IASA DATA SCIENCE CHAMP

ЗАВДАННЯ

відбіркового етапу

GENESIS

1. Про завдання

Для кожної компанії розуміння своїх доходів є дуже важливим щодо вирахування ризиків, планування майбутніх інвестицій, визначення маркетингових бюджетів і бюджетів на інфраструктуру. Усе це має прямий вплив на те, скільки грошей у перспективі може принести продукт, і на вашу привабливість для інвесторів. У холдингу Genesis кожна компанія від самого заснування розв'язує це питання, використовуючи різні інструменти.
На чемпіонаті ваша команда спеціалістів Data Science має виконати прогнозування доходів від користувачів. Крім цього, ваш CEO хотів би також побачити робастну інтероперабельну модель, яка буде готова до використання командою продукту та маркетингу.

Про продукт:

Дані, які ви будете використовувати, взяті з реального проєкту компанії — мобільної гри для платформ iOS і Android. У ній ви можете знайти книги у вигляді візуальних новел. Кожна містить у собі частини, де гравець має робити вибір, що впливає на сюжет і кінцівку.
У грі є дві валюти, що можуть бути отримані за щоденні челенджі або придбані за реальні гроші: кристали, які використовуються для відкриття платних виборів у сюжеті, і тікети для відкриття наступних частин книги (перша завжди є безплатною).
Крім того, користувач може придбати підписку, яка дає нескінченну кількість ресурсів на певний час.
Для гравців, які цим не скористалися, на різних етапах гри може транслюватися реклама, яка допомагає їх монетизувати.

2. Змінні

Кожен рядок в датасеті є унікальним користувачем зі своїми метриками.

Ключові змінні:

target_full_ltv_day30 є головною змінною в цій задачі та являє собою сумарні доходи з користувача на тридцятий день його життя. Саме її ви маєте прогнозувати, і саме по ній ми будемо оцінювати точність вашої кінцевої моделі.
Крім змінної, вказаної вище, ви також отримаєте 3 додаткових, які в сумі дорівнюють ключовій. Тобто **target_full_ltv_day30 = target_sub_ltv_day30 + target_iap_ltv_day30 + target_ad_ltv_day30**, де

- **target_sub_ltv_day30** – доходи з підписок;
- **target_iap_ltv_day30** – доходи з покупки кристалів і тікетів;
- **target_ad_ltv_day30** – доходи з реклами.

Важливо!
Прогнозувати окремо кожну з метрик можна, але використовувати їх як вхідні дані для прогнозування інших (або навіть ключової **target_full_ltv_day30**) – ні.

«Фічі» (або метрики) користувачів:

Усі метрики, описані нижче, можна використовувати для побудови моделей прогнозу. Нотація **{N}** означає конкретний день або сесію користувача. Наприклад, метрика **total_sessions** має в цьому датасеті 4 значення: '**total_sessions_day0**', '**total_sessions_day1**', '**total_sessions_day3**', '**total_sessions_day7**'.
Перелік метрик:

- **total_sessions_day{N}** – кумулятивна кількість сесій користувача на день N;
- **chapters_finished_day{N}** – кількість унікальних глав книги, які було закінчено за N днів;
- **chapters_opened_day{N}** – кількість унікальних глав, які було почато за N днів;
- **chapters_closed_day{N}** – кількість унікальних глав книги, які було закрито за N днів (закриття глави не означає її закінчення);
- **diamonds_received_day{N}** – кількість кристалів, які користувач купив або отримав за N днів;
- **diamonds_spent_day{N}** – кількість кристалів, які було витрачено за N днів;
- **tickets_spent_day{N}** – кількість тікетів, які було витрачено за N днів;
- **retained_day{N}** – булева змінна, означає повернення користувача в гру на день N;
- **app_sub_ltv_day{N}** – кумулятивні доходи з підписок на день N;
- **app_iap_ltv_day{N}** – кумулятивні доходи з купівлі кристалів або тікетів на день N;
- **ad_ltv_day{N}** – кумулятивні доходи з реклами на день N;
- **chapters_finished_session{N}** – кількість унікальних глав книги, які було закінчено за N сесій;
- **chapters_opened_session{N}** – кількість унікальних глав, які було відкрито за N сесій;
- **chapters_closed_session{N}** – кількість унікальних глав, які було закрито за N сесій;
- **diamonds_spent_session{N}** – кількість кристалів, які було витрачено за N сесій;
- **tickets_spent_session{N}** – кількість тікетів, які було витрачено за N сесій;
- **media_source** – закодоване джерело трафіку, з якого прийшов користувач (Facebook, Google тощо);
- **install_date** – дата встановлення гри;
- **country_code** – код країни з якої прийшов користувач;
- **platform** – платформа на яку була встановлена гра (iOS/Android).

Важливо!
Для підрахунку кожної метрики, яка включає день, використовується повний день життя юзера, а не календарний день. Тобто, якщо юзер встановив гру 12.02.2021 о 13:00, то його перший день життя закінчується 13.02.2021 о 13:00.
Крім того, майте на увазі, що дні життя користувачів рахуються, починаючи від дня встановлення гри, а не від часу першого входу до неї.

3. Результати

Що ми чекаємо в результаті?

Посилання на відкритий гіт репозиторій із наступними складовими:
1. Jupyter notebook з EDA процесом і тренуванням моделі.
2. Jupyter notebook з готовим до використання алгоритмом і поясненням вибору основної метрики.
3. Файли серіалізованих моделей.
4. Інші дані за вашим бажанням.

Важливо!
У разі проблем зі зберіганням великих обсягів даних у закритих репозиторіях GitHub, рекомендуємо звернутися до GitLab.

Здача розв'язку завдання відбувається за допомогою форми.

Важливо!
Опис вашого розв'язання задачі є дуже важливим на цьому етапі, оскільки формального захисту роботи перед суддями тут немає. Будь ласка, деталізуйте всі ваші кроки і прийняті рішення. Додаткові візуалізації вітаються.
Зміни в репозиторій дозволяється вносити тільки до останньої хвилини здачі. Команди дискваліфікуються, якщо вони будуть правити щось у ньому після закінчення строку здачі завдання, а також якщо будуть помічені за плагіатом розв'язку інших команд. Для того, щоб такого не відбувалось, рекомендуємо створити закритий репозиторій і після закінчення строку здачі або за 30 хвилин до цього відкрити його й перевіряти, чи доступний він іншим.

Оцінювання

Для оцінювання точностей моделей ми пропонуємо використовувати метрики MAPE, RMSE, MAE. Кожна з них має свою природу, і їх не можна використовувати, замінюючи одна одну. Яку оберете ви як основну і чому?
Оцінювання задачі буде проходити за такими критеріями:
1. EDA і робота з даними.
2. Вибір і тренування моделі.
3. Аналіз моделі та інтерпретація результатів.
4. Метрики ваших моделей (MAPE, RMSE, MAE, — просимо включити кожну).

Дедлайн здачі виконаного завдання — 11.09.2022, 23:59.