КОМАНДА
MAX

ПОИСК LOOK-ALIKE-АУДИТОРИЙ НА ОСНОВЕ АНАЛИЗА ДАННЫХ О КЛИЕНТАХ МАГАЗИНОВ «ПЕРЕКРЁСТОК»

9 марта 2022
DATA SCIENCE

КОМАНДА MAX

АНАЛИЗ ЗАДАЧИ И ДАННЫХ

Задачу рассматриваем как классическую задачу классификации.

Гипотеза: Факт вступления в «Клуб полезных привычек» коррелирует с такими параметрами поведения клиента как товарооборот, количество чеков.

Нулевой этап
Для первоначальной проверки уменьшаем размерность датасета, сокращаем колонки, по месяцам. Суммируем товарообороты и количество чеков. У стандартного отклонения берем среднее арифметическое.

```
rto_n
rto_n_category
cnt_checks_n
cnt_checks_n_category
```

Σ →

```
∑ rto_n
∑ rto_category_n
∑ cnt_checks_n
∑ cnt_checks_n_category
```

```
rto_std_n
rto_std_n_category
```

ср. →

```
ср. арифметическое rto_std_n
ср. арифметическое rto_std_n_category
```

Основные параметры датасета:

- **client_id** — уникальный идентификатор клиента
- **rto** — сумма товарооборота
- **cnt_checks** — количество чеков
- **category** — категория товаров
- **n** — месяц
- **is_in_club** — участие в клубе (1 — да, 0 — нет)
- **rto_std** — стандартное отклонение суммы товарооборота от чека к чеку

Для модели требуются числовые значение, принимаем отсутствие товарооборота клиента как **Nan = 0**

КОМАНДА
MAX

- Добавим средний чек за покупку — **сумму товарооборота / количество чеков**. Выполним это как для общего товарооборота так и для каждой категории отдельно.

```
mean_rto
mean_rto_Крупы и зерновые
mean_rto_Мясная гастрономия
mean_rto_Овощи – Фрукты
mean_rto_Сыры
mean_rto_Рыба и рыбные изделия
mean_rto_Птица и изделия из птицы
```

DATASET

- Удалим всех нецелевых клиентов, которые имели количество нулевых чеков больше **2**.
- Удалим всех нецелевых клиентов, которые имели сумму чеков за все месяцы меньше **20 000** ₽.
- Получили из соотношения 0/1, %:

0	**1**
0.91	0.09
↓	↓
0.84	0.16

- Разбили на **обучающую и тестовую** выборки в соотношении **70%** к **30%**, получив соотношение в числах:

57623	14406

- Применили операции **downsample** и **upsample**, для более равномерных данных, получив соотношение

0	**1**
0.77	0.23
↓	↓
0.63	0.37

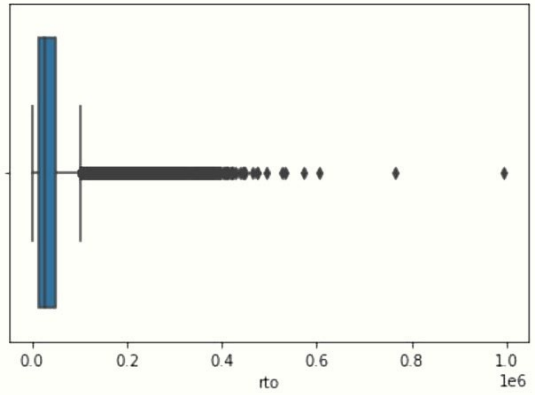
в числах:

