УДК 004:519.6

ПРИМЕНЕНИЕ ПОКАЗАТЕЛЯ ROC AUC ДЛЯ ОБОСНОВАНИЯ НАИБОЛЕЕ ЭФФЕКТИВНЫХ МОДЕЛЕЙ И АЛГОРИТМОВ АТРИБУЦИИ

Хайдаров А.Г., Соловьев А.И., Будко Д.А.

ФГБОУ ВО «Санкт-Петербургский государственный технологический институт (технический университет)», Санкт-Петербург,
e-mail: andreyhaydarov@gmail.com, and74sol@gmail.com, dmitrii.budko21@mail.ru

Проведено сравнение различных методов и алгоритмов атрибуции посредством проведения оценки точности рассматриваемых алгоритмов с использованием показателя ROC AUC. Рассматривается суть эвристических моделей атрибуции, которые используются во многих системах аналитики, в связи с простотой их использования. Для сравнения также рассматриваются алгоритмические модели атрибуции, такие как «Атрибуция на основе цепей Маркова» и «Атрибуция на основе данных». В них анализируется влияние присутствия определенного канала на значение конверсии. Для сравнения этих моделей и алгоритмов атрибуции было использовано обучение алгоритмов XGBoost, Random forest, алгоритма градиентного бустинга GBM и LightGBM, на пользовательских наборах данных, которые содержат информацию о работе реального интернет-магазина электронной коммерции. Для наглядности данные были взяты за 1, 2, 4, 8, 12 месяцев. В дальнейшем данные были нормализованы и приведены к приемлемому формату для используемых алгоритмов обучения. После чего был проведен непосредственно процесс обучения моделей и проверка точности классификации с использованием метода расчёта – AUC. Для выявления более производительного алгоритма для атрибуции был проведен сравнительный анализ времени работы данных моделей. Полученные результаты могут быть использованы для анализа эффективности и расширения возможностей применения этих моделей.

Ключевые слова: цифровая реклама, атрибуция каналов, мультиканальная атрибуция, машинное обучение, ROC AUC

STUDY OF THE MOST EFFICIENT MODELS AND ATRIBUTION ALGORITHMS USING THE ROC AUC INDICATOR

Khaidarov A.G., Soloviev A.I., Budko D.A.

Saint-Petersburg State Institute of Technology (Technical University), Saint-Petersburg,
e-mail: andreyhaydarov@gmail.com, and74sol@gmail.com, dmitrii.budko21@mail.ru

A comparison of various attribution methods and algorithms was made by assessing the accuracy of the considered algorithms using the ROC AUC indicator. The essence of heuristic attribution models, which are used in many analytics systems, is considered in connection with their ease of use. For comparison, algorithmic attribution models are also considered, such as: “Attribution based on Markov chains” and “Attribution based on data”. They analyze the impact of the presence of a particular channel on the conversion value. To compare these models and attribution algorithms, we used the training of XGBoost, Random forest, GBM and LightGBM gradient boosting algorithms on user datasets that contain information about the operation of a real e-commerce online store. For clarity, the data were taken for: 1, 2, 4, 8, 12 months. Subsequently, the data were normalized and brought to an acceptable format for the learning algorithms used. After that, the process of training the models and checking the classification accuracy was carried out using the calculation method – AUC. To identify a more productive algorithm for attribution, a comparative analysis of the running time of these models was carried out. The results obtained can be used to analyze the effectiveness and expand the possibilities of using these models.

Keywords: digital advertising, channel attribution, multi-channel attribution, machine learning, ROC AUC

С 1980-х годов начало приобретать популярность моделирование маркетингового комплекса (Marketing mix modeling) – метод анализа, который позволял маркетологам измерять влияние своих маркетинговых и рекламных кампаний, чтобы определить, как различные элементы способствуют достижению их цели, по повышению конверсии. По мере развития интернет-коммерции росло количество доступных для использования рекламных каналов, помогающих компаниям эффективнее продавать свои услуги. Однако для работы с такими каналами требуется разработка новых стратегий оценки их эффективности.

Развитие технологий позволяет отслеживать и анализировать действия пользователей, привлеченных с помощью различных онлайн-каналов. Полученные таким образом данные необходимо максимально использовать при разработке эффективной стратегии интернет-продвижения. Но в настоящее время существует недостаток знаний о поведении клиентов на рынке.

