Лабораторная работа №2 — Логистическая регрессия

Дано
датасет LR_valentine.csv, содержащий следующие данные о людях
- name — имя
- age - возраст
- gender — пол
- income — годовой доход
- appearance score — уровень внешней привлекательности по мнению окружающих
- interests score — уровень показывающий насколько интересы человека совпадают с интересами окружающих
- education status — уровень образования
- job type — тип работы, которую выполняет человек
- valentine date — флаг, показывающий, ходил ли человек на свидание в День Святого Валентина

Задание
1) Какое поле нерелевантно для анализа? Выявить категориальные поля. Какие из них имеет смысл (и каким образом) перевести в численные?
2) Проанализировать распределения фичей. При необходимости избавиться от выбросов
3) Проанализировать веса классов. Исходя из этого, подобрать оптимальную развесовку классов.
4) Разбить датасет в соотношении 80% - тренировочный, 20% - тестовый
5) На тестовых данных: спрогнозировать вероятность свидания, натренировав логистическую регрессию на численных данных, используя веса классов, подобранные выше
6) Рассчитать ключевые характеристики классификации на тренировочных и на тестовых данных.
7) Какая из метрик precision / recall / accuracy / f1 является по вашему мнению самой релевантной в данной задаче?
8) Проанализировать важность фичей с помощью permutation importance.
9) [BONUS] Улучшить предсказательную силу, добавив категориальные переменные

Система оценивания
Задание будет оценено по 5-балльной шкале.
За позднюю сдачу задания будет снят 1 балл.
Выполнение бонусной части опционально.

Формат выполнения
.ipynb файл, без ошибок прогоняемый сверху вниз с любого компьютера
Файл выслать преподавателю посредством мессенджера Telegram : @evgeny_zavalnyuk

Срок выполнения [время фиксируется исходя из отправки решения преподавателю]
19 апреля 2024г. до 23:59:59 МСК