ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ЦИФРОВАЯ ЭКОНОМИКА

УДК 004.89:339.13

ПОИСК ЗАКОНОМЕРНОСТЕЙ В ИНТЕРНЕТ-РЕКЛАМЕ НА ОСНОВЕ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ

Монастырская Мария Михайловна (182271@edu.fa.ru)
Финансовый университет при Правительстве Российской Федерации

Работа посвящена совершенствованию управления взаимоотношениями с клиентами посредством анализа данных и машинного обучения.В данном исследовании предлагается подход, который заключается в построении модели, основанной на регрессионном и кластерном анализе, для поиска закономерностей в поведении пользователей относительно маркетинговых кампаний с учетом характеристик пользователей и финансово значимых метрик. Для построения модели был проведен сравнительный анализ работы алгоритмов, таких как линейная регрессия, деревья решений, случайные леса, AdaBoost и XGBoost, для предсказания значения метрики эффективности маркетинговых кампаний, и К-средних, EM-алгоритм и иерархическая кластеризация для поиска сегментов в поведении пользователей, а также выполнен поиск ассоциативных правил для анализа выявленных кластеров. С точки зрения практического применения подхода, были проведены сбор и обработка данных по интернет-рекламе двух организаций, была реализована модель по прогнозированию стоимости привлечения CPA (cost-per-action). Для анализа закономерностей в системе был применен кластерный анализ и поиск ассоциативных правил по маркетинговым объявлениям. Такая модель позволит организациям совершенствовать настройки рекламных объявлений для повышения эффективности интернет-рекламы в целом.

Ключевые слова: регрессионный анализ, маркетинговые метрики, цена за действие, интернет-реклама, кластерный анализ, поиск закономерностей.

В середине XX в. рынок потребителя пришел на смену рынку производителя, что привело к высокой конкуренции между компаниями. В ходе борьбы за клиента организации стали более клиентоориентированными, в том числе посредством внедрения технологий для обеспечения персонализации и удобства взаимодействия между компанией и покупателем. Несмотря на значительное количество функциональных возможностей систем для управления взаимоотношениями с клиентами, часть бизнес-проблем до сих пор не решаются функциональными возможностями стандартного программного обеспечения. К таким проблемам можно отнести задачи актуализацию данных о клиентах, обогащение данных, скоринг клиентов, управление маркетинговыми кампаниям и др. Технологии анализа данных широко распространены для решения бизнес-задач, однако в сфере управления взаимоотношениями с клиентами алгоритмы машинного обучения используются достаточно редко.

Значительную долю бюджета компании выделяют на привлечение клиентов. Доля интернет-рекламы существенно выросла за последние 15 лет (с 2% в 2005 г. до 40% в 2019 г.) и стала основным каналом продвижения [15]. Использование интернет-рекламы привело к более высокому уровню конкуренции. На данный момент большинство маркетологов управляют рекламой на основе отчетности, построенной вручную, что позволяет проанализировать только несколько основных разрезов. Такой подход не всегда является эффективным, так как для ручного перебора всех комбинаций требуются слишком большие трудозатраты. Более того, такую активность необходимо проводить на регулярной основе, поэтому оптимизация и совершенствование данных процессов возможна только посредством технологий анализа данных и машинного обучения.

В данной работе рассматривается применение методов машинного обучения для совершенствования процесса привлечения клиентов, а именно поиск паттернов поведения пользователей в интернет-маркетинге. Паттерн – систематически повторяющийся, устойчивый элемент (фрагмент поведения) либо последовательность таких элементов [2]. Выявление закономерностей в поведении пользователей может обеспечить возможность внесения корректировок в маркетинговые кампании путем установления ограничений и затрат на клик, что, в свою очередь, может привести к росту их эффективности, так как рекламные площадки для интернет-рекламы позволяют не только настраивать само рекламное объявление, но также настраивать аудиторию, которой его необходимо показывать.

На сегодняшний день используются различные показатели для оценки эффективности маркетинговой активности, в том числе CPO (costperorder, или стоимость заказа). Показатели метрик необходимо не только измерять, но и уметь предсказывать, так как это влияет на эффективность

маркетинговых активностей и стратегии в целом. Прогнозирование значение метрик реализуется с помощью алгоритмов регрессионного анализа, которые позволяют прогнозировать эффективность рекламных объявлений или значение других маркетинговых метрик. Для поиска закономерностей в поведении объектов системы, а именно, пользователей с точки зрения анализа маркетинговых кампаний, применяются методы кластерного анализа, а также поиск ассоциативных правил.