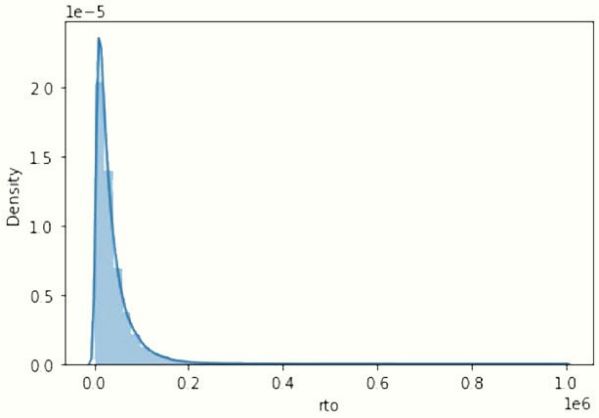
53250	43316

КОМАНДА MAX

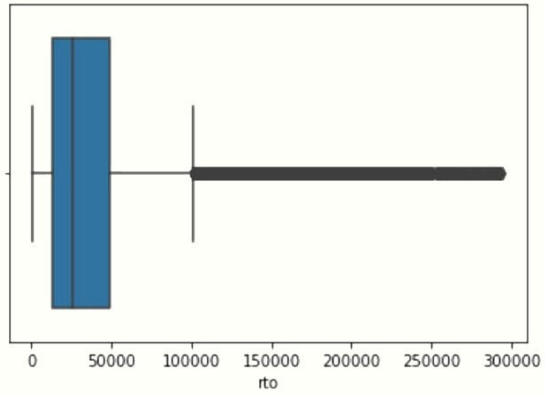
Потеря целевой аудитории при масштабируемости признаков приблизительно 20-30 % от общего количества

