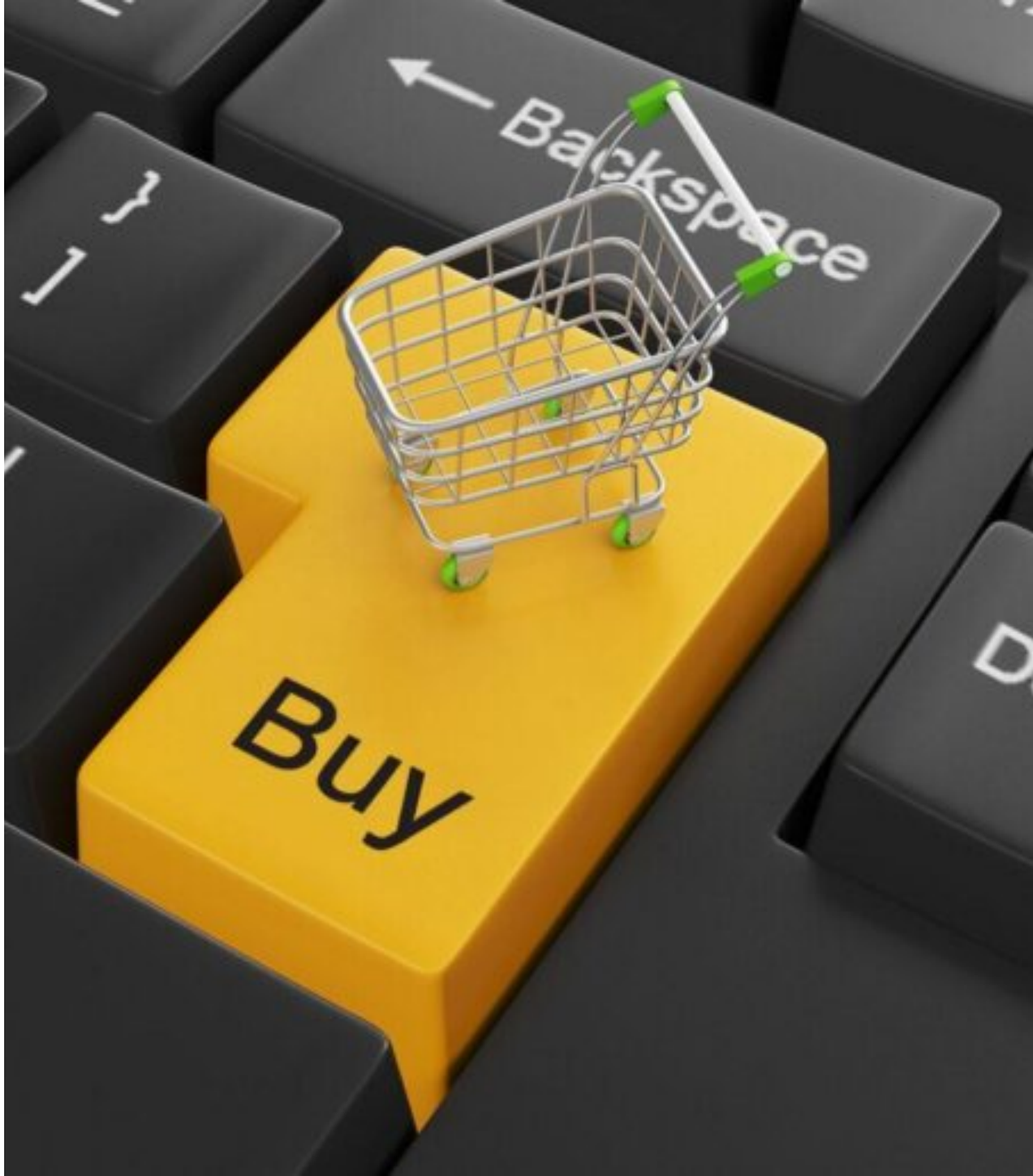
Определение вероятности покупки товара на основе данных о клиенте и его покупательской истории с помощью методов машинного обучения

Итоговый проект по программе
«Специалист по Data Science»

Моренко Антон
DS-16

Август 2025 г.



Содержание

Описание проекта

Цель проекта: Научиться предсказывать вероятность совершения покупки клиентом на основе данных о нем и его покупательской истории

Задачи проекта:

1. Провести предобработку и исследовательский анализ данных полученного датасета

2. Составить портрет покупателя

3. Провести кластеризацию покупателей

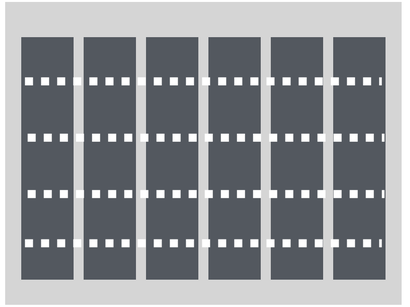
4. Выбрать и обучить модель определения вероятности покупки товара

Основные результаты

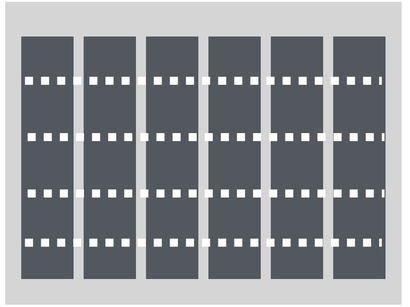
Предобработка данных

1. Получен исходный датасет о 2240 клиентах с 22 признаками
2. Выполнена предобработка данных:
   - Исходный датасет сокращен на 10% за счет следующих операций:
     - Удалено 24 записи с пропусками
     - Удален 201 дубликат
     - Удалены три аномальные записи (по году рождения клиента)
   - Типы данных приведены к целевым

2240 x 22



2012 x 29



3. Добавлены шесть новых признаков:
   - **registration_age** - возраст клиента на момент регистрации в программе лояльности
   - **is_parent** - обобщенный признак наличия детей
   - **mnt_total** - общая сумма покупок по основным товарным категориям
   - **expenses_per_member** - сумма покупок в расчете на каждого члена семьи
   - **total_purchases** - общее количество покупок по разным каналам
   - **avg_check** - средний чек покупки (общая сумма / общее кол-во покупок)

Основные результаты

Исследовательский анализ данных (1/3)

- Выполнена описательная статистика признаков:

- Большая часть клиентов имеют уровень образования Graduation (1014 клиентов) - выпускники средней школы или колледжа / бакалавры. Вторая по популярности категория - PhD (436 клиентов). Меньше всего клиентов с базовым уровнем образования (49).
- В разрезе семейного статуса больше всего клиентов находящихся в браке (781 человек), состоящих в отношениях (509 человек), и не состоящих в отношениях (435 человек). Самые малочисленные категории: Alone, Absurd и YOLO.
- Доходы клиентов варьируются от 1730 до 667 тыс. долларов в год. Медианный доход - 51.5 тыс долл./год, средний доход выше медианного за счет наличия нескольких экстремальных значений. Годовой доход 75% клиентов не превышает 68.6 тыс. долл.
- Больше всего клиентов, имеющих одного ребенка (1023), на втором месте клиенты без детей - 568, три ребенка - самое редкое явление - всего 45 клиентов.
- complain - признак, указывающий на факт жалоб со стороны клиента. Клиентов, которые направляли жалобы всего 19 человек (менее 1%). Такой признак не несет информации, так как по сути является в 99% нулевым столбцом.
- Есть группа клиентов, у которых траты в рассмотренных категориях значительно превышают средний уровень. Это может быть полезным признаком при проведении кластеризации клиентов. Скорее всего это могут быть либо состоятельные клиенты либо клиенты имеющие большую семью.
- Покупка товаров непосредственно в магазине - наиболее популярный способ совершения покупок (около 6 покупок на клиента). Второй по популярности канал - покупки на сайте компании (около 4 покупок на клиента).
- В среднем клиенты посещают веб-сайт компании около 5 раз в месяц.
- Чаще всего регистрировались клиенты в возрасте 34-50 лет и 54-62 года. Средний возраст клиента на момент регистрации - 44 года. Самому младшему клиенту на момент регистрации было 16 лет, самому старшему 73 года.

- только 15% клиентов совершили покупку при получении предложения.