Для решения различных задач кластеризации в интернет-рекламе широко используются метод K-средних и метод максимизации ожиданий (EM), например, для исследования положительной реакции на рекламные объявления [3] и поиска взаимосвязей между рекламными объявлениями и пользователями [11]. Ряд исследований проводится на тему анализа покупательского поведения для предоставления клиентам наиболее релевантных услуг посредством кластеризации на основе посещений клиентов. Например, в работе [5] в основе алгоритма поиска сегментов клиентов лежит поиск ассоциативных правил, которые можно интерпретировать как клиентские сегменты, где вместо набора данных с транзакциями можно использовать статистику по активности посетителей веб-сайта, а в качестве элементов в транзакции можно передавать описательные характеристики пользователей и рекламных объявлений.

Данные могут быть собраны из сервисов веб-аналитики, таких как GoogleAnalytics и Яндекс.Метрика, с помощью пользовательских инструментов отчетности или посредством выполнения запросов через программный интерфейс системы [4]. Набор данных должен состоять из атрибутов, связанных с действием пользователя, рекламной кампанией и самим пользователем, а также значением метрики, например, цены за действие (CPA) (рис. 1).

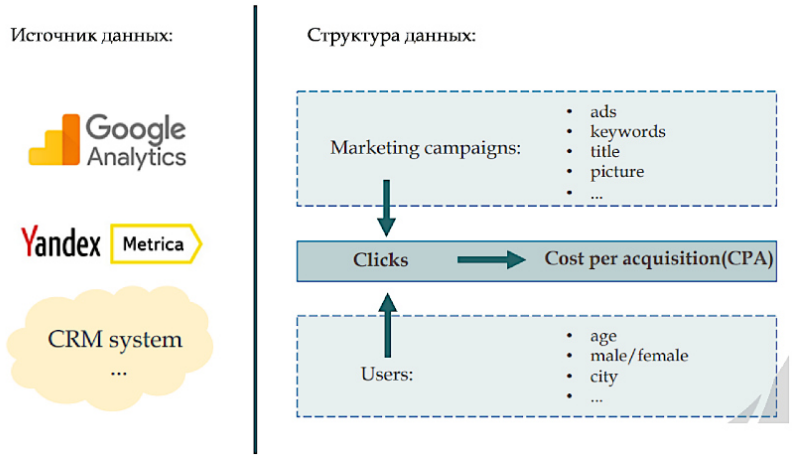

Рисунок 1. Источники данных для совершенствования маркетинговых компаний

Основные данные можно получить из систем сбора статистики о посещаемости сайта, дополнительную информацию можно получить из CRM-системы или внешних источников. Данные о рекламных объявлениях должны содержать описательные признаки, которые позволят выделить различные кластеры объявлений. Данные о клиентах в аналитических инструментах для интернет-маркетинга содержат системную информацию о пользователях, такую как возраст, пол, географические данные, а также информацию об используемой информационной системе и интересах [5]. Помимо системной информации, которую собирают сервисы для создания детальной статистики посетителей веб-сайтов, можно обогатить таблицу дополнительной информацией. Что касается активности клиентов на веб-сайте, то информацию можно получить на основе отчетов аналитических систем для сбора данных о поведении клиентов на веб-сайте. На основе полученного набора данных производится моделирование для поиска закономерностей в интернет-рекламе.

Анализ паттернов может быть реализован путем объединения алгоритмов прогнозирования маркетинговых метрик и кластеризации клиентов для получения достоверной картины о работе маркетинговых кампаний и их эффективности. Прежде всего, алгоритмы регрессионного анализа следует использовать для прогнозирования цены за действие (CPA), чтобы использовать для дальнейшего моделирования эффективности маркетинговых кампаний для заданных параметров [8].

Таким образом, на первом этапе предобработки данных необходимо экспортировать данные из систем анализа веб-сайтов и обогатить данными из внешних

источников, а также преобразовать полученную схему данных в таблицу данных для анализа. На втором этапе моделирования необходимо реализовать модели на основе алгоритмов прогнозирования и кластеризации, применяя кросс-валидацию. На последнем этапе оценки требуется оценить качество моделей для выбора наиболее эффективных вариантов, провести бизнес-анализ на основе полученных результатов и разработать новые правила распределения маркетингового бюджета и настройки активных рекламных объявлений (рис. 2). Результат работы настроенного алгоритма – найденные закономерности поведения пользовательских кластеров при клике по объявлениям в интернете. Паттерны могут быть интерпретированы веб-аналитиком или специалистом в бизнес отрасли, на которой специализируется компания. После настройки новых маркетинговых кампаний и ограничений, процесс определения паттернов может быть воспроизведен повторно через некоторое время. В результате модель может найти новые паттерны или сделать предыдущий прогноз более точным.

