РЕФЕРАТ

Реферат содержит 36 страниц, 30 рисунков, 4 таблиц. Список использованных источников содержит 10 позиций.

UPLIFT, МАШИННОЕ ОБУЧЕНИЕ, ИНКРЕМЕНТАЛЬНЫЙ ОТКЛИК, ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ, РАНЖИРОВАНИЕ КЛИЕНТОВ, ГРАДИЕНТНЫЙ БУСТИНГ, СЛУЧАЙНЫЙ ЛЕС.

Данная посвящен обзору возможных подходов к решению задачи прогноза инкрементального отклика клиента при получении СМС, при планировании рекламной кампании, с помощью UpLift моделирования.

С ростом глобализации и цифровизации появилась возможность работать с потребительскими данными, активно взаимодействовать с потребителями путем разных акций, особых предложений.

Но стоит взять во внимание, что каждая коммуникация стоит денег. Если клиентская база составляет 1 тыс. клиентов, то при стоимости одного СМС в 1 рубль, коммуникация будет не такой дорогой. Но если увеличить масштаб базы до миллиона или нескольких миллионов, то слепая рассылка всем клиентам 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.

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

**Содержание**

**ВВЕДЕНИЕ** 3

**ОСНОВНАЯ ЧАСТЬ** 5

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 6

1.1 Функционалы качества прогноза моделей 6

1.1.1 UpLift на k – процентах выборки 6

1.1.2 Кривая UpLift 7

1.1.3 Кривая QINI 8

1.2 UpLift моделирование методами машинного обучения 9

1.2.1 Постановка задачи UpLift 9

1.2.2 Метод UpLift моделирования с одной независимой моделью 10

1.2.3 Метод UpLift моделирования с двумя независимыми моделями 11

1.2.4 Метод трансформации класса (задача классификации) 11

1.2.5 Метод трансформации класса (задача регрессии) 12

**ЗАКЛЮЧЕНИЕ** 13

**ЛИТЕРАТУРА** 14

## **ВВЕДЕНИЕ**

В данной выпускной квалификационной работе рассматривается проблема ранжирования клиентов для осуществления коммуникации самым убеждаемым клиентам, которые без той самой коммуникации не совершат целевое действие.

В данной работе решается проблема прогноза инкрементального отклика клиента при планировании коммуникаций с помощью UpLift моделирования методами машинного обучения, где на основании полученного значения будет происходить ранжирование клиентов от самых убеждаемых к самым неприкасаемым, для повышения эффективности коммуникации при сохранении объемов затрат на ее проведение.

Результаты данной работы будут использованы в отделе управления взаимоотношений с клиентами в ретейл компании косметики и парфюмерии.

Появление данной задачи обусловлено желанием проводить нативную коммуникацию только с теми людьми, которым это нужно, чтобы не тратить денежный ресурс в пустую на тех, кому коммуникация не нужна или даже вызовет негативные эмоции и заставит уйти к конкуренту.

Объектом исследования являются клиенты ретейл сети косметики и парфюмерии, которых мы хотим ранжировать для выделения наиболее убеждаемых

Предметом исследования выступает сравнение различных алгоритмов ранжирования методами машинного обучения на двух различных источниках данных.

Цель данной работы – исследование подходов к разработке UpLift моделирования методами машинного обучения для планирования проведения рекламной кампании.

Данная работа развивает описанные в [[1]](#_bookmark45) идеи прогнозированию эффекта от коммуникации для каждого клиента при планировании рекламной кампании. С помощью показателей качества обучения из [2] и [3] удалось определить наилучший алгоритм для Uplift моделирования из описанных в [1], [4].

# **ОСНОВНАЯ ЧАСТЬ**

## 1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

### 1.1 Функционалы качества прогноза моделей

#### 1.1.1 UpLift на k – процентах выборки

Так как задача UpLift представляет собой задачу оценки (скор балл) эффекта от коммуникации на реципиента, то нет и истинных ответов. Получается, что не удастся использовать классические метрики, такие как Accuracy и PR AUC, основанные на матрице ошибок, для классификации или среднеквадратичная ошибка для задачи регрессии при трансформации классов.

Самая простая и интуитивно понятная метрика, описанная в [2], особенно для применения в бизнесе и для интерпретации – UpLift на k – процентах выборки.

