



Международный журнал информационных технологий и энергоэффективности

Сайт журнала:

<http://www.openaccessscience.ru/index.php/ijcse/>



УДК 004.891.2

ОПРЕДЕЛЕНИЕ ОПТИМАЛЬНОГО СПОСОБА ВЗАИМОДЕЙСТВИЯ С КЛИЕНТОМ ДЛЯ УВЕЛИЧЕНИЯ ДОХОДНОСТИ МАРКЕТИНГОВЫХ КАМПАНИЙ И СНИЖЕНИЯ ИЗДЕРЖЕК НА НИХ

Обливальный Н.Д.

ФГБОУ ВО "Челябинский Государственный Университет", Челябинск, Россия (454001, г. Челябинск, ул. Братьев Кашириных, д.129), e-mail: rfrepe@gmail.com

В данной работе рассматривается поиск и определение оптимального способа взаимодействия с клиентом для увеличения доходности маркетинговых кампаний и снижения издержек на них. В работе были проделаны следующие задачи: определение универсального подхода к Uplift моделированию для выбора типа коммуникации; формирование требований к структуре данных для достижения максимальной точности; реализация сервиса по работе с данными и получения расчетов для принятия решения о способе коммуникации. Решения данных задач рассматривалось в рамках прикладной области телеком провайдера. Практическим итогом данной работы является сервис, с помощью которого будет определена необходимости и способ коммуникации с абонентами телеком провайдера, имеющими задолженность.

Ключевые слова: Uplift моделирование, машинное обучение, большие данные, коммуникация с клиентами, доходность маркетинговых кампаний, телеком провайдер.

DETERMINING THE OPTIMAL WAY TO INTERACT WITH THE CLIENT TO INCREASE THE PROFITABILITY OF MARKETING CAMPAIGNS AND REDUCE THEIR COSTS

Oblivalny N.D.

FSBEI of HE "Chelyabinsk State University", Chelyabinsk, Russia (454001, Chelyabinsk, Bratya Kashirin str., 129), e-mail: rfrepe@gmail.com

In this paper, we consider the search and determination of the optimal way to interact with the client to increase the profitability of marketing campaigns and reduce their costs. The following tasks were performed in the work: defining a universal approach to Uplift modeling for choosing the type of communication; forming requirements for the data structure to achieve maximum accuracy; implementing a service for working with data and obtaining calculations to make a decision on the method of communication. Solutions to these problems were considered within the scope of the telecom provider's application area. The practical result of this work is a service with which the necessity and method of communication with subscribers of the telecom provider who have debts will be determined.

Keywords: Uplift modeling, machine learning, big data, communication with customers, profitability of marketing campaigns, telecom provider.

Введение

Целью данной работы является определение оптимального способа взаимодействия с клиентом для увеличения доходности маркетинговых кампаний и снижения издержек на них.

Задачами магистерской работы являются: определение универсального подхода к Uplift моделированию для выбора типа коммуникации, формирование требований к структуре данных для достижения максимальной точности, реализация сервиса по работе с данными и получения расчетов для принятия решения о способе коммуникации. Решения данных задач будут рассмотрены на прикладной области телеком провайдера. Практическим итогом данной работы будет являться сервис, с помощью которого будет определена необходимости и способ коммуникации с абонентами телеком провайдера, имеющими задолженность.

Большинство современного бизнеса не может существовать без использования цифровых технологий, особенно в сегменте B2C. Цифровые решения затронули все сферы бизнеса. Начиная от автоматизации внутренних бизнес-процессов, заканчивая различными видами коммуникации с клиентами.

Для любого бизнеса одной из основных целей всегда является увеличение прибыли. И первое, с чего начинается путь привлечения финансов в компанию, это взаимодействие с клиентом. Правильно взаимодействуя с клиентами, бизнес способен мотивировать их на действия. Например, на оформление подписки, покупки товара, рекомендации продукта своим друзьям или на повторную покупку.

