Кейс для X5 Group

Changellenge IT Cup

Команда “Два из десяти”

Захаров Даниил
Савич Александр
Глаголев Александр
Коршунов Иван

План проведённой работы

1. **Статистический анализ данных**: распределения величин, стат. критерии, значимость конкретных факторов

   файл: CupIT2022 – секция DS — Два из десяти.ipynb

2. **Выбор модели,** построение нейронной сети

3. **Результаты** на трейн и тест датасетах

4. **Выводы** и дальнейшее взаимодействие

Статистический анализ датасета

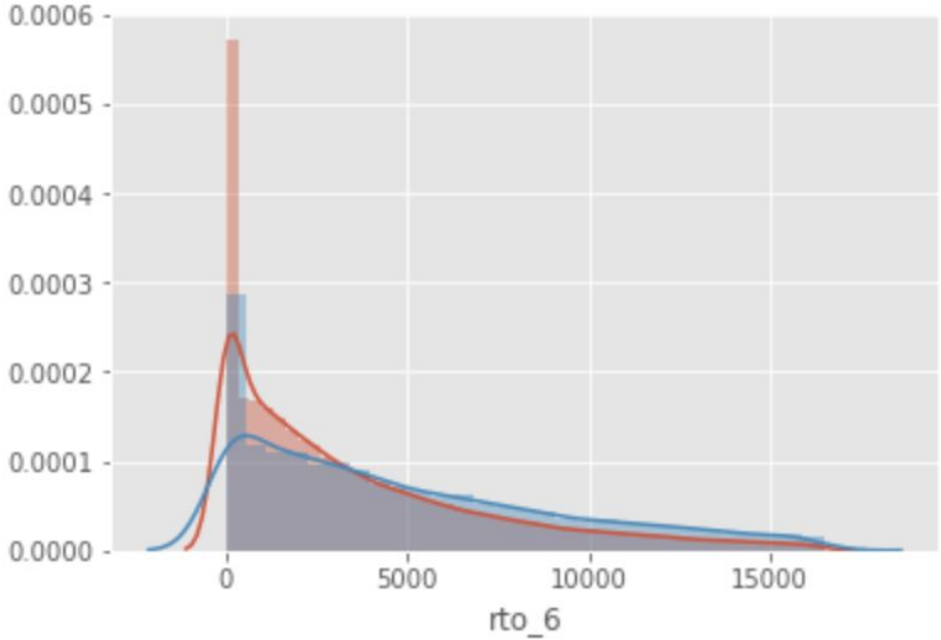
1. **Построение** распределений величин rto_6, rto_7,.., отдельно для членов клуба и обычных покупателей
2. **Проверки на соответствие** полученных распределений известным и часто встречающимся (экспоненциальному, гамме) с помощью статистических критериев
3. **Сравнение** распределений (3х-квартильные значения, средние значения)
4. Аналогичная статистика по **различным категориям товаров**

файл: CupIT2022 – секция DS — Два из десяти.ipynb

Графики распределений по rto

Мы построили графики распределений значений чеков rto_6
отдельно по is_in_club = 0 и 1

Выбросы были исключены из каждого распределения



красный график: is_in_club = 0
синий график: is_in_club =1

Статистические тесты

По критерию Колмогорова-Смирнова можно сравнить распределение с любым заданным.

Мы решили сравнивать с **экспоненциальным**, и, как с обобщением, с **гамма**-распределением.

Предварительно мы убрали из данных “выбросы”, а также произвели нормировку

К сожалению, результаты тестов показали, что наши распределения

не относятся к семейству гамма-распределений

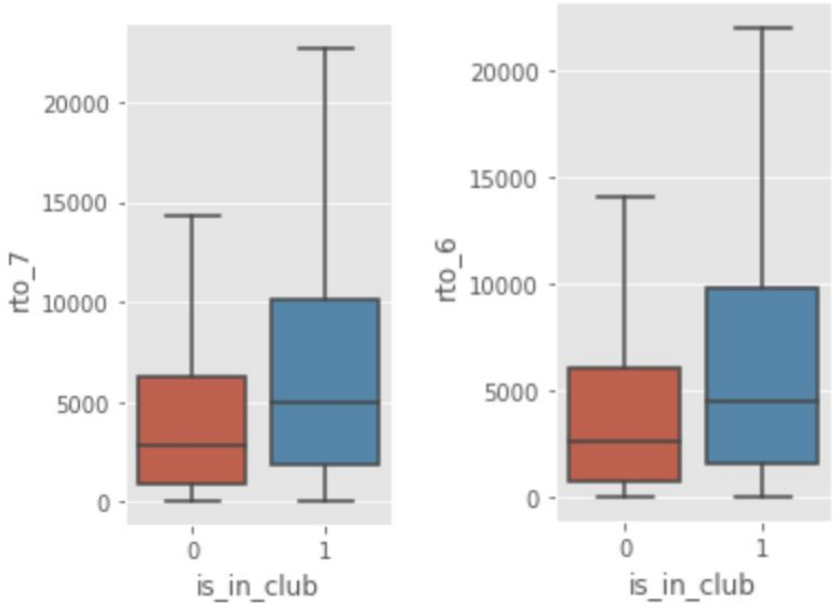
Сравнение распределений по rto

По каждому месяцу (колонки rto_6, rto_7,..) были построены сравнения распределений rto, разделяя по категории is_in_club

В результате легко заметить, что члены клуба **тратят больше** остальных покупателей

Графики справа для rto_6 и 7,

показывают 3х-квартильные значения соответствующих распределений, в том числе средние значения



Сравнение распределений по rto по отдельным категориям товаров

Аналогичные графики можно построить, в частности, по каждой категории товаров.

