**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»

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. / _____ / "04" января 2022 г.
(ФИО) (подпись студента) (дата)

Москва 2022

**1. Место и сроки проведения практики**

*Дата начала практики* *“01” сентября 2021 г.*

*Дата окончания практики* *“04” января 2022 г.*

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

*Название структурного подразделения)* *кафедра 804*

**2. Инструктаж по технике безопасности**

Платонов Е.Н. /_____/ “01” сентября 2021 г.

*(подпись НР)* *(дата проведения)*

**3. Индивидуальное задание студенту**

Исследование задачи прогнозирования оттока клиентов – задача бинарной классификации.

Исследование задачи UpLift моделирования.

## 4. План выполнения индивидуального задания

| |
|---|
| 1 –сентября – получение задания |
| 1 сентября – 1 октября – изучение теоретического материала |
| 2 октября – 21 ноября – формулирование математической постановки задачи |
| 22 ноября – 6 декабря – определение и исследование методов решения |
| 6 декабря – оформление отчета по практике |
| |
| |

*Руководитель практики от МАИ*: Платонов Евгений Николаевич /_____/

(Фамилия, имя, отчество) (Подпись)

___Фейзуллин Кирилл Маратович_____ /_____/ “04” января 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)

“04” января 2022 г.

# Отчет студента

Объектом исследования являются задача прогнозирования оттока клиентов.

Цель работы – постановка задачи и исследование 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 $\eta: \Omega \to$ Y – случайная величина,

представляющая собой случайный правильный ответ из $Y$. Тогда определим случайную величину $(\xi, \eta) : \Omega \to (X, Y)$ 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.

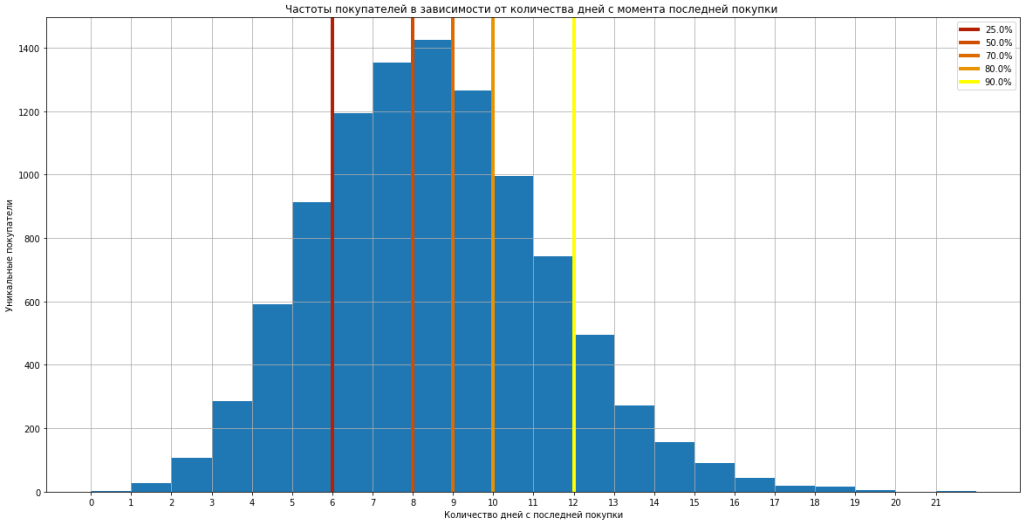


*Рисунок 1. Моделирование потребительского поведения.*

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.

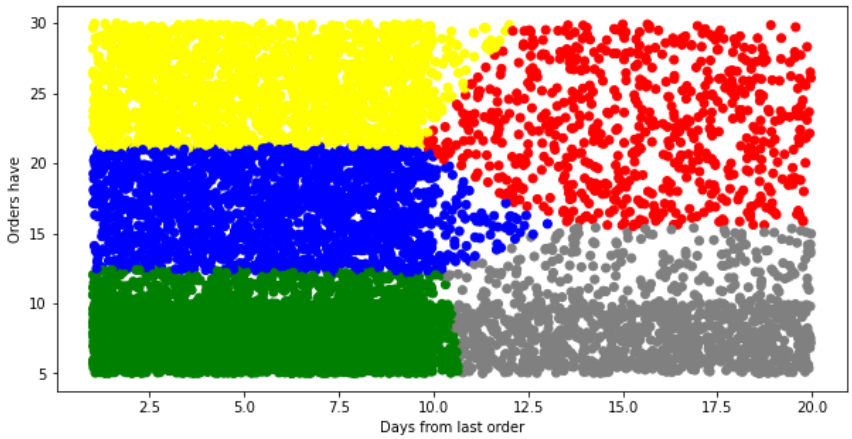


*Рисунок 2. RF сегментация.*

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.

