



Определение цифрового портрета аудитории в мобильной среде

Глобальные игроки начинают ограничивать доступ к идентификатору мобильного устройства, а значит — и данным аудитории.

Необходимо создать решение, которое позволило бы максимально точно определить профиль аудитории в мобильной среде на основе различных косвенных и исторических аудиторных данных.

“ Snap, Facebook, Twitter и YouTube потеряли \$10 млрд выручки после изменения настроек приватности на iPhone”

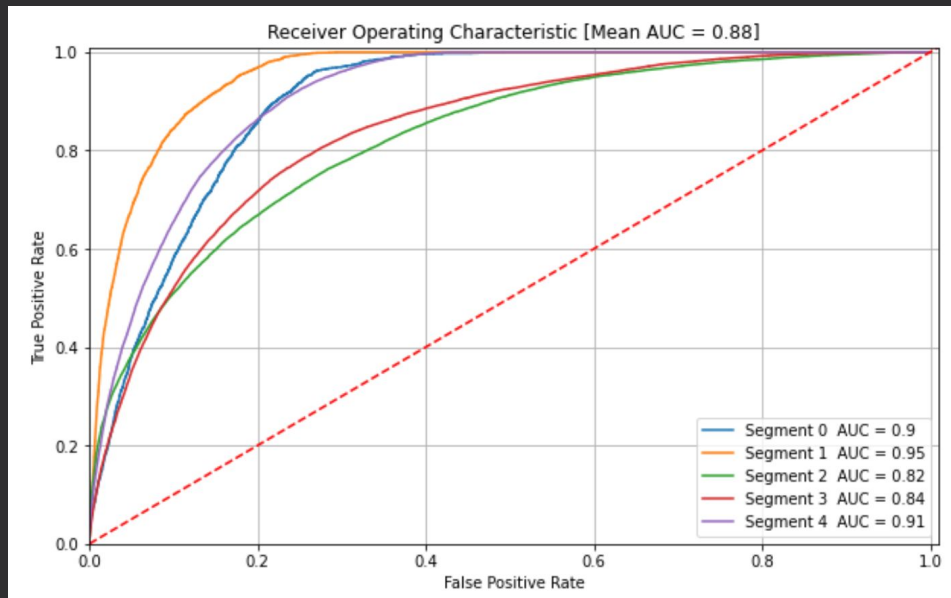
Financial Times, 31 октября 2021

Создана и обучена **модель машинного обучения**, которая не используя идентификатор устройства, на основании данных строки bundle приложения, времени и региона делает предсказание, к какому сегменту отнести пользователя.

Для демонстрации работы модели создан **веб-интерфейс**, который позволяет интерпретировать, на основании каких признаков пользователь был отнесен к тому или иному сегменту.