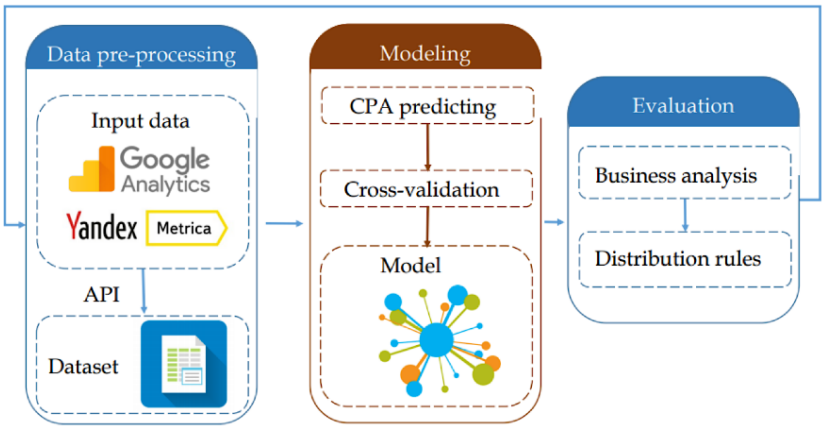

Рисунок 2. Этапы построения модели

Для решения данной задачи предсказания значения ключевой метрики эффективности маркетинговой кампании необходимо построить модель регрессионного анализа. Для получения наиболее точного прогноза предлагается применение следующих алгоритмов: линейная регрессия [1], деревья решений [12], AdaBoost [6], случайный лес [9], а также XGBoost [10]. Характеристики качества построенной модели регрессии отражают, насколько достоверно выбранная модель описывает исходные данные исследования и применима к формированию прогноза для новых значений модели. Средняя абсолютная ошибка (MeanAbsoluteError, MAE) позволяет получить оценку расстояния между фактическимии прогнозными значениями, значение показателя необходимо минимизировать. Основным преимуществом данного показателя является то, что он устойчив к выбросам [13].

Для поиска закономерностей используются данные, полученные в результате прогнозирования значения маркетинговой метрики. Данная задача решается на основе применения методов кластерного анализа на основе алгоритмов K-средних [9], максимизации ожиданий (EM) [16], иерархической кластеризации [16]. Помимо классических методов кластерного анализа, в данной работе предлагается решить задачу посредством алгоритма поиска закономерностей на основе поиска ассоциативных правил [5]. Этот метод машинного обучения находит связи между переменными в исходном наборе данных, в основе которых лежит мера интенсивности. Для поиска ассоциативных правил применяется ряд алгоритмов, из которых наиболее широко используемым является алгоритм Apriori.

Этап поиска паттернов поведения сложен с точки зрения оценки качества работы модели, так как для моделей кластеризации нельзя однозначно определить оптимальный алгоритм кластеризации, также часто в решении прикладных задач алгоритмы кластеризации не могут выявить настоящее количество кластеров для набора данных. В рамках данного исследования применяется метрика, которая сочетает в себе компактность и отделимость кластеров, а именно силуэт, который показывает, насколько элемент кластера похож на свой кластер относительно других.

В данном исследовании разработка модели осуществляется на основе маскированных данных организации XYZ по рекламным кампаниям [7]. Данные содержат информацию об активности пользователей после перехода по рекламному объявлению в разрезе различных описательных характеристик пользователей. Источником данного набора

данных является личный кабинет GoogleAdwards данной организации.

Набор данных содержит 1143 записи с 11 признаками. Для построения модели анализа паттернов поведения пользователей была выполнена предобработка данных. При обработке данных была выполнена проверка на пропущенные значения, в набор данных был добавлен новый признак, отражающий значение цены за действие (CPA), которая рассчитывается как отношение стоимости рекламы к количеству целевых действий. Было выполнено преобразование типов данных, для предобработки числовых параметров с целью улучшения качества модели была получена описательная статистика по набору данных.На основе полученных данных были обработаны выбросы в показателях, для категориальных признаков было проанализировано распределение значений по количеству объектов. Далее был выполнен корреляционный анализ оставшихся признаков, признаки с высоким коэффициентом детерминации были удалены. Категориальные признаки были закодированы методом One-HotEncoding.Таким образом, полученный набор данных содержит 1112 наблюдений по 40 признакам. Также для дальнейшей оценки качества модели выборка была разделена на тестовую и обучающую в отношении 20:80. Для более точной оценки моделей обучающий набор данных был разбит на 10 частей для кросс-валидации.

