model used in retail chain of droggery stores for managing personal offers. It contains the description of the input data and key steps of its processing. The model predicts the probability to make a purchase of goods of specific category during a fixed period of time. The most remarkable advantages of this method are scores calibration, “the best client for the item but choosing the best item for the client from a list” approach and using additional models to increase the economic efficiency of personal communications.

Ключевые слова: Клиентская аналитика, вероятности покупки, персональное предложение, отклик, розничная торговля.

Keywords: CRM, customer relationship management, propensity to buy, personal offer, response, retail.

В условиях, когда большинство задач первостепенной важности в розничной торговле так или иначе нашли свои решения, будь то с использованием новейших технологий или нет, всё более очевидной

становится важность построения доверительных и взаимовыгодных отношений с клиентами для расширения доли рынка и получения дополнительного дохода. Особенно это очевидно в таких высококонкурентных форматах, как сетевые магазины «у дома».

Все крупные игроки рынка розничной торговли инвестируют всё больше денег и ресурсов в развитие наиболее привлекательной для покупателей и прибыльной для сети программы лояльности. Бизнес хочет получить инструмент, способный автоматически определять, что и когда отправить покупателю, чтобы он точно откликнулся.

Задачу релевантных предложений можно рассматривать с двух сторон. Целями программы лояльности являются получение дополнительного дохода в результате проводимых коммуникаций в краткосрочном периоде и обеспечение удержания и стимуляции клиентов в долгосрочном периоде. На основании эмпирически подтверждённой гипотезы о том, что релевантные предложения способствуют повышению лояльности клиентов, возникает задача подбора для клиента самого лучшего товара. Вместе с этим, любая розничная сеть имеет обязательства перед поставщиками, бренды собственной торговой марки и иные товарные группы, которые необходимо продвигать в рамках маркетинговых кампаний. Так возникает задача поиска лучших клиентов для конкретного товара, тех, реклама которым путём персональных коммуникаций будет максимально эффективной. Модель определения вероятности совершить покупку в категории в течение фиксированного периода времени РТВ (Propensity To Buy) позволяет решить обе эти задачи.

Для каждой товарной категории, которая является пригодной для продвижения через используемые каналы коммуникаций, строится отдельная модель. Моделируемые объекты – клиенты. Целевая переменная – покупка в товарной категории в течение фиксированного количества дней от выбранной даты. Поскольку большинство кампаний запускаются на 4-7 дней, период был положен равным 14 дням, чтобы коммуникация могла привлечь клиента чуть раньше и, возможно, увести его у конкурента.

Для обучения модели была собрана большая витрина данных об активности покупателей розничной сети магазинов формата дрогери. Для каждого доступного для коммуникации клиента была выбрана случайная дата за динамический год. На эту дату по нему было рассчитано порядка 2000 показателей, которые можно условно разделить на следующие группы:

- социально-демографические характеристики клиента (пол, возраст, предпочитаемый регион покупки, кластер RFM и т.д.);

- показатели лояльности клиента к сети (транзакции бонусных баллов, стаж, регистрация на сайте и в приложении, откликаемость на коммуникации и т.д.);
- агрегатные статистики покупок в целом и по ключевым товарным категориям (абсолютные агрегаты за различные периоды, отсчитываемые от индивидуальной даты до клиента, их отношения);
- агрегатные статистики покупок по моделируемым товарным категориям (аналогично);
- целевые переменные для моделирования (факт покупки в товарной категории клиентом в течение фиксированного периода от его индивидуальной даты).

В результате анализа истории откликаемости клиентов на предложения был выделен ряд бизнес-правил для их фильтрации. К примеру, клиент должен иметь хотя бы один чек за последние три месяца, купить моделируемый товар хотя бы раз в год, иметь покупки в пределах определённой суммы: не случайный визит, но и не оптовая закупка. Кроме того, из обучающей выборки было решено исключить клиентов, у которых в окно целевой переменной попадает дата так называемого «пика»: резкого всплеска продаж благодаря маркетинговым мероприятиям или иным факторам. Предполагается, что в даты «пиков» вероятность совершить покупку является смещённой, а значит, они создают шум в данных.