AUC
0.89

на тестовой выборке,
очень близок к валидации

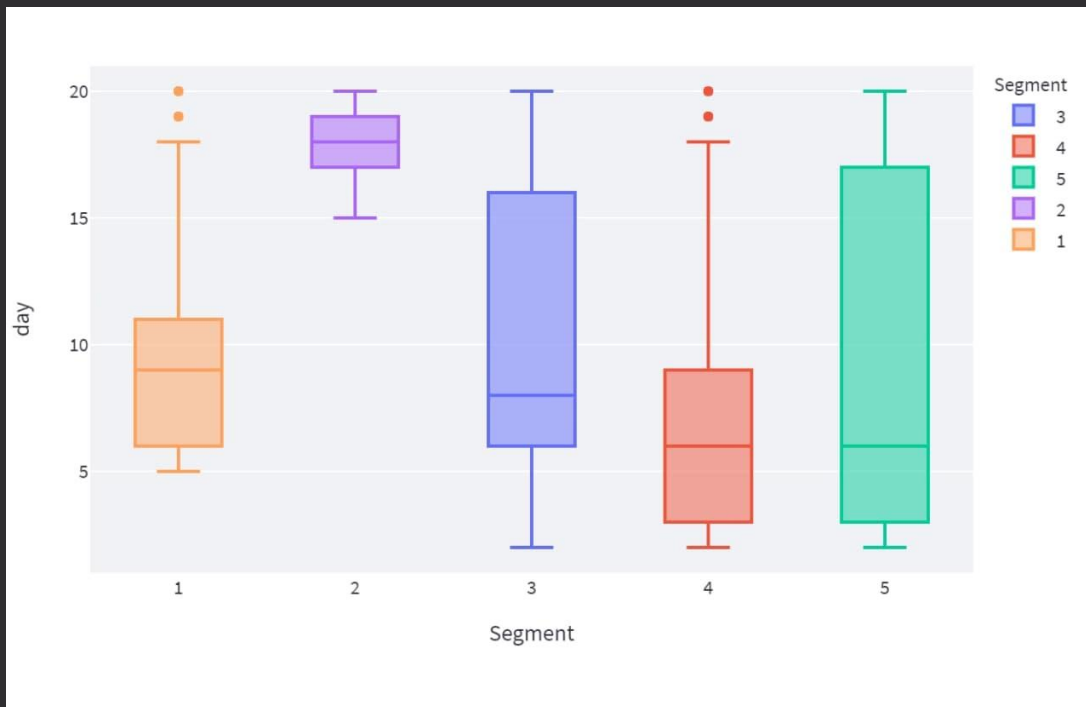


Модель классификации. Признаки

Признаки, сгенерированные из данных, которые дали хороший сигнал для модели:

- **hour, weekday** - час и день недели использования приложения, может отражать разную занятость сегментов (например, кто-то на работе, кто-то учится)
- **nexters, art, water, color** - признаки tf-idf из bundle, может отражать, что определенные сегменты выбирают определенный тип приложения или игры определенного разработчика
- **salary_rank** - место города в топ 100 городов России по зарплате

Особенности данных для обучения



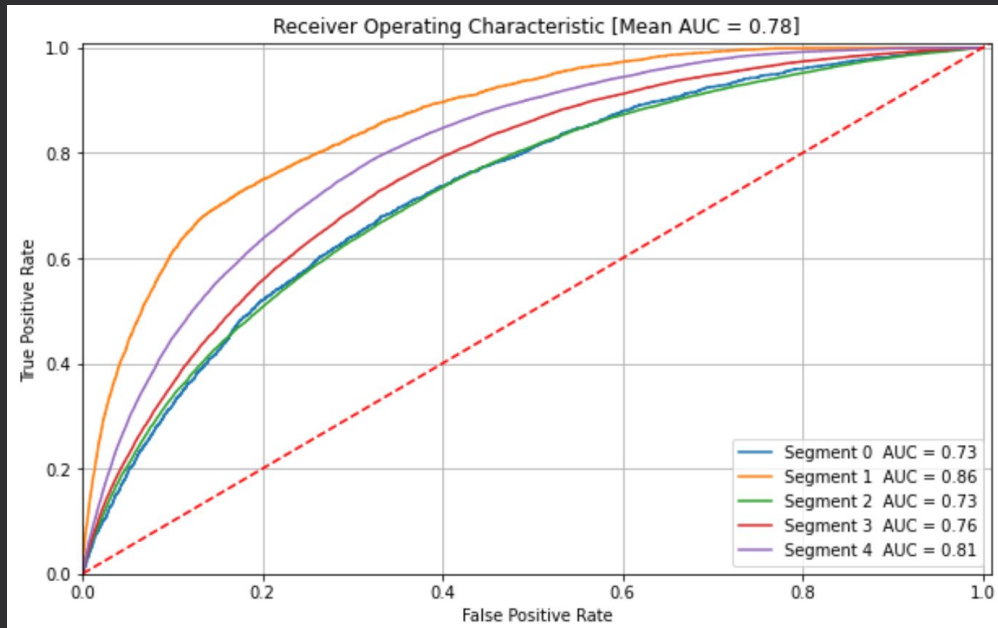
Возникла гипотеза, что высокий скор связан со спецификой конкретной выгрузки данных для обучения (такая же специфика есть и у предоставленных тестовых данных), чтобы при практическом применении не случилось снижение качества модели, этот признак лучше не использовать, без него AUC 0.78. Перепроверено на тестовом наборе данных.

