Завдання кандидату

Загальні рекомендації до виконання завдань:

- зберігайте шлях виконання завдання (формули, код). В першу чергу цікавить, яким чином відбувалось розв`язання, а не отриманий результат;

- розв`язуйте завдання таким, як Ви його зрозуміли. Цілком можливо, що не всі формулювання є чіткими і однозначними. Тим не менше, додаткові питання не приймаються;

- ми не очікуємо, що Ви розв`яжете всі завдання. Зробіть стільки, скільки зможете/встигнете. Можливо, Ваші конкуренти зроблять ще менше.

Дане завдання в тій чи інші мірі дасть відповідь на три запитання – що Ви знаєте, чи здатні швидко вчитися та який у Вас рівень мотивації.

MS Excel

1. Створити інтерактивну діаграму за прикладом https://www.planetaexcel.ru/techniques/4/42/
   Де відобразити кількості виданих кредитів за липень – грудень 2017 року в розрізі віку позичальників (поділити на три групи 18-19 років, 20-25 років, більше 25 років) (використовувати файл appl_data.csv).
2. Показати ризик по регіонам проживання (колонка region у файлі appl_data.csv) позичальників в динаміці по місяцям, тобто створити таблицю, де рядки - регіони, колонки – місяці, ризик – частка позичальників, що не виконала зобов`язання (“bad” в колонці df файлу is_default.csv). Обчислення проводити за допомогою формул, формули не стирати.

SQL

1. Розв’язати задачі № 45, 62, 64, 93, 96, 130, на сайті http://www.sql-ex.ru. Для перевірки надіслати логін та пароль.

Теорія ймовірностей

1. Який середній дохід (виручка) з 1-го «поганого» клієнта, якщо:
   - середня сума першого кредиту 1500 грн
   - середня сума повторного кредиту 1000 грн
   - один клієнт після користування першим кредитом в середньому бере 6 повторних кредитів
   - середній дохід на 1 грн першого кредиту 0.2 грн
   - середній дохід на 1 грн повторного кредиту 0.4 грн
   - ризик неповернення першого кредиту 30%
   - ризик неповернення повторного кредиту 10%.
   Поганий клієнт – клієнт, що не повернув хоча б один з 7 кредитів.
2. Який середній розмір втрат з 1-го «поганого» клієнта при тих же умовах? За можливі витрати беремо лише розмір виданого кредиту.
3. Масштабувати розрахунки для будь-яких можливих значень показників з п.1. (тобто змінюємо кількість повторних кредитів з 6 на 10 – отримуємо відповідь без корегування формул).

Надати формули та відповіді.

Machine Learning

1. Дослідити взаємозв`язки між даними за допомогою EDA.
2. Побудувати модель прогнозування дефолту клієнта на аплікаційних та поведінкових даних за допомогою логістичної регресії, спрогнозувати ймовірність не виконання зобов`язань перед компанією.
3. Те саме, за допомогою моделі на вибір.
4. Надати код на R або Jupiter Notebook, якщо завдання виконувалось на Python.

Для дослідження надано три файли:

appl_data.csv, аплікаційні дані клієнта. Категорійні дані закодовані. Розшифровка колонок:

appl_id – унікальний номер заявки

app_crtime – час та дата створення заявки

client_id – унікальний номер клієнта

birth – дата народження клієнта

gender – стать клієнта

pass_bdate – дата видачі паспорту

fam_status – сімейний статус

quantity_child – кількість дітей

max_age_child – вік старшої дитини

property – майно у власності

lived_since – з якої дати клієнт проживає за місцем проживання

is_same_reg_lived_since – дата реєстрації за місцем реєстрації

region – область проживання

region_reg – область реєстрації

jobworksince – з якої дати працює на останньому місці праці

work_experience – досвід роботи

empl_state – тип зайнятості

empl_type – галузь праці

empl_worker_count – кількість працівників на місці праці (згруповано)

education_area – галузь освіти

education – тип освіти

monthlyincome – місячний дохід

monthlycost – місячні витрати

---

behave_on_site.csv, поведінка клієнта на сайті. Розшифровка колонок:

device_id – унікальний номер пристрою, з якого клієнт відвідує вебсайт

client_id – унікальний номер клієнта

browser – браузер, який використовує клієнт

platform – операційна система, яку використовує клієнт

create_time – час та дата відвідування конкретної сторінки на сайті

id_ref – закодована назва сторінки на сайті (посилання), яку відвідав клієнт в час create_time

---

is_default.csv, цільова функція, чи виконав клієнт зобов`язання перед компанією, чи ні. Розшифровка колонок:

appl_id – унікальний номер заявки

df – ознака дефолту, bad – поганий клієнт, good – хороший клієнт. Ознака проставлена лише для частини клієнтів, для решти потрібно спрогнозувати ймовірність невиконання зобов`язань.