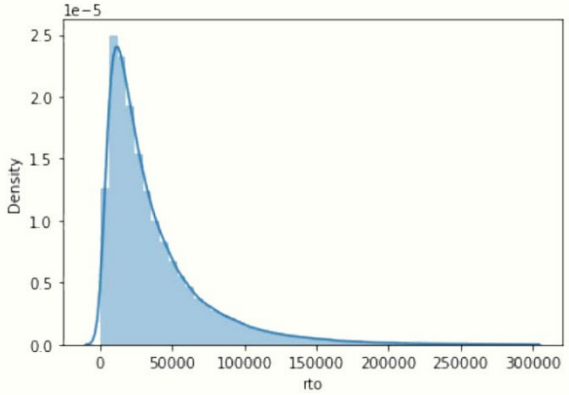
Распределение rto до очистки





Распределение rto после очистки

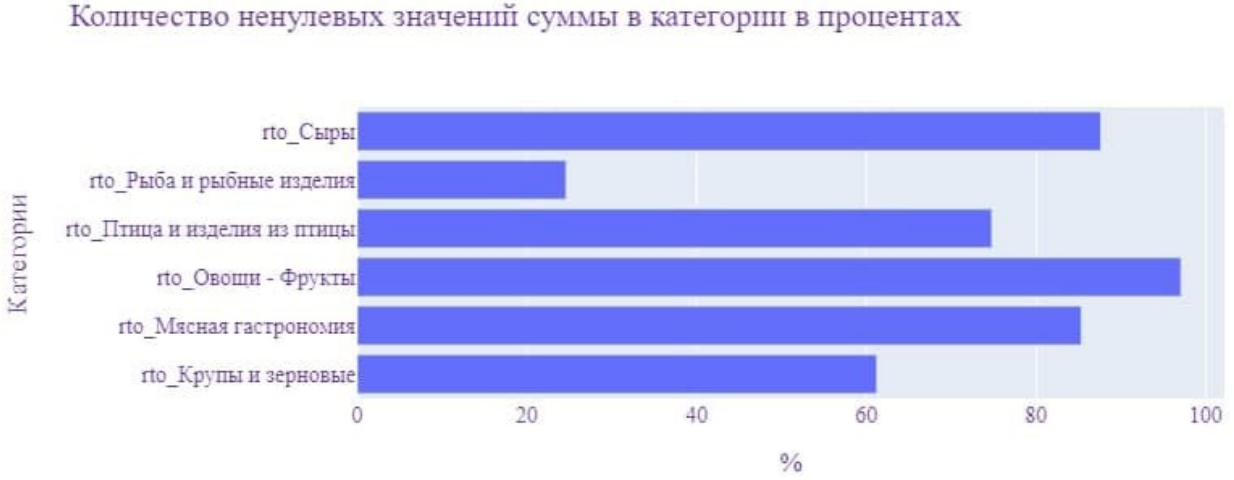


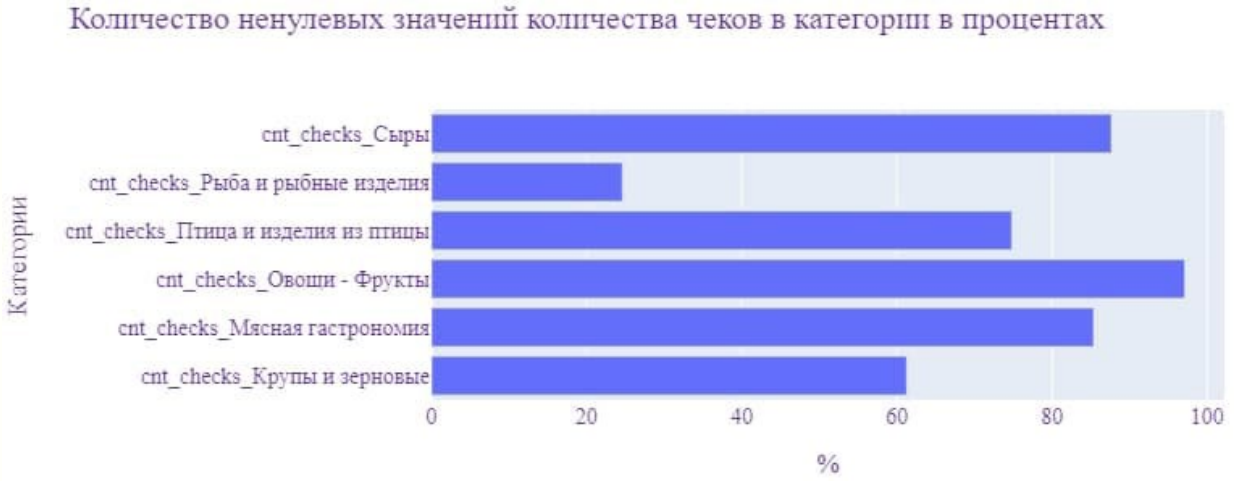


Отсекаем данные по Рыбе по причине малого количества ненулевых значений

КОМАНДА
MAX

Распределение по ценовым диапазонам


Количество ненулевых значений суммы в категории в процентах
Категории
rto_Сыры
rto_Рыба и рыбные изделия
rto_Птица и изделия из птицы
rto_Овощи - Фрукты
rto_Мясная гастрономия
rto_Крупы и зерновые
0
20
40
60
80
100
%


Количество ненулевых значений количества чеков в категории в процентах
Категории
cnt_checks_Сыры
cnt_checks_Рыба и рыбные изделия
cnt_checks_Птица и изделия из птицы
cnt_checks_Овощи - Фрукты
cnt_checks_Мясная гастрономия
cnt_checks_Крупы и зерновые
0
20
40
60
80
100
%

КОМАНДА
MAX

Logistic_reg

	Train	Test
f1	0.53	0.31
Accuracy	0.61	0.61
Roc auc	0.61	0.60
Precision	0.49	0.22
Recall	0.59	0.58

LightGBM

	Train	Test
f1	0.86	0.42
Accuracy	0.90	0.84
Roc auc	0.88	0.65
Precision	0.97	0.49
Recall	0.76	0.37

CatBoost

	Train	Test
f1	0.74	0.42
Accuracy	0.84	0.84
Roc auc	0.79	0.65
Precision	0.92	0.46
Recall	0.62	0.39

КОМАНДА
MAX

Bagging

Classification report

	Precision	Recall	f1
0	0.89	0.94	0.91
1	0.51	0.38	0.43
Accuracy			0.85
Macro avg	0.70	0.66	0.67
Weighted avg	0.83	0.85	0.84