Модель классификации. Метрика

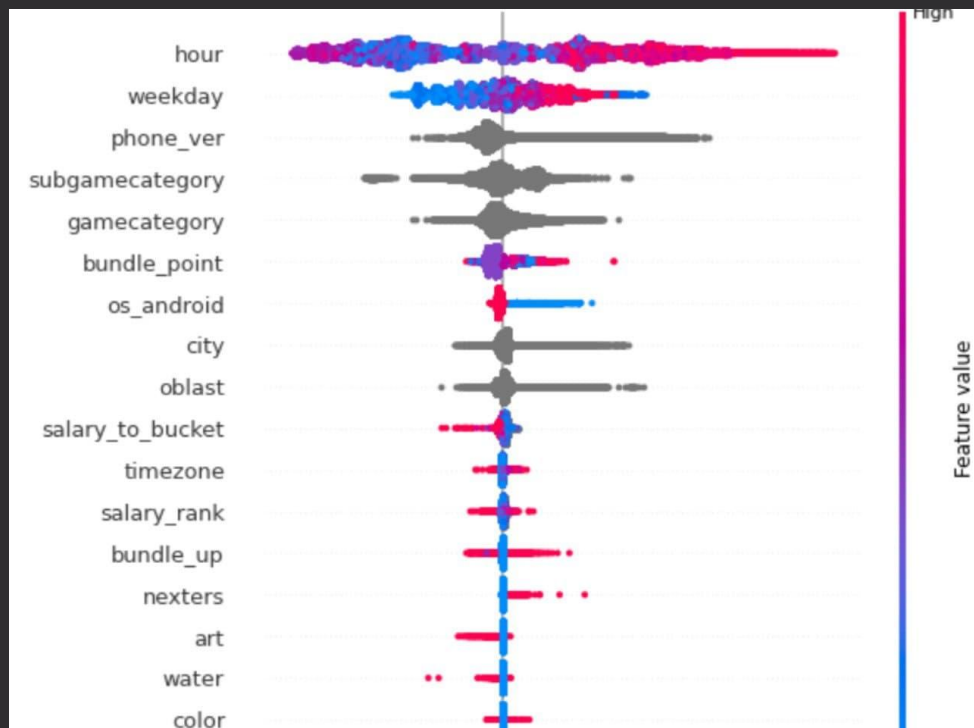
Без признака дня, дающего неоправданно большой скор

AUC
0.78

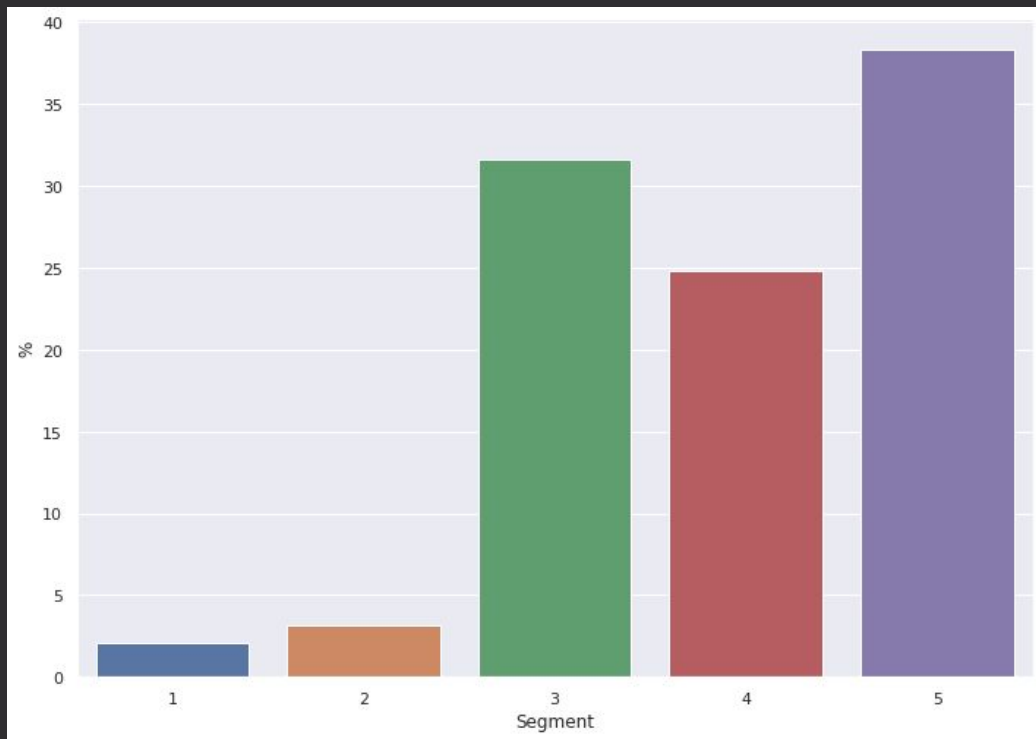
на тестовой выборке,
совпадает с валидацией