Допустим, что на коммуникации в компании имеется скромный бюджет, который может обеспечить связь всего с 30% клиентской базы для побуждения к целевому действию. Тогда целью UpLift моделирования будет найти такой алгоритм, который лучше всех максимизирует эффект от коммуникаций на первых 30% клиентов.

Чтобы получить значение этой метрики, нужно ранжировать результат прогноза по убыванию, чтобы отобрать клиентов, на которых коммуникация оказывает наибольший эффект. Далее берется разница между конверсией целевой группы, с которой осуществлялась коммуникация, и конверсией контрольной группы, которая осталась без коммуникации.

Определяется формулой (1):

, (1)

где .

Как и сам UpLift, имеет область значений [-1, 1].

Причем, данную метрику можно рассчитать двумя способами, в зависимости от ранжирования по прогнозу UpLift:

* Сортировка происходит по прогнозу и далее берется разность рабочей и контрольной группы.
* Сортировка происходит внутри каждой группы обособленно и далее берется разность.

Второй вариант имеет более практическое применение, так для оценки эффективности от коммуникаций при рекламных кампаниях, при планировании проведения мероприятий, образуются две однородные выборки – рабочая и тестовая группа.

Для дальнейшего исследования будем оценивать метрику при k = 30%.

#### 1.1.2 Кривая UpLift

Далее определим кривую, которая строится как функция с нарастающим итогом, где для каждой точки задается соответствующий UpLift.

Определяется формулой (2):

, (2)

где – размер всей рабочей группы при всей выборке выборки размера t, – размер рабочей группы ,совершившей целевое действие, при всей выборке размера t, аналогично и для контрольной группы - control

Так как данный показатель относительный, он может ввести в заблуждение при интерпретации, а также не будет отражать действительность при неравных пропорция target и control. Поэтому далее опишем более интерпретируемый показатель.

Пример кривой UpLift на рисунке (рисунок 1.15).



Рисунок 1.15. – Пример кривой UpLift

#### 1.1.3 Кривая QINI

Следующую функцию, описанную в [2], можно выразить через UpLift кривую и получим определение в формуле (3):

|  |  |
| --- | --- |
|  | (3) |

Данная кривая будет полезна в тех случаях, когда рабочая группа кратно превышает размер контрольной группы, с чем можно столкнуться во время исследования модели при внедрении в бизнес, когда у компании есть бюджет на произведение коммуникаций со всей клиентской базой, и чтобы не упускать потенциальный доход, контрольная группа выделяется как можно меньше.

Таким образом будет получено инкрементальный эффект от коммуникаций в единицах измерения одного клиента.