Первым алгоритмом для предсказания значения маркетинговой метрики является линейная регрессия. Для данного алгоритма изменение значения гиперпараметров не дает улучшения качества модели, так как изменение количества задействованных процессоров не приводит к изменению метрики качества, а нормализация только ухудшает конечный результат. Поэтому оценка качества модели при параметрах по умолчанию является финальной. В данном случае средняя абсолютная ошибка составила 3 598 575.

Для улучшения качества прогнозирования цены за действие (CPA) был применен алгоритм случайного леса. Случайный лес при значениях по умолчанию показывают среднюю абсолютную ошибку, равную 1 960 423, что хуже работы алгоритма деревья решения с оптимизированными гиперпараметрами. После подбора гиперпараметров при глубине деревьев, равной 6, на рассматриваемом наборе данных модель выдала ошибку MAE, равную1 814 407.

Для того, чтобы улучшить качество предсказания, был применен бустинг на основе алгоритма AdaBoost. При значении параметров по умолчанию средняя абсолютная ошибка составляет 1 900 110.После подбора гиперпараметров оптимальное значение глубины дерева оказалось равным 6, минимальное число объектов в листе и объектов, по которым выполняется расщепление – 0,1. В результате MAE составила 1 775 887.

Также для предсказания значения цены за действие (CPA) был применен алгоритм XGBoost. При глубине деревьев 3 и деревьям в лесу 90, данный алгоритм дал ошибку 1 737 275, что меньше, чем все остальные алгоритмы. Результаты сравнения моделей представлены на рисунке 3.

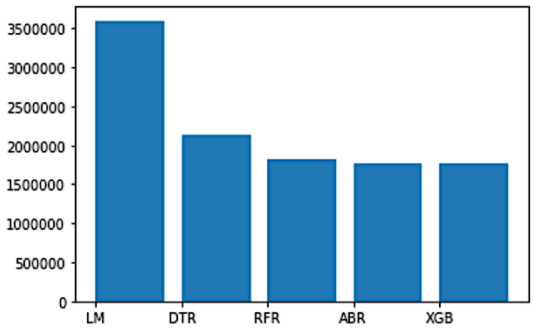

Рисунок 3. Сравнение значение MAE при прогнозировании CPA

Таким образом, наиболее точный прогноз удалось получить на основе алгоритма XGBoost. На основе полученных результатов модель может предсказывать значение Costperaction для заданного сегмента аудитории и определенного рекламного объявления. Набор данных с предсказанным значением маркетинговой метрики позволяет смоделировать эффективность рекламной кампании при запуске без дополнительных расходов на рекламные объявления.

Вторая часть модели по поиску паттернов поведения основана на поиске закономерностей в откликах пользователя на рекламное объявление с учетом финансового анализа показателей для повышения эффективности изменений в маркетинговых кампаниях. В первую очередь для сегментации предлагается использовать кластерный анализ. Для построения модели на основе алгоритма K-средних необходимо определить количество кластеров для модели. На основе метода «локтя» наиболее оптимальным количеством кластеров является 5. Для выбранного количества кластеров была проведена сегментация алгоритмом К-средних и проведен анализ полученных сегментов. Например, наименьшая цена за действие характерна для кластера под номером 3. Этот сегмент свойствен для маркетинговой кампании «xyz_campaign_id_1178» для аудитории мужчин в возрасте от 30 до 39 лет.

Кластерный анализ был также проведен на основе EM-алгоритма, для кластеризации было задано 5 разбиений на основе анализа значений информационных критериев.Для сравнения полученных кластеров был построен график по значению цены за действие (CPA) для каждого найденного сегмента. Значение метрики по каждому сегменту рассчитывалось как среднее. Наименьшая цена за действие наблюдается у сегмента под номером 1 со значением метрики, равной 1.22. Данный сегмент является схожим с наиболее эффективным сегментом, найденным алгоритмом K-средних, он также характерен для маркетинговой кампании «xyz_campaign_id_1178», а аудитория данного сегмента состоит преимущественно из мужчин.