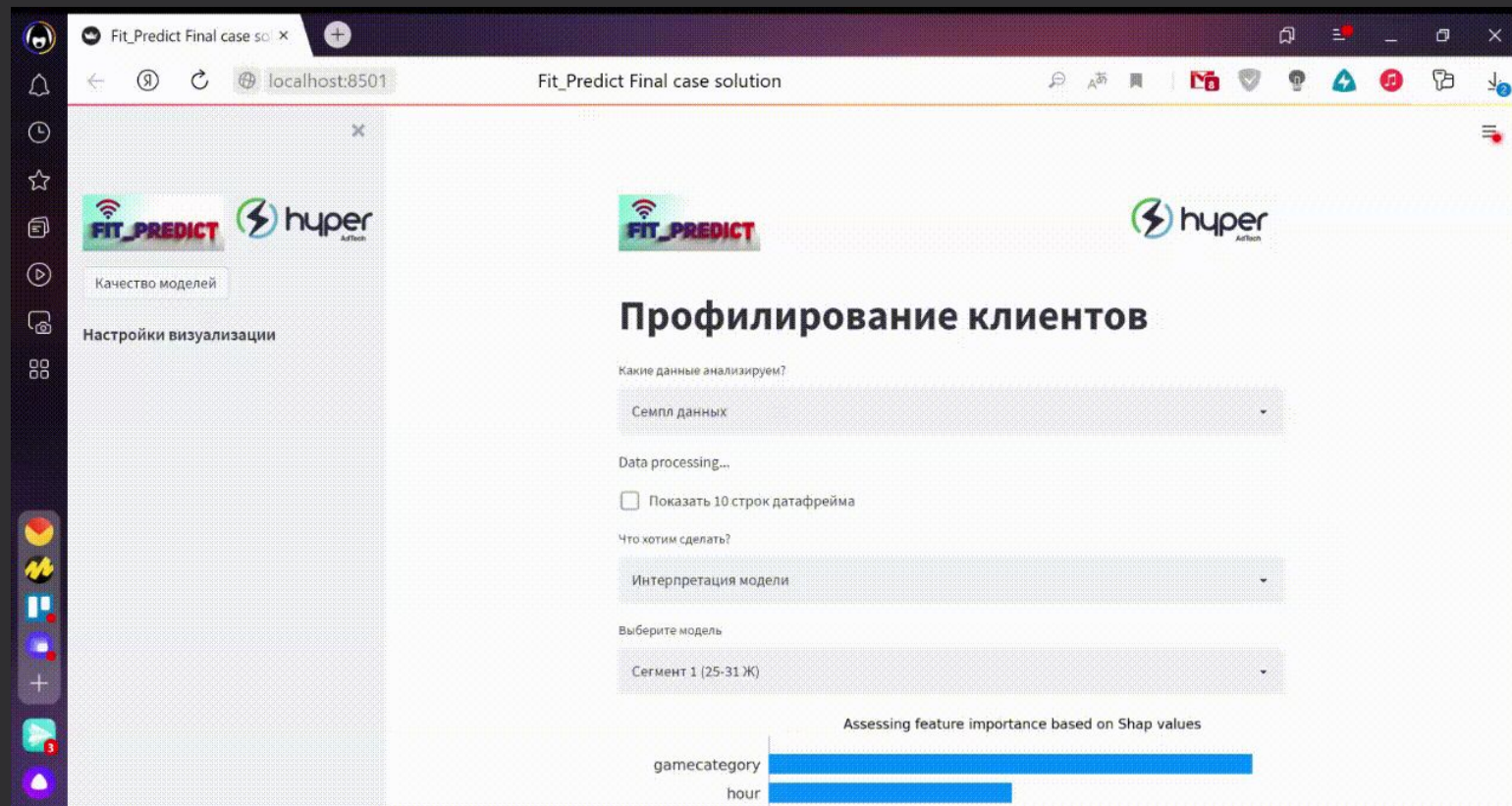
Модель классификации. Признаки



Предсказанные сегменты (test.csv)



