Tinkoff Data Science Challenge

Тренировка по машинному обучению

Яндекс, 11.03.2017

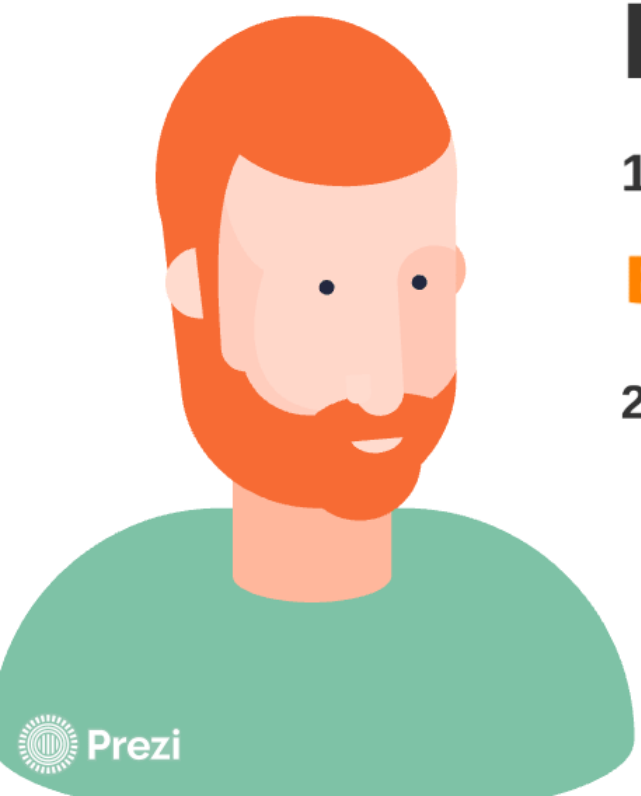
Чернобровов Алексей
к.ф.-м.н.

Jet Retail

Чернобровов Алексей
к.ф.-м.н.

Jet Retail

Релевантный опыт



1. Веб-аналитика и e-commerce

Белый Ветер ЭЛЬДОРАДО BABADÜ Ideas4Web

2. Решал задачи для ЦБ РФ

3. "5" по географии в школе :)

4. Никогда не брал кредиты :)

Tinkoff Data Science Challenge

Задача 1. Выбор кредита Tinkoff.ru

https://boosters.pro/champ_3

Tinkoff.ru работает с сетью магазинов электроники, в которой присутствуют и другие банки. Заявка на кредит от покупателя поступает сразу в несколько банков, часть из них заявку одобряют. После этого покупатель выбирает, в каком банке взять кредит. Датасет содержит данные о кредитах на покупку электроники, которые были одобрены Tinkoff.ru. Необходимо предсказать, **выберет ли** покупатель кредит от Tinkoff.ru.

Задача 1

Метрикой качества в задаче является AUC

Пример решения

```
_ID_,_VAL_
0,1
1,0
...
170745,0
```

Данные

1. Признаки

Категориальные:

- Образование (education)
- Работа (job_position)
- Семейное положение (marital_status)
- Номер предлагаемого тарифа (tariff_id)
- Регион проживания (living_region)
- Пол (Gender)

Вещественные:

- Возраст (Age)
- Сумма кредита (credit_sum)
- Срок кредитования (credit_month)
- Месячный заработок (monthly_income)
- Количество кредитов у клиента (credit_count)
- Количество просроченных кредитов клиента (overdue_credit_count)
- **score_shk** не описана органтзаторами

2. Объем выборки
Тренировочная - 170 тысяч
Тестовая - 91 тысяча

3. Существенный процент пропусков и ошибок в данных

Это были мои первые соревнования

я самбмитил как мог...