Metrics

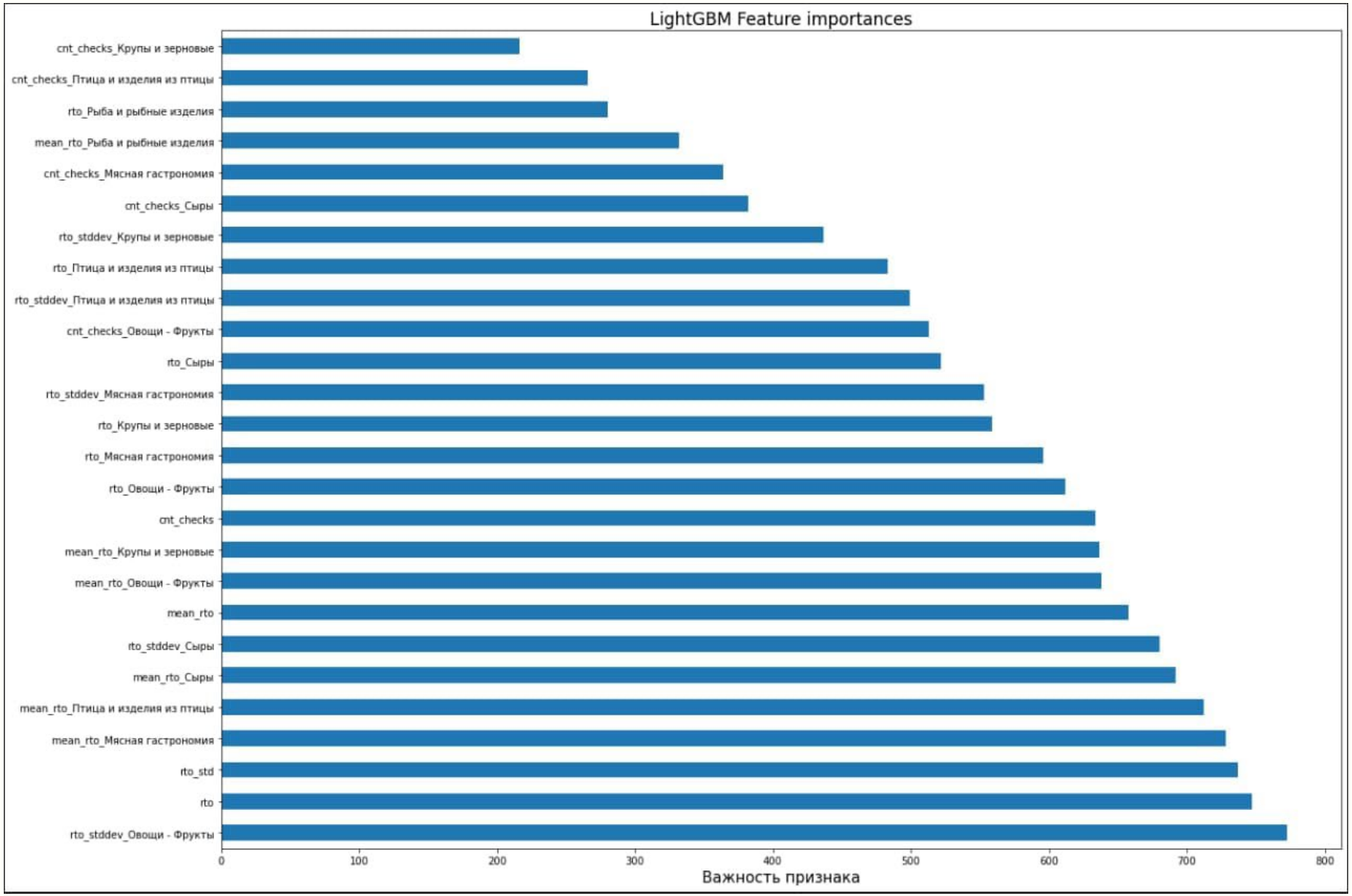
F1_score	Roc_auc
0.433	0.655

Параметры итоговой модели

```
model = BaggingClassifier(random_state=666, n_jobs=-1, base_estimator=LGBMClassifier(max_depth=9, n_estimators=450, learning_rate=0.1), n_estimators=500)
model.fit(features_train, target_train)
```

```
BaggingClassifier(base_estimator=LGBMClassifier(max_depth=9, n_estimators=450),
                  n_estimators=500, n_jobs=-1, random_state=666)
```

КОМАНДА MAX


LightGBM Feature importances
cnt_checks_Крупы и зерновые
cnt_checks_Птица и изделия из птицы
rto_Рыба и рыбные изделия
mean_rto_Рыба и рыбные изделия
cnt_checks_Мясная гастрономия
cnt_checks_Сыры
rto_stddev_Крупы и зерновые
rto_Птица и изделия из птицы
rto_stddev_Птица и изделия из птицы
cnt_checks_Овощи - Фрукты
rto_Сыры
rto_stddev_Мясная гастрономия
rto_Крупы и зерновые
rto_Мясная гастрономия
rto_Овощи - Фрукты
cnt_checks
mean_rto_Крупы и зерновые
mean_rto_Овощи - Фрукты
mean_rto
rto_stddev_Сыры
mean_rto_Сыры
mean_rto_Птица и изделия из птицы
mean_rto_Мясная гастрономия
rto_std
rto
rto_stddev_Овощи - Фрукты
0
100
200
300
400
500
600
700
800
Важность признака

КОМАНДА
MAX

ВЫВОД

Классы не сбалансированы. Отношение целевой аудитории к нецелевой 9% / 91 %

Из 9% целевой аудитории 20-30% являются выбросами. Исходя из дисбаланса и нехватки данных, модели не смогли найти закономерности. Требуется больше данных.

Распределение по суммам в чеках всех категорий (rto) примерно одинаковое.