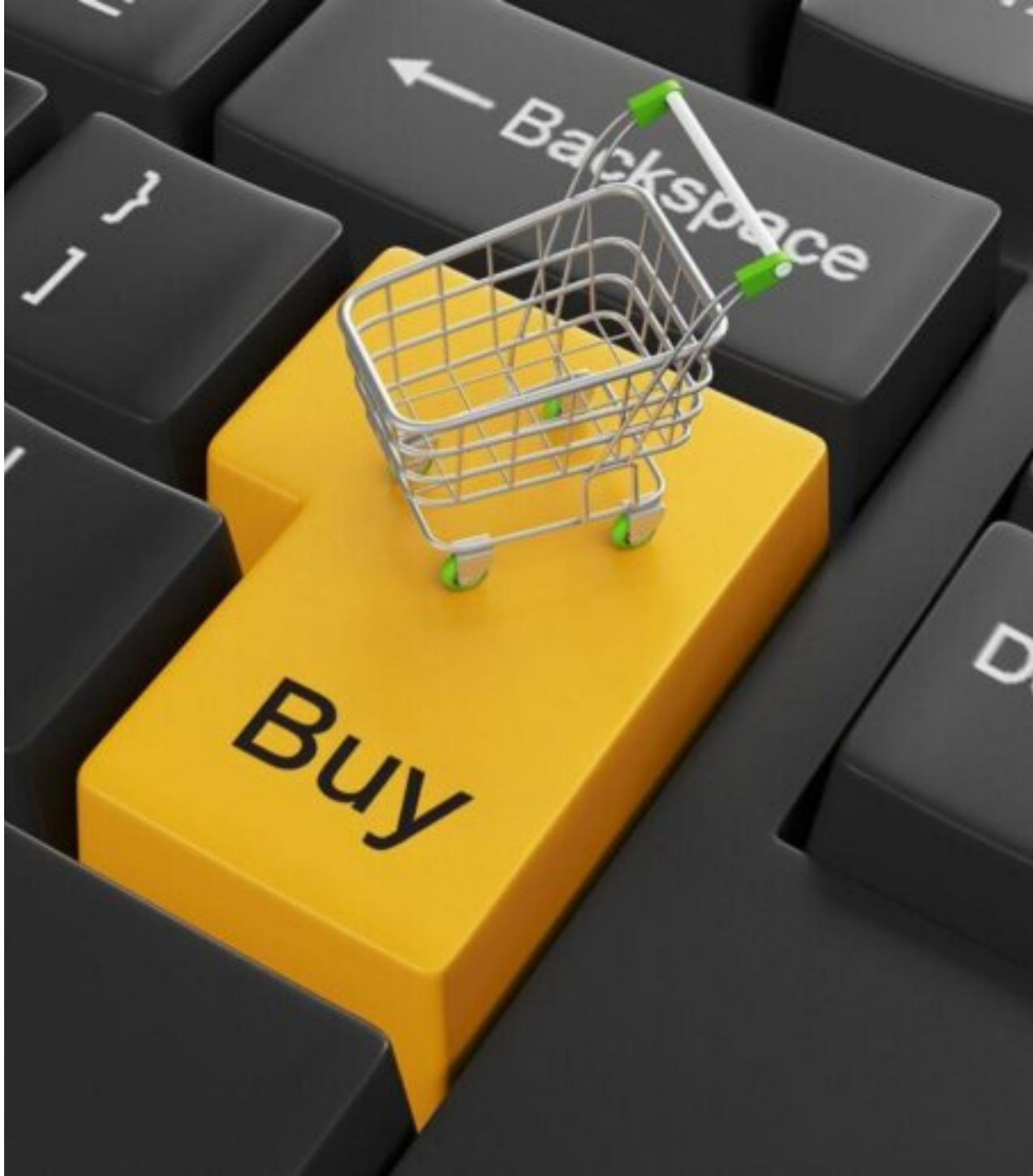
Определение вероятности покупки товара на основе данных о клиенте и его покупательской истории с помощью методов машинного обучения

Итоговый проект по программе
«Специалист по Data Science»

Моренко Антон
DS-16

Август 2025 г.