Для кластеризации поведения пользователей относительно интернет-рекламы также была проведена иерархическая кластеризация с силуэтной метрикой. Для реализации алгоритма был выбран агломеративный подход. В качестве метрики объединения кластеров было использовано расстояние ближайшего соседа. При сегментации данных на 5 кластеров наилучший показатель цены за действие (CPA) показывает сегмент 1 со значением метрики 2.68.

Несмотря на то, что наиболее эффективные для каждого алгоритма кластеры совпадают по основным характеристикам, в целом EM-алгоритм хуже определяет сегменты в соответствии с показателем силуэта, который составляет 0.8 для K-средних, -0.38 для EM-алгоритма. Иерархическая кластеризация работает немного хуже метода K-средних с силуэтом в 0.78. Пересечение кластеров можно видеть на графике визуализации кластеров для алгоритмов (рис. 4).

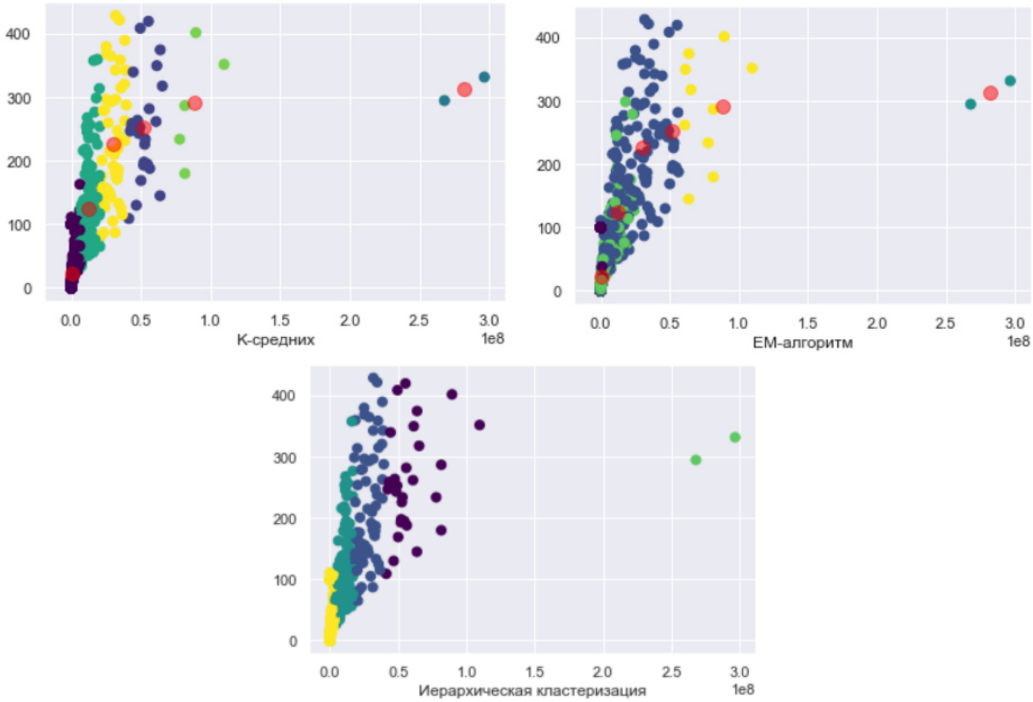

Рисунок 4. Сравнение алгоритмов кластеризации для поиска паттернов в рекламных кампаниях

Наиболее высокого качества сегментации удалось достичь на основе алгоритма K-средних. Для совершенствования рекламных кампаний проведен анализ всех полученных кластеров. Среднее значение стоимости привлечения клиентов варьируется от 1.02 до 7.41. Для повышения эффективности необходимо максимизировать эффективность маркетинговых кампаний с низким показателем стоимости привлечения и отключить убыточные сегменты.

На основе описательной статистики по каждому кластеру и анализа параметров, характерных для полученных кластеров, например, кластеров под номерами «3» и «2», рекомендуется увеличить ставки для рекламной кампании «1178», а также отдать преимущество аудитории мужского пола в возрасте от 30 до 34 лет, у которых отмечен интерес под номером «10». Второй прибыльный сегмент, на который стоит обратить внимание, также относится к маркетинговой кампании «1178» среди женщин в возрасте от 45 до 49 лет с интересами «16» и «18». С точки зрения сегментов с наиболее высокой стоимостью привлечения можно выделить маркетинговую кампанию «936», по рекламным объявлениям которой переходили

женщины в возрасте от 30 до 34 лет. Также необходимо отметить, что кластер под номером «0» схож по всем характеристикам с кластером «2», однако стоимость привлечения различается в несколько раз, основной отличительной чертой является наличие интереса «18»у 2-го кластера.Такие сегменты рекомендуется отключить для данной организации.

