getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Трек: Предсказание конверсий пользователей в рекламном аукционе, используя 3rd-party данные

Команда: «Дайте два»

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Задача

Подготовьте модель машинного обучения, задача которой — предсказать, совершит ли пользователь покупку после клика по рекламе (post-click конверсия). Для улучшения точности прогнозирования вам предоставлены данные о предыдущих рекламных показах и конверсиях, а также информация о покупках, произведенных через сторонние платформы (3rd-party конверсии). Модель должна отличать потенциальных покупателей от тех, кто не совершит покупку, с учетом времени, прошедшего с момента показа рекламы.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Описание продукта

На основе предоставленного датасета «train_views.parquet», состоящего из показов рекламы различным пользователям определить вероятность совершения «post-click» конверсии. Датасет содержит набор признаков, которые могут повлиять на то, совершит ли пользователь конверсию или нет. Также имеется датасет «train_actions.parquet», в котором перечислены действия, которые пользователи совершили после показа рекламы.
В дополнение есть датасет «third_party_conversions.parquet». В нем собраны «third-party» конверсии. «Third-party» конверсии - это конверсии, которые были совершены пользователями со сторонних источников.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Преимущества

Технические требования и ограничения:
- Использование только open-source библиотек.
- Проект укладывается в 120 Гб SSD, 16 Gb RAM.
- Длительность выполнения инференса ограничена и составляет максимум 1 час.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Архитектура продукта

В качестве инструмента предсказания конверсий пользователей выбран классификатор CatBoostClassifier.
После анализа данных выбрана стратегия заполнения пропусков в данных:

- Поле «user_id» заполнено значением «bid_ip».
- При объеме пропусков менее 20% числовых колонки заполнены средним значением, текстовые колонки – модой.
- При объеме пропусков более 20% числовые колонки заполнены значением -127, текстовые колонки «-127».
- Колонки, содержащие идентификаторы преобразованы в категориальные признаки.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Архитектура продукта

После серии экспериментов были подобраны:

- Параметры классификатора CatBoostClassifier.
- Размер валидационной выборки.
- Создан признак для стратификации пользователей на основе частотной встречаемости признака «user_id» и метки класса. В результате получилось 36 групп пользователей. Все записи о пользователе попадали в обучающую или валидационную выборки.
- Итоговый результат был получен при обучении модели на 4-х фолдах с усреднением предсказаний моделей на тестовом датасете.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Перспективы улучшения решения

Нашей командой прорабатывалась двухуровневая модель решения:

1. На первом этапе находились столбцы, которые одновременно были в файле 'third_party_conversions.parquet' и 'train_views.parquet'.
2. После чего по файлу 'third_party_conversions.parquet' обучался многоклассовый классификатор (модель 1-го уровня), в качестве «таргета» использовался столбец 'conversion_name'.
3. Далее модель 1-го уровня, делала предсказания по файлу 'train_views.parquet'.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Перспективы улучшения решения

4. Предсказанные значения отношения каждой строки файла 'train_views.parquet', образовывали новые признаки (15 шт. по количеству уникальных 'conversion_name'), которые подавались для обучения модели второго уровня (наряду с остальными признаками).
При использовании двухуровневой модели удалось получить метрику на ЛБ равную 0.9182.

getintent Газпром-медиа

GPM 30.03 2024 ADTECH CHALLENGE

Безопасность архитектуры приложения

Архитектура безопасности приложения — это набор принципов, политик, процедур и технологий, которые обеспечивают безопасность информации и систем, используемых для разработки, тестирования и эксплуатации приложений:
Защита данных.
Контроль доступа.
Мониторинг активности.
Резервное копирование данных.
Обновление программного обеспечения.
Обучение персонала.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

CJM пользователя продукта

CJM пользователя продукта состоит из нескольких этапов:
Исследование потребностей клиента.
Анализ поведения клиента.
Создание карты путешествия клиента.
Определение ключевых моментов взаимодействия.
Разработка стратегии улучшения опыта использования продукта или услуги.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Любая дополнительная информация

- Для столбцов «user_segments» и «content_category» была опробована векторизация CountVectorizer и TfidfVectorizer с разным количеством выходных признаков. Оба типа векторизации не принесли прироста метрики.
- В итоге столбец «user_segments» был удален, т.к. без него модель лучше обучалась.
- Была попытка передать в CatBoostClassifier столбец «content_category» как text_features, которая тоже не принесла результата.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Любая дополнительная информация

- На предобработку данных затрачивается примерно 6-8 минут, в зависимости от используемых признаков.
- Обучение модели выполнялось на GPU RTX3060 12G.
- Среднее время обучения составило 15-20 минут, т.е. на обучение модели на 4-х фолдах происходит за час-полтора.
- В среднем классификатор обучался за 2000-3000 итераций.
- За время опытов было обучено более 300 моделей.

getintent Газпром_медиа

GPM 30.03 2024 ADTECH CHALLENGE

Демонстрация продукта

Демонстрация продукта — процесс представления продукта потенциальным клиентам или партнерам с целью показать его функциональность, преимущества и уникальность.
Для демонстрации необходимо учитывать несколько важных моментов:
Перед началом демонстрации нужно тщательно изучить продукт, его особенности и возможности.
Важно определить, кому будет интересен продукт и какую проблему он решает.
Убедитесь, что все необходимые технические средства работают корректно.
Старайтесь поддерживать контакт со зрителями, задавайте им вопросы и отвечайте на их комментарии.