Содержание

- Описание проекта
- Основные результаты
- Методология работы
- Дальнейшее развитие проекта

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

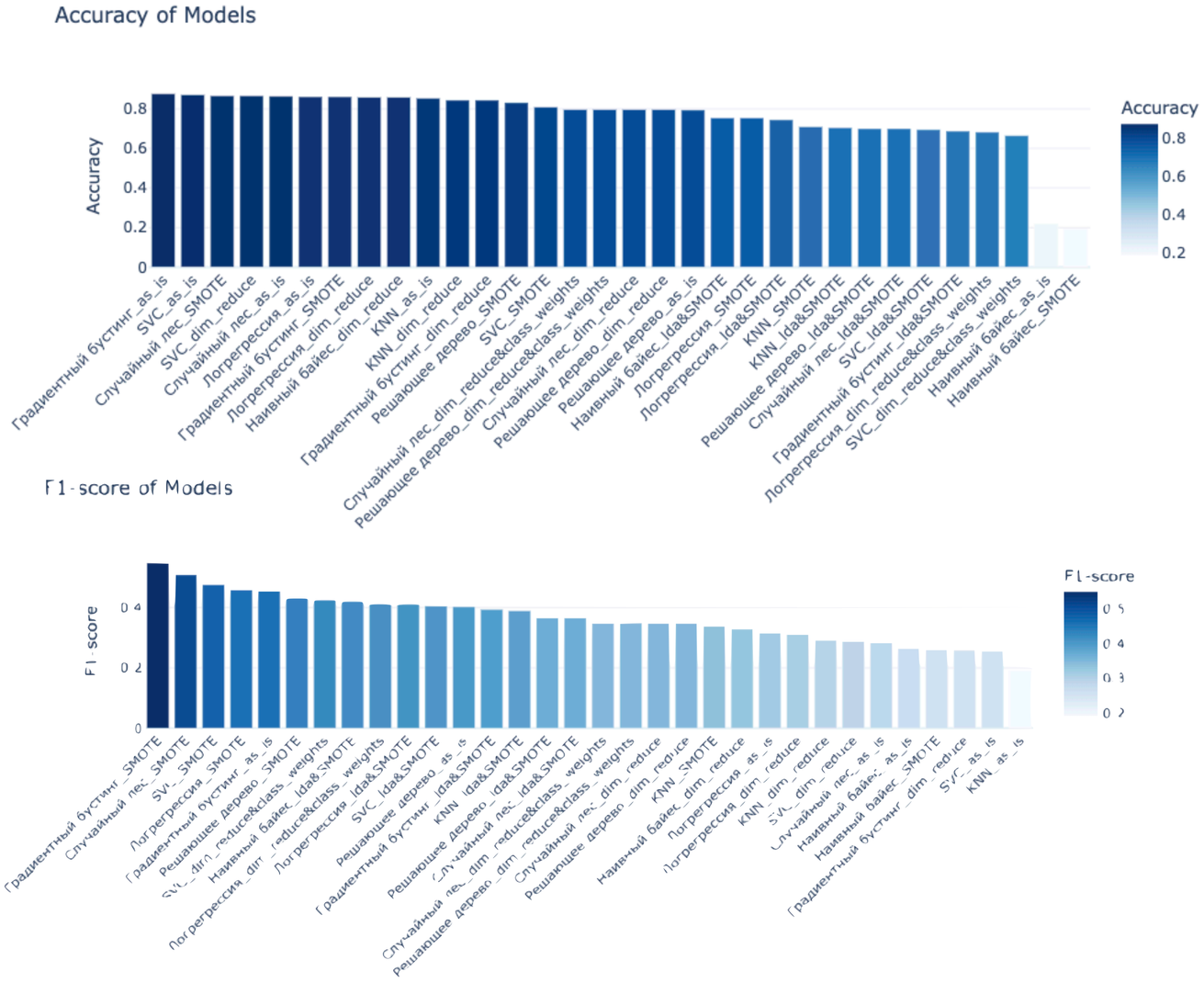
3. Провести кластеризацию покупателей

4. Выбрать и обучить модель определения вероятности покупки товара

Основные результаты

Основные результаты

Выбор и обучение модели для определения вероятности покупки (1/3)

