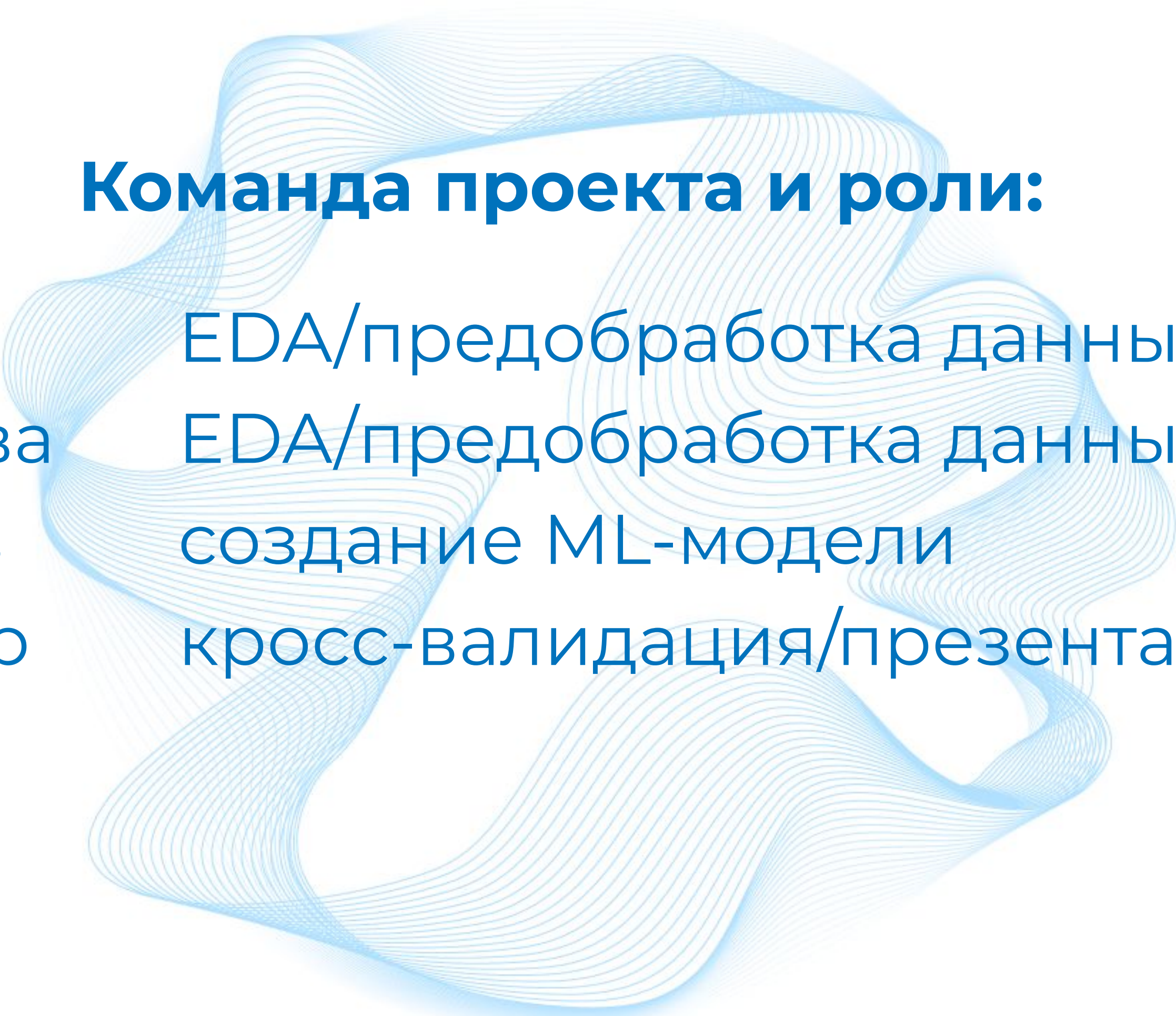


Хакатон (3 семестр)

Проектная задача: ВК (задача 1)
“Предсказание пола пользователей”

Команда проекта и роли:



Комарова Полина	EDA/предобработка данных/документация
Вероника Азмукова	EDA/предобработка данных/документация
Илья Седельников	создание ML-модели
Тухватуллин Динар	кросс-валидация/презентация

EDA (разведывательный анализ данных)

В ходе разведывательного анализа данных выполнены:

- 1 проверка на дисбаланс классов целевой переменной;
- 2 построены гистограммы распределения по странам, регионам, временным зонам, используемым браузерам и ОС;
- 3 построена тепловая карта корреляции признаков;
- 4 построена матрица корреляции признаков с целевой переменной.

Предобработка данных

В ходе предобработки данных выполнены:

- 1 объединение датасетов в единый;
- 2 удаление дублирующих записей;
- 3 удаление записей с пропусками значений;
- 4 удаление неинформативных признаков;
- 5 преобразование данных в требуемый формат;
- 6 нормализация числовых признаков;
- 7 энкодинг категориальных признаков (OneHotEncoder).

Параметры ML модели

В качестве модели выбрана классическая модель RandomForestClassifier. В качестве метрики выбрана точность, результаты приведены в таблице:

№ п.п.	Количество деревьев	Точность на тестовой выборке, %
1	5	78,2
2	10	79,0
3	20	79,6
4	30	79,8
5	50	80,1
6	100	80,2

Кросс-валидация

Для дополнительной оценки качества модели выполнена кросс-валидация методом K-fold (количество фолдов 5). В качестве метрики выбрана точность, результаты приведены в таблице:

№ п.п.	Разбиение по фолдам	Точность на валидационной выборке, %
1	Модель без кросс-валидации	79,8
2	Фолд 1	80,0
3	Фолд 2	79,8
4	Фолд 3	80,1
5	Фолд 4	80,0
6	Фолд 5	79,9

A large, abstract graphic made of many thin, overlapping blue lines that form a complex, wavy, and somewhat circular shape, resembling a stylized wave or a modern logo. It is centered on the page.

СПАСИБО ЗА ВНИМАНИЕ!