Предсказание трат по категориям

Кейс второго отборочного этапа на Молодежную программу FINODAYS

MegaQuant

FINODAYS

Tinkoff Bank

Задача:
Создать **прототип сервиса**, который с помощью алгоритмов машинного обучения сможет **предсказывать траты пользователей** по категориям на следующий месяц

Подход к формированию тренировочной и тестовой выборки:

party_rk	children_cnt	region_flg	AgeGroup	gender_cd_F	gender_cd_M	gender_cd_unknown	marital_status_desc_unknown	marital_status_desc_Вдовец, вдова	mari
61243	0	0	5	1	0	0	1	0	
66535	0	0	1	1	0	0	1	0	
83721	0	0	4	0	1	0	0	0	
88238	0	0	2	1	0	0	1	0	
57179	0	0	2	1	0	0	0	0	
...	...	...	...	...	...	...	...	...	
54994	0	0	2	0	1	0	0	0	
63391	0	0	2	0	1	0	0	0	
5418	0	0	4	1	0	0	0	0	
50273	0	0	2	1	0	0	0	0	
77268	0	0	2	1	0	0	1	0	

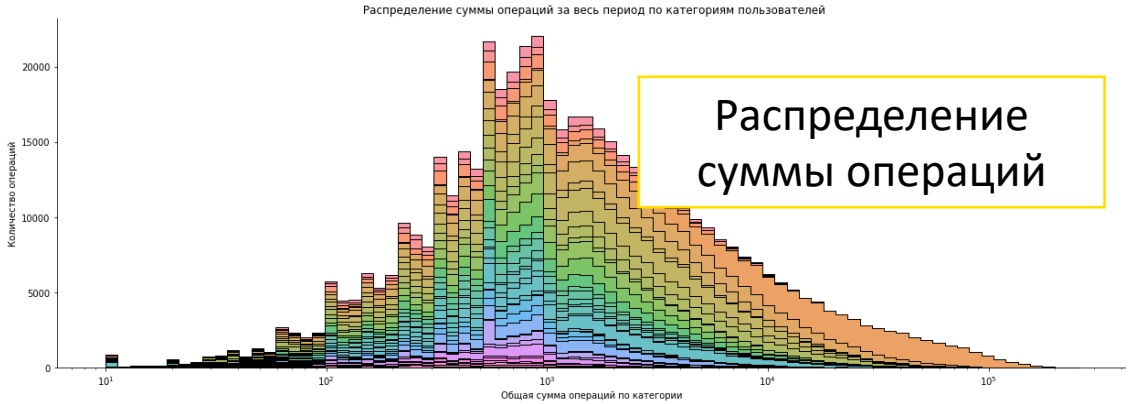
49401 rows × 554 columns

Строки таблицы - пользователи, столбцы - признаки

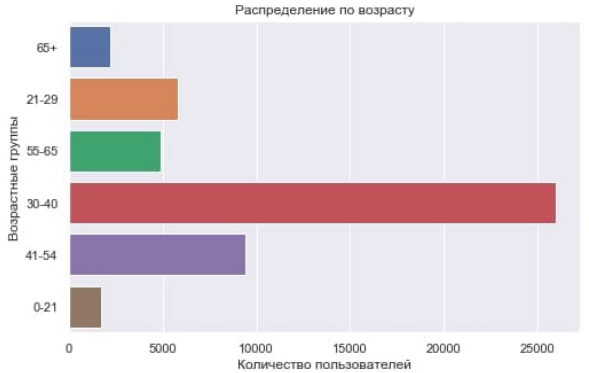
Этапы работы над кейсом:

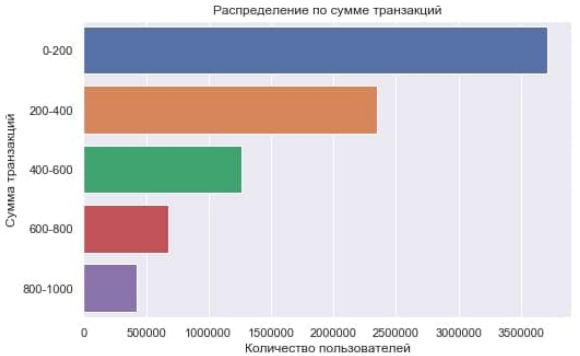
1. Первичный **анализ данных** и **визуализация**
2. Обработка данных:
- работа с **выбросами**
- работа с **категориальными** признаками
- работа с **пропусками**
3. **Разработка** новых признаков
4. **Анализ временных рядов** с помощью ARIMA
5. Создание **итоговой таблицы**
6. Применение **модели предсказания**
7. Разработка **метрик**:
- метрик машинного обучения
- бизнес-метрик
8. Проверка модели на тестовых данных
9. Поиск возможных улучшений сервиса и выводы