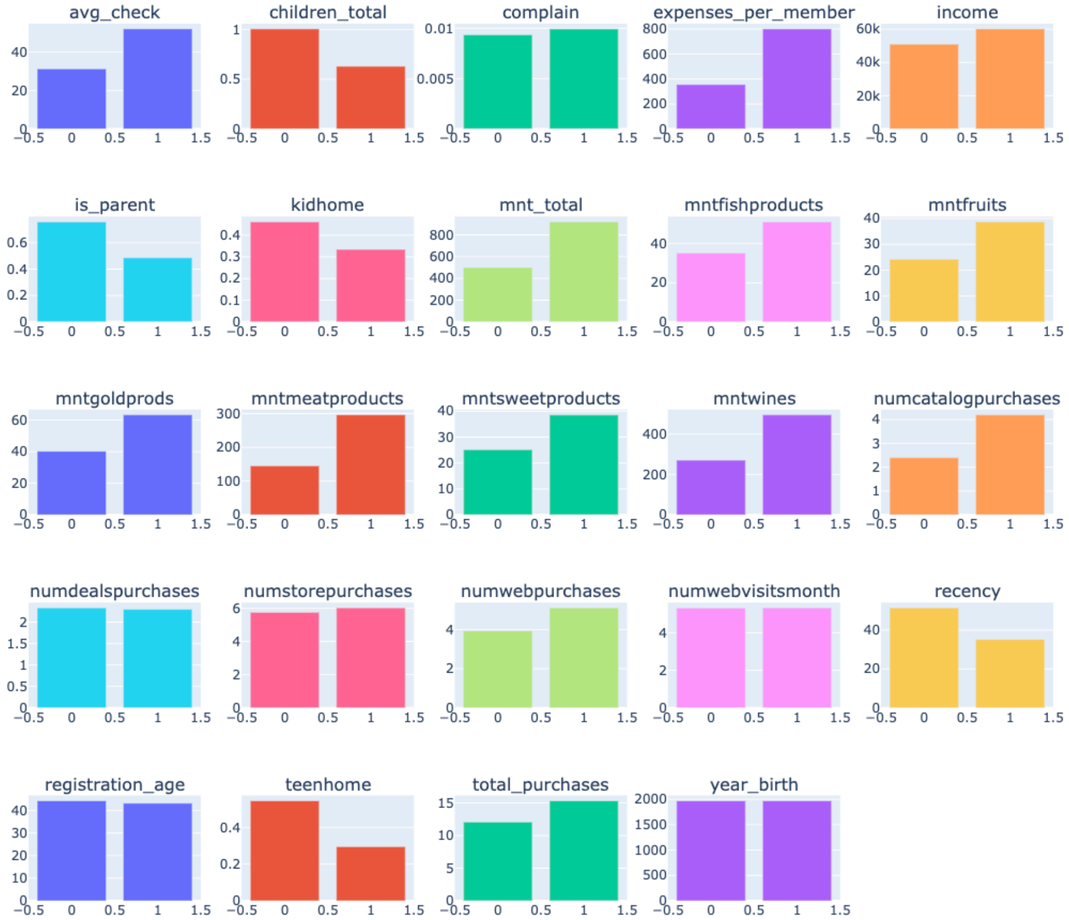
Основные результаты

Исследовательский анализ данных (2/3)

- Выполнен анализ средних значений признаков по группам принявших и не принявших предложение о покупке:

Существенные различия по группам видны у следующих показателей:

- Группа показателей состава семьи:
  - Клиенты являющиеся родителями, имеющие большее количество маленьких детей и подростков чаще попадают в группу 0
- Группа показателей расходов на основные товарные категории:
  - Клиенты из группы 1 в среднем тратят боольше денег на основные товарные категории.
  - Средние расходы в расчете на каждого члена семьи также выше у клиентов из группы 1 (801 против 355 долл.)
- Группа показателей каналов продаж:
  - клиенты из группы 1 чаще заказывают товары по каталогу (4 против 2)
  - клиенты из группы 1 чаще покупают на сайте (5 против 4)
  - клиенты из группы 1 в целом чаще совершают покупки (15 против 12)
- У клиентов из группы 1 средний чек выше (52 долл. против 31)
- Клиенты из группы 1 чаще совершают покупки - 1 раз в 35 дней, у клиентов из группы 0 - 1 р в 51 день.
- Клиенты из группы 1 зарабатывают в среднем на 10 тыс.долл в год больше.



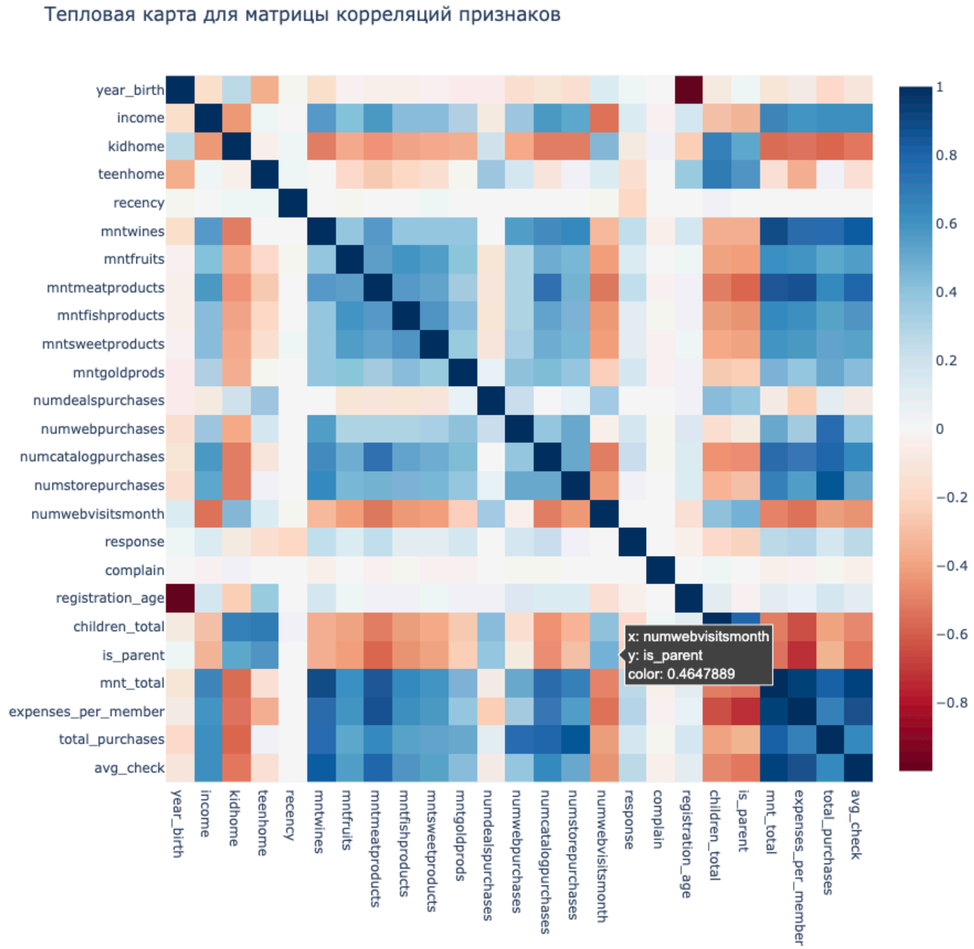
Видимые отличия отсутсвуют или незначительны по следующим признакам:

- жалобы
- покупки со скидкой
- покупки непосредственно в магазине
- количество посещений сайта за последний месяц
- возраст клиента на момент регистрации
- год рождения

Основные результаты

Исследовательский анализ данных (3/3)

- Выполнен анализ матрицы корреляции признаков:



- Из обнаруженных **значимых положительных корреляций** можно сделать следующие обобщающие выводы:
  - Чем выше доходы клиента тем выше траты на все товарные категории и количество совершенных покупок по всем каналам
  - Клиенты с детьми чаще совершают покупки со скидкой
  - Не удалось обнаружить положительной корреляции между целевой переменной и другими признаками в датасете

- Из рассмотрения **значимых отрицательных корреляций** можно сделать основной обобщающий вывод - факт родительства и особенно наличие маленьких детей находится в отрицательной связи с доходами, тратами на основные товарные категории и прочими производными от них показателями. Также не удалось обнаружить значимой отрицательной корреляции целевой переменной с другими признаками датасета.

Основные результаты

Выбор и обучение модели для определения вероятности покупки (1/9)

- **В качестве базовых моделей отобраны следующие алгоритмы:**
  - Логистическая регрессия
  - KNN (K-Nearest Neighbors)
  - Метод опорных векторов (SVM) с вероятностной интерпретацией
  - Наивный байесовский классификатор
  - Дерево принятия решений для классификации
  - Случайный лес для классификации (ансамбль)
  - Градиентный бустинг для классификации (ансамбль)

- **Для оценки качества прогнозов обученных моделей выбраны следующие метрики:**
  - Accuracy
  - F-1
  - ROC-AUC

ROC-AUC хорошо подходит для несбалансированных классов (как в случае с нашим датасетом), AUC позволяет сравнивать различные модели, независимо от их порогов.