Демо модели классификации



The screenshot displays the FIT_PREDICT web application interface. The browser address bar shows 'localhost:8501'. The page title is 'Fit_Predict Final case solution'. The interface includes a sidebar with navigation options: 'Качество моделей' (Model Quality) and 'Настройки визуализации' (Visualization Settings). The main content area is titled 'Профилирование клиентов' (Client Profiling). It contains several interactive elements: a dropdown menu for 'Какие данные анализируем?' (Which data are we analyzing?) set to 'Семпл данных' (Data sample); a 'Data processing...' section with a checkbox 'Показать 10 строк датафрейма' (Show 10 rows of dataframe) which is currently unchecked; a dropdown menu for 'Что хотим сделать?' (What do we want to do?) set to 'Интерпретация модели' (Model interpretation); and a dropdown menu for 'Выберите модель' (Select model) set to 'Сегмент 1 (25-31 Ж)' (Segment 1 (25-31 F)). At the bottom, a horizontal bar chart titled 'Assessing feature importance based on Shap values' shows the importance of 'gamecategory' and 'hour' features. 'gamecategory' has a significantly higher importance value than 'hour'.

Fit_Predict Final case solution

localhost:8501

Fit_Predict Final case solution

hyper

Качество моделей

Настройки визуализации

Профилирование клиентов

Какие данные анализируем?

Семпл данных

Data processing...

☐ Показать 10 строк датафрейма

Что хотим сделать?

Интерпретация модели

Выберите модель

Сегмент 1 (25-31 Ж)

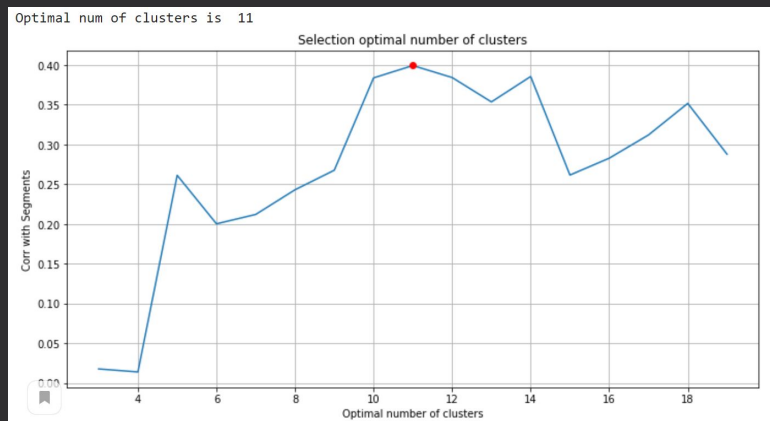
Assessing feature importance based on Shap values

