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. |

# 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**

[РЕФЕРАТ 2](#_Toc91170839)

[Введение 4](#_Toc91170840)

[Основная часть отчета о НИР 5](#_Toc91170841)

[Постановка задачи 5](#_Toc91170842)

[Задача бинарной классификации 5](#_Toc91170843)

[Задача UpLift моделирования 6](#_Toc91170844)

[Анализ области исследования 7](#_Toc91170845)

[Задача бинарной классификации оттока 8](#_Toc91170846)

[ЗАКЛЮЧЕНИЕ 9](#_Toc91170847)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 10](#_Toc91170848)

[ПРИЛОЖЕНИЕ 11](#_Toc91170849)

# 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 которую мы разделим на обучающую подвыборку и тестовую подвыборку так, что 𝑋 = 𝑋′ ∪ 𝑋′′ и 𝑋′ ∩ 𝑋′′ = ∅. Так же мы делим множество правильных ответов на 𝑌′ и 𝑌′′ так, что 𝑌 = 𝑌′ ∪ 𝑌′′ и 𝑌′ ∩ 𝑌′′ = ∅. Итак, есть выборка пользователей 𝑋 и выборка правильных ответов . Пусть 𝜉: 𝛺 → X – случайная величина, представляющая собой случайного покупателя 𝑋. И пусть 𝜂:𝛺 → Y – случайная величина, представляющая собой случайный правильный ответ из 𝑌. Тогда определим случайную величину (𝜉, 𝜂) ∶ 𝛺 → (𝑋, 𝑌) c распределением 𝑝(𝑦|x), которое является совместным распределением объектов и их классов. Тогда размеченная выборка – это элементы из распределения . Определим, что все элементы независимо и одинаково распределены. Тогда задача классификации будет сведена к задаче нахождения 𝑝(𝑦|x) и заданном наборе элементов . 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*:

*,*

где - реакция i – го человека, если коммуникация была, - реакция, если коммуникации не было.

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

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:

,

Где – наблюдаемая реакция клиента в результате маркетинговой кампании:

, если объект попал в *целевую* (threatment) группу, в которой была коммуникация,

, если объект попал в *контрольную* (control) группу, в которой коммуникации не было,

, если объект совершил целевое действие,

, 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.

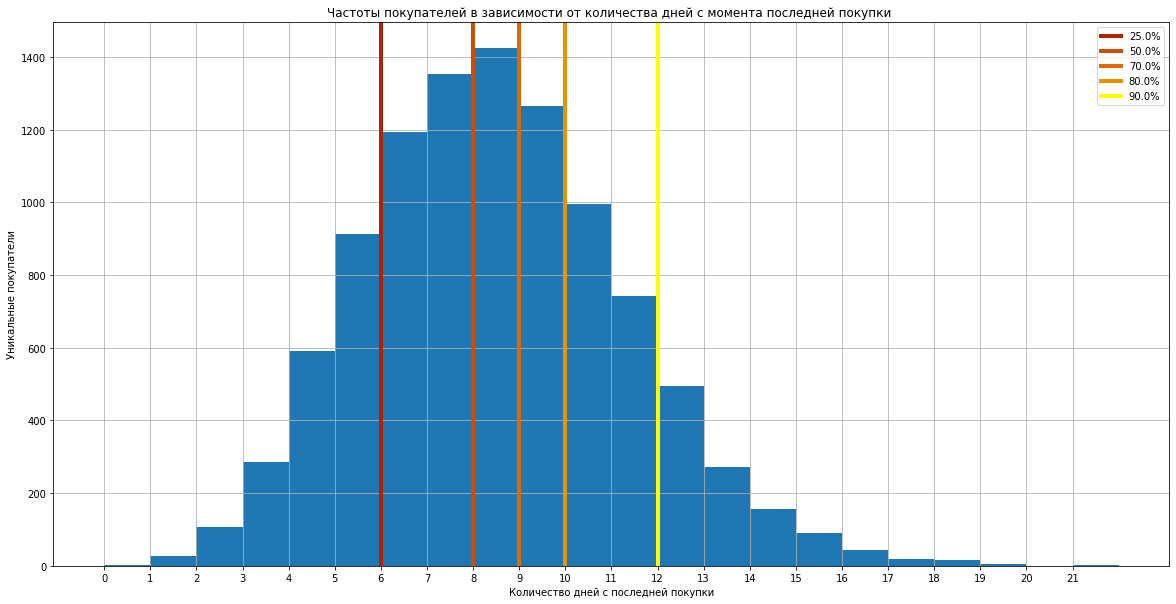


Рисунок 1. Моделирование потребительского поведения.

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.

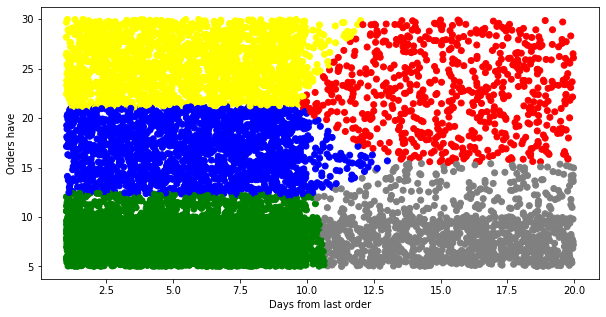


Рисунок 2. RF сегментация.

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