Несмотря на то, что наиболее очевидным объяснением «пиков» являются крупные акции, они не всегда корректно заводятся в базу данных и не всегда имеют сильный эффект, поэтому был разработан эвристический алгоритм для обнаружения выбросов во временном ряде суммарных продаж по всем клиентам. «Пики» по каждой из моделируемых товарных категорий анализируются отдельно. Вдохновением для алгоритма послужил один из подходов к определению числа кластеров: по графику «каменистой осыпи». Его автоматизированная версия была применена к абсолютным и относительным ошибкам мультилинейной полиномиальной регрессии, используемой для приближения трендовой и сезонной компонент. При добавлении новой моделируемой категории скрипт выводит аналитику 8 изображений с наиболее удачными вариантами «пиковых» дат, а далее решение принимается экспертно.

Очищенная в соответствии с бизнес-правилами обучающая выборка проходит ещё один этап фильтрации: она подаётся на вход модели изолирующего леса (Isolation Forest) – популярному алгоритму для

обнаружения аномальных значений в данных. Изолирующий лес находит наблюдения, легко отделяемые от всей остальной выборки, и удаляет их.

В качестве основного алгоритма рассматривалось несколько классических моделей: логистическая регрессия, случайный лес, реализации градиентного бустинга решающих деревьев в XGBoost, CatBoost, LightGBM. На ряде предварительных тестов лучше всего себя показала модель LightGBM, поэтому впоследствии для моделей РТВ было решено использовать именно её.

Поскольку решалась задача именно определения вероятности, сходство распределения прогнозируемых вероятностей с фактическими долями пришедших было важнее качества классификации. Именно поэтому в качестве основной метрики оценки точности моделей была выбрана площадь под ROC-кривой (ROC-AUC). Для подбора оптимальных гиперпараметров был использован поиск по сетке допустимых вариантов для максимизации ROC-AUC на кросс-валидации в 5 слоёв за 30-50 эпох.

Первые обученные модели имели неплохую точность на тестовой выборке (среднее значение ROC-AUC около 0.7), однако распределение предсказанных вероятностей по некоторым из них имели неестественное распределение. Для обеспечения в будущем анализировать эффективность коммуникаций в зависимости от вероятности покупки предсказанные значения по таким моделям были откалиброваны с помощью изотонической регрессии. На рисунке 1 представлен пример сравнения фактического распределения с моделируемым. По оси абсцисс расположены предсказанные значения вероятностей, по оси ординат – реальные доли совершивших покупку в персентиле.

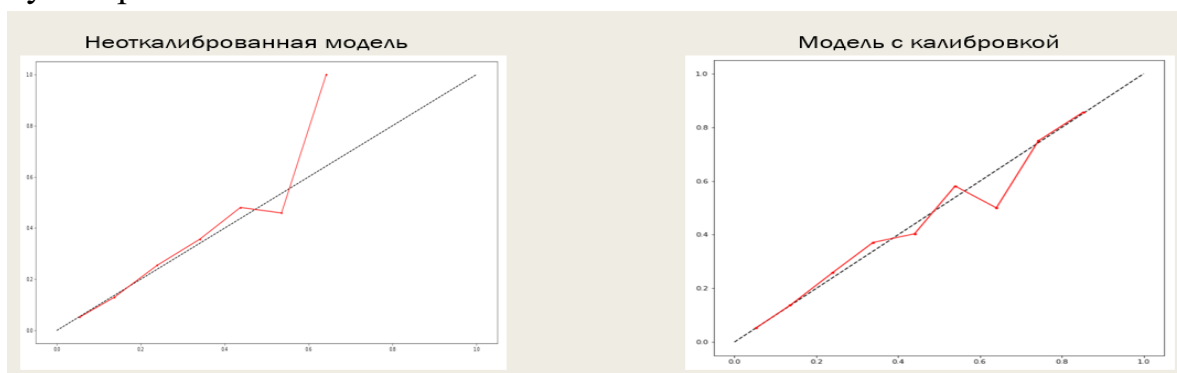


Рис. 2. – Сравнение фактического распределения с моделируемым

Модель после 0.6 начинает занижать вероятности, но калибровка помогает решить эту проблему и приближает полученное распределение к фактическому.

