## Модуль 3. Фидбек

#### *Результат модели*

Вам удалось превзойти baseline на валидации - хороший старт. Различие в результате на валидации и на тесте скорее всего объясняется тем, что вы перемешали строки в тестовом датафрейме при merge. Надо проследить, чтобы порядок предсказаний при заполнении файла submission.csv соответсвовал тому, который в файле sample\_submission. Дальнейшее улучшение результата проще всего достичь включением внешних источников данных и постобработкой предсказаний.

*Комментарии*

Для one-hot encoding признаков, содержащих множественные значения можете попробовать в дальнейшем MultiLabelBinarizer,   
<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html>

*Как сделать лучше*

Возможно, стоило, попробовать PCA (либо другое снижение размерности) - может какие-то признаки удалось бы сформировать таким образом - при этом можно как переходить в новое пространство признаков, так и добавлять полученные новые признаки к старым (это делать аккуратно - не совсем математически обосновано, но иногда работает). Возможно стоит попробовать методы feature selection для уменьшения количества используемых признаков (примеры методов можно посмотреть тут <https://scikit-learn.org/stable/modules/feature_selection.html>). Для категориальных признаков, вроде типов кухонь, у которых большое количество различных значений можно было попробовать объединить по какому-то принципу значения в одно, например по географической принадлежности - Asia, European, либо по более широкому типу кухни (национальная, фаст-фуд, и т.д.). Можно было попробовать добавить признаки на основе кластеризации объектов обучающей выборки - значением признака выступало бы отношение объекта к тому или иному кластеру. Также, очевидным направлением дальнейшего улучшения является анализ текстов отзывов на тональность (положительную и отрицательную), при этом не обязательно сразу использовать для этого все state of the art достижения в области natural language processing - возможно даже просто подсчёт количества положительных/отрицательных слов мог бы дать прирост. Кроме того, прирост качества, возможно, было бы получить добавив дополнительные данные (именно строки, а не признаки) - они могут быть получены либо из других датасетов/парсинга, либо (так делать очень аккуратно) сгенерированы искусственно, повторяя распределение обучающей выборки.

Помимо этого, в некоторых задачах (особенно это актуально для улучшения score модели на kaggle) может помочь post processing результатов модели - в данном случае ваш результат был бы гораздо выше, если бы вы округлили предсказания модели до 0.5 - потому что итоговая переменная именно такие значения и принимает.

Отзыв подготовил ментор Леонид Саночкин.  
Если возникнут вопросы, обращайтесь ко мне в канал проекта **#0\_project\_3-о\_вкусной\_и\_здоровой\_пище** в Slack. Постараюсь ответить на ваши вопросы и разобраться с моментами, которые вызывают трудности.

Буду ждать вас на итоговый вебинар по данному проекту, который я провожу по четвергам в 19:00 МСК (<https://us02web.zoom.us/j/82874351759?pwd=WEhGcDliUlVLbUNYM0Nsem9QMGpaZz09>). На нём мы разберём основные ошибки и рассмотрим как улучшить решение. До встречи на вебинаре!

Удачи в обучении!