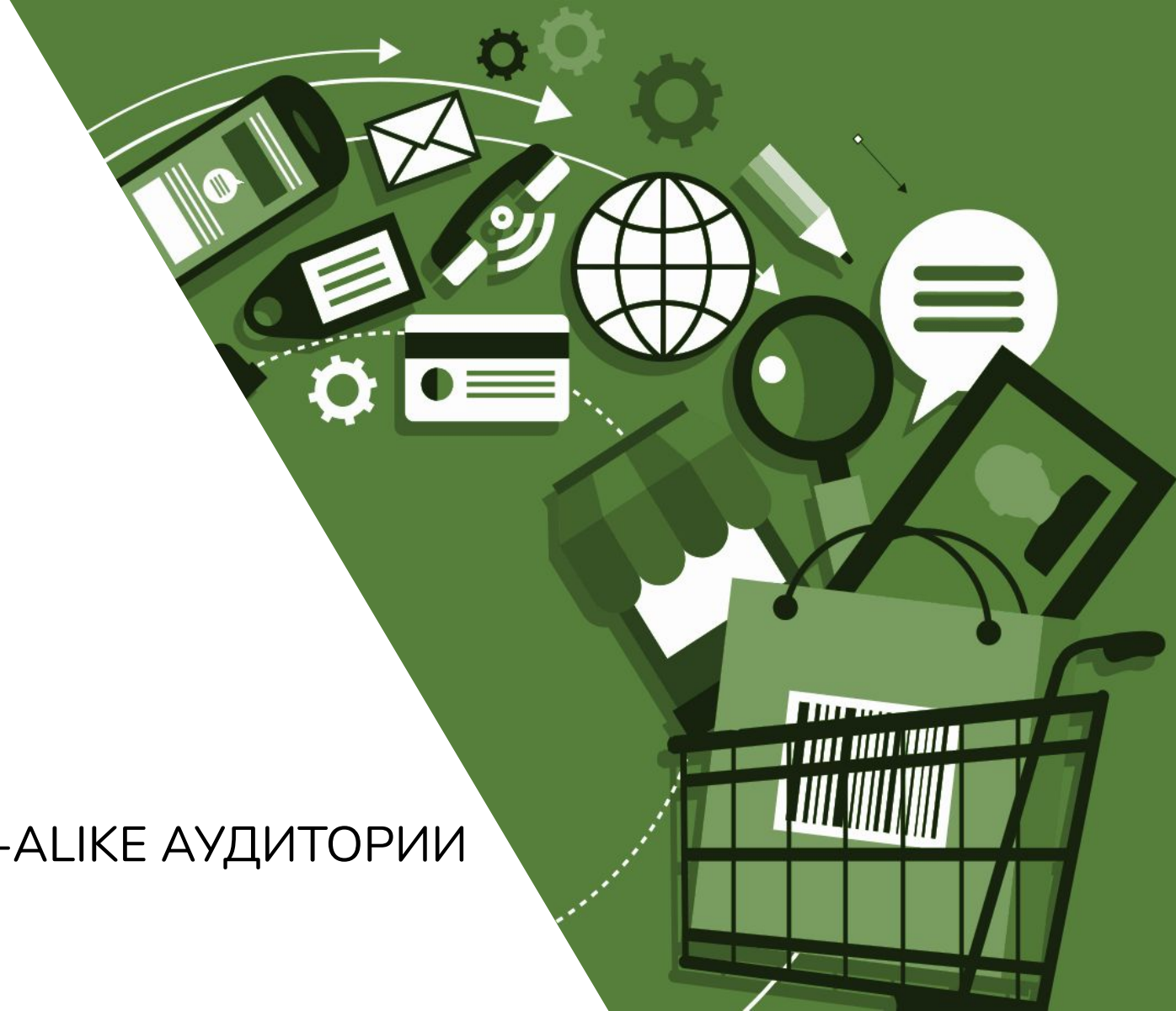


Суп IT 2022

ПОИСК LOOK-ALIKE АУДИТОРИИ



Резюме

Задачи:

🎯 Провести более глубокий анализ данных. Посмотреть по каким признакам различаются клиенты, вступившие в клуб от тех, которые не вступили. Включить внешние данные. На основе этой информации построить новые фичи для улучшения качества модели.