Основные результаты

Выбор и обучение модели для определения вероятности покупки (2/9)

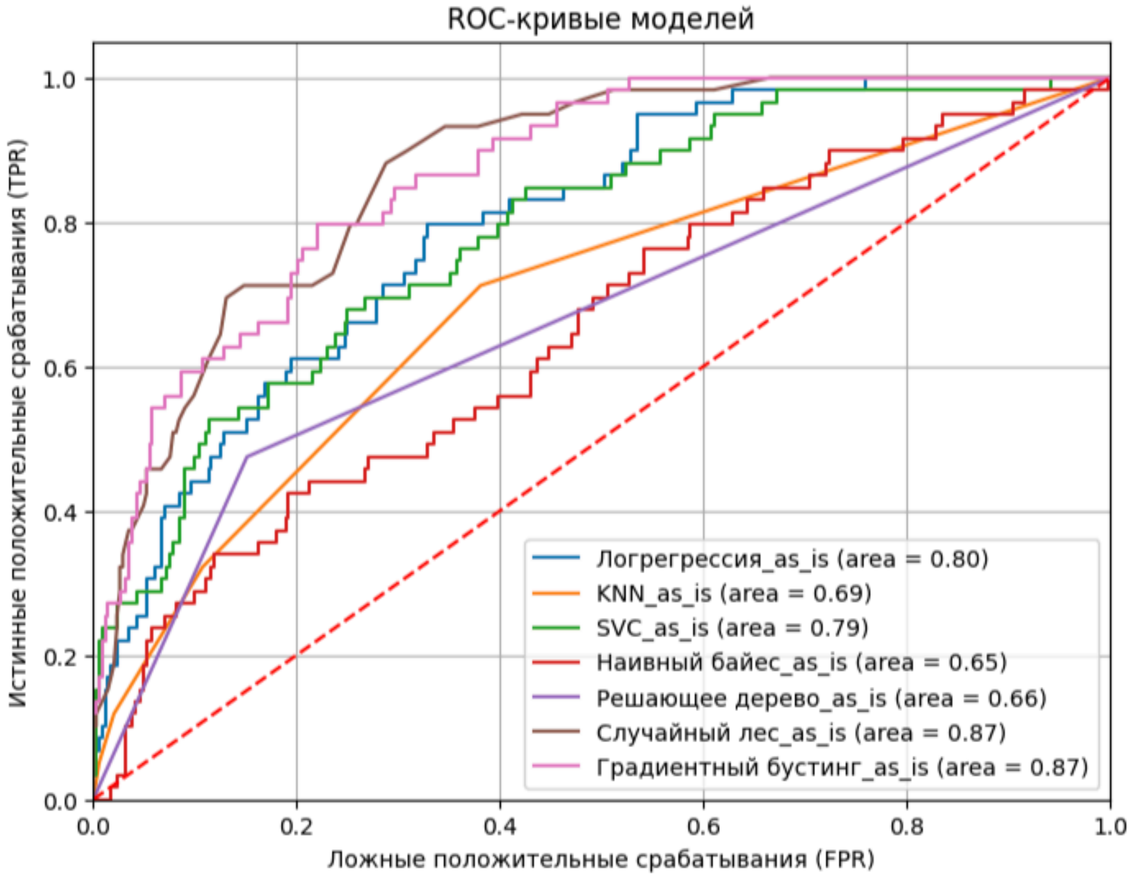
Проведенные эксперименты:

Обучение на данных «как есть»	Обучение базовых моделей с учетом снижения размерности и без обработки дисбаланса классов	Обучение базовых моделей с учетом снижения размерности и с обработкой дисбаланса классов	Обучение базовых моделей с с обработкой дисбаланса классов методом оверсэмплинга
	• LinearDiscriminantAnalysis (LDA)	• LDA + изменение весов классов • LDA + SMOTE	• SMOTE • ADASYN • SMOTEK

Основные результаты

Выбор и обучение модели для определения вероятности покупки (3/9)

Обучение на данных «как есть»

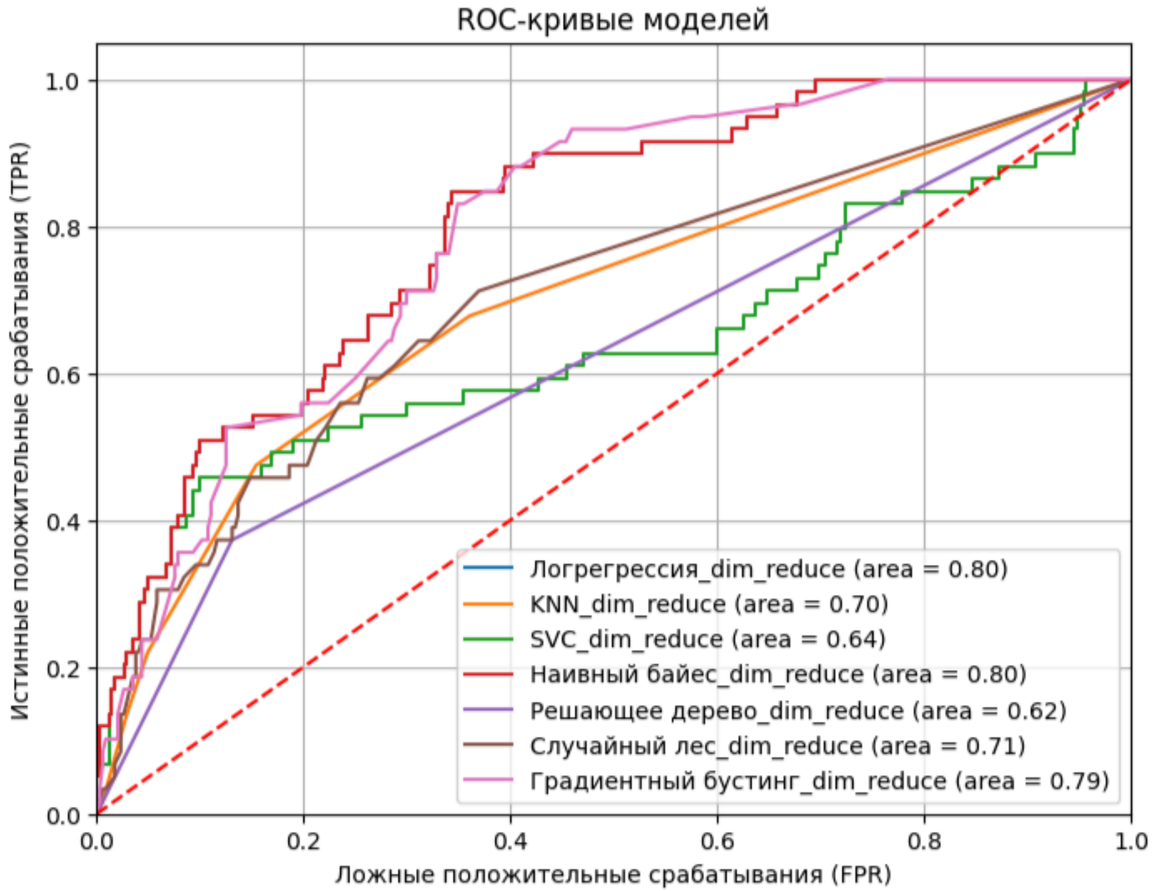


- По интегральной оценке хуже всего получилась модель наивного байесовского классификатора. Модель градиентного бустинга показала самый высокий результат - 0.79.
- Наибольший ROC-AUC получен для случайного леса (0.87), наименьший - для наивного байеса (0.65)
- Наибольший Accuracy получен для градиентного бустинга (0.88), наименьший для наивного байеса (0.22)

Основные результаты

Выбор и обучение модели для определения вероятности покупки (4/9)

Обучение базовых моделей с учетом снижения размерности и без обработки дисбаланса классов



Снижение размерности позволило:

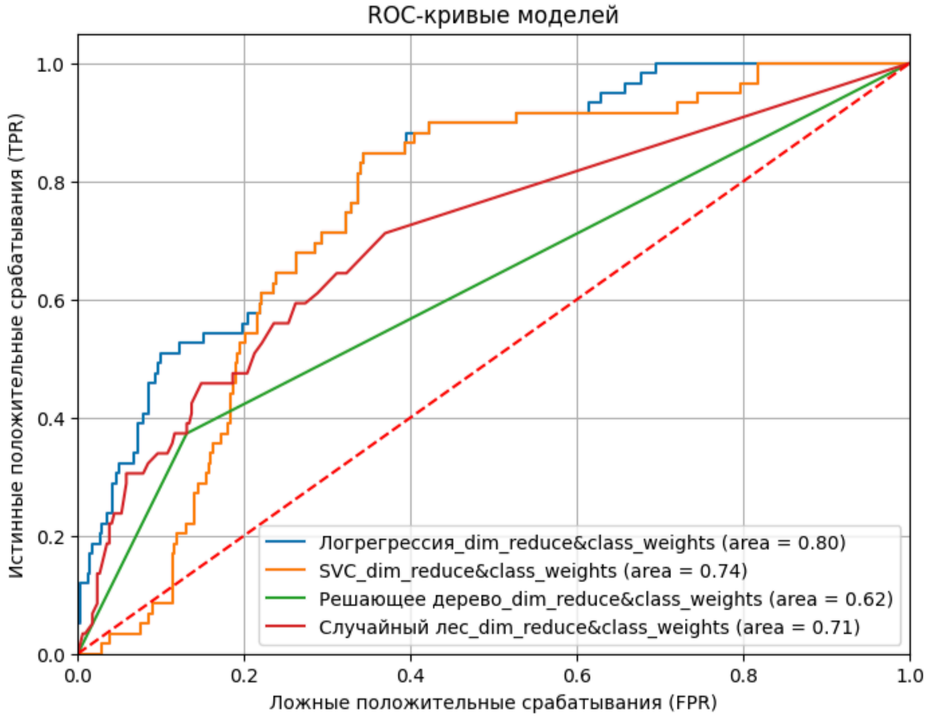
- улучшить качество моделей KNN и наивного байеса,
- слабо повлияло на модель логистической регрессии
- ухудшило предсказательную способность моделей SVC, градиентного бустинга, решающего дерева и случайного леса.

