Драфт проекта python.

Тема:

Hotel Reservations Dataset

“Can you predict if customer is going to cancel the reservation?”

Краткое описание проекта.

Цель проекта заключается в предсказании отмены бронирования отеля, используя данные из набора данных "Hotel Reservations Dataset" на платформе Kaggle. Этот проект может быть полезен для отелей и других компаний в сфере гостеприимства, чтобы оптимизировать свою деятельность и повысить удовлетворенность клиентов.

В рамках проекта я проведу базовый анализ данных (EDA), включающий в себя описательные статистики и построение графиков, чтобы понять, какие признаки влияют на отмену бронирования. (Data engineering + feature engineering)

Затем обучим не менее трех различных моделей машинного обучения для предсказания отмены бронирования. Одной из основных метрик для оценки качества модели будет полнота (recall). Это связано с тем, что ложные отрицательные результаты (то есть случаи, когда отмена бронирования не была предсказана, хотя на самом деле произошла) могут привести к серьезным проблемам для бизнеса в сфере гостеприимства, таким как потеря прибыли и ухудшение репутации компании. Поэтому важно, чтобы модель максимизировала полноту, т.е. минимизировала число ложных отрицательных результатов. Кроме того, также можно использовать F1-score, который учитывает и точность, и полноту, а также accuracy для общей оценки качества модели.

В заключение сделаем выводы о полученных скорах моделей, EDA и графиках, и предоставим рекомендации для оптимизации процесса бронирования и уменьшения отмены бронирования в будущем. Кроме того, попытаемся улучшить скор любой из обученных моделей путем перебора гиперпараметров.

Примерная структура.

1) Загрузка данных из датасета с помощью библиотеки Pandas.
2) Предварительный анализ данных:
    - Распределение классов в целевой переменной (отмена бронирования)
    - Анализ пропущенных значений
    - Базовые статистики по признакам
3) Визуализация данных:

- Построение диаграмм рассеяния для корреляционного анализа признаков
- Построение гистограмм и ящиков с усами для оценки распределения признаков
- Построение столбчатых диаграмм для категориальных признаков

4) Предобработка данных:
    - Обработка пропущенных значений
    - Кодирование категориальных признаков
    - Масштабирование признаков
    - Разделение данных на обучающую и тестовую выборки
5) Обучение моделей:
    - Обучение трех различных моделей, например, логистической регрессии, решающего дерева и случайного леса
    - Оценка качества моделей на обучающей и тестовой выборках с использованием выбранных метрик (например, полноты, F1-score и accuracy)
6) Подбор гиперпараметров модели (бонус):
    - Подбор оптимальных гиперпараметров для одной из обученных моделей с использованием кросс-валидации и поиском по сетке
    - Оценка качества модели с подобранными гиперпараметрами
7) Выводы:
    - Сравнение качества трех моделей и выбор наилучшей модели для предсказания отмены бронирования отеля
    - Выводы по результатам предварительного анализа данных и визуализации
    - Рекомендации для бизнеса на основе полученных результатов.

Дополнительные инструменты.

Pandas – для работы с данными и объектами типа dataframe.

Matplotlib – для простой и быстрой визуализации.

Seaborn – для красивой визуализации.

Scikit-learn – библиотека машинного обучения предназначенная для обучения с учителем и не очень хорошо работает в обучении без учителя.

Numpy – для работы с многомерными массивами.