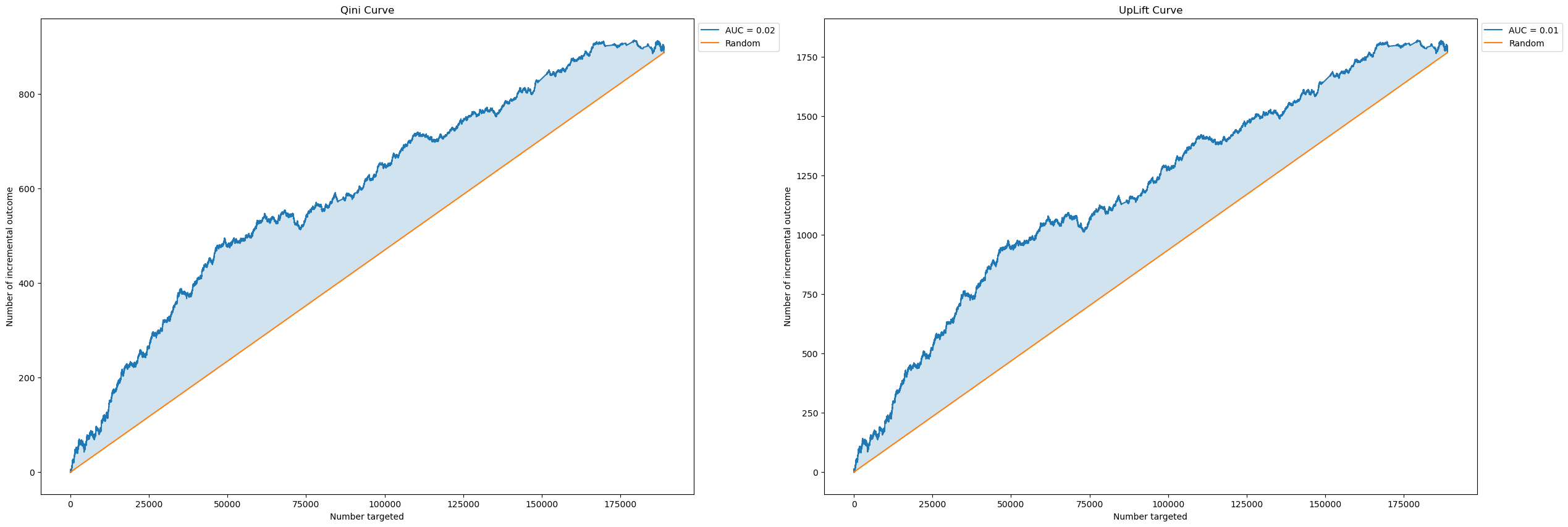


Рисунок 1.16 – Приме кривой QINI

### 1.2 UpLift моделирование методами машинного обучения

#### 1.2.1 Постановка задачи UpLift

Суть UpLift моделирования в том, чтобы определить, на каких клиентов коммуникация сработает, а на каких нет. Воспользовавшись [1], определим базовые понятия.

Эффект от коммуникации определим как casual effect:

, (4)

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):

(5)

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:

, (6)

Где – наблюдаемая реакция клиента в результате маркетинговой кампании:

(7)

, 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

#### 1.2.2 Метод UpLift моделирования с одной независимой моделью

Данный вариант решения из [1] использует переменную W как признак. Тогда обучающий набор данных имеет вид, приведенных в таблице 1.1.

Таблица 1.1 - Пример обучающего набора данных

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Обучающие признаки | | | | Целевая переменная |
| X11 | … | X1n | W1 | Y1 |
| X21 | … | X2n | W2 | Y2 |
| ………………………………………………… | | | | … |
| Xm1 | … | Xmn | Wm | Ym |

С помощью логистической регрессии или подобной модели классификации обучаем модель на данных и после обучения находим разность вероятностей на тестовой выборке, где в переменной W задаем везде единицы – будто бы была коммуникация, и на той же выборке обрабатываем данные, где в переменной W задаем нули – будто бы единицы не было. Тогда Uplift будет иметь вид:

, (8)

где P – вероятность целевого действия

#### 1.2.3 Метод UpLift моделирования с двумя независимыми моделями

Второй подход из [1] требует уже обучения двух моделей, одна модель для экспериментальной группы – , где была коммуникация, вторая модель для контрольной группы P где коммуникации не было. После обучение моделей на тренировочных выборках, совершается обработка тестовой выборки для каждой модели и за UpLift берется так же разность двух вероятностей:

, (9)

где – вероятность целевого действия первой модели, а – вероятность целевого действия первой модели

#### 1.2.4 Метод трансформации класса (задача классификации)

В данном методе из [1] мы вернемся снова к единой модели, но теперь преобразуем коммуникационную переменную и целевую переменную в одну следующим образом:

(10)

где -целевая переменная, -коммуникационная переменная.

Тогда трансформированный класс будет иметь следующие значения:

(11)

Тогда UpLift будет определяться следующим образом по формуле (12):

, (12)

где P – вероятность выполнения закодированного целевого действия

#### 1.2.5 Метод трансформации класса (задача регрессии)

В данном методе мы преобразуем коммуникационную переменную и целевую переменную в одну следующим образом:

, (13)

где – целевая переменная, коммуникационная переменная, – вероятность принадлежности к целевой группе.

В нашем случае, . Тогда трансформированный класс будет иметь следующие значения:

(14)

Тогда UpLift будет определяться следующим образом по формуле (15):

, (15)

где R – регрессионное значение закодированного целевого действия.

### 1.3 Модели машинного обучения

#### 1.3.1 Логистическая регрессия

Логистическая регрессия — это метод машинного обучения, который используется для предсказания вероятности отнесения объекта к определенному классу.

Он основан на логистической функции (16):

(16)

, которая принимает на вход линейную комбинацию признаков — объекта и выдает вероятность его принадлежности к классу — , а – параметры логистической регрессии .