Мы построили следующие графики, чтобы отразить особенности поведения пользователей и составить среднестатистический профиль клиента



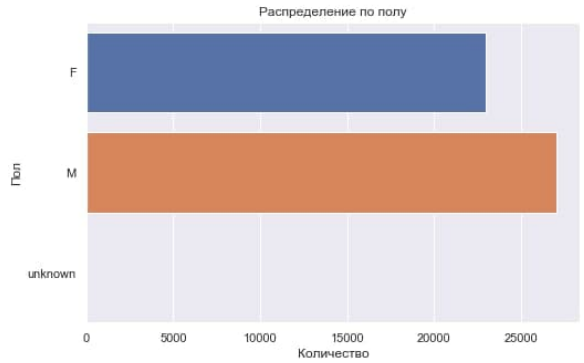
Распределение суммы операций

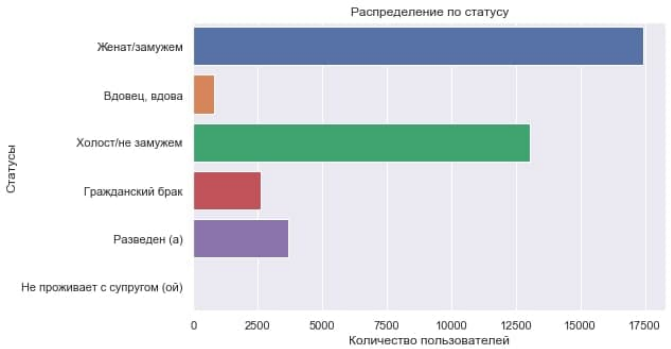




Распределение объема транзакций

Распределение клиентов по социально-демографическим показателям





Мы разработали 7 новых признаков для тренировочной и тестовой выборок. Они характеризуют частоту, объем и количество совершаемых транзакций, временные ряды и данные о действиях пользователей в Интернете.

Признак №1

$$\frac{\text{(Количество лайков у человека в конкретной категории в данном месяце)}}{\text{(Общее количество лайков человека в конкретном месяце)}}$$

Признак №2

$$\frac{\text{(Количество дизлайков у человека в конкретной категории в данном месяце)}}{\text{(Общее количество дизлайков человека в конкретном месяце)}}$$

Признак №3

Средняя сумма в месяц, потраченная пользователем на конкретную категорию (подсчет по последним трем месяцам)

Признак №4

Стандартное отклонение суммы траназкций в месяц по категориям (посчитано для месяцев с транзакциями)

Признак №5

Частота повторных покупок у одного человека по каждой из категорий.