gamecategory

hour



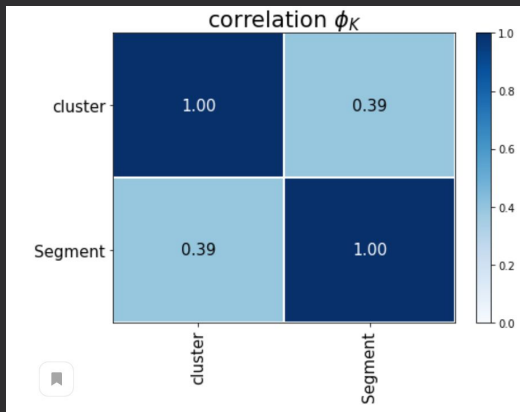
Модель кластеризации



KMeans, 11 кластеров
на признаках hour, dayofweek, day, oblast

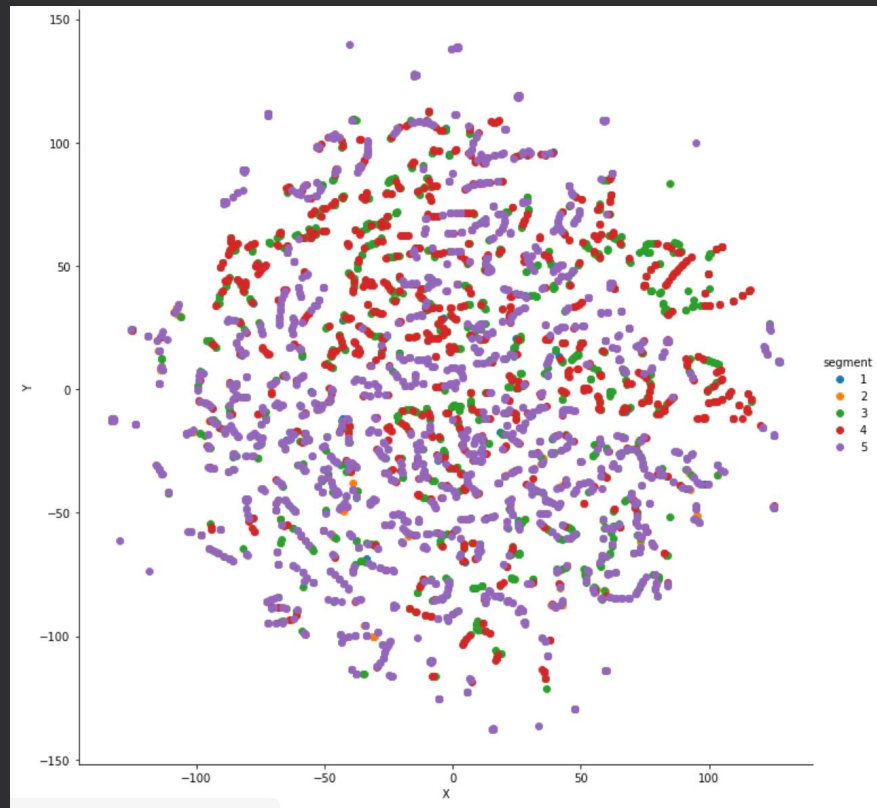
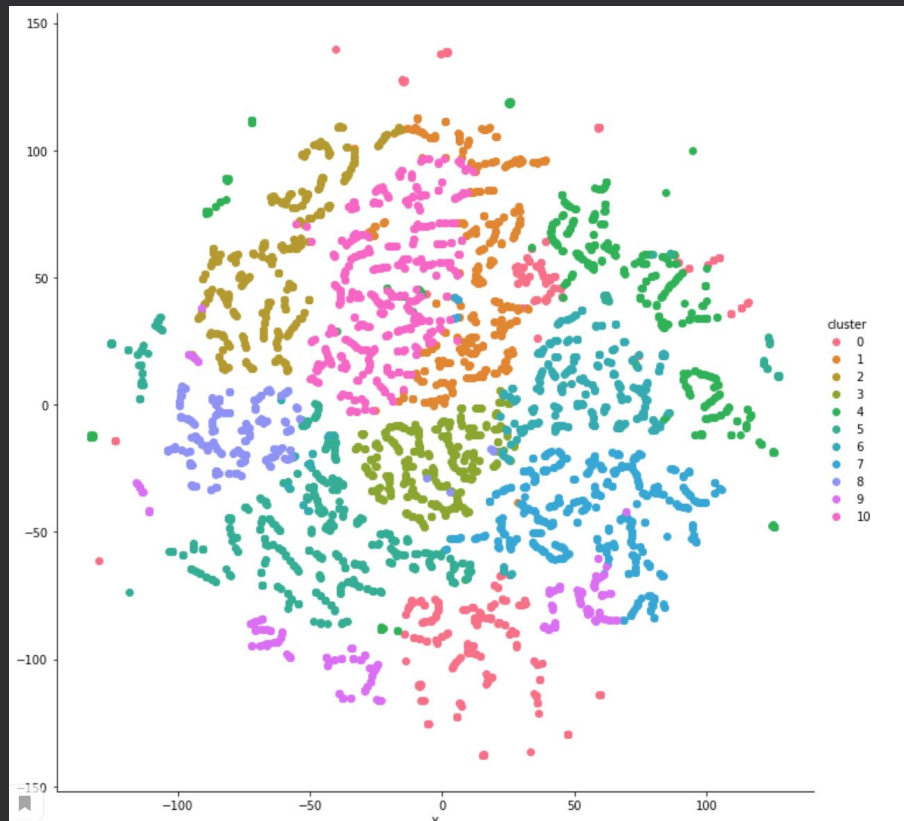
Результат кластеризации показывает
корреляцию $\Phi_K > 0.39$ с размеченными
сегментами.