При обработке больших объемов информации, которые могут находиться в неструктурированном виде, исследователи используют модели на основе различных алгоритмов машинного обучения: интеллектуальный анализ текста; методы обра-

ботки естественного языка (NLP) и другие. Моделирование атрибуции – один из способов использования данных. Многоканальная атрибуция предназначена для оценивания вклада различных рекламных каналов в итоговую конверсию интернет-продукта и отвечает на вопрос: какой из рекламных каналов оказал влияние на клиента.

Вопросам атрибуции посвящено значительное количество работ. Так, например, в исследовании [1] проведен обзор существующих методов атрибуции, представлена их таксономия на основе современных научных публикаций. В работе [2] с опорой на структурированный процесс исследования литературы были оценены существующие подходы с точки зрения их применимости в многоканальной среде. По результатам исследования научных публикаций в области применения машинного обучения для атрибуции ценности по разным источникам трафика было выявлено отсутствие исследований сравнения моделей и алгоритмов атрибуции.

Целью данного исследования является сравнительный анализ различных методов и алгоритмов атрибуции, а также оценка точности рассмотренных алгоритмов по показателю ROC AUC и выявление наиболее эффективных и практически применимых моделей на основе пользовательского набора данных.

Для достижения поставленных целей были решены следующие задачи:

− исследование существующих методов и моделей атрибуции каналов;

− анализ цепочек взаимодействия для рекламных каналов при помощи эвристических и базовых многоканальных алгоритмических моделей атрибуции;

− создание и обучение моделей машинного обучения, проверка точности классификации и сравнительный анализ времени работы всех рассмотренных алгоритмов;

− визуализация полученных результатов расчета физического времени работы эвристических, алгоритмических моделей атрибуции и моделей машинного обучения в виде графиков.

Научная новизна заключается в предложенной методике использования машинного обучения для выбора наиболее эффективных моделей и алгоритмов атрибуции, которые позволяют с использованием показателя ROC (Receiver operating characteristic) AUC (Area Under Curve) получить оптимальный метод работы фирмы для привлечения клиентов. Предложенный в статье анализ моделей атрибуции помогает понять, какой из каналов продвижения товара на рынке является максимально эффективным. Список каналов продвижения товара на рынке будет приведён ниже.

Материалы и методы исследования

В работе использован пользовательский набор данных [3], который содержит информацию о работе реального интернет-магазина. Информация была обработана таким образом, чтобы можно было использовать для её анализа программные модули из библиотеки «MTA» [4], которые позволяют сравнить цепочки взаимодействия клиента с каналами продвижения, а также проверить эффективности всех рассмотренных алгоритмов.

Многие системы аналитики, используемые современными маркетологами, предлагают эвристические модели для атрибуции ценности канала.

Начнем с моделей атрибуции, которые основаны на распределении ценности по одному каналу:

− по первому взаимодействию (First Interaction);

− по последнему взаимодействию (Last Interaction);

− по последнему непрямому клику (Last Non-Direct Click).

Несмотря на то что упомянутые выше модели тривиальны в использовании и расчете, а также часто игнорируют всю дополнительную информацию, они тем ни менее часто используются. Необходимо учитывать, что каждый канал в более длинных воронках продаж может как положительно, так и отрицательно влиять на решение клиентов продолжить движение по воронке, а само воздействие промежуточных точек различного характера может иметь нарастающий эффект.

При необходимости проанализировать несколько каналов следует объединить несколько источников из различных рекламных сервисов, а также применять наиболее комплексные модели атрибуции. Таким образом, можно получить информацию, какие пары или связки рекламных каналов эффективно взаимодействуют в совокупности и на каких этапах.

Рассмотрим модели атрибуции для нескольких каналов одновременно. К ним относятся:

− линейная модель атрибуции (Linear model);

− с учетом давности взаимодействия (Time Decay);

− на основе позиции (Position Based или U-образная).

В многоканальных моделях учитывается большее количество взаимодействий с клиентом. Однако все они работают по за-

данным траекториям и не учитывают отклонений в сторону. Ещё одна проблема таких моделей связана с тем, что точки взаимодействия, инициированные клиентами, и точки взаимодействия, инициированные фирмой, не всегда совпадают. Поэтому оценку атрибуции следует выполнять с осторожностью.