При сегодняшнем многообразии каналов коммуникации: смс-сообщение, звонок, пуш-уведомление, электронное письмо, баннер, контекстная реклама и т.д. - выбор способа взаимодействия стал актуальной проблемой. При планировании бюджета нужно грамотно распределять его между всеми способами взаимодействия, при этом необходимо достичь максимального увеличения доходности.

Практика показала, что у клиентов есть собственные привычки взаимодействия с бизнесом. Кто-то отслеживает смс-сообщения, но при этом никогда не читает электронную почту. И сегодня большинство компаний стараются не экономить на коммуникации с клиентом и пользуются всеми доступными средствами. Из-за чего делают только хуже. Информационный шум, который они создают, понижает эффективность каналов коммуникации из-за снижения внимания у клиентов.

Так, например, по результатам исследования компании «Мета» (запрещённой в Российской Федерации) можно сделать вывод о том, что пользователи, получающие меньшее количество уведомлений в приложении, больше используют приложение (Рисунок 1).

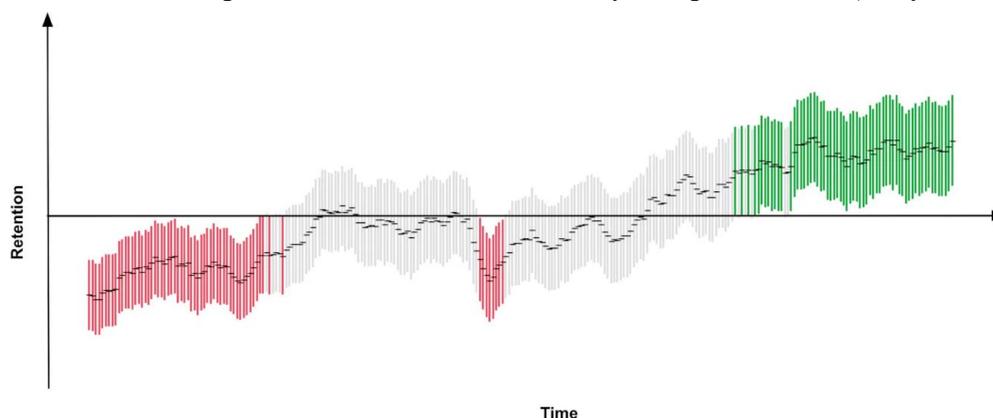


Рисунок 1 – График возвращения пользователей, получивших меньшее количество уведомлений.

Источник: исследование компании «Мета» (запрещённой в Российской Федерации)

На графике показан уровень вовлеченности пользователя в приложение. По данным исследования уменьшение числа уведомлений в начале привело к кратковременному снижению активности пользователя, но в конечном итоге спустя время увеличило его вовлеченность в приложение.

Таким образом, все вышеперечисленное доказывает актуальность данной работы.

Методы

Для принятия решений о выборе взаимодействия с клиентом рассмотрим несколько подходов к Uplift моделированию [1].

Первый подход основан на использовании одной модели классификации. Модель обучалась на данных клиентов, с которыми взаимодействовали и не взаимодействовали. Для определения эффекта от взаимодействия тип взаимодействия выводят в отдельный признак. После обучения модели при получении прогноза необходимо сделать прогноз для каждого типа взаимодействия. После чего определить тип, при котором вероятность совершения целевого действия максимальная.

Второй подход заключается в использовании двух независимых моделей. Для каждого типа взаимодействия формируется отдельный датасет. Далее на сформированных датасетах обучаются модели классификации. В результате мы получаем набор моделей, которые оценивают вероятность совершения ключевого действия для каждого типа взаимодействия. Получив прогноз от каждой модели, мы можем определить наилучший тип взаимодействия.

У второго подхода есть различные вариации реализации зависимостей между полученными моделями. Например, с зависимым представлением данных. Идея состоит в том, что при наличии различных меток можно построить столько же различных классификаторов, каждый из которых решает задачу бинарной классификации. В процессе обучения каждый следующий классификатор использует предсказания предыдущих в качестве дополнительных признаков. Авторы данного метода предложили использовать ту же идею для решения проблемы uplift моделирования в два этапа. Вначале мы обучаем классификатор по контрольным данным. Затем выполним предсказания в качестве нового признака для обучения второго классификатора на тестовых данных, тем самым вводя зависимость между двумя наборами данных.