Есть смысл исследовать дальше и попробовать
использовать результаты кластеризации как
дополнительный признак модели
классификации



* Φ_K is a new and practical correlation coefficient based on
several refinements to Pearson's hypothesis test of independence of
two variables. <https://phik.readthedocs.io/en/latest/>

Визуализация кластеров и сегментов

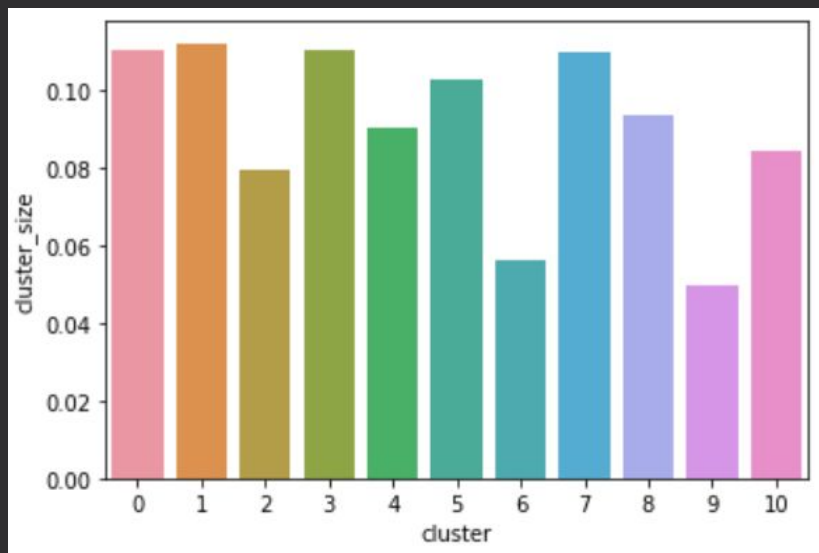


Процент охвата по сегментам

Процент каждого сегмента в выявленных кластерах

Segment	1	2	3	4	5	sum_row
cluster						
0	2.09	0.02	28.77	27.67	41.44	38898.0
1	2.25	0.00	27.51	28.54	41.70	45972.0
2	2.69	0.00	31.31	32.46	33.54	40773.0
3	2.17	0.00	29.87	31.84	36.12	36392.0
4	2.45	0.05	39.75	33.09	24.67	46486.0
5	1.28	8.96	28.35	11.75	49.66	54989.0
6	2.48	0.41	33.50	29.81	33.80	37277.0
7	1.48	9.22	28.91	11.90	48.48	50382.0
8	1.80	9.42	29.82	12.79	46.17	21705.0
9	1.46	9.53	29.75	11.46	47.80	22314.0
10	2.52	0.00	37.49	34.05	25.94	53357.0

Процентное распределение сегментов в кластерах и доли каждого кластера



Программные инструменты

- Jupyter notebook, Python, Scikit-learn



Машинное обучение

- Shap (отбор значимых признаков)
- CatBoost (модель классификации)



Веб-интерфейс для демонстрации:

- Streamlit



Дальнейшее развитие



Для промышленного использования предлагается

1. провалидировать модель на других наборах данных (чтобы исключить переобучение на специфику выборки, на которой обучались),
2. попробовать ещё улучшить её качество признаками полученными unsupervised обучением (кластеризация)
3. “упаковать” обученную модель в веб-сервис, который будет по входящим данным возвращать предсказания.

Стек: Docker + FastApi + Catboost

Оценка реализации :

4 месяца,

1 млн 200 тыс. руб

Команда

Пермь



Data Science

Олег Черемисин

Москва



Data Science

Альбина
Ахметгареева

Москва



Data Science

Дима Васькин