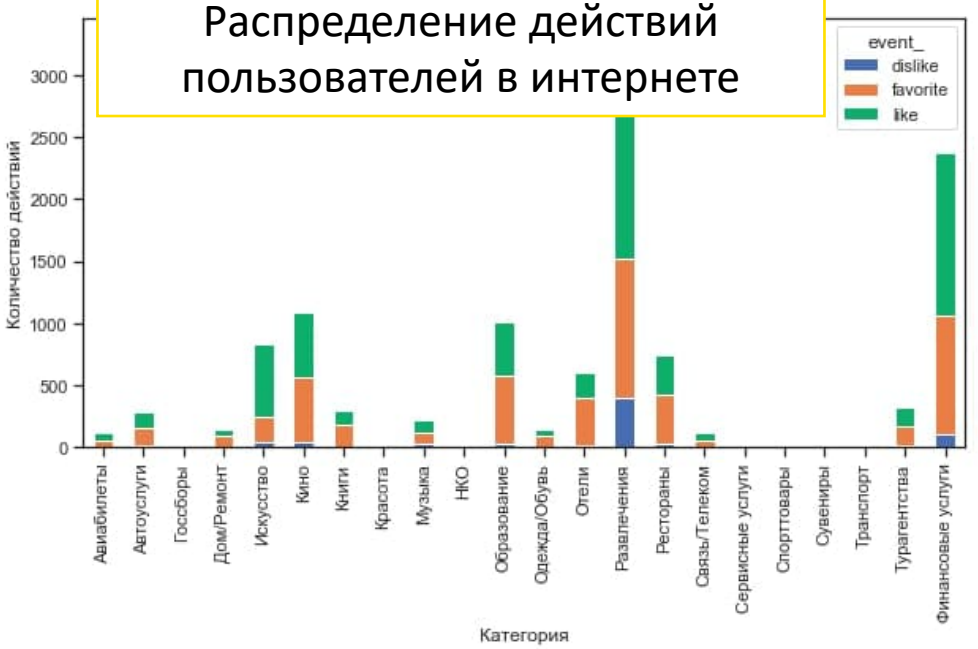
Признак №6

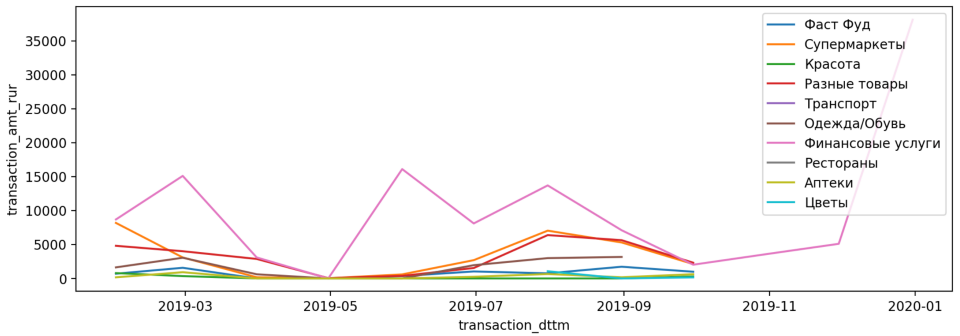
$$\frac{\text{(Количество покупок у человека в конкретной категории в данном месяце)}}{\text{(Общее количество транзакций человека в конкретном месяце)}}$$