Основные результаты

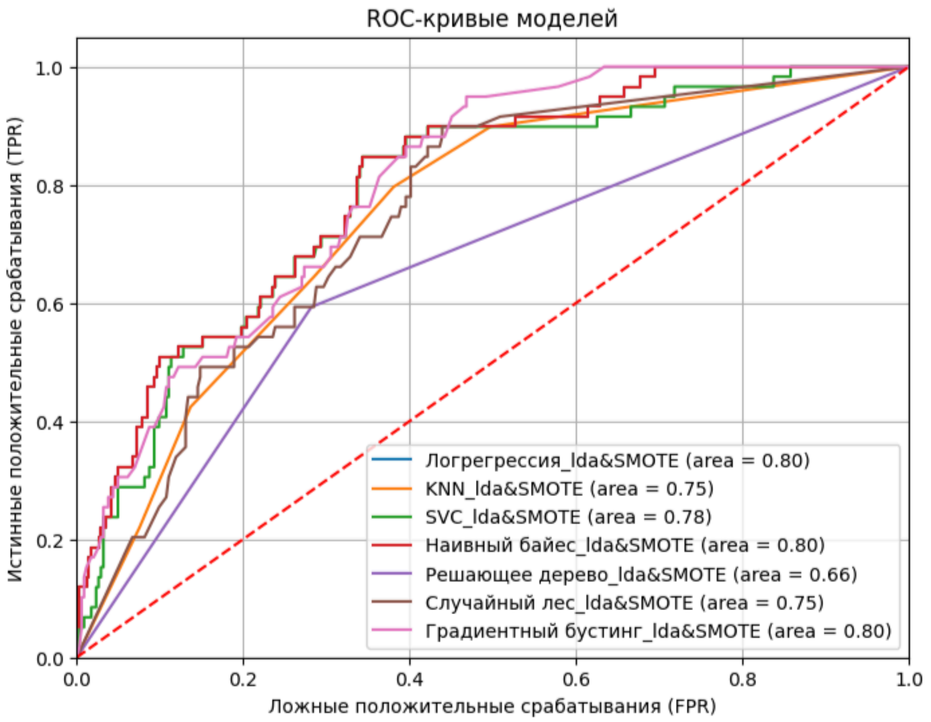
Выбор и обучение модели для определения вероятности покупки (5/9)

Обучение базовых моделей с учетом снижения размерности и с обработкой дисбаланса классов

- LDA + изменение весов классов



- LDA + SMOTE



Применение оверсэмплинга (SMOTE) к данным с пониженной размерностью позволило:

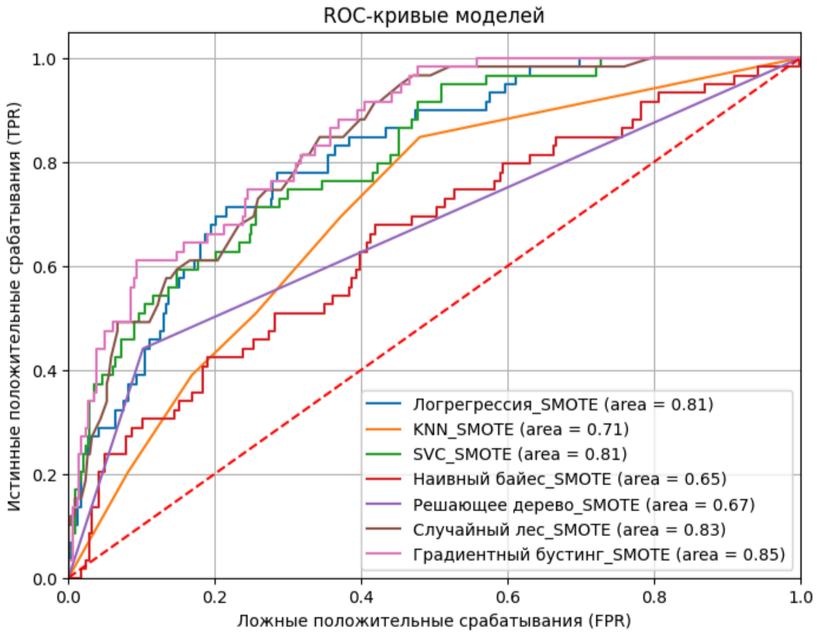
- повысить качество моделей KNN, SVC,
- на модели решающего дерева, случайного леса и логистической регрессии повлияло слабо.

Основные результаты

Выбор и обучение модели для определения вероятности покупки (6/9)

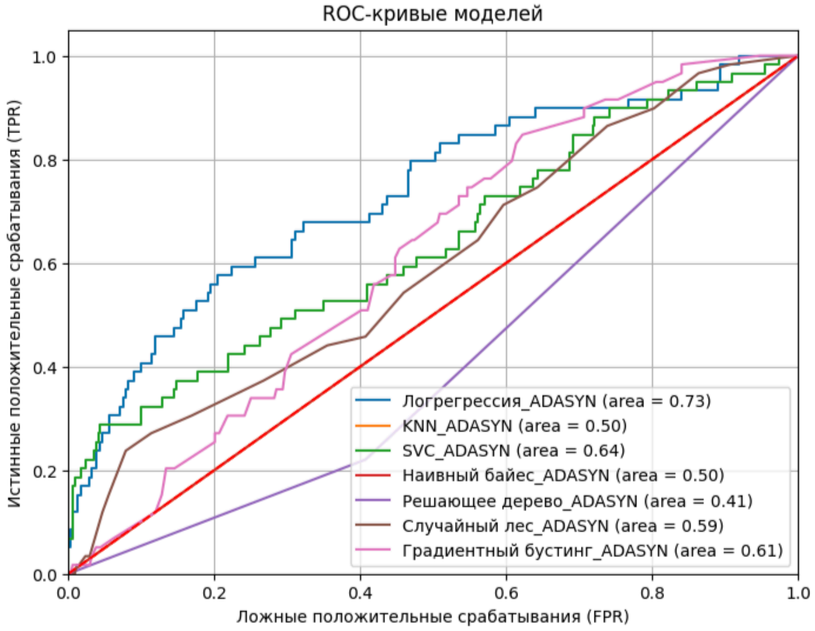
Обучение базовых моделей с с обработкой дисбаланса классов методом оверсэмплинга

- SMOTE



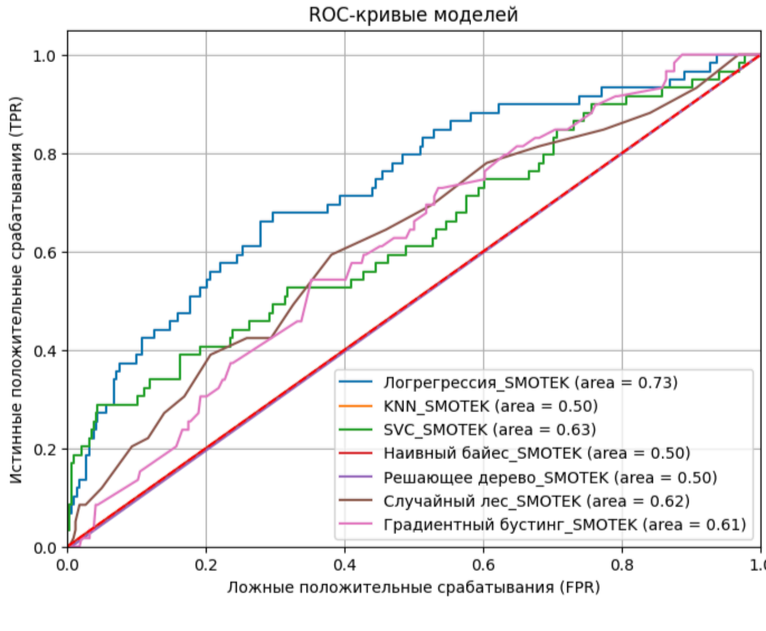
Применение метода оверсемплинга в случае несбалансированности классов в исходном датасете позволило повысить качество работы всех моделей кроме наивного байеса.

- ADASYN



Все модели получились достаточно слабыми, модель решающего дерева выполняет классификацию хуже случайного угадывания. Исходя из полученных данных можно сделать вывод, что ресэмлинг методом ADASYN не подходит для нашего набора данных.