#	Оценка	Дата и время
1	0.5000	2017-01-31 23:22:51
2	ошибка	2017-02-01 23:09:03
3	ошибка	2017-02-04 00:45:05
4	0.5000	2017-02-04 00:46:57
5	0.5000	2017-02-04 00:48:18
6	0.7032	2017-02-05 00:59:03

200 место

Траектория на Public LB

LightGBM

Реализация градиентного бустинга над решающими деревьями от Microsoft

https://github.com/Microsoft/LightGBM

Скорость в 6-10 раз выше, чем у XGBoost

Категориальные фичи

```
import lightgbm as lgb
gbm = lgb.LGBMClassifier()
gbm.fit(X, y, categorical_feature = ['job_position','tariff_id'])
#If 'auto' and data is pandas DataFrame, use pandas categorical columns
```

Качество немного хуже, чем у XGBoost

Предварительная работа

1. "Физика" задачи
2. Особенности кредитного скоринга
3. **score_shk (загадочная фича)**
4. Тарифы
5. Регионы РФ

Кредитный скоринг

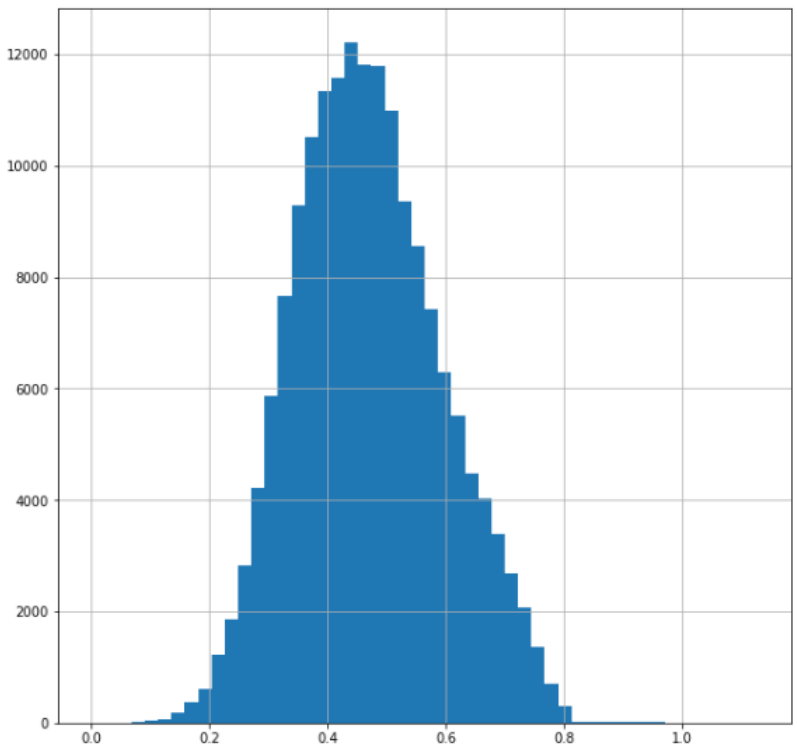
ПРИМЕР СКОРИНГОВОЙ КАРТЫ №2

Показатель	Значение	Балл
Возраст	< 20 лет	15
	20-25	34
	25-30	55
	30-35	90
	35-50	114
	50-60	97
	> 60 лет	15
Семейное положение	Холост (не замужем)	87
	Женат (замужем)	115
	Женат (замужем), но живет раздельно	30
	В разводе	70
	Вдовец (вдова)	65
Наличие детей	Нет детей	87
	Один	64
	Два	52
	Три	14
	Более трех	4
Сфера деятельности	Госслужба	93
	Коммерческая структура	124
	Пенсионер	19
	Другие	47
Квалификация	Нет квалификации	3
	Обслуживающий (вспомогательный) персонал	17
	Специалист	72
	Служащий	83
	Руководитель	122

Особенности:

1. Очень простая модель.
2. Должна расчитываться онлайн.
3. Устойчива к выбросам.

Модель кредитного скоринга Тинькофф банк?



1. Гаусовский шум?

При удалении переменной было только хуже

2. Оценка скора клиентов?

Если использовать перменную как таргет, то можно решить задачу регресии достаточно хорошо...