Признак №7

ARIMA

Распределение действий пользователей в интернете



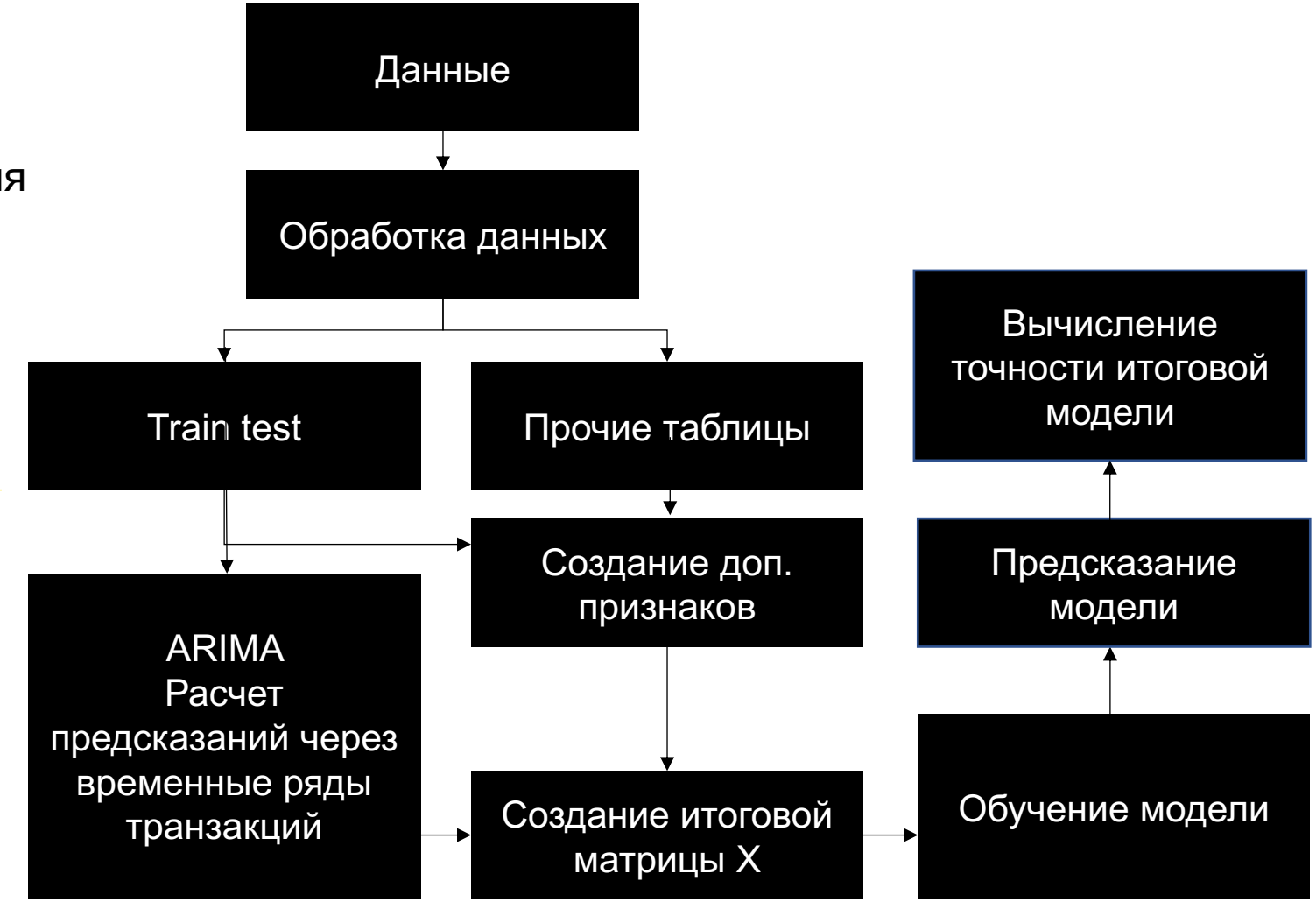


Преимущества ARIMA:

- ✓ ARIMA создана специально для работы с временными рядами
- ✓ Часто используется для предсказания финансовых показателей
- ✓ Может работать с короткими временными рядами
- ✓ Учитывает стационарность
- ✓ Учитывает шумы

Почему используем не только ARIMA?

- ✓ Хотим учитывать характеристики и предпочтения пользователей, а не только их историю транзакций
- ✓ Прогнозы ARIMA могут варьироваться в зависимости от выбранных параметров, поэтому модель может выявлять ложные тенденции



Возможные метрики машинного обучения, применимые к нашей задачи:

1. MSE
2. MAE
3. RMSE
4. Huber-Loss
5. R-квадрат

Почему мы выбрали R-квадрат?

- ✓ Его значения лежат в интервале от 0 до 1, поэтому его результаты интерпретируемы
- ✓ У используемых нами моделей есть метод score(X, y[, sample_weight]), который возвращает коэффициент детерминации прогноза

Модели и параметры, которые мы использовали, и результаты их применения

Модель	Время Обучения	Accuracy	Параметры модели
Random Forest	00:39	0.060	Depth=1, n_est=25
Random Forest	03:40	0.073	Depth=3, n_est=100
Random Forest	09:50	0.071	Depth=1, n_est=600
Random Forest	02:27	0.103	Depth=1, n_est=100
Random Forest	10:10	0.290	Depth=3, n_est=100
Gradient Boosting	12:00	0.176	
Net Random Forest	20:00	0.176	Depth=3, n_est=300
Random Forest	00:30	0.340	Depth=5, n_est=100
Random Forest	06:00	0.400	Depth=10, n_est=500

0,4

R-квадрат

Бизнес-метрики:

- ✓ Прирост средней суммы затрат по всем категориям у пользователя за месяц (среднее значение по всем пользователям)
- ✓ **ARPU** - Средняя выручка на одного пользователя

$$M_2 = ARPU_i \text{ (за 2 мес.)} = \frac{\text{общая выручка со всех продаж в i-ой категории}}{\text{количество покупателей в i-ой категории}}$$

- ✓ **LTV** (Lifetime Value) - пожизненная стоимость клиента

$$LTV_{ij}(\text{ по месяцу}) = (\text{средний чек по i-ой категории}) \cdot (\text{количество месяцев, где клиент пользуется сервисом на данный момент}).$$

- ✓ Прирост количества «like» или «favorite» на истории со словом «Кэшбек»

Увидимся в Сочи!

НИУ ВШЭ ‘2023 /Бизнес-информатика

Камран Абдулхаев **Алина Мусина** **Елизавета Петяева** **Филимошина Екатерина** **Антон Сметанин**