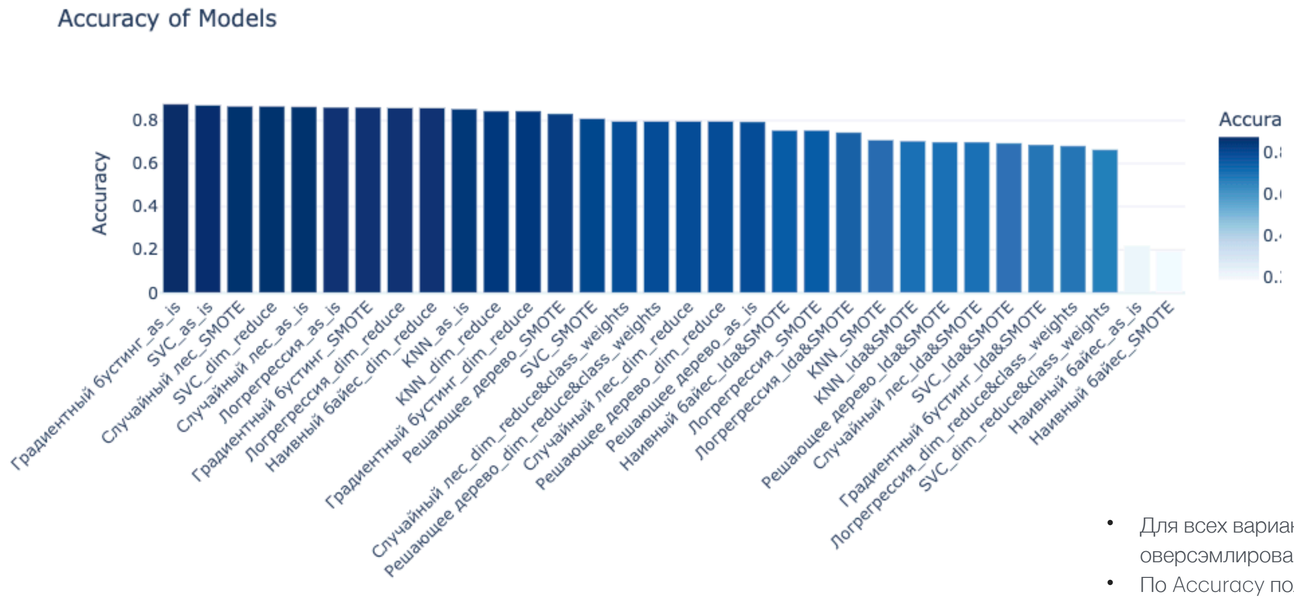
- SMOTEK

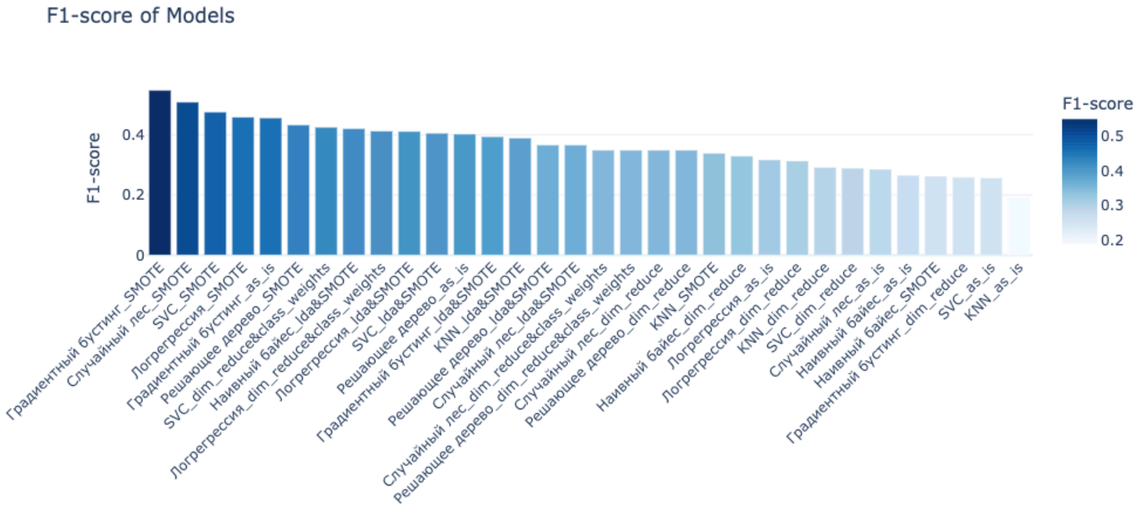


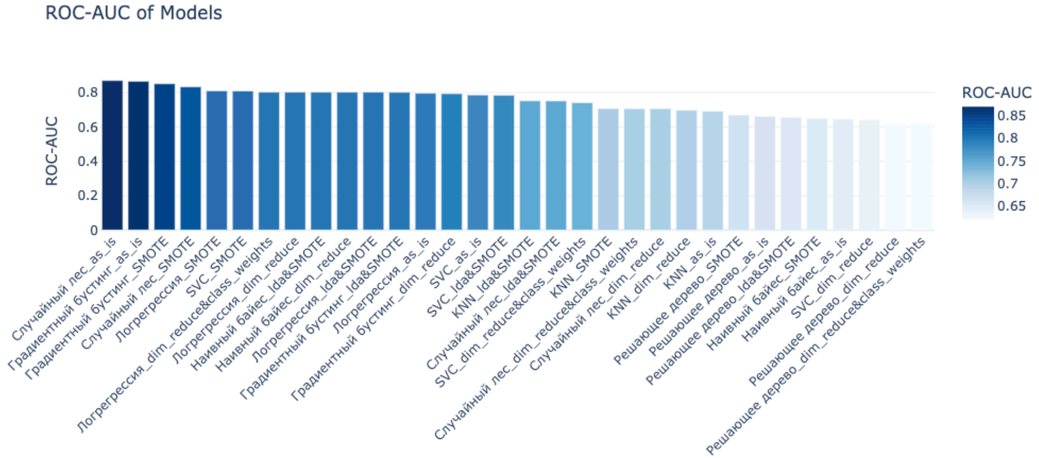
Площадь под кривыми в данном эксперименте не превышает 0.73 (самая высокая у логистической регрессии), остальные модели ближе к линии случайного угадывания.

Основные результаты

Выбор и обучение модели для определения вероятности покупки (7/9)







- Для всех вариантов обученных моделей метрика F1 не превысила 0.55, самое высокое значение данной метрики получилось для оверсэмлированного градиентного бустинга (0.55)
- По Accuracy получились следующие топ-3 модели:
  - Градиентный бустинг (as is) - 0.88
  - SVC (as is) - 0.87
  - Случайный лес (SMOTE) - 0.87
- топ-3 модели по ROC-AUC:
  - Случайный лес (as is) - 0.87
  - Градиентный бустинг (as is) - 0.87
  - Градиентный бустинг (SMOTE) - 0.85
- топ-3 модели по суммарной оценке:
  - Градиентный бустинг (SMOTE) - 0.79
  - Градиентный бустинг (as is) - 0.79
  - Случайный лес (SMOTE) - 0.78

Таким образом, лучше всего из базовых моделей себя показали градиентный бустинг и случайный лес. Наиболее слабыми для нашего набора данных оказались наивный байес, решающее дерево и KNN.

В качестве базовой модели выбралиградиентный бустинг с оверсэмплингом, так как он получил самую высокую суммарную оценку, а также градиентный бустинг показал высокие результаты по Accuracy, F1 и ROC-AUC.

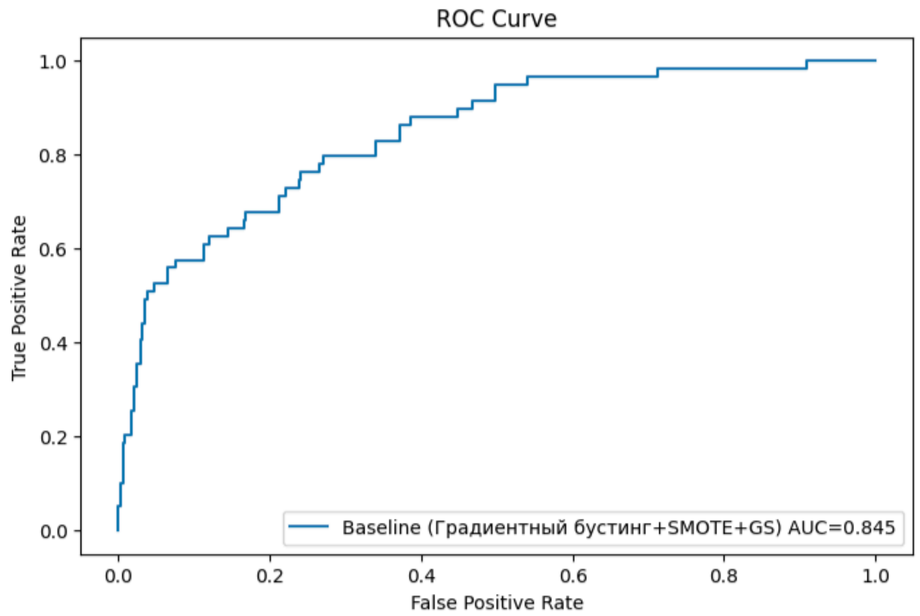
Основные результаты

Выбор и обучение модели для определения вероятности покупки (8/9)

Финальное обучение базовой модели:

Градиентный бусины с оверсэмплингом (base line)

Подбор параметров (Grid Search)