Для реализации поиска закономерностей был применен алгоритм Apriori. Гиперпараметры функции позволяют настраивать, насколько сильными должны быть связи между объектами для определения ассоциативных правил, а именно минимальный уровень поддержки, минимальное значение достоверности для правил, минимальный уровень зависимости, а также максимальное количество элементов в правиле. Были подобраны следующие оптимальные значения гиперпараметров: 0.003, 0.2, 4 и 2 соответственно. Так как для поиска ассоциативных правил требуется набор данных с транзакциями с участвовавшими в них объектами в качестве набора данных для анализа, между которыми необходимо установить связь, признак CPA был преобразован в категориальный посредством разбиения значения на интервалы. Названия значений нового категориального признака были названы в соответствии со значением метрики, то есть самая низкая цена за действие у категории «CPA1», в том время как самым дорогим сегментом является «CPA2». Для визуализации найденных правил был построен граф (рис. 5).

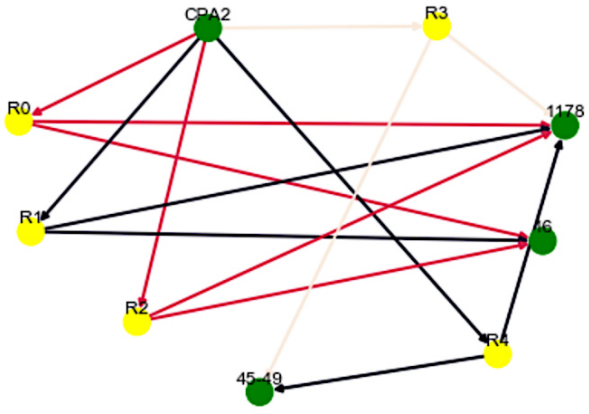

Рисунок 5. Визуализация найденных закономерностей в маркетинговых активностях

При добавлении маркетинговой метрики в набор данных мы получаем, что достаточно низкая цена за действие CPA характернадля рекламной кампании«1178», но среди пользователей с интересом «16», а также среди пользователей возрастной категории «45-49». Всего алгоритм определил 36 ассоциативных правил. Наблюдается высокая зависимость значения метрики в интервале «CPA2» для кампании «1178» для возрастной категории «40-49» и категории интереса«16». Высокая доля мужчин в кластерах, найденных алгоритмом K-средних, объясняется высокой заинтересованностью этого сегмента в кампании «1178» без привязки к эффективности рекламы. Категория интереса«16» также характерна для значений цены за действия в интервале«CPA1».

Для внесения изменений в настройки рекламных кампаний критически важно понимать не только общие тенденции среди пользователей, но и то, как найденные закономерности связаны с их финансовой эффективностью, так как наиболее активный сегмент не всегда является самым прибыльным. Чтобы проверить эффективность предложенных изменений для совершенствования маркетинговых кампаний, также предлагается рассчитать для нихпрогнозное значение цены за действие. Описанные значения характеристик были заданы в новой тестовой выборке, где для двух правил были указаны равные бюджеты по 1000, для первой гипотезы указана маркетинговая кампания«1178» с интересами «10», «16» и «18» и двумя возрастными категориями «30-34» и «45-49». Вторая гипотеза идентична первой, но только для аудитории мужчин. Третья гипотеза относится к кампании «936» без привязки к возрастным группам и полу. Действительно, разницы между первой и второй гипотезой не наблюдается, стоимость привлечения составляет 2.24 и 2.33 соответственно и является достаточно низкой, т. е. пол пользователя не влияет на финансовую эффективность, что показали ассоциативные правила. Третья гипотеза подтвердила, что кампания «936» не является эффективной и имеет высокую цену за действие.

На основе полученных результатов можно определить новые правила распределения маркетингового бюджета и перенастроить маркетинговые кампании. После внесения изменений можно повторить процедуру с учетом новых данных и изменения поведения пользователей при новых настройках рекламы и построить процесс непрерывного улучшения онлайн-рекламы организации.