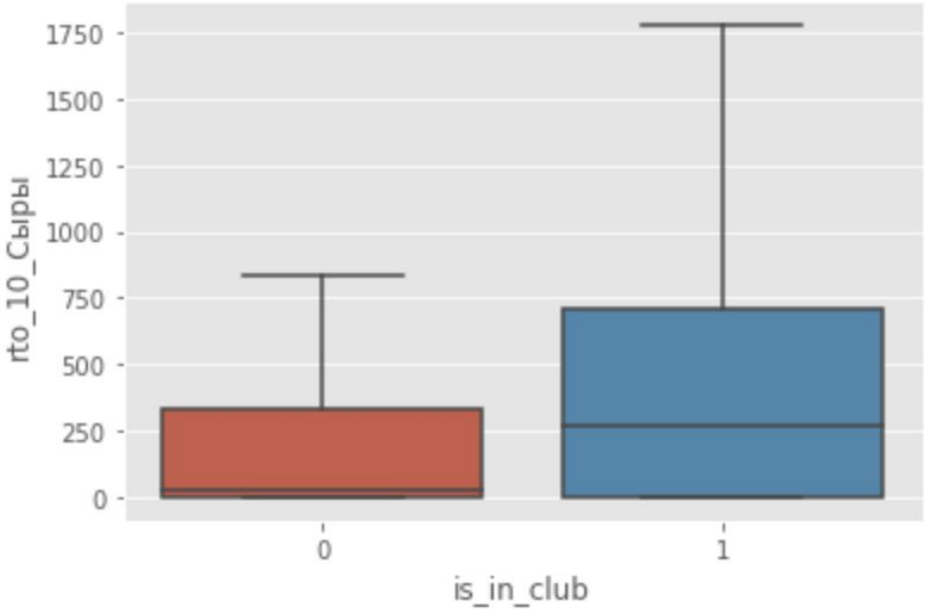
Это позволит определить наиболее “затратные” категории для членов клуба (по сравнению с обычными покупателями).

В дальнейшем это позволит, например, найти категории, наиболее востребованные для членов клуба.

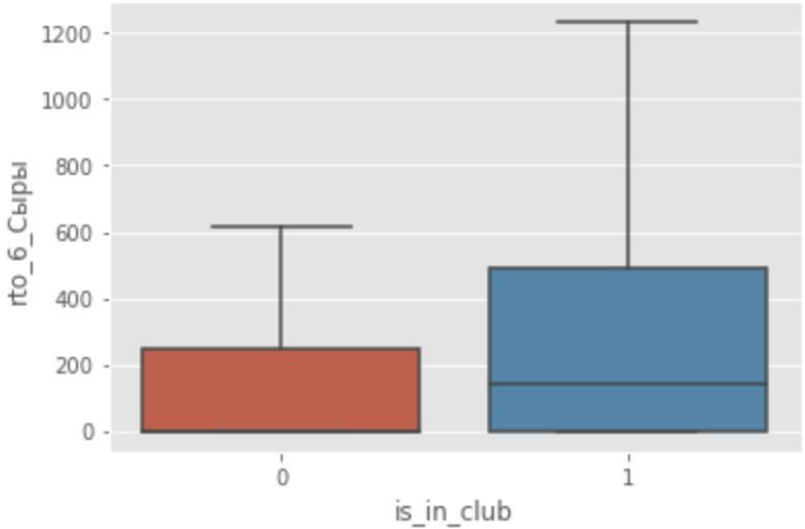
Выделяя на такие категории специальные предложения и скидки, мы сможем привлечь в клуб больше покупателей, которые, возможно, колеблются над покупкой

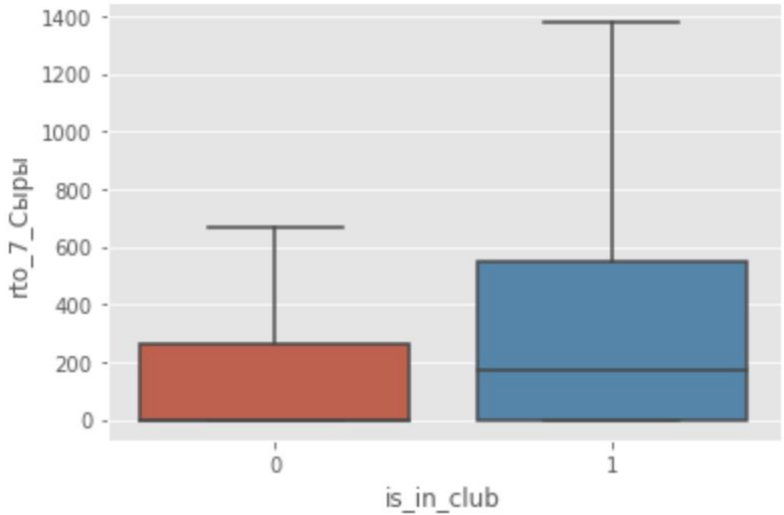
Сравнение распределений по rto по отдельным категориям товаров