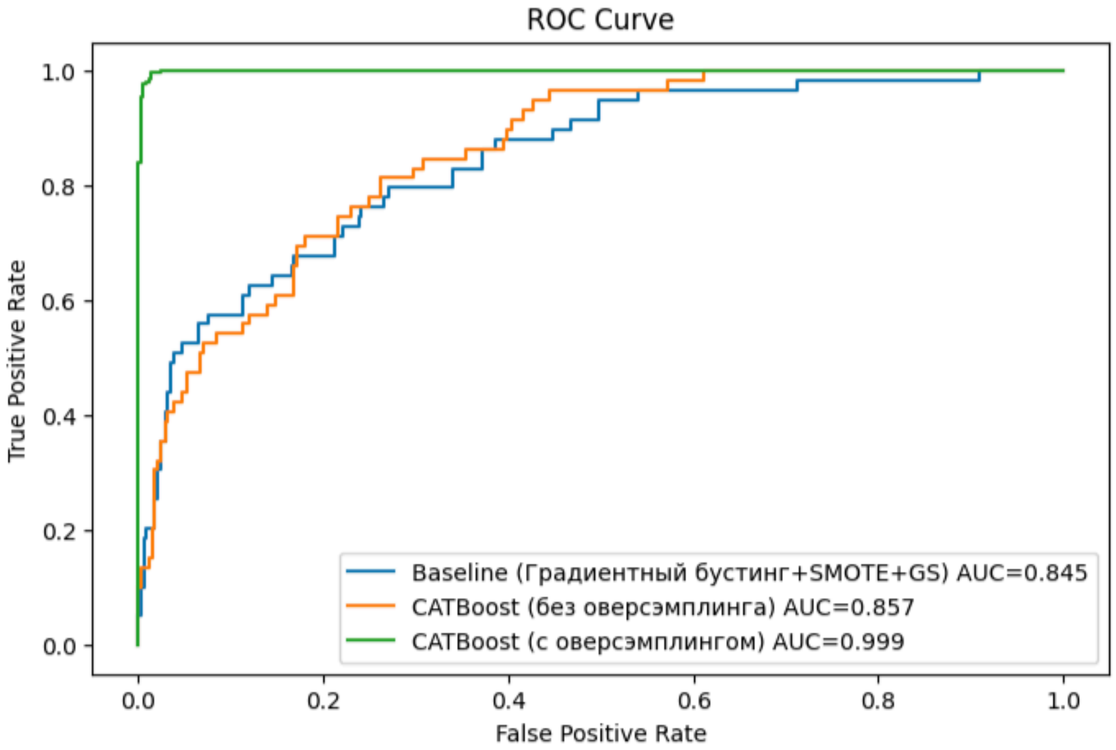


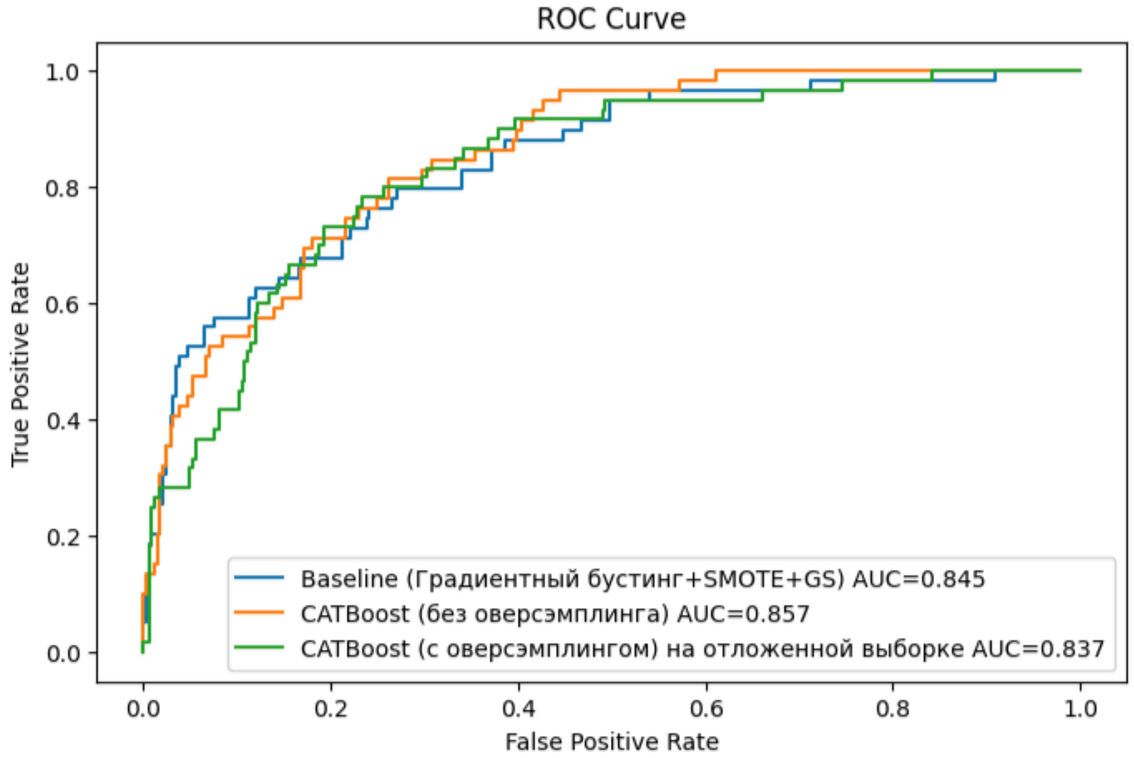
Значимого прироста метрик в результате подбора гиперпараметров не произошло - приняли решение о использовании более продвинутой модели - CATBoost

Основные результаты

Выбор и обучение модели для определения вероятности покупки (9/9)

Обучили CATBoost на данных без оверсэмплинга и с оверсэмплингом, проверили на отложенной выборке:



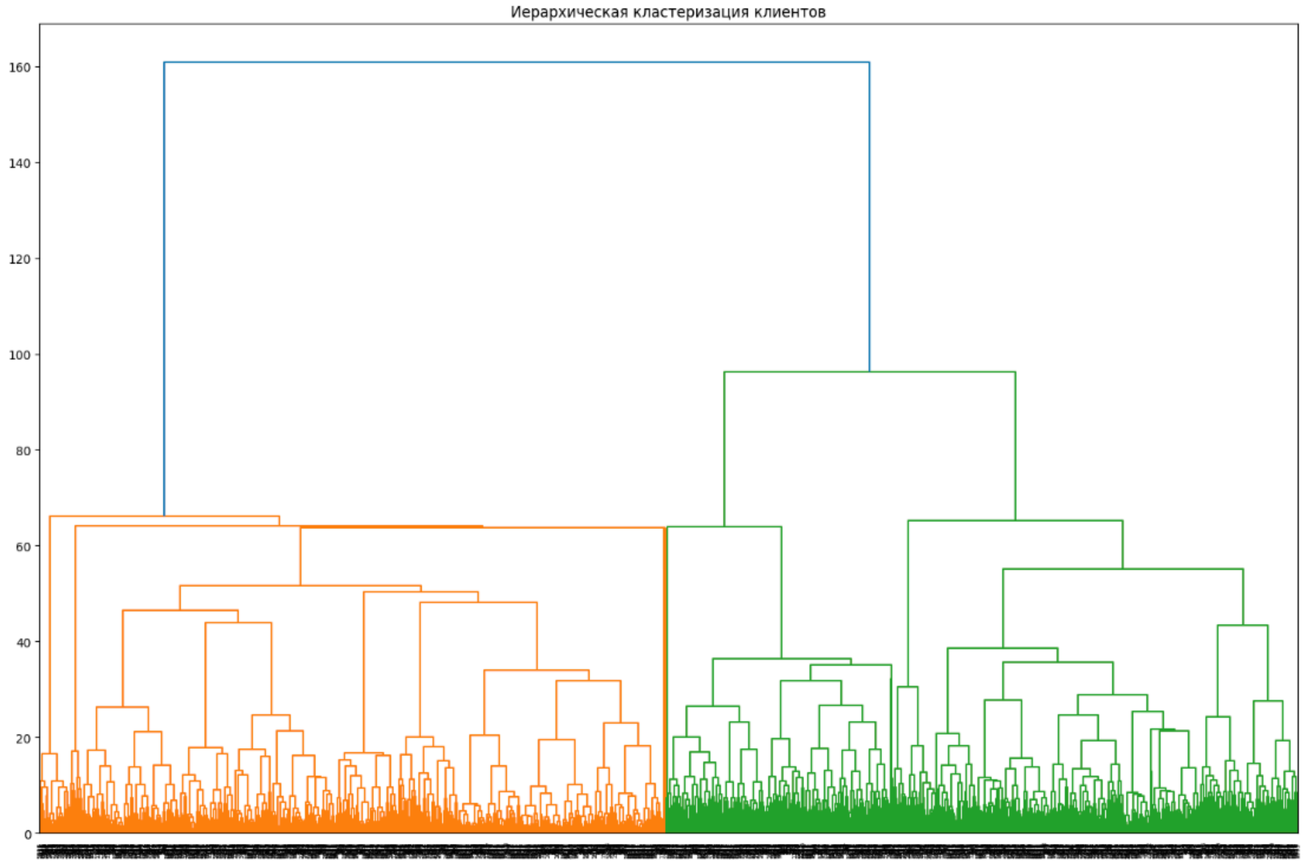


Модель CATBoost с оверсэмплингом, с метриками стремящимися к 1, была переобучена и на отложенной выборке показала результат хуже чем CATBoost без оверсэмплинга.

В целом для финальной модели можно использовать CATBoost без дополнительной обработки датасета по размерности и дисбалансу классов.

Основные результаты

Кластеризация клиентов



- Для определения количества кластеров построили матрицу расстояний на основе нашего датасета с дополнительными признаки и с учетом стандартизации данных. После этого построили дендрограмму для иерархической кластеризации клиентов.
- На дендрограмме получили два кластера.
- Обучили модель на основе алгоритма K-Means, получили метки кластеров. Получили значение метрики силуэта 0.22, что говорит о приемлемом уровне разделения кластеров.
- В результате кластеризации получили 2 класса с распределением 39% - класс 1, 61% - класс 0., что значительно отличается от распределения по целевой переменной (response), где положительный класс составлял 15%, отрицательный - 85%.