Атрибуция на основе цепей Маркова – это модель, использующая вероятности перемещений по шагам воронки, которая дает оценку влиянию шагов на конверсию и позволяет определить наиболее значимый вклад в общую конверсию. Идея рассматриваемой модели состоит в определении набора состояний клиента и оценке вероятности перехода между различными состояниями [5].

Кроме того, существуют наиболее точные алгоритмы атрибуции, основанные на применении данных, собранных заранее. Главное преимущество моделей атрибуции на основе таких данных состоит в том, что в них не учитывается порядок канала в цепочке, а анализируется влияние на конверсию конкретного канала.

В библиотеке MTA имеются модули, позволяющие провести анализ рекламных каналов с помощью таких моделей, как модели Shao&Li, которые предлагают несколько статистических моделей атрибуции ценности на основе заранее собранных данных: bagged logistic regression model и simple probabilistic model [6]. Кроме модели Shao&Li, авторами также был выбран для работы подход с использованием вектора Шепли [7].

При изменении порядка сессий ценность каналов по вектору Шепли не меняется. Вектор Шепли – это метод, изначально изобретенный для назначения выигрыша игрокам в зависимости от их вклада в общий выигрыш [8].

Ценность канала по вектору Шепли рассчитывается по формуле 1:

$$\Phi(v)_i = \sum_{K \ni i} \frac{(k-1)!(n-k)!}{n!} \left(v(K) - v(K \setminus i)\right), \quad (1)$$

где n – количество игроков (в нашем случае это рекламные каналы); v – ценность, которую принес источник; k – количество участников коалиции K.

Кроме того, было уделено внимание системам машинного обучения. С помощью машинного обучения система обрабатывает некоторое количество заданных примеров, идентифицирует похожие цепочки взаимодействия клиентов с каналами и использует их для прогнозирования новых данных [9].

В то же время становится всё более очевидным, что «ретроспективный подход» к атрибуции перестал обладать достаточной эффективностью, а саму концепцию атрибуции следует рассчитывать, используя для расчётов вероятностную характеристику [10].

Гибридный подход к атрибуции будет помогать эффективно предсказывать вероятность покупки.

Результаты исследования и их обсуждение

Прежде всего, для выполнения процесса распределения ценностей по каналам загрузим заранее обработанные данные и выполним их анализ с помощью модулей из библиотеки «MTA» [4].

Для моделирования эвристических и алгоритмических моделей атрибуции использовались следующие методы: mta.first_touch(); mta.last_touch(); mta.time_decay(); mta.markov(); mta.shapley(); mta.shao().

Были проанализированы следующие каналы: Affiliates (Партнеры); Direct (Прямой трафик); Display (Медийная реклама); Organic Search (Органический поиск); Paid Search (Платная реклама); Referral (Реферальный трафик); Social (Социальные сети); Other (Прочее).

На рисунке 1 показаны результаты выполнения анализа цепочек взаимодействия с помощью модулей из библиотеки «MTA».

Как видно из рисунка 1, большинство моделей для расчёта атрибуции присвоили максимальное значение каналу Direct (Прямой трафик).

Перейдём непосредственно к обучению методов машинного обучения. Попробуем создать и обучить модель градиентного бустинга, а затем произвести оценку сложности работы данного алгоритма, подсчитаем время его работы.

Процесс обучения модели содержит в себе следующие этапы: подготовка данных, создание наборов данных для обучения, создание классификатора, обучение классификатора, составление прогнозов и оценка сложности, времени работы классификатора.

В процесс предварительной обработки данных по маркетинговым каналам для классификатора входит следующее: преобразование данных в форму, подходящую для классификации; обработка аномалий в этих данных, например аномалий отсутствия некоторых значений в подготавливаемых данных и т.п. Если не убрать аналогичные аномалии, они в конечном итоге могут негативно повлиять на производительность алгоритма классификации и снизить качество полученных результатов.