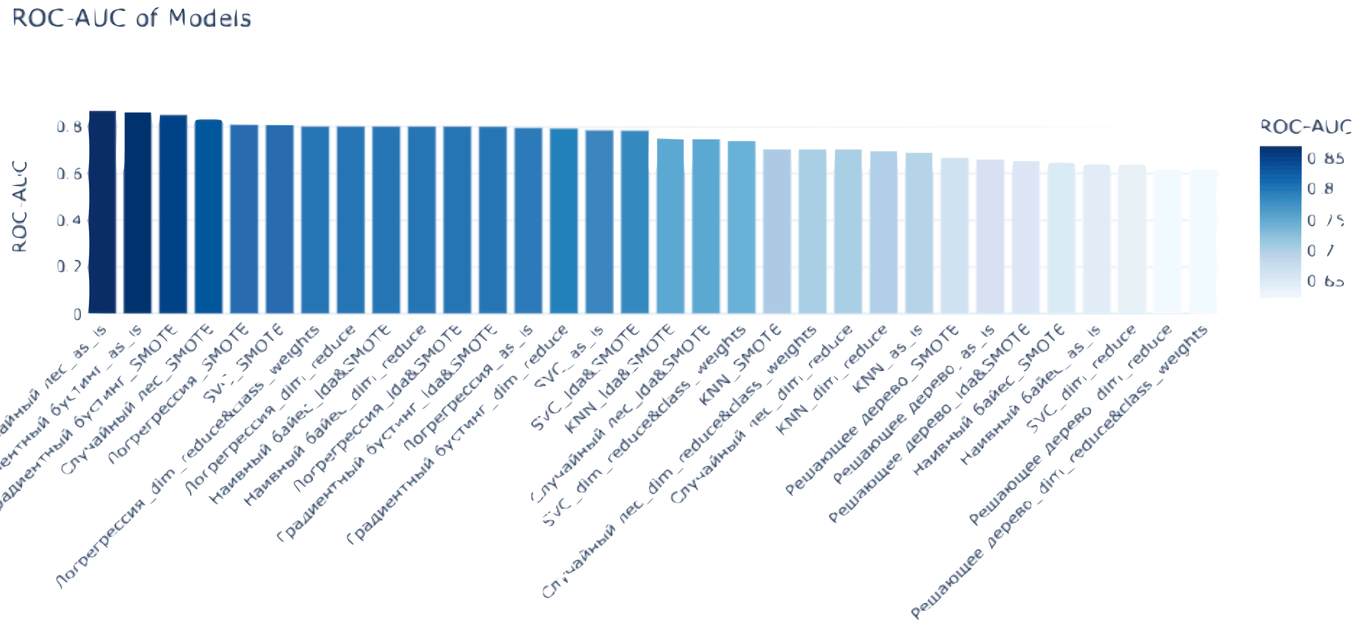


Для подбора базовой модели проведено 32 эксперимента:

- Обучение на данных «как есть»
- Обучение со снижением размерности
- Обучение со снижением размерности и обработкой дисбаланс классов
- Обучение с обработкой дисбаланс классов

Основные результаты

Выбор и обучение модели для определения вероятности покупки (2/3)



- Для всех вариантов обученных моделей метрика F1 не превысила 0.55, самое высокое значение данной метрики получилось для оверсэмлированного градиентного бустинга (0.55)

- **топ-3 модели по Accuracy:**
  - Градиентый 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
  - Градиентый бустинг (as is) - 0.87
  - Градиентый бустинг (SMOTE) - 0.85
- **топ-3 модели по суммарной оценке:**
  - Градиентый бустинг (SMOTE) - 0.79
  - Градиентый бустинг (as is) - 0.79
  - Случайный лес (SMOTE) - 0.78

