Лабораторна робота №2 “EDA and Supervised Learning”

Мета: навчитись робити розвідку даних (Exploratory Data Analysis - EDA) та використовувати інструменти для моделювання при навчанні з вчителем (Supervised Learning).

Завдання:

1. Прочитати опис та викачати дані з Kaggle змагання https://www.kaggle.com/c/home-credit-default-risk/data. Опис колонок знаходиться у файлі HomeCredit_columns_description.csv.
2. Для лабораторної роботи достатньо працювати лише з таблицями “application_{train|test}.csv”. З’єднувати інші таблиці не потрібно.
3. Провести EDA на таблицях “application_{train|test}.csv”: про що ці дані, які розподіли колонок, скільки missing values, наскільки збалансовані дані, яким чином обробити missing values, чим відрізняються train/test, ...?
4. Побудувати класифікатор на основі проаналізованих даних. Дозволяється використати будь-який з відомих Вам методів (наприклад за допомогою https://scikit-learn.org/stable/supervised_learning.html). Для оцінки якості натренованої моделі використати train-val-split, тобто відкласти частину “application_train.csv” для валідації. Метрика - ROC_AUC.
5. Класифікувати дані з файла application_test.csv, записати результат в submission файл за прикладом “sample_submission.csv”.
6. Залити submission файл на платформу в “Leaderboard”->“Late Submission”, отримати результат - оцінку, зробити знімок екрану і додати його до фінального ноутбука.

Лабораторна повинна бути оформлена у вигляді Jupyter Notebook, який можна виконати локально, або Kaggle Notebook, або Google Collab. У всіх випадках для здачі потрібно надіслати посилання на ноутбук на GitHub/Kaggle/Collab.

Теоретичний базис для виконання лабораторної - **лекції №1-№6**, але можна використати і методи з наступних лекцій - **№7, №10**.