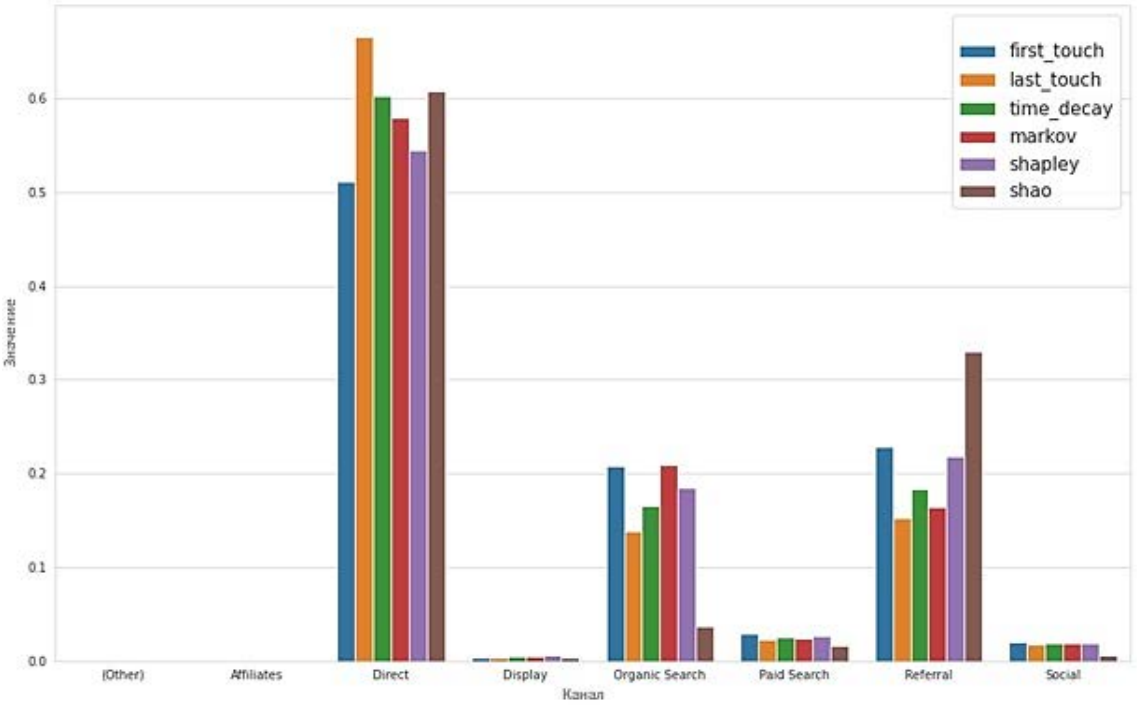

Рис. 1. Результаты выполнения анализа цепочек взаимодействия

```
@show_time # Оценка временных затрат
def mod_gbm(): # Настройка модели
  model_gbm = GradientBoostingClassifier(n_estimators=5000, # Количество этапов бустинга для выполнения
                                         learning_rate=0.05, # Скорость обучения
                                         max_depth=3, # Максимальная глубина
                                         subsample=0.5, # Доля выборок
                                         validation_fraction=0.1, # Доля обучаемых данных
                                         n_iter_no_change=20, # Необходимость ранней оставновки
                                         max_features='log2', # Количество признаков
                                         verbose=1) # Подробный вывод
  model_gbm.fit(X_train, y_train) # Обучение настроенной модели
  return model_gbm

model_gbm = mod_gbm() # Вывод результата
```

Рис. 2. Настройка модели машинного обучения

Сравнивая показания классификатора с фактически известными данными, можно сделать вывод о точности классификатора.

После процесса обучения можно проверить точность классификации, используя площадь ROC-кривой (AUC). Таким образом, процесс обучения представляет собой оценку способности классификатора различать подходящие и не подходящие какому-либо классу объекты. Для обучения модели градиентного бустинга [11] и оценки ее производительности был написан модуль на языке Python. Текст модуля показан на рисунке 2.

Таким образом, были получены следующие результаты оценки производительности. Затраченное время на обучение – 1,040 сек., рассчитана площадь под ROC-кривой и построен график (рис. 3).

Полученный результат AUC свидетельствует о приемлемом классификаторе.