Еще один подход основан на построении перекрестно зависимых моделей [2]. Его рекомендуется применять тогда, когда целевая группа достаточно маленькая. В этом случае есть риск, что модель, построенная на целевой группе, будет обладать недостаточной обобщающей способностью. Поэтому создается перекрестная зависимость двух моделей, чтобы усилить одну модель данными другой. Сначала обучаем параллельно две модели: одну на контрольной группе, другую — на целевой. Затем преобразуем обе целевые переменные, используя предсказания контрольной модели на данных целевой группы и предсказания целевой модели на данных контрольной группы. Полученные величины называются *вменяемым эффектом от воздействия*. Обучим две новые модели на преобразованных таргетах. Взвешенная с некоторым коэффициентом сумма предсказаний этих моделей и будет uplift.

Данные подходы являются базовыми в Uplift моделировании. Также существуют методы, основанные на деревьях, трансформации классов и много классовых моделях.

Метод подготовки данных

Предобработка данных и подготовка датасета - один из важнейших этапов построения моделей машинного обучения. От них напрямую зависит качество прогнозов моделей. Поэтому данному этапу нужно всегда уделять большое внимание.

Для нашей задачи всегда нужно определиться с ключевой метрикой. Как правило, ей является совершение определённого действия, категориальный признак. Например, покупка или оформление заказа. В таком случае лучше всего использовать модели классификации.

Но в ряде случаев ключевой метрикой может стать количественный признак. Например, сумма заказа или средний чек. В таком случае лучше использовать регрессионные модели в качестве основы для uplift моделирования.

Не зависимо от выбора базовых моделей необходимо провести стандартную предобработку данных. Традиционные EDA анализ поможет нормализовать данные, избавиться от выбросов и ошибок в датасетах.

В качестве данных для Uplift моделей может использоваться различная информация о клиенте. От текущих статусов клиента до агрегированной исторической информации[3].

Основными требованиями к датасету будут являться:

- Сбалансированность выборок с различными типами взаимодействия. Для каждого типа взаимодействий должно быть достаточно примеров для получения достоверных прогнозов.
- Идентификация конкретного клиента. Все данные должны быть привязаны к конкретному клиенту, по которому будут армироваться данные и в итоге формироваться прогноз.

Главной задачей на данном этапе становится построение универсального способа подготовки данных для формирования датасета. Так как данные можно разделить на 2 типа. Описывающие объект (человека) и описывающие событие (взаимодействие), то необходимо собрать 2 датасета.

В первом будет находиться вся информация по клиенту. Те признаки, которые его описывают в конкретный момент времени.

Структура первого датасета:

- Дата события
- Идентификатор клиента
- Признак
- Значение

Второй содержит информацию о взаимодействиях или отсутствии взаимодействия с клиентом и информацию о результате взаимодействия.

Структура второго датасета:

- Дата взаимодействие
- Идентификатор клиента
- Тип взаимодействия
- Результат

Для того чтобы получить итоговую обучающую выборку, нам необходимо агрегировать всю информацию о клиенте на даты взаимодействия. Для этого нужно преобразовать первый датасет.

Сперва необходимо определить типы данных у всех признаков. Затем для категориальных признаков найти значения: часто встречающиеся, медианное и среднее значение промежутков времени между событиями. Для количественных признаков: среднее значение, медианное значение, сумму, дисперсию и медианное и среднее значение промежутков времени между событиями. Все признаки считаются на момент взаимодействия для каждого клиента.

Затем нужно объединить агрегированные данные из первого датасета с данными из второго. Добавить в первый датасет признаки “Тип взаимодействия” и “Результат”.

Таким образом, мы поделили датасет с информацией о клиенте на конкретный момент времени с указанием взаимодействия и итогом его взаимодействия. Дальнейшая его преобработка будет зависеть от выбранного подхода к построению Uplift модели.

