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 15 с., 7 рис., 1 табл., 5 источн., 1 прил.

Объектом исследования являются задача прогнозирования оттока клиентов

Цель работы – постановка задачи и исследование методов решения.

В результате работы определены методы решения задачи. Дальнейшее исследование может включать в себя сравнительное исследование решений задачи.

**Оглавление**

[РЕФЕРАТ 2](#_Toc91149225)

[Введение 3](#_Toc91149226)

[Основная часть отчета о НИР 4](#_Toc91149227)

[Постановка задачи 4](#_Toc91149228)

[Задача бинарной классификации 4](#_Toc91149229)

[Задача UpLift моделирования 6](#_Toc91149230)

[Анализ области исследования 7](#_Toc91149231)

[ЗАКЛЮЧЕНИЕ 7](#_Toc91149232)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 9](#_Toc91149233)

[ПРИЛОЖЕНИЕ 11](#_Toc91149234)

# 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**

# ЗАКЛЮЧЕНИЕ

Таким образом, в ходе данной работы было произведено исследование …

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Димитрия Князь. Анализ основных алгоритмов кластеризации многомерных данных. – М.: LAP Lambert Academic Publishing, 2014. – 64 с.

The Query-flow Graph: Model and Applications

Segmenting User Sessions in Search Engine Query Logs Leveraging Word Embeddings

Towards Semantic Query Segmentation

Attention Is All You Need

Unsupervised Query Segmentation Using Monolingual Word Alignment Method

Query Segmentation via RNNs Encoder-Decoder Framework

An Unsupervised Neural Attention Model for Aspect Extraction

Unsupervised Identification of the User’s Query Intent in Web Search

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Изображение выглядит как текст

Автоматически созданное описание

Примеры на стр 23-24 Госта

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата Подпись

# ПРИЛОЖЕНИЕ

Код программы, реализующей алгоритм «случайный лес»:

import numpy as np #linear algebra