Выполнил студент НИУ ВШЭ МИЭМ
3 курса образовательной программы
Прикладаная математика -
Серебров Борис Алексеевич

Отчет по проекту: №725 Предсказание пола пользователей соцсетей

Цель: Обучить модель машинного обучения для предсказания пола пользователей соц. сетей по определенным признакам.

Задачи:

1. Изучение данных, выделение целевой переменной и основных признаков.
2. Предобработка данных, сочетание всех представленных данных в месте и преобразование их для дальнейшего обучения модели..
3. Изучение зависимостей и разделение данных на более явные и лучше интерпретируемые.
4. Обучение модели.
5. Проверка результатов обучения.
6. Предсказание пола для данного списка пользователей.

Реализация:

1. **Изучение исходных данных.** Были исследованы следующие файлы с исходной информацией:
    - Наборы train.csv и train_labels.csv, содержащие данные для обучения
    - Файлы test.csv и test_users.csv с тестовыми данными
    - Дополнительные данные: referer_vectors.csv (векторные представления URL) и geo_info.csv (географические параметры)
    - Установлены взаимосвязи между таблицами через ключевые поля: user_id, referer и geo_id
2. **Генерация признаков.** Выполнено преобразование исходных данных в признаки:
    - Временные характеристики: день месяца, часть суток
    - Анализ URL: выделение домена, пути
    - Разбор user-agent: определение браузера, ОС.
    - Геоданные: идентификаторы страны, региона, временная зона
    - Векторные компоненты component0 - component9 из referer_vectors
3. **Формирование наборов данных.** Создан объединенный датасет с меткой тестовых записей (test).
    - Получен df_comb с уникальными записями пользователей. Обучающая выборка X_train сформирована объединением с train_labels.
4. **Построение и обучение модели.** Для решения задачи классификации использован алгоритм градиентного бустинга по решающим деревьям CatBoostClassifier. Основные гиперпараметры модели:
    a. iterations=2000 — количество деревьев в ансамбле;
    b. learning_rate=0.03 — скорость обучения.

c. depth=10 — максимальная глубина деревьев;
d. eval_metric='Accuracy' — метрика качества для мониторинга во время обучения;
e. random_seed=81 — фиксированное зерно генератора случайных чисел;
f. verbose=80 — вывод прогресса каждые 80 итераций.

Модель была обучена на тренировочной выборке (X_train, y_train).

5. **Предсказание и оценка качества.** Предсказание выполнено на тестовой выборке (test). Качество работы модели оценивалось с использованием следующих метрик:
   a. Accuracy — доля верных предсказаний;
   b. AUC — площадь под ROC-кривой;
   c. classification_report — подробный отчёт с precision, recall и F1-score для каждого класса.
6. **Итоговый результат.** Обучена модель машинного обучения и сохранена в файл mdl_catboost.joblib; Сохранены результаты предсказания в файл submission.csv; Весь код предобработки и обучения в файле jupiter Notebook main.ipynb.