Площадь под ROC-кривой – AUC (Area Under Curve) – является характеристикой качества классификации, чем больше охватываемая область, тем больше значение AUC и тем лучше модель классификации. Так принято, что AUC = 1 – лучший случай, при AUC = 0.5 – алгоритм случайного гадания, который соответствует худшему случаю. В нашем случае AUC = 0.78, что говорит о том, что данную модель классификации можно использовать для оценки вероятности конверсии.

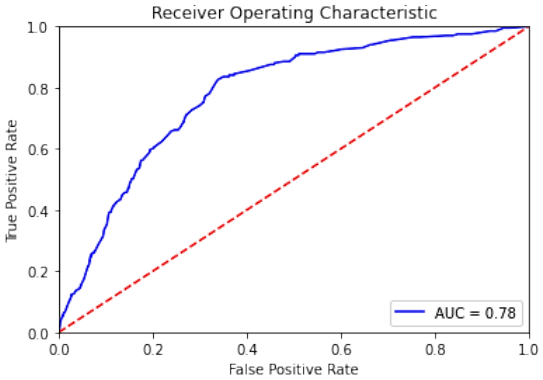

Рис. 3. Визуализированная кривая ROC

Для комплексной оценки производительности указанных алгоритмов и сравнения времени работы указанных алгоритмов с данными по точкам касания клиентов и маркетинговых каналов, было проведено обучение алгоритмов Random forest, градиентного бустинга GBM и нескольких его реализаций: LightGBM, XGBoost на наборах данных, содержащих сведения за 1, 2, 4, 8, 12 месяцев. Результаты расчета AUC после обучения каждого классификатора показаны в таблице.

Результаты расчета AUC
после обучения каждого классификатора

№	Алгоритм	Train_AUC	Valid_AUC
0	Random Forest	0,75	0,75
1	GBM	0,75	0,75
2	LightGBM	0,74	0,75
3	XGBoost	0,73	0,74
4	All	0,76	0,75

По результатам расчёта AUC после обучения каждого классификатора можно сделать вывод, что все модели показали примерно одинаковый результат, без большого разброса с приемлемым качеством классификации.

Перейдем к сравнению физического времени работы эвристических, алгоритмических моделей атрибуции и моделей машинного обучения.

На рисунке 4 показана визуализация полученных результатов расчета физического времени работы эвристических, алгоритмических моделей атрибуции и моделей машинного обучения в виде графиков, на которых показана зависимость физического времени работы алгоритма (в секундах) по оси Y от количества данных (за 1, 2, 4, 8, 12 месяцев) в обрабатываемом наборе по оси X. Контрольное количество данных по оси X (за 1, 2, 4, 8, 12 месяцев) для сравнения на графиках обозначено метками.

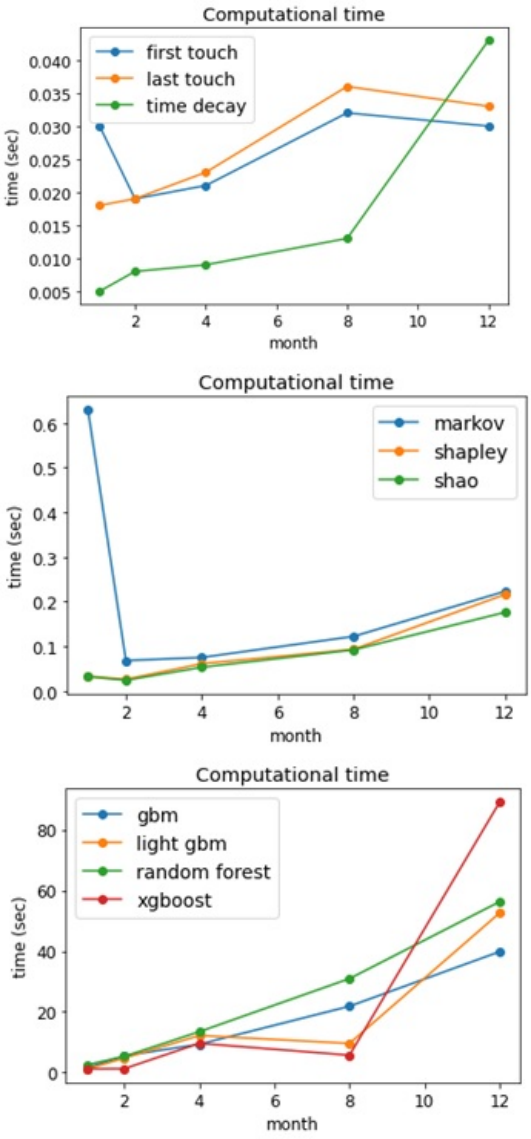

Рис. 4. Сравнение времени работы эвристических, алгоритмических методов и моделей машинного обучения