Как кодировать тарифы?

1. Использовать как вещественную перменную

2. Таргет-кодирование

3. Категориальная фича для LGBM

Много "странностей" в регионах

	living_region
0	74
1	98
43	Г. ЧЕЛЯБИНСК
45	Г.ОДИНЦОВО МОСКОВСКАЯ ОБЛ
46	ГОРЬКОВСКАЯ ОБЛ
47	ГУСЬ-ХРУСТАЛЬНЫЙ Р-Н
49	ДАЛЬНИЙ ВОСТОК
117	МОСКОВСКИЙ П
120	МЫТИЩИНСКИЙ Р-Н
192	ОРЁЛ
202	ПЕРМСКАЯ ОБЛ
205	ПРИВОЛЖСКИЙ ФЕДЕРАЛЬНЫЙ ОКРУГ
250	РЕСПУБЛИКАТАТАРСТАН
251	РОССИЯ
304	ЧИТИНСКАЯ ОБЛ

Частый паттерн:

ЛЕНИНГРАДСКАЯ
ЛЕНИНГРАДСКАЯ ОБЛ
ЛЕНИНГРАДСКАЯ ОБЛАСТЬ

Скорее всего, это распознавание скана паспортных данных клиента или разные магазины по-разному обрабатывают данные.

Как бороться с этим и как кодировать?

1. Использовать как категориальную перменную
2. HotOne-кодирование
3. Нормализация
4. Таргет кодирование
5. 1.+3.
6. 3.+4.

0.7721

1. Обогащение данных о регионах
2. Какой магазин предлагает взять кредит в Тинькофф?

3. Дополнительные фичи:
   - наличие копеек в сумме
   - отношение зарплаты к среднемесячной выплате
   - среднемесячная выплата, умноженная на тариф

...

1. Кодирование регионов соц.демом

~~http://www.gks.ru/~~ http://ya.ru/

- средняя зарплата
- население
- продолжительность жизни

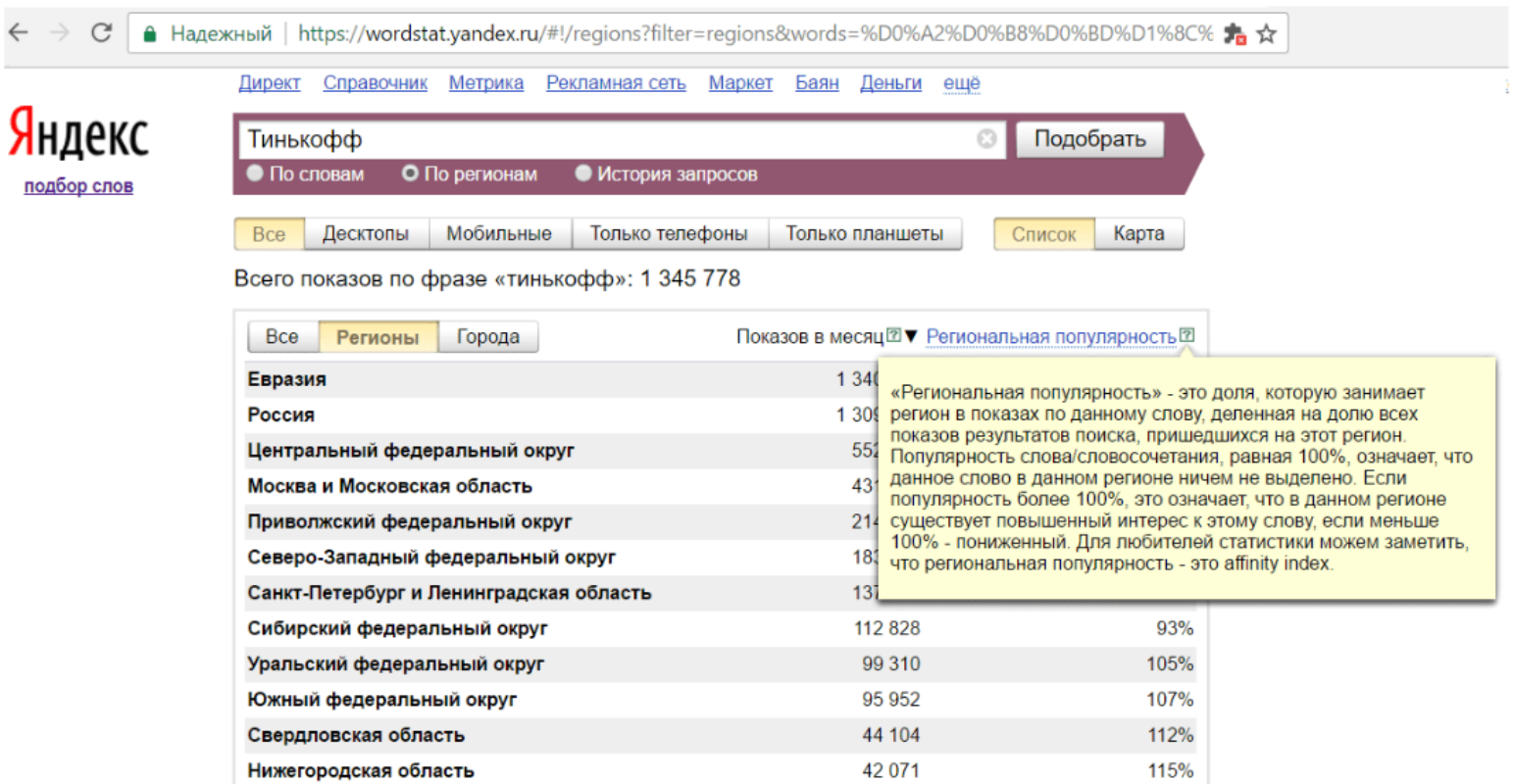
2. Данные для задачи http://www.svyaznoy.ru/store/kredit