Для экономической оценки эффективности проводимых кампаний используется процент отклика. Он равен разности между долей пришедших в целевой группе (ЦГ, участники коммуникаций) и долей пришедших в контрольной группе (КГ, коммуникаций не производилось). Хотя математическая точность первых моделей и была неплохой, извлечь дополнительную выгоду с их помощью не удалось: отклик по тестовым кампаниям составил меньше 1%.

Первым потенциальным объяснением было разнообразие ассортимента сети и потребностей клиентов: поскольку было смоделировано лишь несколько товарных категорий, основная масса покупателей получила максимально релевантное для них предложение лишь из узкого перечня. В связи с этим было решено масштабировать модель РТВ на все ключевые категории и отказаться от автономных рассылок по одной модели: для каждого клиента определяется вероятность покупки в каждой из доступных для продвижения в определённую дату товарных категорий, и он получает предложение по товару с самой высокой вероятностью. Это было бы трудно осуществимо без калибровки, но при её использовании вероятности получают достаточно однородными.

Второе объяснение состоит в том, что коммуникации сами по себе могут быть неэффективными даже при очень хорошем прогнозе. Дополнительный доход возникает только тогда, когда коммуникации производятся только с клиентами, которые пришли в магазин исключительно благодаря рассылке. Это наглядно представлено в таблице 1.

Таблица 1

Экономическая эффективность коммуникации в зависимости от склонности клиента к отклику

	Откликнется без коммуникации	Не откликнется без коммуникации
Откликнется с коммуникацией	Убыток = рассылка + скидка	Прибыль
Не откликнется с коммуникацией	Убыток = рассылка + упущенная прибыль	Убыток = рассылка

Так возникла идея создать дополнительную модель склонности клиента откликаться на коммуникацию и приносить дополнительный доход. Первая её итерация строилась на информации по прошедшим коммуникациям и принесла первые плоды: отклик более 3%, что раньше считалось очень хорошим показателем на гораздо меньшей выборке.

Тем не менее, поскольку история коммуникаций начала вестись в удобном формате относительно недавно, данных для моделей почти не было, однако было замечено, что в выборке малоэффективных клиентов преобладают те, кто пришёл бы и так, и те, кто точно не пришёл бы. Так родилась вторая итерация модели: вероятность совершения любой покупки в течение следующих 5 дней. По своей логике она почти полностью идентична модели РТВ на товарную категорию. В результате анализа эффективности прошедших кампаний были определены оптимальные отсечки для модели (4% и 35%). Распределение прогнозируемых вероятностей прихода представлено на рисунке 2.

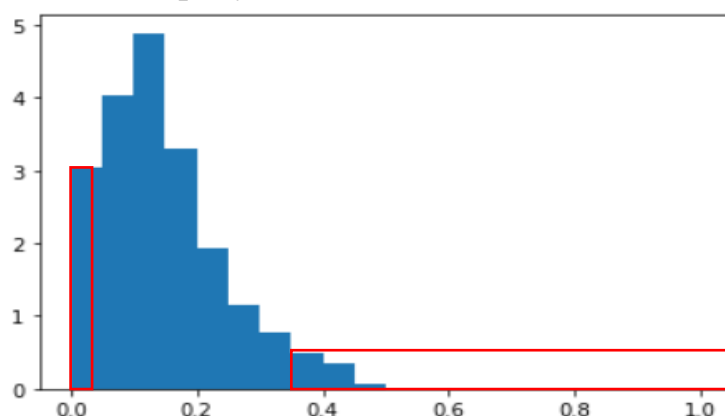


Рис. 3. – Распределение вероятностей совершить покупку в течение следующих 5 дней

С использованием дополнительной модели и одновременным запуском кампаний по нескольким товарным категориям удалось достичь высоких значений прибыли для каждой кампании (запускаются почти ежедневно). На текущий момент построены модели по 40 категориям, из которых подвергнуто калибровке 29. Среднее значение ROC-AUC на обучающей выборке составляет 0.7884, на тестовой – 0.7379. Средний отклик по кампаниям ноября-декабря 2019 года составил почти 4%, и это очень хороший результат при регулярных запусках на больших выборках.

