### Имя студента: Andrew Solovei

### Ник студента в Slack: Andrew Solovei

### Почта студента: a\_solovey@ukr.net

## Модуль 0. Фидбек

#### 

## Модуль 1. Фидбек

## Модуль 2. Фидбек

Отличная работа, отдельно можно отметить качественную работу с категориальными признаками в ключе сглаживания редких значений, а также последовательные и обоснованные выводы по данным.

Оценка согласно критериям:

1. Качество кода и соответствие стандартам оформления *pep-8*.

Полностью соответствует, 3 балла.

1. Для каждого столбца найдено количество пустых строк, найдены и убраны выбросы.

Пропущенные значения найдены во всех столбцах. Корректно обработаны выбросы и ошибки, дополнительно для некоторых переменных объединены некоторые значения в одно, 4 балла.

1. Для количественных переменных построены гистограммы распределений и сделаны выводы.

Требуемый анализ полностью проведен, выводы последовательны и обоснованы, 3 балла

1. Проведён корреляционный анализ и анализ номинативных переменных, корректно отобраны переменные для построения модели.

Корреляционный анализ и t-test выполнены корректно, сформулированы правильные выводы, 3 балла.

1. Сформулированы промежуточные и итоговые выводы.

Сформулированные выводы последовательны и обоснованы, предложены интересные гипотезы по данным, 4 балла.

Замечания/предложения:

* выдвинута интересная гипотеза по поводу пропусков - можно было её проверить имеющимися статистическими инструментами - сформировать две выборки значений score где было бы “много пропусков” и “мало пропусков” и проверить гипотезу о равенстве их средних;
* про образование и работу матери, на мой взгляд спорный вывод, мне не хватило обоснования. Можно, например, утверждать, что отсутствие работы у матери приводит к большему её погружению в процесс образования ребенка, но такой довод также нуждается в проверке на данных;
* абсолютно правильным решением было удалить строки с пропусками в score. Пропуски в признаках так или иначе заполнить теоретически можно, пропуске в оцениваемой переменной заполнять ни в коем случае нельзя. В то же время, иногда такие данные (без известных ответов) могут быть полезны, в некоторых случаях, например на них можно получить начальные веса для нейронной сети (это так - информация на будущее);
* можно предложить также рассмотрение следующих гипотез: различимы ли студенты, набравшие 0 баллов и остальные; различимы ли среди студентов, набравших 0 баллов, те, кто не пришел на экзамен и те, кто его не сдал; имеет ли смысл переход к новой шкале для score (заменить её значения, например на A - {85-100}, B - {70-84} и т.д.)

В заключении, я считаю, что проект подготовлен на высоком уровне, результаты могут быть использованы на практике, надеюсь увидеть такую же качественную работу в дальнейших проектах!

## Модуль 3. Фидбек

#### *Результат модели*

Неплохой результат! Вам удалось превзойти результат baseline модели - это хорошо. Дальнейший прирост качества проще всего было получить за счёт дополнительного использования внешних источников данных и анализа информации об отзывах.

*Как сделать лучше*

Вы не использовали признаки, полученные путём комбинации исходных с помощью алгебраических операций - иногда такие признаки могут сработать. В дополнении к вашим данным о городах, можно было расширить информацию о туризме в этих городах. Возможно, стоило, попробовать PCA - может какие-то признаки удалось бы оттуда достать (либо другое снижение размерности) - 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. Также, очевидным направлением дальнейшего улучшения является анализ отзывов, 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 (полученных также из внешнего источника) мог бы дать прирост. Также скорее всего стоило снизить количество различных типов кухонь за счёт объединения нескольких редких значений в одно.

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.

## Модуль 4. Фидбек

…

## Модуль 5. Фидбек

#### *Результат модели*

Отличный результат, поздравляю!

#### *Комментарии*

В данном случае мне видится перспективнее для валидации использовать кросс валидацию со стратификацией - поскольку если разбивать на фолды случайно, то можно получить фолд, где ещё больше дисбаланс классов, и как предельный случай - что совсем не будет представлен один из классов.

*Как сделать лучше*

В целом, грамотно провели исследование, могу предложить следующие моменты для потенциального улучшения:

* на kaggle часто пользуются такой техникой - для k фолдов кросс-валидации обучаются соответственно k моделей - а на этапе инференса (предсказаний для тестового множества) используются все эти модели вместе, и итоговым значением выступает усредненное предсказание всех k моделей.
* для настройки гиперпараметров можно попробовать в дальнейшем более продвинутые библиотеки типа optuna.
* можно попробовать oversampling и undersampling для балансировки классов.
* как рискованный подход можно попробовать pseudo labeling.
* можно было попробовать другие модели - в правилах соревнования вас ограничивали только запретом использования нейронных сетей, а кроме этого, неплохо в соревнованиях себя проявляют различные ансамбли моделей (stacking, blending, etc.)

Это все небольшие советы, Вы отлично справились с заданием!

## Модуль 6. Фидбек

…

## Модуль 7. Фидбек

…

## Модуль 8. Фидбек

…

## Модуль 9. Фидбек

…

## Модуль 10. Фидбек

…

## Финальный фидбек и общие комментарии

...