Работа велась в Google colab.

В качестве моделей были выбраны логистическая регрессия (логрег) и дерево решений. Выбор обуславливается тем, что в рамках курса ЧелГУ пока что были пройдены только эти модели классификаций.

1. Подготовка датасета. Были добавлены названия признаков, и 2 датасета были объединены в один. Также были проверены пропуски в данных и исследована корреляция между числовыми признаками и целевым.
2. Был лемматизирован текст.
3. Проведена TF-IDF векторизация (токенизация по буквам дала лучший результат, чем по словам). Вся выборка была разбита на учебную и тестовую выборки.
4. При помощи GridSearchCV были найдены оптимальные гиперпараметры для логистической регрессии (логрег) и для дерева решений.
5. Были выведены классификационные отчеты и построены ROC-AUC кривые для оценки качества моделей.

Лучший результат показала логистическая регрессия, что видно по ROC-AUC кривой.