Логистическая регрессия широко используется в задачах классификации, например, для определения того, является ли электронное письмо спамом или нет, или для диагностики заболеваний на основе медицинских данных.

В данном случае, логистическая функция будет использоваться для прогноза вероятности – где в нашем случае это вероятность целевого действия при параметрах = {X, T}, где X – это параметры объекта, а T – признак воздействия на объект.

Так как логистическая регрессия – модель машинного обучения с учителем, нужно оценивать качество прогноза и корректировать параметры модели для нахождения оптимального набора весов, при котором вероятность целевого действия будет иметь оценку, наиболее приближенную к истинному значению.

Функция потерь для логистической регрессии называется логистической функцией потерь (log loss) или кросс-энтропией (cross-entropy loss). Она измеряет разницу между предсказанными вероятностями классов и фактическими метками классов. Логистическая функция потерь минимизируется при обучении модели. Формула логистической функции потерь выглядит так (17):

(17)

, где - фактическая метка класса (0 или 1), – вероятность совершения целевого действия.

Далее, для минимизации функции ошибки логистической регрессии, определим метод оптимизации.

При обучении логистической регрессии часто используется метод градиентного спуска для оптимизации параметров модели. Он позволяет находить минимум функции потерь путем итеративного изменения параметров модели в направлении антиградиента функции потерь.

Существуют различные варианты метода градиентного спуска, такие как стохастический градиентный спуск (SGD), мини – пакетный градиентный спуск (mini – batch GD) и т.д.

SGD является наиболее распространенным методом оптимизации для логистической регрессии, так как он работает быстрее и требует меньше вычислительных ресурсов, чем другие методы оптимизации. В SGD параметры модели обновляются на каждом шаге для каждого объекта в обучающей выборке, что позволяет быстрее сойтись к оптимальному решению.

Метод оптимизации SGD (стохастический градиентный спуск) работает следующим образом:

1. Инициализируются параметры модели случайными значениями.

2. Выбирается случайный объект из обучающей выборки.

3. Вычисляется градиент функции потерь по параметрам модели на основе выбранного объекта.

4. Обновляются параметры модели в направлении антиградиента функции потерь с помощью формулы (18):

(18)

, где – вектор параметров модели, – шаг изменения весов (learning rate), – градиент функции потерь.

5. Повторяются шаги 2 – 4 для каждого объекта в обучающей выборке.

6. Повторяются шаги 2 – 5 до тех пор, пока не будет достигнут критерий остановки, например, определенное количество эпох обучения или достижение минимального значения функции потерь.

SGD обновляет параметры модели на каждом шаге, что позволяет быстрее сойтись к оптимальному решению.

В методе стохастического градиентного спуска, градиент находится следующим образом – он определяется как вектор частных производных функции потерь по каждому параметру модели. Градиент показывает направление наискорейшего убывания функции потерь и используется для обновления параметров модели в каждой итерации оптимизации.

Для логистической регрессии градиент функции потерь вычисляется следующим образом (19):

(19)

где – вектор параметров модели, – матрица объектов обучающей выборки, y – вектор фактических меток классов, логистическая функция активации, ∇ - оператор градиента, T - оператор транспонирования матрицы, N - количество объектов в обучающей выборке.

Для каждого объекта в обучающей выборке градиент вычисляется по формуле (20):

(20)

, где - вектор признаков объекта.

Таким образом, для каждого объекта в обучающей выборке вычисляется градиент функции потерь, и параметры модели обновляются в направлении антиградиента.

#### 1.3.2 Линейная регрессия

Так как для моделирования UpLift используется и метод трансформации класса с переходом к задаче регрессии – рассмотрим один из основных и известных методом машинного обучения – линейная регрессия.

Линейная регрессия - это статистический метод, используемый для оценки связи между непрерывными переменными. Это модель, которая пытается установить линейную зависимость между зависимой переменной (таргетом) и одной или несколькими независимыми переменными (факторами).

Линейная регрессия предполагает, что зависимость между переменными может быть описана линейной функцией. Линейная функция представляет собой уравнение прямой линии в двумерном пространстве, которая может быть расширена на более высокие размерности.

Для построения модели линейной регрессии необходимо определить коэффициенты, которые лучше всего соответствуют уравнению линейной функции, связывающей зависимую и независимые переменные. Эти коэффициенты могут быть определены с помощью метода наименьших квадратов (МНК), который минимизирует сумму квадратов расстояний между фактическими значениями таргета и предсказанными значениями.