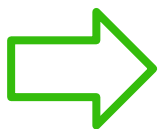
🎯 Попробовать различные state-of-the-art классификаторы. Выбрать лучшие и объединить их предсказания для финальной модели.

Итоговая модель: CatBoost Classifier + LightGBM

$$Final_score = \gamma * CatBoost + (1 - \gamma) * LightGBM$$

Параметр γ подбираем на валидационной выборке так, чтобы максимизировать F-score.

F1-score val: 0.7
F1-score test: 0.64



F1-score val: 0.71
F1-score test: 0.7



LightGBM

Основные параметры:



bagging_fraction	0.9
lambda_l1	7.0
lambda_l2	3.0
learning_rate	0.025
min_data_in_leaf	3700
num_leaves	106

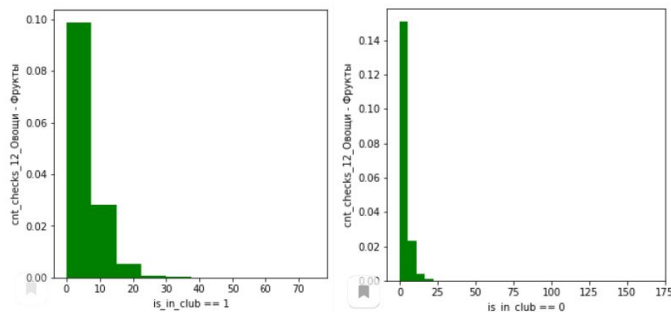
CatBoost

Основные параметры:



boosting_type	Ordered
depth	3
l2_leaf_reg	4.758
learning_rate	0.056
max_ctr_complexity	6
n_estimators	306

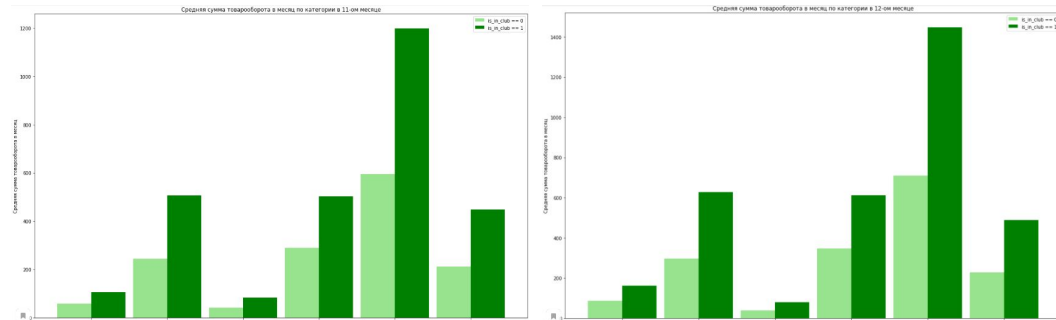
Анализ данных и Feature engineering



Клиенты клуба чаще продолжают покупать в полезных категориях



Бинарный признак: Больше ли число покупок, чем пороговое значение (Достаточно ли покупок)



Клиенты клуба в среднем покупают на большие суммы



Бинарный признак: Больше ли сумма покупок за месяц в целом или по категориям, чем 80 перцентиль



Из второго графика можно сделать вывод, что средняя сумма по всем категориям растет. Это просто объясняется наступающим Новым Годом. Следовательно, последний месяц наименее показательный и мы можем далее давать меньший вес таким признакам. Кроме того, средний товарооборот в 'Фрукты-Овощи' имеет больший перевес у клиентов клуба.

Работа с выбросами

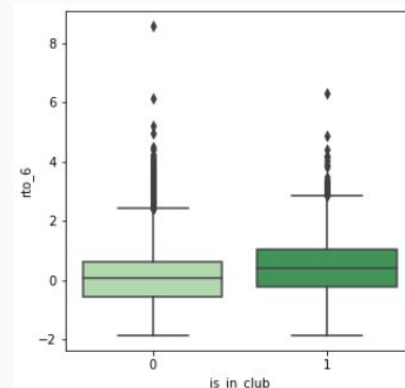
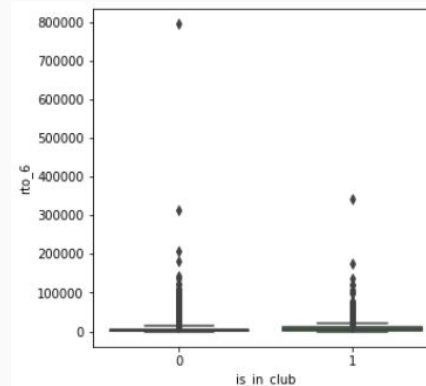


Стандартная обработка выбросов при использовании метода IQR или Z-score в нашем случае выявляют более 50% выбросов в некоторых переменных.