На рисунке 3 представлена полная схема построения модели РТВ на конкретную категорию.

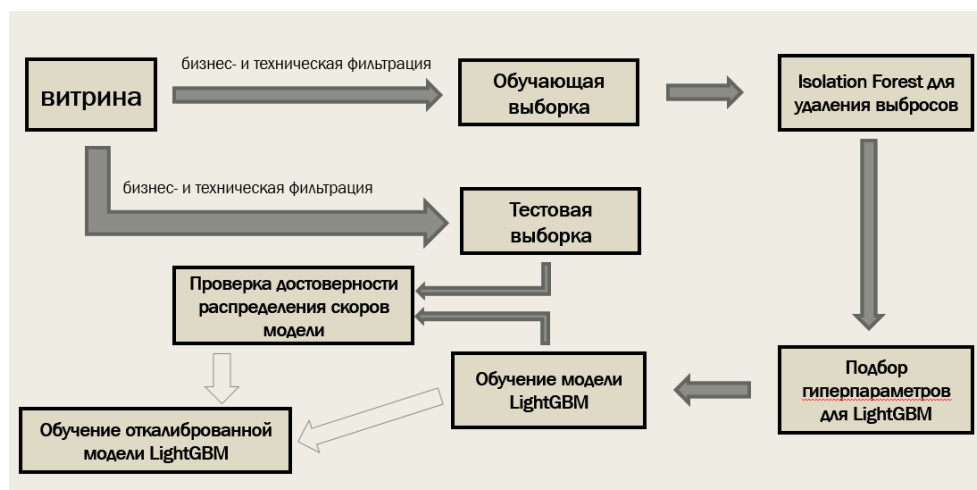


Рис. 4. – Схема обучения модели РТВ

Модель находится в промышленной эксплуатации, а кампании по ней уступают по прибыльности только поддержке массовых промо-акций.

Литература

1. Положение о структуре и оформлении письменных работ обучающимися по программам среднего профессионального и высшего образования в ФГБОУ ВО «Санкт-Петербургский государственный экономический университет» [Электронный ресурс]. – Санкт-Петербург: СПбГЭУ, 2019. – Режим доступа: https://unecon.ru/sites/default/files/s_zamenoy_27_list1.pdf, свободный (дата обращения 07.01.2020).
2. Liu F.T. Isolation Forest [Электронный ресурс]/ Liu F.T., Ting K.M., Zhou Z.H. – Режим доступа: <https://cs.nju.edu.cn/zhoush/zhoush.files/publication/icdm08b.pdf>, 2008, свободный (дата обращения 09.01.2020)
3. Niculescu-Mizil A. Predicting Good Probabilities With Supervised Learning [Электронный ресурс]/ Niculescu-Mizil A., Caruana R. – Режим доступа: <https://www.cs.cornell.edu/~alexn/papers/calibration.icml05.crc.rev3.pdf>, свободный (дата обращения 09.01.2020)
4. Royston-Webb T. Propensity Modelling for Business // A Data Science Foundation White Paper. – 2018.

Literature

1. The reglament about the structure and formalization of written papers of students of secondary vocational and higher education programs in “Saint Petersburg State University of Economics” [Electronic resource]. – Saint Petersburg: SPBSUE, 2019. – Access mode: https://unecon.ru/sites/default/files/s_zamenoy_27_list1.pdf, free (access date 07.01.2020).

2. Liu F.T. Isolation Forest [Electronic resource]/ Liu F.T., Ting K.M., Zhou Z.H. – Access mode: <https://cs.nju.edu.cn/zhoush/zhoush.files/publication/icdm08b.pdf>, 2008, free (access date 09.01.2020)
3. Niculescu-Mizil A. Predicting Good Probabilities With Supervised Learning [Electronic resource]/ Niculescu-Mizil A., Caruana R. – Access mode: <https://www.cs.cornell.edu/~alexn/papers/calibration.icml05.crc.rev3.pdf>, free (access date 09.01.2020)
4. Royston-Webb T. Propensity Modelling for Business // A Data Science Foundation White Paper. – 2018.