Добрый день, Андрей Иванович!
Практически по всем пунктам критериев Вы отлично справились с заданием, замечательная и красивая работа! :)

Требования к оформлению:
+ структура оформления ноутбука: всё последовательно разделено на логические части, описаны цели и задачи, каждый этап работы, подробные сопровождающие пояснения и содержательные выводы о проделанной работе;
+ общие правила оформления: читаемый понятный код с грамотными названиями функций и переменных, графики построены по всем правилам визуализации, соблюдение стандартов PEP-8, отформатированные выводы в отдельных ячейках типа MarkDown;
+ общие правила оформления: читаемый понятный код с грамотными названиями функций и переменных, графики построены по всем правилам визуализации, соблюдение стандартов PEP-8, отформатированные выводы в отдельных ячейках типа MarkDown.

Анализ и обработка данных:

> Отлично оформили этап первичного исследования данных: здорово, что посмотрели на наличие пропусков, дубликатов и выбросов.
> Здорово, что построили графики распределений признаков, упростить этот процесс может использование функции pairplot. Плюс за карту корреляций, это очень полезный и информативный график, который может дать возможность вовремя детектировать мультиколлинеарность признаков и избежать переобучения.
> Можете ещё потестировать статические гипотезы (самое простое - проверить какие-нибудь некоторые распределения на нормальность).
> Предобработка данных во feature engineering замечательная; единственное - можно попытаться автоматизировать процесс обработки данных, загнав все уникальные значения в список, и затем обрабатывать их в цикле.
> Также для увеличения качества модели можно попробовать убрать шум в виде лишних признаков, сделав их отбор это может дать небольшой прирост в метриках.
> Качественно сделали очистку данных и непосредственную подготовку их на вход модели, здесь нечего добавить.

Применение ML и DL:
Здорово, что последовательно “по всем правилам” протестировали все известные методы от простых к сложным. Зачастую случается так, что классические алгоритмы для некоторых задач показывают гораздо лучшие метрики - но, как видим, не всегда:)
Как вариант, можете ещё попробовать использовать CatBoost - это сейчас самый популярный алгоритм, дающий, как правило, метрики чуть лучше, чем другие бустинги.
Также можете попробовать инструменты подбора гиперпараметров для моделей обучения с учителем - самая популярная библиотека для этого - optuna, из примеров в материалах курса можете посмотреть GridSearchCV.
Кластеризация тоже получилась достаточно успешной. В качестве модели можете посмотреть DBSCAN - она нечувствительна к выбросам, поэтому иногда может работать лучше.

Отдельно хочется отметить:
Хорошо, что пояснили весь код комментариями.
Отличная реализация варианта решения в продакшене в docker.
Серьёзно подошли к выполнению проекта, это очень здорово! Ничего не упустили из основного, замечательная работа

Из советов и пожеланий:
Можете также посмотреть про модуль tqdm - с его помощью легко следить за прогрессом выполнения операций в цикле, обучения моделей и применения apply.

Спасибо за выполненное задание! Если возникнут вопросы, можете обратиться ко мне в канал с окончанием 08_final в пачке, постараюсь ответить на все вопросы и разобраться с моментами, которые вызывают трудности. Удачи в обучении!

Проверял ментор

Мария Жарова

zharova.ma@phystech.edu