Такая обработка приведет к потере большей части данных.



Power Transformer преобразовывает числовые переменные для получения более гауссовского распределения вероятностей, выбросы становятся не такими критичными.



Избавление от асимметрии в данных



Борьба с выбросами

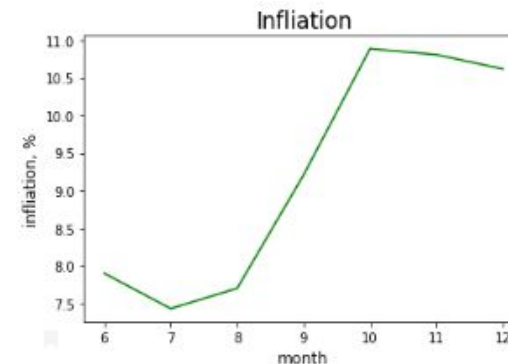


Многие Алгоритмы ML работают лучше, когда переменные близки к гауссовскому распределению

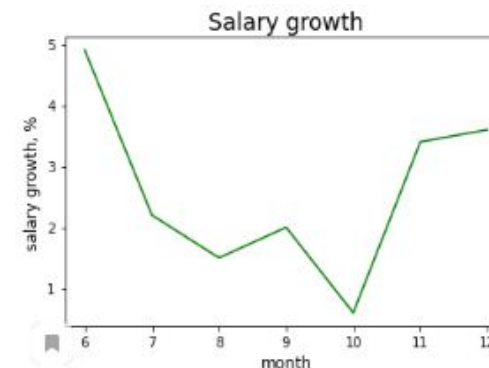
Внешние данные

Продовольственная инфляция в 4 кв. 2021 г. составила 10.8%.

- Рост цен на сырье продовольственных товаров
- Ослабленное предложение и высокий спрос
- Рост стоимости рабочей силы



- Рост заработной платы населения в течение 2021 г. Повышается спрос на полезные товары
- Высокий спрос на рабочую силу поддерживал рост заработной платы (на 88% в октябре)
- Исторический минимум безработицы во второй раз в ноябре



Несмотря на многие экономические предпосылки, данные об инфляции и доходах не сильно повлияли на качество модели

Выбор дополнительной модели

Для построения look-alike-модели мы обучили еще несколько state-of-the-art классификаторов.

TabNet - нейросеть для структурированных данных

- ✓ Интерпретируемость
- ✓ Повышение производительности за счет архитектур глубокого обучения.

F1-score test: 0.14

F1-score val: 0.25



LightGBM — градиентный бустинг на деревьях

- ✓ Высокая скорость обучения и эффективность.
- ✓ Низкое использование памяти.
- ✓ Поддержка параллельного обучения.
- ✓ Возможность обработки больших объемов данных.

F1-score test: 0.68

F1-score val: 0.71



TabNet показал плохое качество на нашем наборе данных. Зато LightGBM хорошо справился с задачей, поэтому мы взяли его, как часть финальной модели.

Модель

Предобработка данных

Заменяем нулями пропуски

Корректируем отрицательные значения в колонках товарооборота

Зануляем отрицательные отклонения

Добавляем новые признаки и учитываем макропоказатели

Подбор гиперпараметров

Использование кросс-валидации с совместно с класса StratifiedKFold из sklearn

Оптимизация на основе функции, приближающей показатели на тесте и валидации

Применение библиотек Hyperopt и Optuna

Обучение модели



Yandex
CatBoost

Финальная модель


Результаты предыдущей модели:

F1-score val:	0.70
F1-score test:	0.64



Результаты итоговой модели:

F1-score val:	0.71
F1-score test:	0.70



6 %



Серебрякова Софья

ВМК МГУ

Магистратура, 1 курс



Абдуллаева Ума

РГУНГ им. Губкина

Бакалавриат, 3 курс