Модуль 3. Фидбек

Результат модели

Вам удалось превзойти baseline - это хороший старт. Дальнейшее улучшение результата проще всего достичь включением внешних источников данных и постобработкой предсказаний. Для one-hot encoding признаков, содержащих множественные значения можете попробовать в дальнейшем MultiLabelBinarizer
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html
Вероятно по ошибке, вы не используете дополнительные признаки, которые генерировали в начале в функции df_preproc() - из-за этого у вас они не участвуют в итоговом обучении модели.

Как сделать лучше

Возможно, стоило, попробовать PCA (либо другое снижение размерности) - может какие-то признаки удалось бы оттуда достать - 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 некоторые признаки, которые Вы используете не так важны? Можно ли убрать какие-то признаки? Возможно какие-то признаки можно сжать в один? Такие вопросы можно задать себе на досуге и снова вернуться к заданию, либо использовать в дальнейшем - 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 объединять по какому-то принципу значения в одно (например редко встречающиеся кухни выделить в одно значение, либо сделать разделение по географической принадлежности - Asia, European и т.д.) и также включать в модель, например в виде one-hot векторов. Можно было попробовать расширить информацию о городах, добавив признаки численности населения, близости ресторана к центру города или к основным достопримечательностям и т.д. 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. Также 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 в бы 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.

Удачи в обучении!