Для проверки эффективности работы модели был проведен анализ поведения клиентов относительно объявлений интернет-рекламы на основе данных GoogleMerchandiseStore [14]. Данные интернет-магазина с брендированной продукцией Google доступны в рамках доступа к демо-аккаунту в кабинетеGoogleAnalytics и содержат информацию о трафике на сайт, каналах

привлечения клиентов (органический поиск и платный трафик), поведении пользователей на сайте (просмотр страниц, покупка товаров и т.д.), а также данных о транзакциях (заказах и информации о них). В качестве целевой метрики выбрана цена за действие, так как для данного сайта настроено отслеживание целевых действий до момента оформления заказов. Для анализа были выгружены данные за период с 1 января по 1 апреля 2020 г. Набор данных состоит из 1072 наблюдений. После предобработки данных в наборе данных осталось 1018 наблюдений и 46 переменных.

Для предсказания значения CPA были построены модели на основе алгоритмов, рассмотренных ранее. Для каждого алгоритма были подобраны оптимальные значения гиперпараметров. Для выбора оптимальной модели для прогнозирования цены за действие в GoogleMerchandiseStore для каждого из алгоритмов было определено значение средней абсолютной ошибки предсказания, которое варьируются от 2.603 для линейной регрессии до 1.740 для алгоритма XGBoost. Дальнейший поиск закономерностей проводится на основе модели XGBoost, для которой заданы следующие гиперпараметры: $\alpha = 0.1, \gamma = 4,$, число деревьев в лесу равно 50, глубина деревьев равна 3, а минимальное количество наблюдений в листе дерева задано на уровне 0.5.

Для поиска закономерной в поведении пользователей был проведен кластерный анализ на основе алгоритмов К-средних, EM-алгоритма и иерархической агломеративной классификации. Выбор оптимального количества кластеров был выполнен методом «локтя» на основе оптимизации метрик качества сегментации. Для сравнения полученных результатов для каждого алгоритма была рассчитана метрика силуэта. Наилучший результат показал алгоритм К-средних, для которого силуэт оказался равен 0.52 для данного набора данных, в то время как значение силуэта для EM-алгоритма и иерархической классификации составило–0.23 и 0.35 соответственно.

Был проведен анализ кластеров, выделенных алгоритмом К-средних. Наиболее эффективные с точки зрения минимизации значения цены за действие кластеры представлены по рекламной кампании «id_campaign_1686803890» для пользователей, которые искали название интернет-магазина, или тех, кто искал товары с мобильного телефона. Самый неэффективный сегмент относится к маркетинговым кампаниям «id_campaign_1686803890» и «id_campaign_1687828286» для запросов товаров, по ключевым словам, относящимся к YouTube.

На основе полученных результатов также был выполнен поиск ассоциативных правил для выявления закономерностей в поведении пользователей, которые пришли на сайт с платных каналов привлечения. Алгоритм нашел 51 ассоциативное правило при минимальных уровне поддержки 0.01 и уровне достоверности 0.8.

Анализ ассоциативных правил подтвердил, что низкая цена за действие характерна для маркетинговой кампании «id_campaign_1687828286» по ключевым словам, связанным с названием интернет-магазина и профилем реализуемой продукции. Для данного интернет-магазина рекомендуется проанализировать контент рассматриваемых кампаний по рекламным объявлениям бренда YouTube, так как высокая стоимость привлечения может быть связана с доступностью товаров или нерелевантными предложениями. Если проблема не связана с контентом и ассортиментом, то по данным ключевым словам рекомендуется отключить рекламные кампании.

Таким образом, разработанные рекомендации позволяют на основе стандартных инструментов систем для веб-анализа строить отчеты, которые могут быть использованы в качестве исходных данных для разработанной модели. Модель может быть применена в различных организациях для совершенствования управления взаимоотношениями с клиентами, например, для привлечения целевых пользователей на сайт.

Задачи, поставленные в рамках данной работы, выполнены, цель достигнута. Перспективными направлениями дальнейших исследований являются применение поиска закономерностей в других процессах по управлению взаимоотношениями с клиентами, например, в продажах и послепродажном обслуживании, разработка новой модели на основе более глубокого финансового анализа, в том числе на основе анализа маржинального дохода, а также автоматизация внесения изменений в процесс на основе полученных результатов.