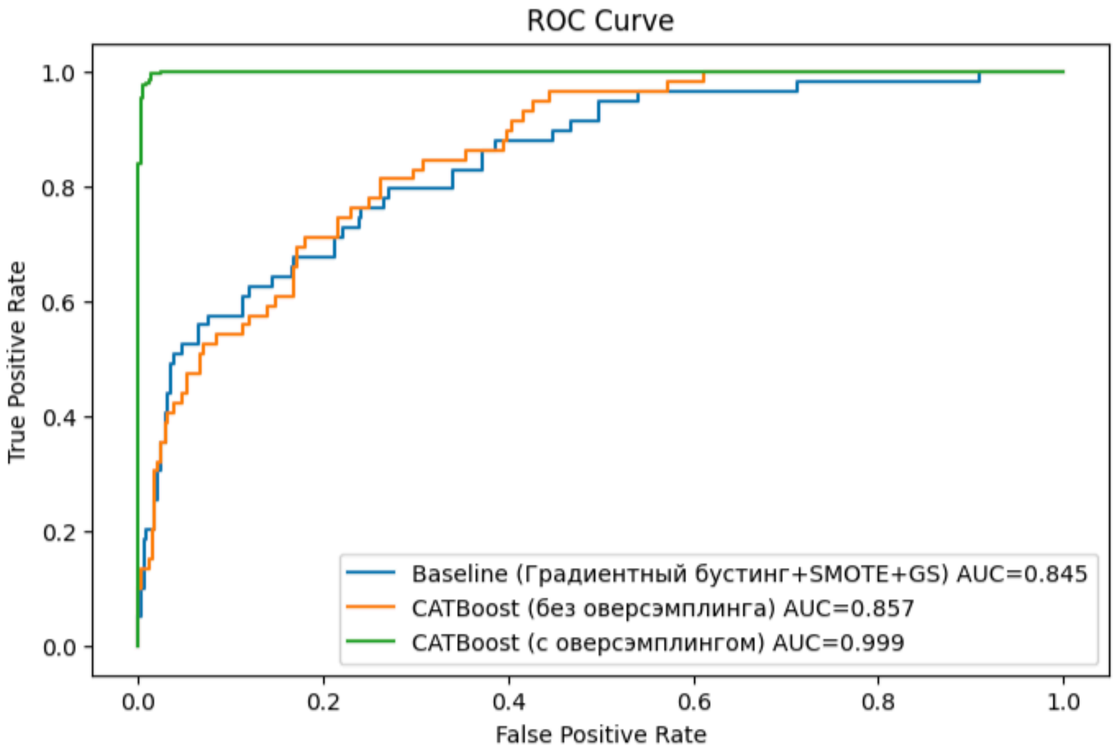
Таким образом, лучше всего из базовых моделей себя показали градиентный бустинг и случайный лес. Наиболее слабыми для нашего набора данных оказались наивный байес, решающее дерево и KNN.

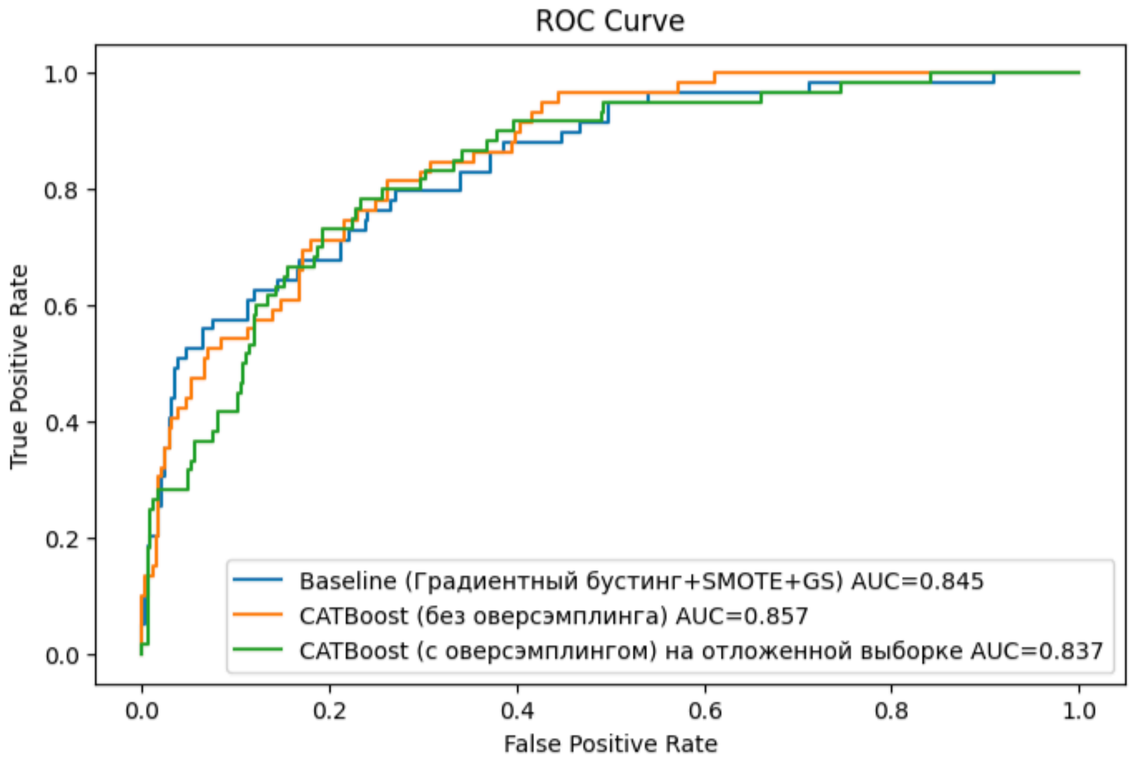
В качестве базовой модели выбрали градиентный бустинг с оверсэмплингом, так как он получил самую высокую суммарную оценку, а также градиентый бустинг показал высокие результаты по Accuracy, F1 и ROC-AUC.

