ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ЖУРНАЛ ПРАКТИКИ

Студента 1 курса Фейзуллина Кирилла Маратовича

(Фамилия, имя. отчество)

Институт №8 «[Информационные технологии и прикладная математика](https://mai.ru/education/fpmf/)»

Кафедра 804 «Теория вероятностей и компьютерное моделирование»

##### Учебная группа М8О-101М-21

Направление 01.04.04. \_\_\_\_\_Прикладная математика\_\_\_\_\_\_

(шифр) (название направления)

Вид практики \_\_\_\_\_учебная (исследовательская)\_\_

в Московском авиационном институте (НИУ)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(наименование предприятия, учреждения, организации)

Руководитель практики от МАИ \_\_Платонов Е.Н.\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(ФИО) (Подпись)

\_\_\_Фейзуллин К.М.\_\_\_\_\_\_ /\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ “07” июня 2022 г.

(ФИО) (подпись студента) (дата)

1. **Место и сроки проведения практики**

Дата начала практики “09*” февраля 2022 г.*

Дата окончания практики “07*” июня 2022 г.*

*Наименование предприятия* МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Название структурного подразделения) \_\_\_\_\_кафедра 804\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

1. **Инструктаж по технике безопасности**

\_\_\_Платонов Е.Н.\_\_\_\_ /\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ “09” февраля 2022 г.

*(подпись НР)* *(дата проведения)*

1. **Индивидуальное задание студенту**

|  |
| --- |
| Определение метрик для оценки качества UpLift моделирования |
| Поиск источника данных для машинного обучения |
| Реализация UpLift моделирования методами машинного обучения |
|  |
|  |
|  |
|  |

1. **План выполнения индивидуального задания**

|  |
| --- |
| 9 – февраля – 28 февраля – анализ метрик оценки качества UpLift моделирования |
| 1 марта – 31 марта – поиск данных |
| 1 апреля – 30 апреля – анализ и агрегирование данных |
| 1 мая – 31 мая – реализация UpLift моделирования методами машинного обучения |
| 7 июня – оформление отчета по практике |
|  |

*Руководитель практики от МАИ*: Платонов Евгений Николаевич /\_\_\_\_\_\_\_\_\_\_\_\_\_/

(Фамилия, имя, отчество) (Подпись)

\_\_\_Фейзуллин Кирилл Маратович\_\_\_\_\_\_ /\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ “07” июня 2022 г.

(ФИО) (подпись студента) (дата)

**5.Отзыв руководителя практики**

|  |
| --- |
| Запланированная работа выполнена. Материалы, изложенные в отчете студента, |
| полностью соответствуют индивидуальному заданию. |
| Оценка за практику «отлично». |
|  |
|  |
|  |
|  |
|  |
|  |

*Руководитель*  \_\_\_Платонов Е.Н.\_\_\_ /\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/

(Фамилия, имя, отчество) (Подпись)

“07” июня 2022 г.

Отчет студента

Объектом исследования являются задача прогнозирования оттока клиентов.

Цель работы – первые эксперименты реализации UpLift моделирования методами машинного обучения.

## **Определение метрик для оценки качества UpLift моделирования**

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.

### UpLift на первых k – процентах выборки

Самая простая и интуитивно понятная метрика, особенно для применения в бизнесе и для интерпретации.

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.

Формула имеет следующий вид:

,

где .

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

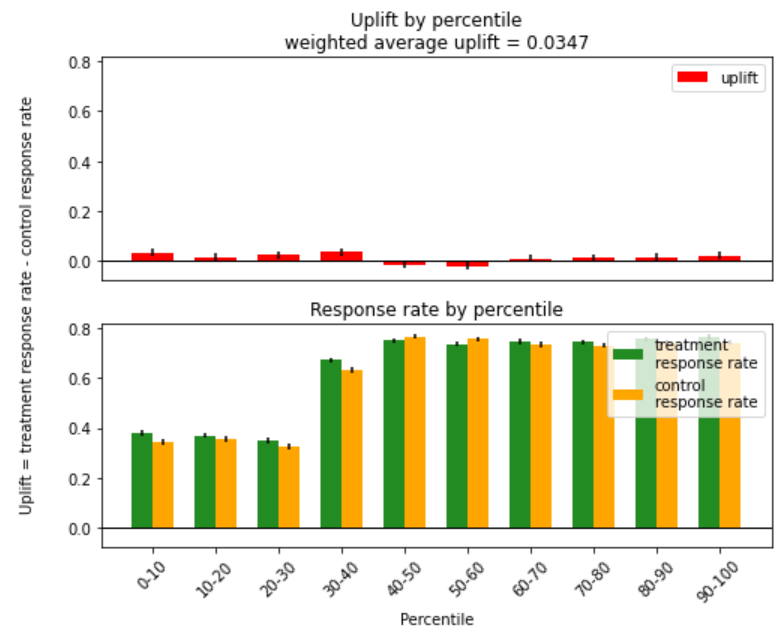
Причем, данную метрику можно рассчитать двумя способами, в зависимости от ранжирования по прогнозу UpLift:

* 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.

### UpLift по процентилям

Данная метрика представляется в виде таблицы или графика для общего понимания качества работы модели при разных долях выборки K%, где для каждого K% определяется . Пример отображения на рисунке 1.



Рисунок

Алгоритм расчета схож с предыдущей метрикой:

1. Выборка сортируется по прогнозу UpLift.
2. Отсортированные данные делятся на интервалы – обычно берется 10 интервалов.
3. Для каждого интервала оцениваем и берем разность.

## Средний взвешенный UpLift (Weighted Average UpLift)

Данная метрика представляет собой оценку UpLift по всей выборки и определяется следующий образом:

,

где

## UpLift кривая (UpLift Curve)

Данная кривая строится как функция с нарастающим итогом, где для каждой точки задается соответствующий UpLift.

Определяется следующим образом:

, где

,

.

Аналогично и для контрольной группы.

Пример данной кривой на рисунке 2.



Рисунок

### Qini кривая

Данную функцию можно выразить через UpLift кривую следующим образом:

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.

## **Источник данных**

За источник данных было взято уже завершенное соревнование по UpLift моделированию от российской мега-корпорации X5 Retail Group (ныне X5 Group) на платформе Open Data Science (ODS)[[1]](#footnote-1). Этот набор данных имеет преимущество над ныне существующими в открытом доступе благодаря тому, что это фактически моментальный снимок базы данных компании, во временном интервале за четыре месяца, хранящий в себе транзакции клиентов за соответствующий период, их обезличенные анкетные данные, обезличенный продуктовый справочник с данными по каждому товару сети.

Данное преимущество позволяет самому смоделировать и выделить важные признаки, и получить релевантный опыт работы с живыми, а не синтетическими или уже агрегированными данными.

Опишем набор данных детальнее. Он состоит из:

* Общей информации о клиентах:

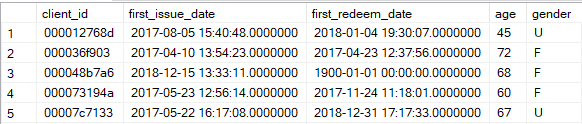


Рисунок 3

* Общая информация о товарах на складе:

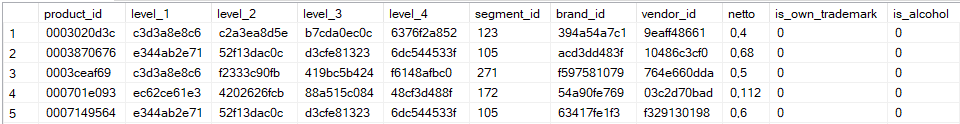


Рисунок 4

* История покупок клиента до коммуникаций:

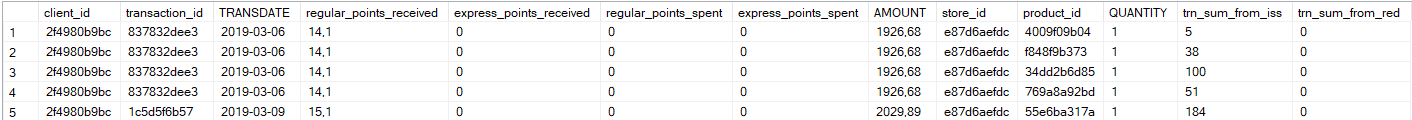


Рисунок 5

* Целевые переменные для обучения:

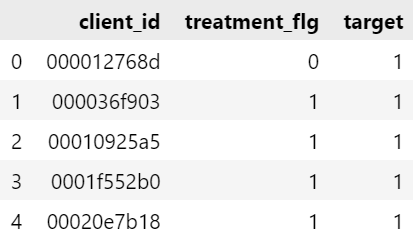


Рисунок 6

* Данные для теста:

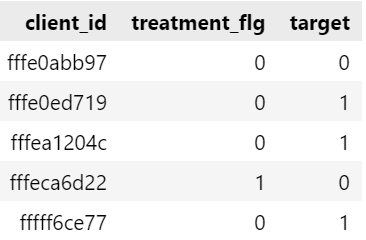


Рисунок 7

## **Анализ и агрегирование данных**

Так как данные для UpLift моделирования составляют 4 Гб. в формате csv, что достаточно много для табличных данных самом экономном формате, то было решено взаимодействовать с ними через реляционный язык запросов SQL. Для этого был развернут локальный SQL Server на СУБД MSSQL и с помощью SQL Management Studio были загружены табличные данные.

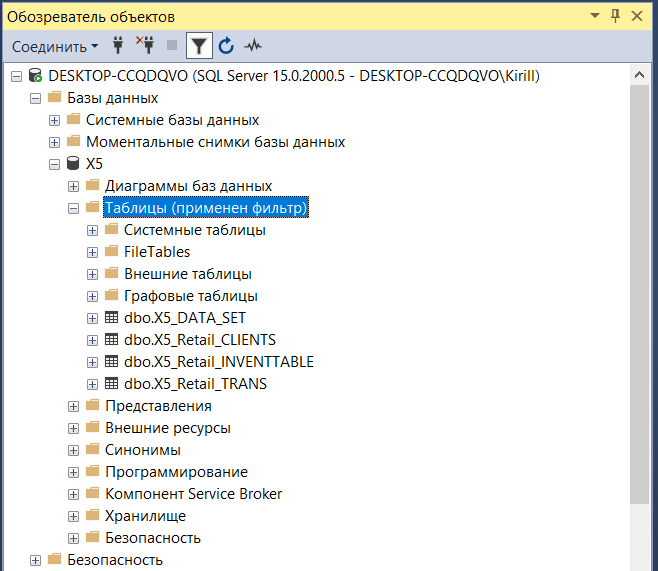


Рисунок 8

Через транзакции были выделены наиболее часто покупаемые товары для агрегации их в признаки.

Таким образом были выделены наиболее продаваемые:

* Уровни в иерархии товаров – рисунок 8.

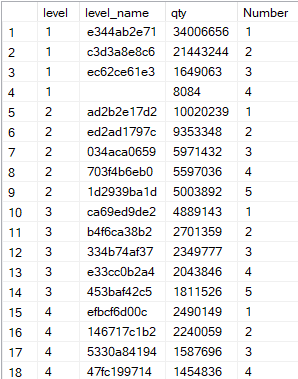


Рисунок 8

* Бренды – рисунок 9.

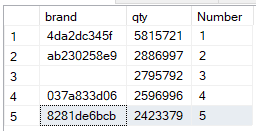


Рисунок 9

* Поставщики – рисунок 10.

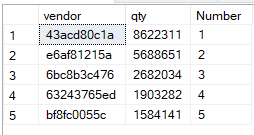


Рисунок 10

* Сегменты товаров – рисунок 11.

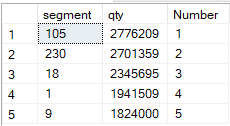


Рисунок 11

Для моделирования основных обучающих признаков был использован принцип RFM - сегментации[[2]](#footnote-2). То есть, по покупкам клиентов были определены следующие параметры:

* Частота покупок – количество покупок за расчетный период.
* Период с момента последней покупки.
* Сумма товарооборота с клиента за расчетный период - в нашем случае возьмем средний чек, так как это стратифицировать клиентов явным образом.

Также была собрана статистика по доле алкогольных товаров в чеке, доля внутренних брендов, среднее время между покупками и сопутствующая статистика по уровням товаров, брендам, поставщикам и сегментам. Вдобавок к этому были учтены и анкетные данные.

Таким образом было получено пространство из 30-ти обучающих признаков.

# **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

5. Курс лекций анализа данных. Лекция 7 – UpLift моделирование / Платонов Е.Н. // Московский авиационный институт

1. https://ods.ai/competitions/x5-retailhero-uplift-modeling [↑](#footnote-ref-1)
2. RFМ – сегментация // <https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/> [↑](#footnote-ref-2)