МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Московский авиационный институт
(национальный исследовательский университет)» (МАИ)

УТВЕРЖДАЮ
Заведующий кафедрой
«Теория вероятностей и компьютерное моделирование»
д.ф.-м.н., профессор

_____ А.И. Кибзун
«04» января 2022 г.

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по теме:
Исследование задачи прогнозирования оттока клиентов

Научный руководитель

к.ф.-м.н., доцент _____ Платонов Е.Н.

Исполнитель

магистр группы М8О-101М-21 _____ Фейзуллин К.М.

Москва 2022

Реферат

Отчет 11 с., 2 рис., 1 табл., 4 источн.

Объектом исследования являются задача прогнозирования оттока клиентов

Цель работы – постановка задачи и исследование методов решения.

В результате работы определены методы решения задачи бинарной классификации оттока покупателей. Дальнейшее исследование может включать в себя исследование решений задачи UpLift моделирования и сравнительное исследование решений задач.

Оглавление

Введение

В данной научно-исследовательской работе проводится исследование возможных подходов к решению задачи прогнозирования оттока клиентов.

С ростом глобализации и цифровизации появилась возможность работать с потребительскими данными, активно взаимодействовать с потребителями путем разных акций, особых предложений. Чтобы клиент не забывал о поставщике потребительских услуг, производитель может напомнить о себе посредством коммуникации.

Но стоит взять во внимание, что каждая коммуникация стоит денег. Если клиентская база составляет 1 тыс. клиентов, то прислать всем SMS стоит не дорого. Но если увеличить масштаб базы до миллиона или нескольких миллионов, то слепая коммуникация со всеми подряд станет очень дорогой. Даже если у компании большой оборот выручки, каждая такая коммуникация будет ощутимо сказываться на общем бюджете.

Поэтому коммуникацию можно использовать гораздо более оптимальным способом. Например, совершать коммуникацию с потенциально ушедшим пользователем.

Однако с ростом клиентской базы даже выборочная коммуникация с потенциально потерянными клиентами будет затратной и следующей задачей является прогнозирование, повлияет ли коммуникация на пользователя.

Основная часть отчета о НИР

В данной работе производится первичный анализ области исследования, определяются возможные подходы к решению задачи и основные этапы работы, приводятся примеры разработок в сходных областях и их возможные модификации в терминах текущей задачи, проводится анализ предлагаемых методов решения.

Постановка задачи

Задача бинарной классификации

Как было описано выше, у продуктовых или ретейл компаний появилась потребность в прогнозировании оттока покупателей для применения мер предотвращения. Для оптимального распределения бюджета нельзя осуществлять коммуникацию со всеми пользователями сразу, так как это будет очень дорогая коммуникация. Тогда будем осуществлять коммуникацию с теми пользователями, от которых мы получим наибольший отклик, наибольшую пользу. В современном мире пользу нельзя рассматривать только как прибыль, теперь пользу для компании несет сам покупатель, уделяя ей внимание. Тогда главной целью коммуникации определим сохранение внимания и наибольшую пользу такая коммуникация принесет с потенциально ушедшим пользователем. Формализуя, задача будет классификации выглядеть следующим образом. Дана выборка пользователей с одинаковым набором признаков $X = \{x_i \,|i \in \{1, 2, \dots, n\}\}$, которую мы разделим на обучающую подвыборку $X' = \{x_i' \,|j \in \{1, 2, \dots, m\}\}$ и тестовую подвыборку $X'' = \{x_i'' \,|k \in \{1, 2, \dots, j\}\}$ так, что $X = X' \cup X''$ и $X' \cap X'' = \emptyset$. Так же мы делим множество правильных ответов $Y = \{y_i \,|i \in \{1, 2, \dots, n\}\}$ на Y' и Y'' так, что $Y = Y' \cup Y''$ и $Y' \cap Y'' = \emptyset$. Итак, есть выборка пользователей X и выборка правильных ответов $Y \in \{0, 1\}$. Пусть $\xi: \Omega \to$ X – случайная величина, представляющая собой случайного

покупателя X. И пусть η:$\Omega \rightarrow$ Y – случайная величина, представляющая собой случайный правильный ответ из Y. Тогда определим случайную величину $(\xi, \eta) : \Omega \rightarrow (X, Y)$ с распределением $p(y|x)$, которое является совместным распределением объектов и их классов. Тогда размеченная выборка – это элементы из распределения $(x_i , y_i) \sim p(y|x)$. Определим, что все элементы независимо и одинаково распределены. Тогда задача классификации будет сведена к задаче нахождения $p(y|x)$ и заданном наборе элементов $D = \{(x_i , y_i) \sim p(y|x), i = \overline{1,N}$. С помощью обучающей выборки X' и правильных ответов Y' будем находить распределение $p(y|x)$, а уже на тестовой выборке X'' и наборе правильных ответов Y'' для нее, будем смотреть, как хорошо тот или иной метод решения с помощью машинного обучения работает с контрольной выборкой.