Основные результаты

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

Обучили альтернативную модель на базе CATBoost на данных без оверсэмплинга и с оверсэмплингом, проверили на отложенной выборке и сравнили с базовой моделью:

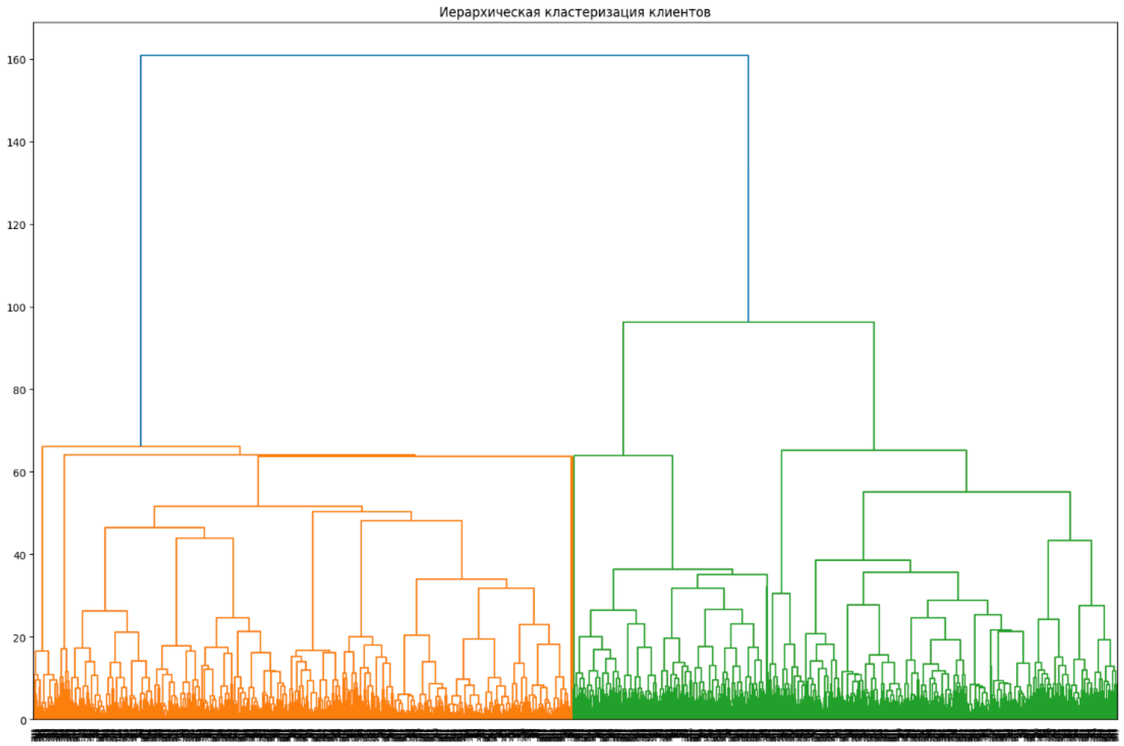




Для финальной модели используем CATBoost без дополнительной обработки датасета по размерности и дисбалансу классов.