Онлайн-кредит можно оформить в следующих банках:

- Хоум Кредит Банк;
- Ренессанс Кредит;
- ОТП Банк;
- Тинькофф Банк;
- Почта-банк.

3. Популярность банков

https://wordstat.yandex.ru/



0.7736

1. Тестирование разных pipelin'ов решений
   - для замены пропусков
   - для обработки категориальных переменных

2. "Жадный" отбор признаков.

3. Тюнинг параметров с помощью Hyperopt

0.7742

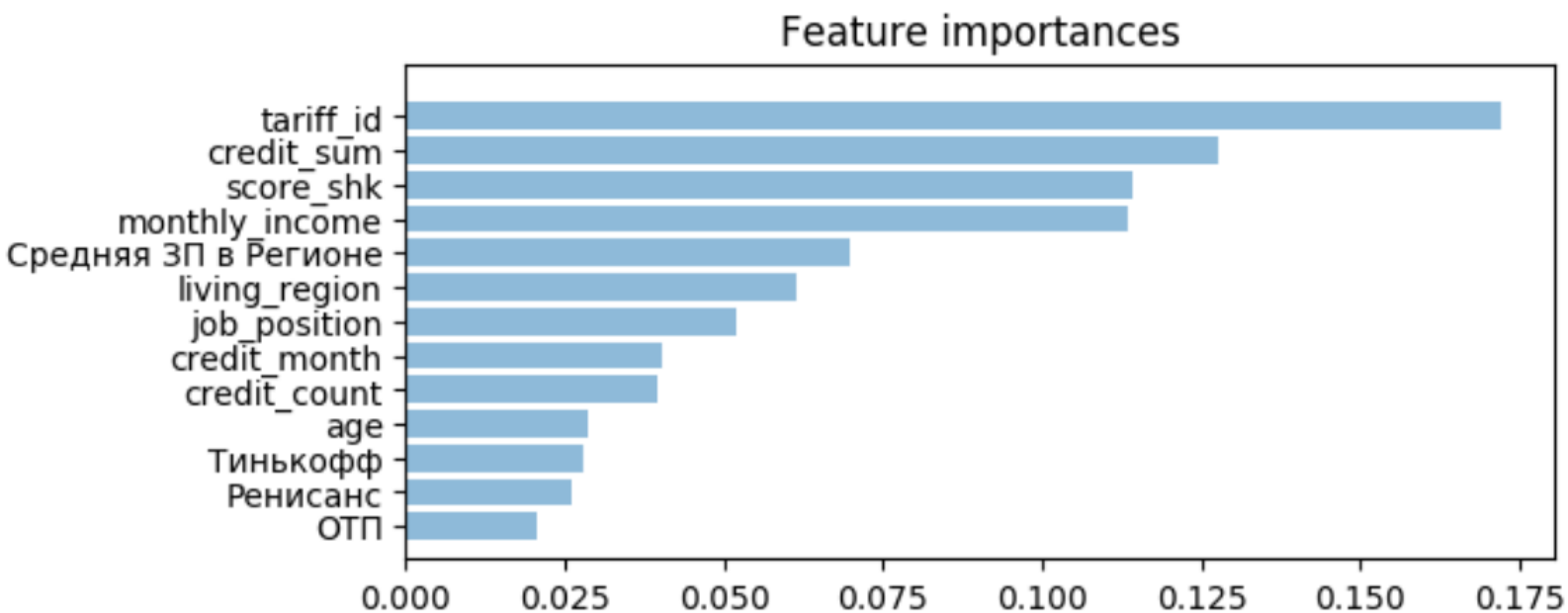
1. Нахождение лучших "разных" pipelin'ов.

2. Финальный тюнинг параметров Hyperopt для моделей.

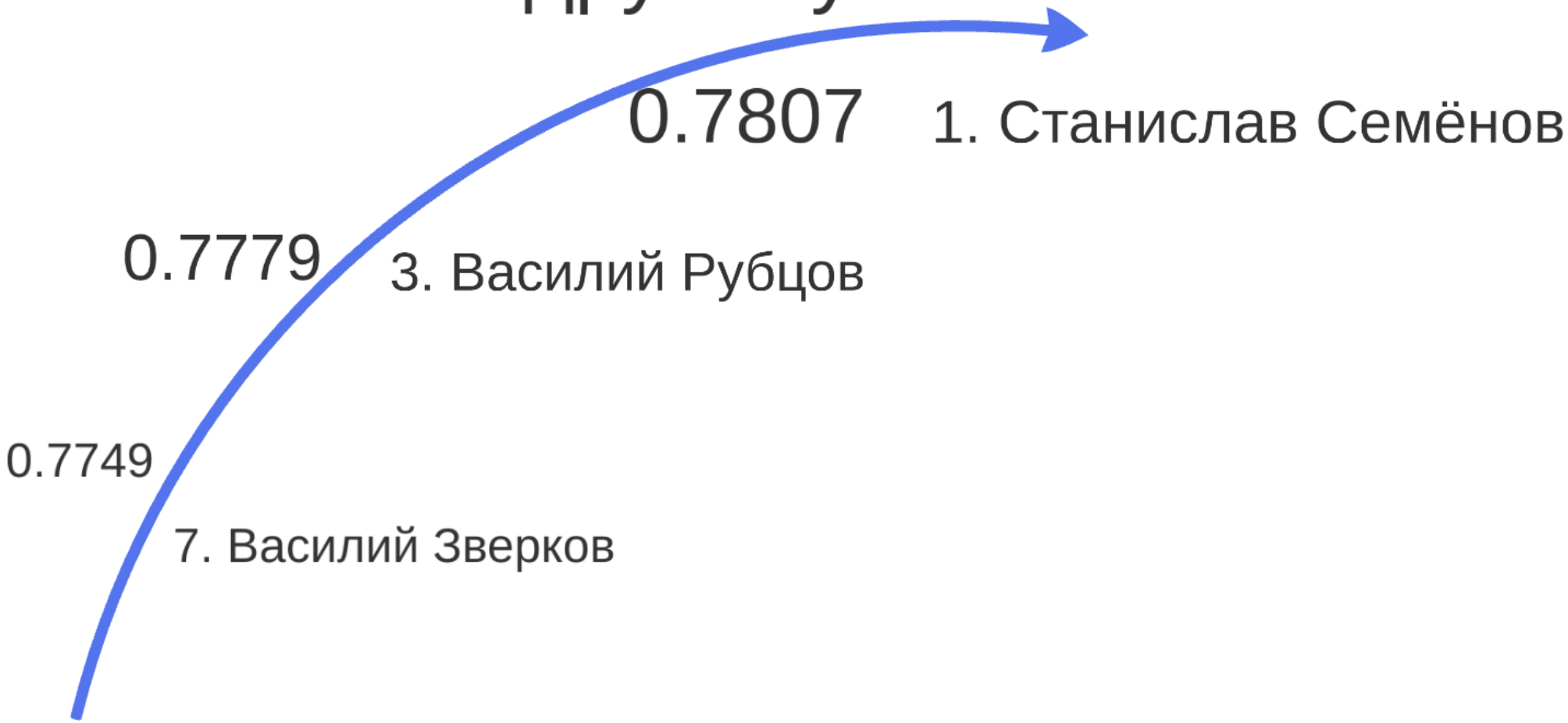
3.1. Блендинг. (+0.0005)
3.2. Стэкинг лучших pipelin'ов (5LGBM+2 XGB) с линейной моделью на 2-ом уровне. (+0.0006)