Задача UpLift моделирования

При росте клиентской базы мало знать, какой клиент может вскоре от нас уйти. Для минимизации затрат нужно определить, на каких клиентов коммуникация сработает, а на каких нет.

Эффект от коммуникации определим как *casual effect*:

$$\tau_i = Y_i^1 - Y_i^0,$$

где Y_i^1- реакция i – го человека, если коммуникация была, Y_i^0- реакция, если коммуникации не было.

Зная признаковое описание i – го объекта X, можно ввести условный усредненный эффект от воздействия *Conditional Average Effect* (CATE):

$$CATE(x) = M[Y_i^1 | X_i] - M[Y_i^0 | X_i]$$

Casual effect и CATE можно только оценить, так как одновременно невозможно провести коммуникацию с человеком и не провести. Оценка CATE и является UpLift. Тогда для конкретного объекта он имеет следующее определение:

$$UpLift(x) = M[Y_i|X_i = x, W_i = 1] - M[Y_i|X_i = x, W_i = 0],$$

Где Y_i – наблюдаемая реакция клиента в результате маркетинговой кампании:

$$Y_i = W_I Y_i^1 + (1 - W_i)Y_i^0 = \begin{cases} Y_i^1, \text{если } W_I = 1 \\ Y_i^0, \text{если } W_i = 0 \end{cases}$$

$W_i = 1$, если объект попал в *целевую* (threatment) группу, в которой была коммуникация,

$W_i = 0$, если объект попал в *контрольную* (control) группу, в которой коммуникации не было,

$Y_i = 1$, если объект совершил целевое действие,

$Y_i = 0$, если объект не совершил целевое действие (произошел отток)

Анализ области исследования

Исходя из задач, при решении которых могут быть использованы результаты данной работы, могут принципиально отличаться как алгоритмы решения, так и подходы к нему в целом. Решения могут быть эвристическими, могут включать в себя построение более сложных алгоритмов, в том числе с использованием моделей машинного обучения. От выбора подхода к решению зависят существование базовых решений, набор используемых атрибутов запроса, определение методов извлечения эвристик и построения правил, способы оценки параметров алгоритма,

необходимость в наличии разметки данных, методы оценки качества и многие другие факторы.

Задача бинарной классификации оттока

Решение данной задачи возможно как аналитически, с помощью анализа исторических данных, так и с помощью машинного обучения.

Одно из аналитических решений предполагает анализ «выживаемости». Находится период с момента последней покупки до настоящего времени всей пользовательской базы. Для каждой сферы продаж распределение будет отличаться.

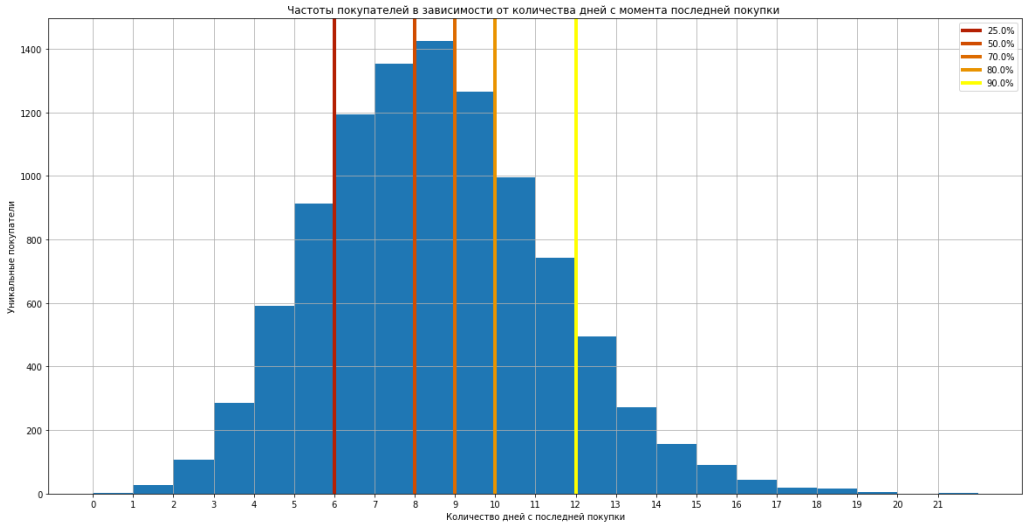

Рисунок 1. Моделирование потребительского поведения.