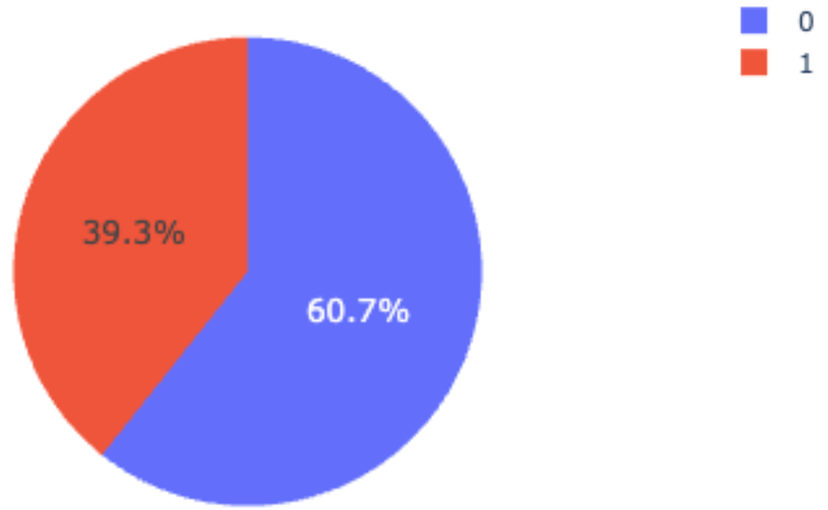
Основные результаты

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.

Распределение полученных 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 (1/2)

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.

Состав семьи

- 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

- 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.**

Основные результаты

Портреты клиентов (2/2)

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 семьи 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 работы

Методология работы

Основные этапы исследования

1. Сбор данных:

Данные о клиентах: Год рождения клиента, уровень образования, состав семьи, семейное положение, история покупок в различных товарных категориях, история покупок по разным каналам продаж, жалобы, метка совершения / несовершенная покупки

2. Предобработка данных:

- Очистка данных: Обработка пропущенных значений, удаление дубликатов.
- Трансформация данных: Нормализация/стандартизация, преобразование категориальных признаков (One-Hot Encoding), снижение размерности, ресэмплинг
- Feature Engineering: Создание новых информативных признаков (средний чек, частота покупок).

3. Выбор модели:

- 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)
- CATBoost

4. Обучение и оценка модели:

- Разделение данных на обучающую и тестовую выборки.
- Обучение выбранной модели на обучающей выборке.
- Оценка качества модели на тестовой выборке с использованием метрик, таких как AUC-ROC, precision, recall, F1-score.
- Проверка метрик на отложенной выборке

Методология

Метрики качества моделей (1/2)

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 (2/2)

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:

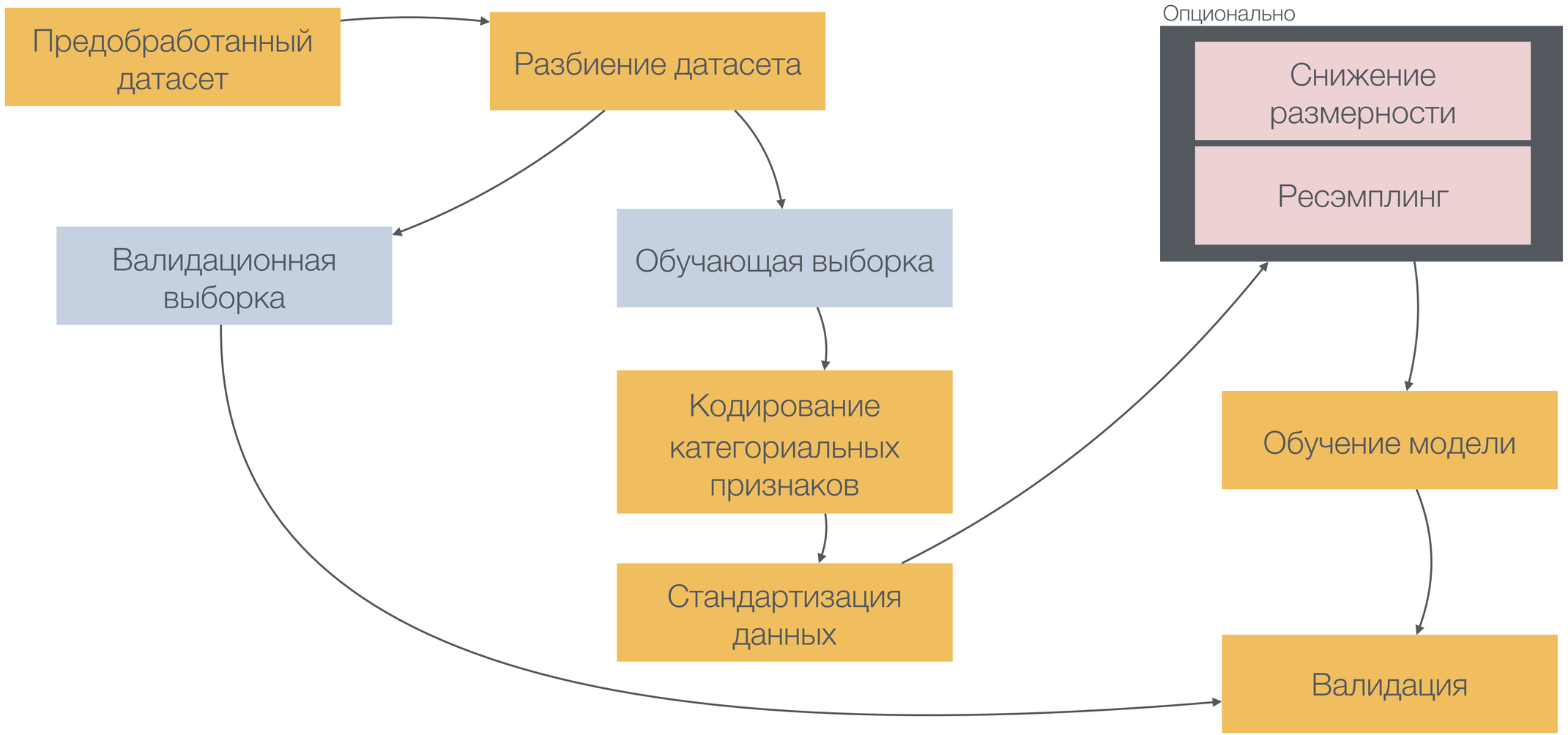
- Accuracy
- F-1
- ROC-AUC

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



Дальнейшее развитие проекта

Дальнейшее развитие проекта (1/2)

Совершенствование предсказательной способности модели:

1. Работа с данными:

- **Feature Engineering:** Самый эффективный способ улучшить модель. Попробуйте создать новые признаки, которые могут быть информативными для предсказания вероятности покупки. Примеры:
  - Соотношение покупок разных категорий товаров.
  - Взаимодействие с маркетинговыми активностями (открытия писем, клики по ссылкам).

- Собрать больше данных - чем больше данных, тем лучше модель может обобщаться.

2. Тюнинг гиперпараметров CatBoost:

- **Bayesian Optimization**: Более продвинутый метод оптимизации гиперпараметров, который может быть эффективнее Grid Search/Randomized Search.
- **Использование кросс-валидации при тюнинге гиперпараметров**, чтобы получить более надежную оценку производительности модели.

3. Ансамблирование моделей:

- **Stacking/Blending**: Объединение предсказаний CatBoost с предсказаниями других моделей (например, LightGBM, XGBoost, логистическая регрессия).
- **Bagging:** Обучение нескольких CatBoost моделей на разных подвыборках данных и усреднение их предсказаний.

5. Анализ важности признаков:

- **Feature Importance**: Анализ важности признаков, чтобы понять, какие признаки наиболее важны для модели.

Улучшение модели – итеративный процесс, требующий экспериментов с разными подходами

Дальнейшее развитие проекта (1/2)

Встраивание модели в бизнес-процесс:

- Таргетированное взаимодействия с клиентами на основе аналитических данных.
- Оптимизация маркетинговых затрат за счет фокусировки на клиентах с высоким уровнем вероятной конверсии

Ожидаемые результаты:

- Модель предсказывающая вероятность совершения покупки клиентов
- Наличие такой модели позволит таргетированно работать с клиентами, предлагать специальные акции, скидки и пр. промо активности.

- 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.