Лучший одиночный LGBM

250 деревьев, без регуляризации
Категориальные признаки: tariff_id, job_position, living_region



Решения других участников



7. Василий Зверков

1. Регионы практически не чистил.
2. Для категориальных переменных считал простые статистики: mean, std, count.
3. Считал преобразования (Log, x^2, exp, x^3) для числовых переменных и квантили.
4. **Использовал копейки, единицы, десятки, сотни и тысячи из суммы кредита как признаки**.
5. Отбирал признаки по feature importance.
6. В качестве модели использовал xgboost. Параметры настраивал с помощью hyperopt.

3. Василий Рубцов

vasiliy 10:54 AM

Коротко напишу о своей первой задаче. Делал стандартную обработку признаков, генерил новые (например категории отображал в их частотность, делил сумму кредита на количество месяцев и тому подобное, некоторые категориальные признаки обозначал в порядке среднего ключевого признака внутри категории) на них запускал хгбуст. Потом делал признаки так чтобы они были удобны для линейной модели - категории one hot, непрерывные - разбить на квантили, потом one hot и все такое. На этих признаках обучал простую нейросеть с двумя полносвязными слоями. И наконец, обучал хгбуст которому помимо изначальных признаков подавал выходы из нейросети. Потом все модели сблендил. Вот мой код: https://github.com/VasiliyRubtsov/Tinkoff

> github.com
> **GitHub - VasiliyRubtsov/Tinkoff**
> Contribute to Tinkoff development by creating an account on GitHub.

1. Станислав Семёнов

stasg7 Feb 28th at 11:26 AM
in #kaggle_crackers

Но если вкратце - в первой я опять тупо запустил скрипт, немного подправив. Сами данные не смотрел

stasg7 Mar 1
Расчет среднего значения таргета для переменных.
Подробную реализацию см. в видео BNP Paribas

stasg7 Feb 28
У меня там были пары и тройки признаков

Вместо заключения

1. Добиваться хороших результатов можно и с простыми моделями.

2. Понимание предметной области +0.05 к AUCROC и +1 к карме.

3. Участие в соревнованиях помогает узнать много нового.

4. В жизни переобучение не так страшно, как в ML :)

Спасибо за внимание!

Вопросы?

vk.com/chernobrovov

OpenDataScience @alex4er

Алексей Чернобровов