Линейная регрессия широко используется в различных областях, включая экономику, финансы, бизнес, медицину, науку о материалах и многие другие.

Регрессионная модель представляет собой следующее (21):

(21)

, где – вектор весов модели, – случайная составляющая модели, а определяется как (22):

(22)

, где k – количество факторов линейной регресии, а – факторы модели.

Наиболее распространенной функцией потерь для линейной регрессии является среднеквадратичная ошибка (Mean Squared Error, MSE), которая определяется следующим образом (23):

(23)

где N - количество наблюдений, - фактическое значение таргета для i-го наблюдения, - предсказанное значение таргета для i-го наблюдения.

Для оптимизации параметров модели линейной регрессии с использованием функции потерь в виде среднеквадратичной ошибки (MSE), наиболее распространенным методом является метод градиентного спуска, который работает следующим образом:

1. Инициализируются параметры модели случайными значениями.

2. Выбирается случайный объект из обучающей выборки.

3. Вычисляется градиент функции потерь по параметрам модели на основе выбранного объекта.

4. Обновляются параметры модели в направлении антиградиента функции потерь с помощью формулы (24):

(24)

, где – вектор параметров модели, – шаг изменения весов (learning rate), – градиент функции потерь вида среднеквадратического отклонения.

5. Повторяются шаги 2 – 4 для каждого объекта в обучающей выборке.

6. Повторяются шаги 2 – 5 до тех пор, пока не будет достигнут критерий остановки, например, определенное количество эпох обучения или достижение минимального значения функции потерь.

## **ЗАКЛЮЧЕНИЕ**

В данной выпускной квалификационной работе магистра предлагается исследование подходов к UpLift моделированию методами машинного обучения на исходных данных ретейл компании в сфере косметики и парфюмерии.

Были выбраны и описаны структуры с одной моделью машинного обучения, с двумя независимыми моделями машинного обучения и два вида трансформации класса для обучения одной модели машинного обучения классификации и регрессии.

Численные результаты эксперимента показали, что наилучшего UpLift по показателям качества обучения можно добиться с помощью автоматического подбора моделей задачи классификации и последующим применением ее в алгоритме с одной независимой моделью.

Найденный алгоритм, возможно, будет наилучшим только для рассматриваемых в задаче данных, так как в зависимости от скрытой природы зависимостей обучающих признаков, различные структуры могут показывать наилучшие результаты на одних данных и наихудшие на других.

В работе приведены обзоры на различные способы решения проблемы и полученные результаты в перспективе могут быть аналогичны и для остальной клиентской базы ретейл компании косметики и парфюмерии.

## **ЛИТЕРАТУРА**

1. Gutierrez P., G´erardy J. Causal Inference and Uplift Modeling A review of the literature // PMLR – 2016 - URL: <https://proceedings.mlr.press/v67/gutierrez17a/gutierrez17a.pdf>
2. Weijia Zhang, Jiuyong Li, Lin Liu A unified survey of treatment effect heterogeneity modelling and uplift modelling // arXiv – 2021 – URL: <https://arxiv.org/pdf/2007.12769>
3. Devriendt F., Guns T., Verbeke W. LEARNING TO RANK FOR UPLIFT MODELING // arXiv – 2020 – URL: <https://arxiv.org/pdf/2002.05897>
4. Nyberg O., Kussmierczyk T., Klami A. Uplift Modeling with High Class Imbalance // PMLR – 2021 – URL: <https://proceedings.mlr.press/v157/nyberg21a/nyberg21a.pdf>
5. RF – сегментация – URL: <https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/>
6. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучениX`е // пер. с анг. А. А. Слинкина. – 2-е изд., испр. – М.: ДМК Пресс – 2018. – 652
7. Туториал по uplift моделированию. Часть 1 – URL: <https://habr.com/ru/companies/ru_mts/articles/485980/>
8. Курс по uplift моделированию – URL: <https://ods.ai/tracks/uplift-modelling-course>
9. Введение в Uplift моделирование – URL: <https://newtechaudit.ru/vvedenie-v-uplift-modelirovanie/>
10. Продвинутые методы Uplift-моделирования – URL: https://habr.com/ru/companies/glowbyte/articles/686398/