Данные телеком провайдера

Для постановки эксперимента была выбрана область телеком провайдера. Необходимо собрать датасет для абонентов компании, попавших в статус должников.

Для первого датасета была собрана информация о всех услугах и сервисах, которыми пользуется или пользовался должник. В признаки попала следующая информация: трафик, смена тарифов, установление автоплатежей, заявки на подключение, сессии в мобильном приложении, платежи, блокировки услуг, взаимодействие с домофоном, обращения в техническую поддержку, информация о жилом фонде клиента и еще около 100 признаков.

Для каждого признака были подсчитаны агрегированные значения на момент запланированного взаимодействия. Итоговая размерность датасета составила 861 признаков.

Оценка результатов

Для оценки результатов работы uplift модели существуют специальные метрики.

Например, $uplift@k$. С помощью обученной uplift модели мы хотим отобрать какое-то количество клиентов, с которыми будем коммуницировать. Пусть бюджет рассчитан на $k\%$ клиентов. Тогда нам интересно оценить качество прогноза не на всей тестовой выборке, а только на объектах с наибольшими предсказаниями при отсечении по порогу в k процентов [4].

Для расчета $uplift@k$ нужно отсортировать выборку по величине предсказанного uplift и посмотреть разницу средних значений таргета Y (в англоязычных статьях использую термин response rate, мы его тоже будем использовать в дальнейшем) в целевой и контрольной группах. Целевая группа - группа, которая получила коммуникацию. Контрольная группа - которая не получила (1).

$$\begin{aligned} uplift@k &= response\ rate@k_{(treatment)} - response\ rate@k_{(control)} \\ response\ rate@k &= mean(Y@k) \\ Y@k &= \text{таргет на } k\% \end{aligned} \quad (1)$$

Есть еще одна метрика, называющиеся кривая qini. Физический смысл qini кривой в том, чтобы не давать модели поднимать вверх в ранжировании только целевую (treatment) группу,

штрафуя ее за это множителем NT/NC , который уменьшает итоговое значение, если NT сильно больше, чем NC (2).

$$gini\ curve(t) = Y_t^T - \frac{Y_t^C N_t^T}{N_t^C} \quad (2)$$

Y_t^T, Y_t^C – таргет в *treatment* группе, таргет в *control* группе

N_t^T, N_t^C – размер *treatment* группы, размер *control* группы

Для перехода от кривых к числам используется коэффициент AUQC.

Для оценки результатов экспериментов будут использоваться две представленные метрики. Также для оценки качества классификаторов будет использовать метрика f1-score.

Результаты экспериментов

В рамках эксперимента использовался тестовый набор данных.

Сперва было необходимо определиться с подходом в методе, использующим две модели. Для этого были обучены модели с использованием одинаковых классификаторов. Результаты качества получившихся моделей представлены в Таблице 1.

Таблица 1 – Сравнение подходов с использованием двух моделей.

Подходы	f1 (целевой группы)	uplift@20%	AUQC
Две независимые модели	0.5007	0.1744	0.1495
Две зависимые модели (зависимое представление данных)	0.5303	0.0597	0.0903
Две зависимые модели (перекрестная зависимость)	0.0514	0.0368	0.0846

Источник: анализ автора

В итоге максимальную точность имеет подход с использованием двух независимых моделей.

Следующим этапом стало определение подходящего классификатора. Также было произведено сравнение самых популярных моделей классификации. Результаты сравнения классификаторов представлены в таблице 2 [5-7].

Таблица 2 – Оценка модели при применении подхода с не зависимыми моделями

Модели	Контрольная f1-score		Целевая f1-score	
	Не оплатил	Оплатил	Не оплатил	Оплатил
XGBoost	0.811	0.742	0.806	0.555
CatBoost	0.814	0.72	0.807	0.522
LogReg	0.79	0.7	0.801	0.48

Источник: анализ автора

Лучший результат показала модель XGBoost. При подборе гиперпараметров удалось достичь следующих показателей. Результаты представлены в Таблица 3.

Таблица 3 – Лучшая модель подхода с использованием двух независимых моделей.