Литература

1. Соловьев В.И. Анализ данных в экономике. Теория вероятностей, прикладная статистика, обработка и визуализация данных в MicrosoftExcel. М.: КНОРУС, 2018. 500 с.
2. Ценёв В. Словарь нейролингвистического программирования (НЛП). URL:

http://psyberia.ru/zip/lingvonlp.zip (дата обращения: 10.05.2020).
3. Boonjing V., Pimchangthong D. Data mining for customers' positive reaction to advertising in social media // Proceedings of the Federated Conference on Computer Science and Information Systems, Annals of Computer Science and Information Systems. 2017. V. 11. P. 945–948. DOI: 10.15439/2017F356
4. Brendan McMahan H., Holt G., Sculley D. Ad click prediction: A view from the trenches // KDD'13: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, P. 1222–1230. DOI: 10.1145/2487575.2488200.
5. Griva A., Bardaki C., Pramatari K., Papakiriakopoulos D. Retail business analytics: Customer visit segmentation using market basket data // Expert Systems with Applications. 2018. V. 16. P. 10–16. DOI: 10.1016/j.eswa.2018.01.029.
6. Freund Y., Schapire R.E. A decision-theoretic generalization of on-line learning and an application to boosting // Journal of Computer and System Sciences. 1997. V. 55. P. 119–139. DOI: 10.1006/jcss.1997.1504.
7. Hu Y., Shin J., Tang Z. Incentive problems in performance-based online advertising pricing: Cost per click vs cost per action // Management Science. 2016. V. 62. P. 2022–2038. DOI: 10.1287/mnsc.2015.2223.
8. Richardson M., Dominowska E., Ragno R. Predicting Clicks: Estimating the Click-Through Rate for New Ads // WWW '07: Proceedings of the 16th international conference on World Wide Web. 2007. P. 521–530.DOI:10.1145/1242572.1242643.
9. Steinhaus H. Sur la division des corps materiels en parties // Bulletin de l'Académie Polonaise des Sciences. 1956. V. 4. P. 801–804.
10. Xinran H., Junfeng P., Ou J., et al. Practical lessons from predicting clicks on ads at Facebook //ADKDD'14: Proceedings of the Eighth International Workshop on Data Mining for Online Advertising. 2014. P. 1–9. DOI: 10.1145/2648584.2648589.
11. Yang H., Zhu Y., He J. Local algorithm for user action prediction towards display ads //KDD'17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017. P. 2091–2099. DOI: 10.1145/3097983.3098089.
12. Импортданных: руководство для разработчиков. URL: https://developers.google.com/analytics/devguides/config/mgmt/v3/mgmtDataImport (дата обращения: 10.05.2020).
13. Начало работы с Google Аналитикой. URL: https://support.google.com/analytics/answer/1008015?hl=ru (дата обращения: 10.05.2020).
14. Google Analytics Spreadsheet Add-on. URL: https://developers.google.com/analytics/solutions/ google-analytics-spreadsheet-add-on (дата обращения: 10.05.2020).
15. Sales Conversion Optimization. URL: https://kaggle.com/loveall/clicks-conversion-tracking (дата обращения: 15.04.2020).
16. Wu M.-L., Chang C.-H. Aggregate two-way co-clustering of ads and user data for online advertisements // Journal of Information Science and Engineering. 2012. V. 28. P. 83–97.

УДК 004.89:368

ДВУХЭТАПНАЯ МОДЕЛЬ МАШИННОГО ОБУЧЕНИЯ С ИСПОЛЬЗОВАНИЕМ ДЕРЕВЬЕВ LIGHTGBM ДЛЯ ПРОГНОЗИРОВАНИЯ СТРАХОВЫХ РЕЗЕРВОВ

Соловьев Владимир Игоревич (VSoloviev@fa.ru)
Феклин Вадим Геннадьевич
Жукова Анастасия Сергеевна
Финансовый университет при Правительстве Российской Федерации

Рассматривается проблема прогнозирования страховых резервов на микроуровне без агрегирования данных для анализа с использованием деревьев решений с бустингом. Предлагается использование деревьев LightGBM в двухэтапной модели. На первом этапе определяется, есть ли претензии по контракту, которые возникли, но не были представлены (IBNR), или нет, а на втором этапе прогнозируется страховой резерв для случаев IBNR. Показано, что предложенная методика более эффективна, чем традиционные методы.

Ключевые слова: страхование, расчет страховых резервов, оценка размера отдельных страховых выплат, LightGBM.

Введение

Прогнозирование страховых резервов является крайне важной актуарной задачей, поскольку важно быть уверенными в достаточности резервов для покрытия убытков. При этом наиболее сложной проблемой является оценка претензий, которые возникли, но не были представлены (*Incurred But Not Reported, IBNR*) [1].

Традиционно для прогнозирования резервов используются метод цепной лестницы, метод Борнхьюттера – Фергюсона,