Например, проследив, на сколько каждый месяц в среднем покупали сыры члены клуба и обычные клиенты, нетрудно заметить, что данная категория товаров **пользуется большим спросом в клубе,** а значит, скорее всего является более привлекательной для потенциальных новых членов клуба



Сравнение распределения rto для сыров





Выбор модели

Цель искомой модели - распознавание клиентов, похожих на участников программы лояльности в исходном датасете, т.е. мы стремимся получить достаточно большое число ложноположительных результатов на клиентах, похожий на истинных участников.

В результате анализа датасета было установлено:

- Порядка 90% наблюдений - не члены клуба.
- Прогноз обученными классификаторами логистической регрессии и байесовским классификатором состоит из всех нулей, что не является удовлетворительным результатом.
- Прогноз при помощи CatBoost тоже распознавал слишком мало положительных результатов, что приводило и к низкому показателю ложноположительных результатов.

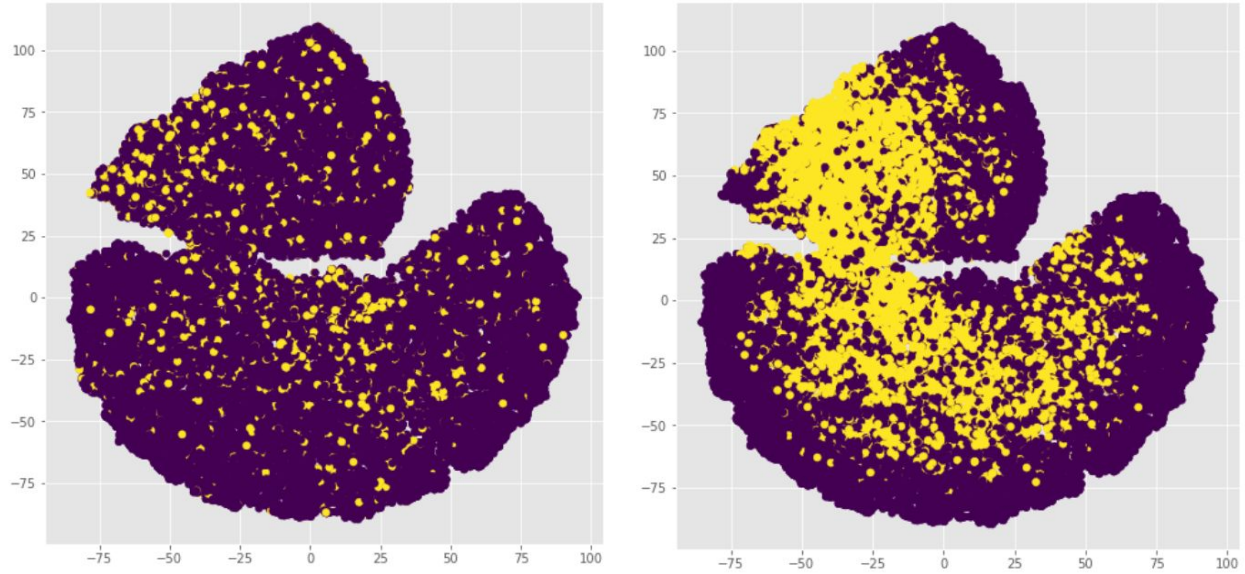
По итогам жарких дебатов было решено обучить нейронную сеть.

Нейронная сеть

Для простоты было решено использовать полносвязную нейронную сеть с функцией активации RELU и BatchNorm, чтобы избежать переобучения. Путем балансировки классов был получен True Positive Rate > 50%, при котором нейронная сеть определила 26757 потенциальных клиентов на части трэйн и 11391 на валидационном.

Для проверки пространственной близости предсказаний нейронной сети к истинным участникам программы лояльности и ее визуализации был использован алгоритм T-distributed Stochastic Neighbor Embedding. На иллюстрациях ниже можно заметить, что потенциальные участники программы (желтые точки на втором изображении) сконцентрированы пространственно близко к истинным участникам (первая диаграмма)

T-distributed Stochastic Neighbor Embedding



Выводы

1. Полученная нейронная сеть **хорошо** показала себя на тестовых данных
2. При необходимости данную модель **можно доработать**, при наличии дополнительной информации о клиентах
3. Для расширения клубной базы следует, воспользовавшись моделью на **“новых” данных**, определить возможных **“будущих” членов клуба**
4. Для полученного списка клиентов использовать различные схемы привлечения в клубную базу (помимо простой рассылки)
5. Например, по (статистически) выбранным категориям товаров **предложить скидки** при вступлении в клуб