## Задача UpLift моделирования

Перед решением самой задачи UpLift моделирование следует описать предшествующие шаги, так как при расчете данной величины

$UpLift(x) = M[Y_i|X_i = x, W_i = 1] - M[Y_i|X_i = x, W_i = 0]$ мы уже должны работать с выборками, где была коммуникация и где ее не было.

Опишем шаги:

1. Формируется клиентская база с разделением на две части – контрольная и экспериментальная.
2. Проводится коммуникация.
3. Строим UpLift модель.

Если первые два шага напоминают АВ – тестирование и имеют понятную природу, то третий шаг разберем подробнее дальше.

Существует несколько основных методов UpLift моделирования:

1. Метод с 1 моделью.
2. Метод с 2 моделями.
3. Трансформация целевой переменной.
4. Решающие деревья с UpLift критерием разбиения.

Начнем с первого метода. Данный вариант решения использует переменную W как признак. Тогда обучающий набор данных имеет вид, приведенных в таблице 1.

*Таблица 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:

$$UpLift = P\left(\begin{bmatrix} x_1^1 & \cdots & x_1^n & 1 \\ \vdots & \ddots & & \vdots \\ x_m^1 & \cdots & x_m^n & 1 \end{bmatrix}\right) - P\left(\begin{bmatrix} x_1^1 & \cdots & x_1^n & 0 \\ \vdots & \ddots & & \vdots \\ x_m^1 & \cdots & x_m^n & 0 \end{bmatrix}\right)$$

Второй подход требует уже обучения двух моделей, одна модель для экспериментальной группы – $P[Y|X = x, W = 1]$, где была коммуникация, вторая модель для контрольной группы $\mathrm{P}[Y|X = x, W = 0]$ 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:

$$UpLift = P[Y|X = x, W = 1] - \mathrm{P}[Y|X = x, W = 0]$$

Первые два метода имеют простоту реализации как позитивный фактор, но отрицательным фактором является то, что признак коммуникации W является не целевой переменной, а лишь признаком.

Решает данную проблему трансформация целевой переменной следующим образом:

$$Y_i^* = Y_i \frac{W_I}{p(X_i)} - Y_i \frac{1-W_I}{1-p(X_i)} = Y_i \frac{W_I - p(X_i)}{p(X_i) * (1-p(X_i))},$$

Где $p(X_i) = P(W_i = 1|X_i)$ – вероятность принадлежности к целевой группе. Причем, в практике это просто доля экспериментальной выборки, что обычно составляет половину, то есть $p(X_i) = 0.5$.

Тогда область значений новой целевой переменной и ее область определения имеет вид:

$$\begin{pmatrix} Y & w \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} \rightarrow^{p=0.5} \begin{pmatrix} Y^* \\ 2 \\ 0 \\ -2 \\ 0 \end{pmatrix}$$

Далее на тренировочной выборке обучается модель регрессии с среднеквадратичной функцией потерь, так как она обеспечивает связь между UpLift и новой целевой переменной $Y^*$ при обучении, что доказано в [5].

Последний метод решения основан на деревьях решений, в которых изменен критерий разделения на дочерние узлы. Критерий изменяется для максимизации разброса UpLift, так как в данной задаче требуется найти максимальную разность между контрольной и экспериментальной выборкой. Тогда критерий разбиения будет иметь следующий вид на примере евклидового расстояния:

$$D_E\,(P,Q) = \sum(p_i - q_i)^2 \text{ , где}$$

$$p = \frac{\sum Y_i W_i}{\sum W_i},$$

$$q = \frac{\sum Y_i(1 - W_i)}{\sum(1 - W_i)}$$

Так как в нашей задаче стоит максимально эффективная коммуникация, то есть пользователи с прогнозируемым максимальным значением UpLift, список прогнозов ранжируется по убыванию и выбираются первые N пользователей для коммуникации. Размер выборки N определяется исходя из бюджета, заложенного на коммуникацию

# 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