Best model XGboost	f1(целевой группы)	uplift@20%	AUQC
Две независимые модели	0.553	0.6594	0.3574

Источник: анализ автора

Данную модель взяли для проведения А/В теста в реальных условиях. Для подтверждения качества модели на практике необходимо было получить статистическую значимость у двух групп: «Убеждаемые» и «Не беспокоить». Эти две группы имеют противоположный по своему характеру смысл. Хорошая модель должна уметь максимально разводить их и иметь статистическую значимость.

Эксперимент с использованием А/В теста был поставлен на должниках телеком провайдера. Его результаты представлены в Таблице 4.

Таблица 4 – Результаты А/В теста подхода с использованием двух независимых моделей.

Группы	Конверсия в оплату		Доверительный интервал		Значимость
	А (звонок)	В (без звонка)	1	2	
Убеждаемые	30,65%	17,81%	-21,97%	-3,72%	Значимое
Потерянные	14,02%	10,57%	-8,08%	1,18%	Незначимое
Лояльные	35,66%	36,82%	-1,31%	3,63%	Незначимое
Не беспокоить	20,45%	25,48%	0,4%	10,03%	Значимое

Источник: анализ автора

По итогам А/В теста можно сделать вывод о том, что получившаяся модель действительно имеет практический эффект. Она определит группу людей, с которыми коммуникация будет наиболее эффективна, и группу людей, с которыми нет практической ценности взаимодействовать.

Заключение

За счет использования uplift модели, основанной на подходе с двумя независимыми моделями, удалось добиться:

1. Сохранения текущей конверсии должников в оплату. По данным эксперимента, процент вырос на группе «Убеждаемых», но в контексте всех должников данный прирост не является значительным. Поэтому точнее утверждать, что процент конверсии не снизился.
2. Снизил издержки на исходящих звонках должникам. За счет прекращения звонков оставшимся группам удалось высвободить большое количество ресурсов операторов, которые

были перераспределены на другие задачи. В рамках процесса обзвона должников процент экономии ресурсов составил более 80%.

Данный подход является универсальным за счет унификации работы с данными и имеет практическую ценность.

Список литературы

1. Artem Betlei, Eustache Diemert, Massih-Reza Amini. Uplift Modeling with Generalization Guarantees, 2021.
2. Robin M. Gubela, Stefan Lessmann. Interpretable Multiple Treatment Revenue Uplift Modeling, 2021.
3. Henrik Karlsson, Linda Wanstrom. Uplift Modeling: Identifying Optimal Treatment Group Allocation and Whom to Contact to Maximize Return on Investment, 2019.
4. Ta-Wei Huang, Eva Ascarza. When Less is More: Using Short-term Signals to Overcome Systematic Bias in Long-run Targeting, 2022.
5. Xgboost documentation. - Режим доступа: <https://xgboost.readthedocs.io/en/stable/>, свободный.
6. Catboost documentation. - Режим доступа : <https://catboost.ai/en/docs/>, свободный.
7. LogisticRegression documentation. - Режим доступа:: <https://scikit-learn.org/stable/index.html>, свободный.

References

1. Artem Betlei, Eustache Diemert, Massih-Reza Amini. Uplift Modeling with Generalization Guarantees, 2021.
 2. Robin M. Gubela, Stefan Lessmann. Interpretable Multiple Treatment Revenue Uplift Modeling, 2021.
 3. Henrik Karlsson, Linda Wanstrom. Uplift Modeling: Identifying Optimal Treatment Group Allocation and Whom to Contact to Maximize Return on Investment, 2019.
 4. Ta-Wei Huang, Eva Ascarza. When Less is More: Using Short-term Signals to Overcome Systematic Bias in Long-run Targeting, 2022.
 5. Xgboost documentation. - Режим доступа: <https://xgboost.readthedocs.io/en/stable/>, свободный.
 6. Catboost documentation. - Режим доступа : <https://catboost.ai/en/docs/>, свободный.
 7. LogisticRegression documentation. - Режим доступа:: <https://scikit-learn.org/stable/index.html>, свободный.
-