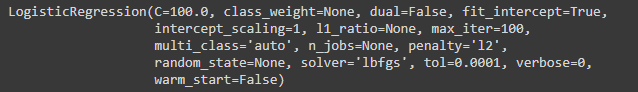
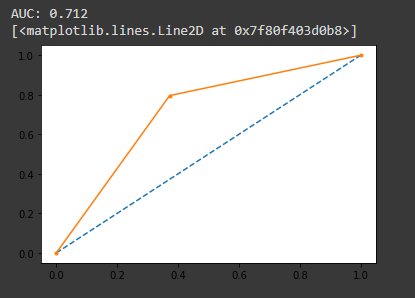
Работа велась в Google colab.

1. Подготовка датасета. Были добавлены названия признаков, и 2 датасета были объединены в один. Также были проверены пропуски в данных.
2. Был лемматизирован текст.
3. Проведена TF-IDF векторизация, данные были разбиты на учебную и тестовую выборки.
4. При помощи GridSearchCV были найдены оптимальные гиперпараметры для логистической регрессии (логрег).



1. Была построена ROC-AUC кривая для оценки качества модели.



Также в задумке было обучить не только логрег, но также попробовать обучить случайный лес и беггинги логрегов и деревьев, и сравнить их результаты. Но для этого не хватило мощности моего компьютера, а google colab имеет свойство, ничего не сказав, крашиться на полпути. Учитывая объем данных, краш после 5 часов компилирования – это очень большая грусть.

Еще хотелось бы добавить, что все знания, используемые в данной работе, я получил за два месяца обучения в челгу, а с текстами начал работать буквально на этой неделе. В университет я пришел с нулевыми знаниями. А в Интерсвязь хочу попасть для получения опыта в прикладном machine learning.