

Яндекс



ML Boot Camp IV

С 1 в паблице на 35 в прайвате: кто виноват и можно ли было что-то сделать

Бабушкин Валерий

Постановка задачи и исходные данные

Постановка задачи и исходные данные



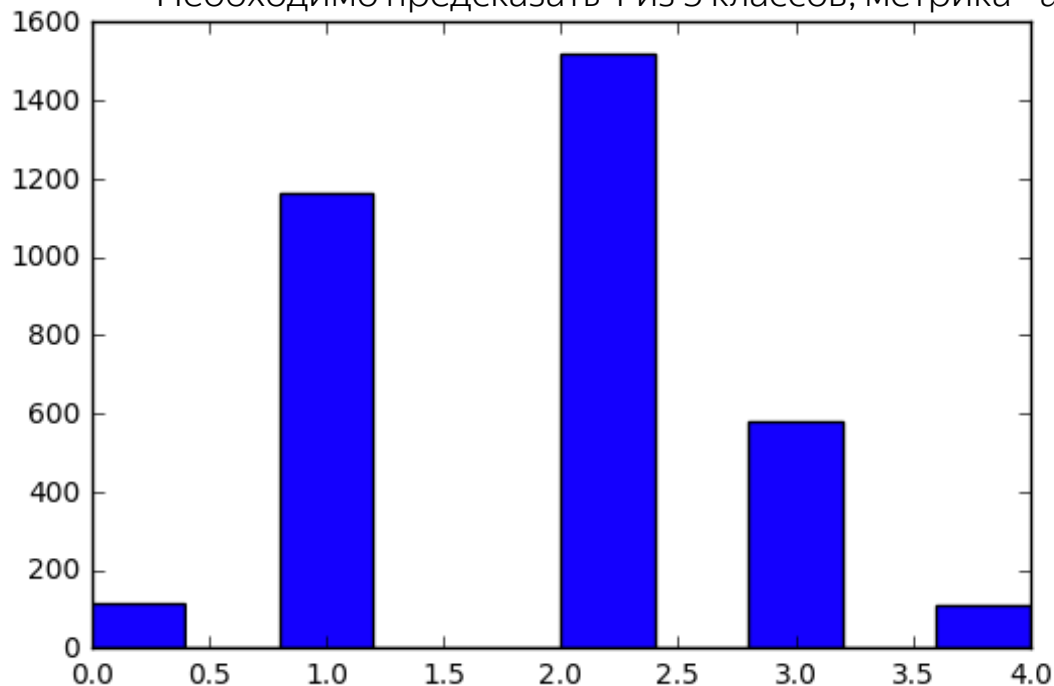
Постановка задачи и исходные данные

Обучающая выборка 3489 x 223

Для предсказания 2327 x 223

Дубликатов и пропусков нет

Необходимо предсказать 1 из 5 классов, метрика - accuracy



Постановка задачи и исходные данные

После объединения получаем массив 5816 x 223

Есть 12 колонок с количеством уникальных значений от 89 до 1425, остальные > 5000

У 10 колонок уникальных значений от 89 до 309

Если посмотреть на среднее то 11 колонок со средним ~0

10 с средним 25 + 1 со средним 17

10 с средним 3

2 с средним 0.9 и 0.8

1 с средним 0.12

188 с средним 0

Построим классификатор (RF) на каждой из фич

На пересечении обнаружатся следующие колонки:

11, 76, 79, 96, 97, 115, 131, 138, 156, 182, 200

Постановка задачи и исходные данные





Базовый классификатор(RF) – на всех фичах ~ 0.55

Базовый классификатор(RF) – на отобранных фичах ~ 0.665 (топ 10 на момент нахождения)

Как оказалось исходный набор данных тоже включал в себя 11 фичей – они были зашумлены

Так же в виде шумов добавили еще 212 признаков

Здесь то и надо было остановиться

		RF_6258_750tr_12_f... 26.04.17 00:44:06 скачать		Статус: ОК	0,6611842	0,6636042
1		Александр Иванов			0,6622807	
2		Святослав Ковалёв			0,6611842	
3		Иван Черданцев			0,6600877	



MULTI-CLASSING

Because wizards run out of spells

Причина падения

Причина падения



Причина падения

Все шло довольно неплохо в последний день соревнований

1



Валерий Бабушкин

0,7010601

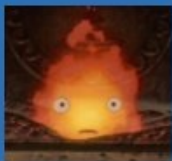
2



Артем Дубинич

0,6925795

3



Александр Киселев

0,6918728

Причина падения

Сгенерированы новые фичи деление, умножение, сложение, вычитание

Сверху добавлена кластеризация

Небольшой стекинг

Постепенно качество на пабlike выросло

Причина падения

Качество на кросс-валидации было довольно нестабильным, даже при 20-кратном усреднении, причина оказалось довольно прозаичной

Если отбросить шумы и еще раз посмотреть на данные, то мы увидим следующее

```
shape of train+test (5816, 377)
shape of train+test without duplicates (4852, 377)
shape of train (3489, 377)
shape of train without duplicates (3118, 377)
shape of test (2327, 377)
shape of test without duplicates (2148, 377)
```

Причина падения

Всего уникальных дублей 450, некоторые записи продублированы 3, 4 и более раз

Алгоритм научился четко видеть дубли между обучающей выборкой и тестовой – этим и вызван наибольший прирост

Если смотреть на вероятности, то довольно много случаев когда 2 класса имеют очень схожую вероятность

Один и тот же алгоритм, при прочих равных, но разном сиде давал разницу в ~100-150 ответов

Причина падения

1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6

Салатовый
Паблик

ассигасу красного 0,733 ассигасу зеленого 0,266

Серый Приват

ассигасу красного 0,266 ассигасу зеленого 0,733

Паблик и приват были перепутаны местами, как результат, разбиение было не 60 на 40, а 40 на 60
В итоге приват состоял из 931 образца с учетом дублей

Можно ли было что-то сделать

Можно ли было что-то
сделать



Можно ли было что-то сделать

Разница между 1-м и 50-м местом – 13 предсказаний, что с учетом дублей сводится к 3-4-5-6 уникальным наблюдениям. На первые 48 мест приходится 12 уникальных результатов

Три стратегии:

Усреднять модели – надежно для попадания в топ 20, но снижает вероятность выиграть

Надеяться на случайный сид – есть случаи неожиданного подъема в топ-5

Придумать киллер фичу

Как быть

Как быть



Как быть

Заранее смотреть на метрику и количество данных

Понимать причины и трезво оценивать шансы

Получать наслаждение

https://github.com/VENHEADs/ML_boot_camp_Secret_task