Таким образом, сравнительный анализ времени работы эвристических моделей показал, что разброс во времени работы моделей – незначителен и зависит от количества данных, а сравнение времени работы алгоритмических моделей показало, что разброс во времени работы разных алгоритмов – не существенный и также зависит от количества данных. Было выявлено, что времени на выполнение было затрачено больше, чем в эвристических моделях. Сравнение времени работы моде-

лей машинного обучения также выявило, что присутствует зависимость скорости обработки моделей от количества данных. Наиболее быстрым методом является – метод gbm с точностью 0,74. Самый медленный алгоритм – метод xgboost с точностью 0,73. Контрольное количество данных было взято за 12 месяцев.

В данной работе была проведена комплексная оценка точности всех указанных алгоритмов по показателю ROC AUC, а также были выявлены наиболее эффективные и практически применимые модели. Было проведено исследование существующих методов распределения ценности рекламных каналов. Проанализировано взаимодействие рекламных каналов. Создано программное обеспечение для оценки точности классификации по показателю ROC AUC. Проведен сравнительный анализ времени работы рассмотренных алгоритмов. Была построена визуализация полученных результатов расчета физического времени работы эвристических, алгоритмических моделей атрибуции и моделей машинного обучения.

Несмотря на незначительное количество времени работы как эвристических, так и алгоритмических моделей, точность данных методов крайне мала. Результаты точности моделей машинного обучения можно проверить различными способами, в том числе методом вычисления площади под кривой ошибок, что дает основное преимущество при выборе модели многоканальной атрибуции.

Список литературы

1. Buhalis Dimitrios, Volchek Katerina. Bridging marketing theory and big data analytics: The taxonomy of marketing attribution. International Journal of Information Management. 2021. Vol. 56. P. 1-14.

2. Nass Ole, A.G. José, Gil Gomez, Klaus-Peter Hermenegildo. Attribution modelling in an omni-channel environment – new requirements and specifications from a practical perspective. International Journal of Electronic Marketing and Retailing. 2020. Vol. 11. P. 81-111.

3. Google BigQuery. Google Analytics Sample. [Электронный ресурс]. URL: https://www.kaggle.com/bigquery/google-analytics-sample (дата обращения: 11.07.2022).

4. Igor Korostil. Multi-Touch Attribution. Find out which channels contribute most to user conversion. [Электронный ресурс]. URL: https://github.com/eeghor/mta (дата обращения: 11.07.2022).

5. Katsov I. Introduction to Algorithmic Marketing: Artificial Intelligence for Marketing Operations, 2017. 492 p.

6. X. Shao, L. Li. Data-driven multi-touch attribution models. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011. Vol. 17. P. 258-264.

7. Singal Raghav, Besbes Omar, Desir Antoine. Shapley Meets Uniform: An Axiomatic Framework for Attribution in Online Advertising. WWW '19: The World Wide Web Conference. San Francisco CA USA: Association for Computing Machinery, New York, NY, United States, 2019. Vol. 28. P. 1713–1723.

8. Cano Sebastian Berlanga, Cori Vilella. Attribution models and the Cooperative Game Theory. Col·lecció "Documents de treball del departament d'economia creip". 2017. 20 p.

9. Zanker Markus, Rook Laurens, Dietmar Jannach. Measuring the impact of online personalisation: Past, present and future. International Journal of Human-Computer Studies. 2019. Vol. 131. P. 160-168.

10. Constantine Yurevich. Behavior-Based Attribution Using Google BigQuery ML [Электронный ресурс]. URL: https://cxl.com/blog/attribution-google-bigquery-ml/ (дата обращения: 24.05.2022).

11. Pedregosa F., Varoquaux G., Gramfort A., Michel V. Sklearn.ensemble.GradientBoostingClassifier. [Электронный ресурс]. URL: https://clck.ru/gvtFF (дата обращения: 11.07.2022).