Итоговый проект по курсу “Машинное обучение” (DS-7)

В итоговом домашнем задании по курсу вам предлагается поучаствовать в одном из соревнований на платформе Kaggle (или платформе Zindi) и описать свои результаты.

Решать задачу можно индивидуально или в небольших группах 2-3 человека.

Шаг 1: зарегистрируйтесь на www.kaggle.com

Шаг 2: выберите соревнование из списка ниже:

- https://www.kaggle.com/c/house-prices-advanced-regression-techniques
- https://www.kaggle.com/c/new-york-city-taxi-fare-prediction
- https://www.kaggle.com/c/tmdb-box-office-prediction
- https://www.kaggle.com/c/ghouls-goblins-and-ghosts-boo/leaderboard
- https://www.kaggle.com/c/whats-cooking/leaderboard
- https://zindi.africa/competitions/expresso-churn-prediction
- любое другое соревнование на ваш выбор (в случае выбора этой опции необходимо согласовать выбранную задачу с преподавателем)

Подтвердите своё участие в соревновании, скачайте данные и начните работать над задачей.

Шаг 3: создайте любой алгоритм, делающий предсказания в данной задаче, сделайте предсказания на тестовых данных и отправьте посылку на Kaggle. Цель этого шага – сделать первую успешную посылку.

После того, как сделаете этот шаг, запишите в текстовый файл краткое описание вашей первой модели (ваш baseline), затем запишите в файл качество, полученное на кросс-валидации и качество, которое вы увидели на leaderboard на Kaggle.

Также напишите, под каким именем искать вас на leaderboard в соревновании.

Шаг 4: поработайте над улучшением модели. Подумайте, какие признаки можно добавить (и добавьте) в данные, как очистить данные от выбросов, попробуйте снизить размерность. Применяйте любые известные вам алгоритмы и методы. Когда новая модель получилась – отправляйте её на Kaggle (но помните, что во многих соревнованиях стоит ограничение на количество посылок в день).

Обязательно применить модели решающих деревьев, случайного леса, градиентного бустинга (из sklearn) и одной из имплементаций бустинга (XGBoost, CatBoost, LightGBM).

Для каждой из этих моделей необходимо по кросс-валидации (GridSearchCV) подобрать оптимальные гиперпараметры.

Другие известные вам модели, а также их смеси, тоже рекомендуется попробовать.

Создайте в вашем текстовом файле таблицу по шаблону:

Номер модели	Краткое описание модели	Качество модели на кросс-валидации	Качество модели на leaderbord
1	…	…	…
2	…	…	…
…	…	…	…

Помните, что нельзя оценивать качество модели только по leaderboard, это ведет к переобучению! Следите и за качеством на leaderbord, и за качеством на кросс-валидации!

Шаг 5: выберите модель, которая с вашей точки зрения лучше всего себя показала в данной задаче (качество на leaderboard высокое, и на кросс-валидации тоже хороший результат). Подробно опишите вашу модель в текстовом файле. Для описания вашей модели вы можете ответить на следующие вопросы:

1) Какая проводилась обработка признаков?
2) Были ли удалены выбросы и как?
3) Были ли заполнены пропуски и как?
4) Какие новые признаки были добавлены (если были)?
5) Было ли проведено снижение размерности и каким образом?
6) Какая модель или какие модели были использованы?

Также можете добавить другую информацию по вашему усмотрению.

Наконец, сделайте скриншот качества выбранной модели в списке ваших посылок на Kaggle и вставьте его в файл с описанием модели (выбранная модель не обязана совпадать с моделью, дающей наилучший результат на leaderboard).

Готовый текстовый файл вместе с jupyter notebook-ом, содержащим наилучшую по вашему мнению модель, отправьте на проверку в anytask. Также необходимо подготовить презентацию с описанием вашего решения.

Дедлайны:

1) **Дедлайн по выбору команды: 9 июня 23:59.**
   Команду (или просто себя, если работаете индивидуально) записывайте в этот файл https://docs.google.com/spreadsheets/d/1Phpyox2qdii0tuVJ8ivRbWUDvv_oJ7AVMI8APDmV-AE/edit?usp=sharing
2) **Дедлайн по отправке решения в anytask: 24 июня 23:59.**
   На последнем занятии 25 июня будем слушать презентации ваших решений (требования к презентации будут выложены позднее).