Основные результаты

Портреты клиентов

Демографические признаки

- Клиенты из кластера 1 в среднем старше на 4 года клиентов из кластера 0.
- Средний возраст регистрации клиента из кластера 0 - 43 года, из кластера 1 - 46 лет.

Доходы клиентов и уровень образования

- Клиенты из кластера 1 зарабатывают в среднем в 1.8 раза больше, чем клиенты из кластера 0.
- Уровень образования у клиентов из кластера 1 выше чем у клиентов из кластера 0. Так доля клиентов выпускников средней школы / колледжа / бакалавриата на 4% выше в кластере 1. Доля клиентов с докторской степенью в кластере 1 выше на 4%. При этом в структуре клиентов кластера 0 заметная доля клиентов с базовым образованием - около 4%, в кластере 1 таких клиентов 0.1%.

Состав семьи

- У клиентов из кластера 0 чаще бывает один маленький ребенок, у клиентов из кластера 1 чаще нет маленьких детей.
- У клиентов из кластера 0 чаще есть один ребенок-подросток, у клиентов из кластера 1 это встречается реже.
- У клиентов из кластера 0 чаще есть более одного ребенка в семье, у клиентов из кластера 0 чаще нет детей или есть только один ребенок.
- Структура пользователей по семейному положению для разных кластеров похожа, можно отметить, что среди пользователей кластера 1 доля клиентов в браке на 3% ниже, чем у клиентов из кластера 0, а доля разведенных на 1.5% выше.

Активность клиентов

- По среднему количеству дней с момента последней покупки разница между кластерами не существенная - 49 и 50 дней.
- В среднем количество покупок совершенных со скидкой у клиентов из кластера 1 незначительно меньше по сравнению с клиентами из кластера 0 - 2 и 2.5 соответсвенно.
- Клиенты из кластера 1 в среднем в 2 раза чаще совершают покупки на сайте чем клиенты из кластера 0.
- Клиенты из кластера 1 в среднем чаще в 5 раз совершают покупки по каталогу чем клиенты из кластера 0.
- Клиенты из кластера 1 в среднем чаще в 2 раза совершают покупки непосредственно в магазине чем клиенты из кластера 0.
- Клиенты из кластера 1 реже посещают сайт по сравнению с клиентами из кластера 0 - 4 против 6.
- Вероятность совершения покупки среди клиентов, попавших в сегмент 1 выше чем у клиентов из сегмента 0 в 2.3 раза.

Траты клиентов по категориям

- Клиенты из кластера 1 в среднем тратят на вино в 6 раз больше, чем клиенты из кластера 0.
- Средние расходы на фрукты у клиентов из кластера 1 в 8 раз выше, чем к улиентов из кластера 0.
- Средние траты на мясные продукты у клиентов из кластера 1 почти в 10 раз выше, чем у клиентов из кластера 0.
- Средние траты клиентов из кластера 1 на рыбные продукты в 8 раз выше, чем у клиентов из кластера 0.
- По общим средним расходам клиенты из кластера 1 также обгоняют клиентов из кластера 0 примерно в 6 раз.
- У клиентов из кластера 1 средние траты в расчете на одного члена сеьми превышают аналогичные у клиентов из кластера 0 более чем в 10 раз.
- Средний чек клиентов из кластера 1 примерно в три раза выше чем у клиентов из кластера 0.

Клиентов из кластера 1 можно охарактеризовать как активных образованных и состоятельных, тратят значительно больше на все основные товарные категории по сравнению с клиентами из кластера 0. Активно пользуются всеми каналами продаж - сайт, каталог, офлайновый магазин. Клиенты из данного кластера редко имеют более одного ребенка.

Клиенты из кластера 0 менее активны, меньше зарабатывают, меньше тратят и реже совершают покупки. Соответственно ниже средний чек, средние траты на члена семьи. Клиенты из данного кластера часто имеют несколько детей.

Методология

Пайплайн подготовки данных

Загрузка данных

Приведение типов данных

- Дата / время
- Целочисленные признаки
- Вещественные признаки
- Строки
- Бинарные признаки

Обработка пропусков

- «Заглушки»
- Заполнение
- Удаление

Обработка дубликатов

- Явные дубликаты
- Неявные дубликаты

Дообогащение данных

Методология

Исследовательский анализ данных

- **Описательная статистика численных признаков**
  - Меры центральной тенденции
  - Описание распределений
- **Распределение категориальных признаков**
- **Средние значения признаков по целевой переменной**
- **Матрица корреляции признаков**

Методология

Модели машинного обучения

Исходя из задачи исследования требуется выполнить бинарную классификацию клиентов на основе имеющихся признаков - предсказать класс 0 - клиент не совершит покупку, либо 1 - клиент совершит покупку, а также получить показатель вероятности совершения покупки.

Для бинарной классификации подходят следующие алгоритмы:

- **Логистическая регрессия**
  - Линейный алгоритм, используемый для бинарной классификации, который предсказывает вероятность принадлежности объекта к тому или иному классу. Она использует логистическую (сигмоидальную) функцию для преобразования линейной комбинации признаков в значение от 0 до 1, что удобно для интерпретации как вероятностей.
  - Применение: Широко используется для задач, где результаты могут быть классифицированы на два класса
- **KNN (K-Nearest Neighbors)**
  - Основывается на принципе, что объекты, находящиеся близко друг к другу в пространстве признаков, имеют схожие характеристики.
  - Применение: классификация текстов, рекомендательные системы: помогает в создании рекомендаций на основе схожести пользователей или товаров, медицинская диагностика: используется для классификации заболеваний на основе симптомов и медицинских данных, а также применяется для кластеризации и сегментации данных.
- **Метод опорных векторов (SVM) с вероятностной интерпретацией**
  - Основная идея SVM заключается в нахождении гиперплоскости, которая максимально разделяет классы в пространстве признаков.
  - Эффективен для высокоразмерных данных.
  - Может работать с различными ядрами для нелинейной классификации.
  - Градиентный бустинг – это ансамблевый метод, который строит модели последовательно, добавляя новое дерево решений, которое минимизирует ошибку предыдущих моделей. Каждая новая модель обучается на ошибках (остатках) предыдущих, что позволяет улучшать качество предсказания на каждом этапе.
  - Применение: Используется в задачах, требующих высокой точности

Методология

Модели машинного обучения

- **Наивный байесовский классификатор**
  - Этот алгоритм основан на теореме Байеса и предполагает независимость признаков.
  - Быстрый и эффективный, особенно для текстовых данных.
  - Хорошо работает с категориальными признаками.
- **Дерево принятия решений для классификации**
  - Дерево решений – это алгоритм, который строит модель предсказаний в виде дерева, где каждый узел представляет собой условный тест на признак, а каждый лист представляет собой класс. Алгоритм работает путем последовательного деления данных на подмножества, основываясь на значениях признаков, что делает его интерпретируемым и наглядным.
  - Применение: Используется в задачах, где важна интерпретируемость, и может быть применён в медицине, финансах и анализе данных.
- **Случайный лес для классификации (ансамбль)**
  - Случайный лес – это ансамблевый метод, который строит множество деревьев решений на случайных подвыборках данных и усредняет предсказания для повышения точности и уменьшения переобучения. Каждое отдельное дерево обучается на случайной подвыборке обучающих данных с использованием случайного подмножества признаков для каждой разделяющей точки.
  - Применение: Широко используется для более сложных задач классификации и регрессионного анализа, особенно когда есть много признаков и сложные зависимости между ними.
- **Градиентный бустинг для классификации (ансамбль)**
  - Градиентный бустинг – это ансамблевый метод, который строит модели последовательно, добавляя новое дерево решений, которое минимизирует ошибку предыдущих моделей. Каждая новая модель обучается на ошибках (остатках) предыдущих, что позволяет улучшать качество предсказания на каждом этапе.
  - Применение: Используется в задачах, требующих высокой точности

Для выбора бэйзлайна обучили каждую из перечисленных моделей и определили наиболее эффективную для дальнейшей работы с ней, а также для сравнения с результатами работы более сложных моделей.

Методология

Метрики качества моделей

Для задачи оценки качества модели классификации могут применяться следующие метрики:

Матрица ошибок

- Матрица ошибок отражает количество наблюдений в каждой группе (TN, FP, FN, TP)
- У хорошей модели бóльшая часть прогнозов должна попадать в группы TP и TN.