На рисунке 1 отображена гистограмма зависимости количества покупателей от количества прошедших дней с момента последней покупки. По рисунку 1 можно сказать, что если пользователь не закупался в течении 12 дней, то скорее всего, мы его потеряли, так как данное количество дней соответствует перцентилю в 90%.

Вариантом сложнее является RF[1] сегментация покупателей на основе частоты и давности покупки. Пример моделирования такой сегментации на рисунке 2.

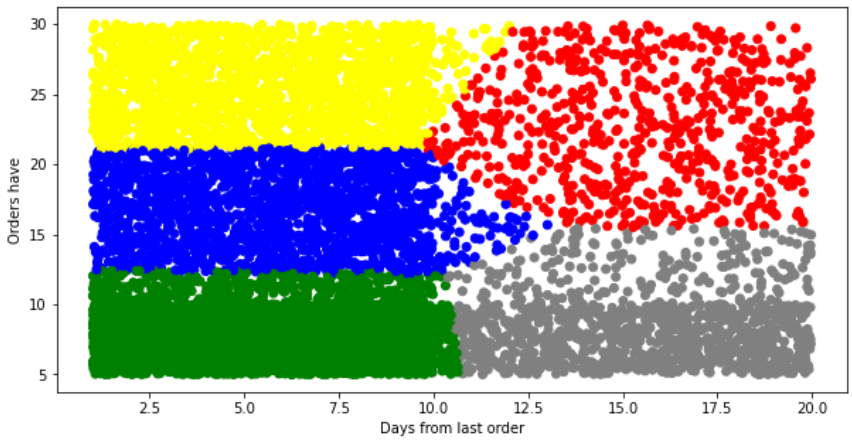

Рисунок 2. RF сегментация.

В нашем случае покупателей в ручную разобьем на пять сегментов на основе нашей экспертной оценки. На основе рисунка 1 было предположение, что если человек не совершил покупку в течении 12 дней, то скорее всего, он для нас потерян. Тогда можно сказать, что сегмент точек, отмеченных серым, можно считать оттоком покупателей, так как это множество давно не совершало покупки и в общей сложности совершило их малое количество.

Данные подходы можно использовать как с размеченными данными, так и с не размеченными.

Следующим вариантом решения задачи прогноза оттока клиентов является машинное обучение. Данного рода решений существует огромное количество, начиная классической логистической регрессией[2] и заканчивая нейронными сетями[2][3].

Эффективность стандартных методов решения задачи бинарной классификации[4] отразим в таблице 1.

Метод классификации	Верно классифицированных объектов	Ошибочно классифицированных объектов
Случайный лес	69%	31%
Градиентный бустинг	73.3%	26.7%
Наивный Байесовский классификатор	75%	25%
Дискриминантный анализ	75.7%	24.3%
Логистическая регрессия	74%	26%

Однако, стоит взять во внимание, что в зависимости от задачи, точность классификации может варьироваться для одних и тех же методов. Из чего сделаем вывод, что придется исследовать некоторые модели для нашей задачи самим.

Заключение

Таким образом, в ходе данной работы было произведено исследование задачи оттока в двух формах – бинарная классификация пользователей и UpLift моделирование. Был определен список алгоритмов решения задачи бинарной классификации задачи оттока клиентов и в будущем будут рассмотрены способы решения задачи UpLift моделирования.

Список использованных источников

1. RF – сегментация // https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/

2. Глубокое обучение / Ян Гудфеллоу, Иошуа Бенджио, Аарон Курвилль // ДМК Пресс, 2018г., второе цветное издание, исправленное

3. Глубокое обучение. / Николенко С., Кадурин А., Архангельская Е. // СПб: Питер, 2018. — 480 с.: ил. — (Серия «Библиотека программиста»).

4. Анализ методов бинарной классификации / Ю.С. Донцова // Известия Самарского научного центра Российской академии наук, том 16, No 6(2), 2014

_____ _____

Дата Подпись