Доля правильных ответов (acuracy)

- Это доля верно угаданных ответов из всех прогнозов.
  Чем ближе значение accuracy к 100%, тем лучше

$$Accuracy = \frac{TP + TN}{n}$$

Точность (precision)

- Precision показывает долю правильных ответов только среди целевого класса
- В бизнесе метрика precision нужна, если каждое срабатывание (англ. alert) модели — факт отнесения к классу "1" — стоит ресурсов.

$$Precision = \frac{TP}{TP + FP}$$

Полнота (recall)

- Показывает, сколько реальных объектов "1" класса вы смогли обнаружить с помощью модели.
- Эта метрика полезна при диагностике заболеваний: лучше отправить пациента на повторное обследование и узнать, что тревога была ложной, чем прозевать настоящий диагноз

$$Recall = \frac{TP}{TP + FN}$$

F1-score

- Сводная метрика, учитывающая баланс между precision и recall

$$F_1 = \frac{2 * precision * recall}{precission + recall}$$

Методология

Метрики качества моделей

Площадь под кривой (ROC AUC)

- ROC-кривая — это график, который отображает соотношение между True Positive Rate и False Positive Rate
- ROC-кривая строится путем изменения порога классификации и вычисления TPR и FPR для каждого порога. Это позволяет увидеть, как меняется качество классификации при различных значениях порог
- AUC — это площадь под ROC-кривой. Она принимает значения от 0 до 1:
  - AUC = 0.5: Модель не лучше случайного угадывания. Это означает, что модель не может различить положительные и отрицательные классы.
  - AUC < 0.5: Модель работает хуже случайного угадывания, что может указывать на проблемы с данными или моделью.
  - AUC = 1: Модель идеально различает положительные и отрицательные классы.
- Хорошо подходит для несбалансированных классов
- AUC позволяет сравнивать различные модели, независимо от их порогов

При оценке качества моделей в данной работе использованы следующие метрики:

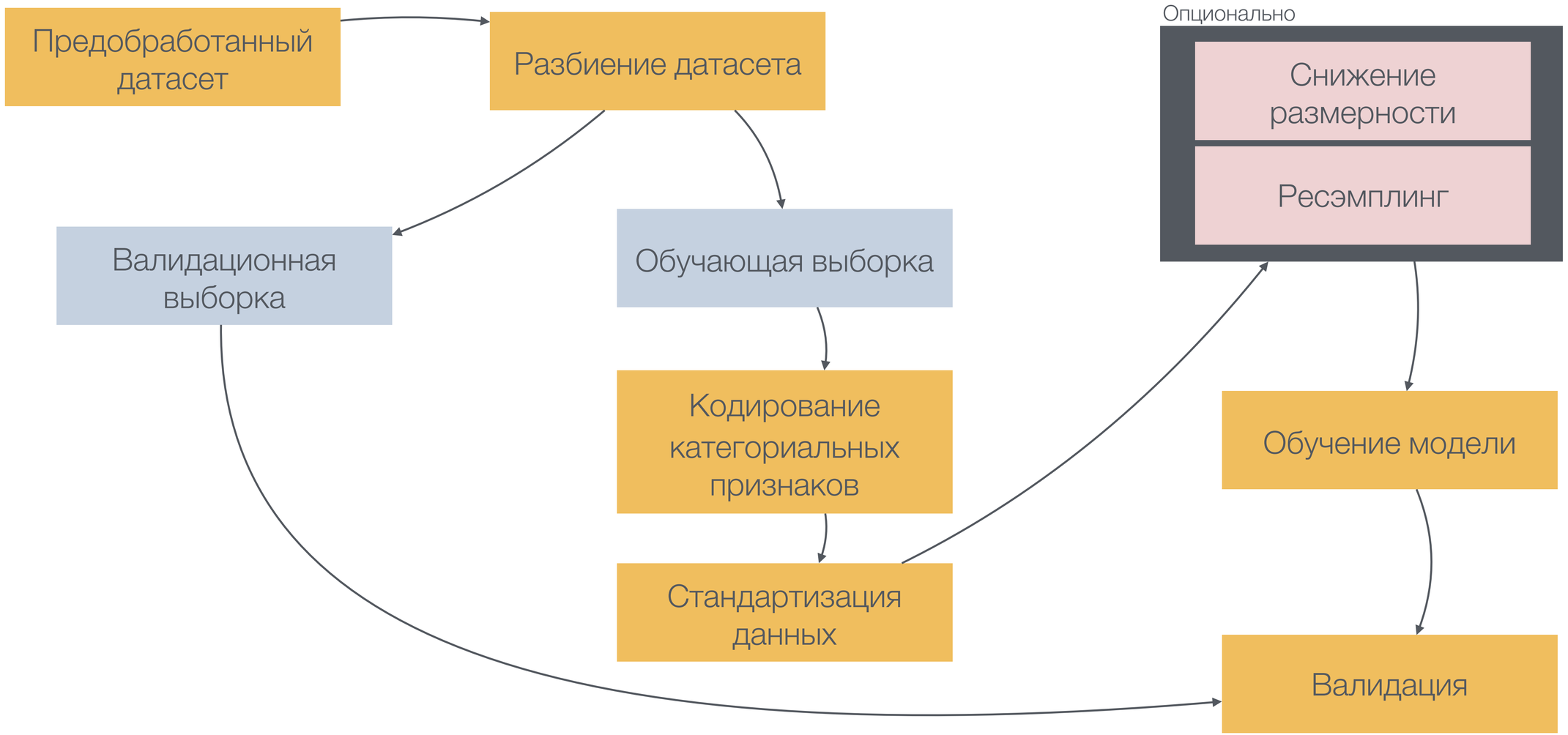
- Accuracy
- F-1
- ROC-AUC

Для получения агрегированной метрики с использованием весовых коэффициентов использована интегральная оценка (IS - Integral Score) по всем метрикам:

$$IS = 0.5 * (ROC - AUC) + 0.3 * Accuracy + 0.2 * F1$$

Методология

Пайплайн обучения модели



Методология

Снижение размерности данных

Снижение размерности данных — процесс преобразования данных с высокой размерностью в данные с более низкой размерностью, сохраняя при этом как можно больше информации. Снижение размерности дает следующие преимущества:

- Устранение избыточности:
  - В данных может быть много признаков, которые содержат избыточную или коррелированную информацию. Снижение размерности помогает устранить такие избыточные признаки.
- Улучшение производительности моделей:
  - Модели машинного обучения могут работать медленнее и менее эффективно с высокоразмерными данными. Снижение размерности может ускорить обучение и улучшить производительность модели, так как уменьшает количество вычислений.
- Снижение риска переобучения:
  - Высокая размерность может привести к переобучению модели, когда она слишком хорошо подстраивается под обучающие данные и плохо обобщает на новых данных. Снижение размерности помогает уменьшить сложность модели и улучшить ее обобщающую способность.
- Улучшение интерпретируемости:
  - Модели с меньшим количеством признаков легче интерпретировать. Это может быть важно в областях, где объяснение результатов модели имеет значение.

В задачах классификации один из рекомендуемых методов снижения размерности - Линейный дискриминантный анализ (LDA). LDA позволяет снижать размерность и учитывает классы - ищет линейные комбинации признаков, которые максимизируют разделение между классами.

Методология

Обработка дисбаланс классов

- **Изменение весов классов** — метод, который позволяет модели уделять больше внимания недопредставленным классам при обучении

- **Oversampling** - это метод, который позволяет увеличить количество примеров для недопредставленного класса в наборе данных
    - **SMOTE** (Synthetic Minority Over-sampling Technique), который создает синтетические примеры для недопредставленного класса
    - **ADASYN** (Adaptive Synthetic Sampling) — это метод, используемый для обработки дисбаланса классов в задачах классификации. Он создает синтетические примеры для меньшинственного кла
    - **SMOTEK** от SMOTE заключается в способе выбора соседей для генерации синтетических примеров. SMOTEK рассматривает только соседей внутри миноритарного класса, в то время как SMOTE рассматривает всех соседей миноритарного класса. Это делает SMOTEK более чувствительным к локальной структуре данных миноритарного класса и может привести к созданию более реалистичных синтетических примеров, особенно когда миноритарный класс имеет сложную форму или состоит из нескольких кластеров.

Методология

Кластеризация

- **K-Means**
  - Описание: Разбивает данные на k кластеров, минимизируя сумму квадратов расстояний от каждой точки до центра её кластера.
  - Плюсы: Простой, быстрый, хорошо масштабируется на большие наборы данных.
  - Минусы: Требует задать количество кластеров (k) заранее, чувствителен к выбросам и начальному расположению центроидов, предполагает сферическую форму кластеров.

- **Агломеративная иерархическая кластеризация**
  - Описание: Строит иерархию кластеров, объединяя или разделяя их на основе расстояния между ними. Может быть агломеративной (объединяющей) или дивизимной (разделяющей).
  - Плюсы: Позволяет визуализировать иерархию кластеров с помощью дендрограммы, не требует заранее задавать количество кластеров (хотя нужно определить критерий остановки).
  - Минусы: Вычислительно затратен для больших наборов данных.

Сначала строим дендрограмму на основе матрицы расстояний полученной методом `linkage()`, а затем обучаем модель кластеризации на основании алгоритма K-means.

Метод `linkage()` используется в библиотеке SciPy для иерархической кластеризации данных. Он позволяет объединять объекты в кластеры на основе расстояний между ними. Метод принимает в качестве входных данных матрицу расстояний или набор данных и возвращает иерархическую структуру, которая может быть использована для построения дендрограммы.

Дальнейшее развитие проекта

- Целевым сегментом для компании являются активные обеспеченные клиенты, необходимо фокусироваться на их удержании и поддержании активности за счет формирования уникальных предложений, премиального обслуживания, специальных акций. Также необходимо привлекать новых клиентов, которые по характеристикам соответствуют данному сегменту.

- По клиентам из менее активного сегмента необходимо повышать их активность за счет предложения скидок, специальных акций, возможно сделать акцент на предложениях семейного формата и товары для детей и подростков.