Sberbank

Data Science Journey

Роман Шеховцов

О себе

Sberbank
Artificial Intelligence

- Роман Шеховцов, работаю в Сбербанк-Технологии архитектором.

- AI / DS / ML начал изучать в июле
- Это мой первый опыт участия в соревнованиях

Наивный, но не Байес

Sberbank
Artificial Intelligence

Подсмотрел, что хорошо работало у авторов auto-sklearn:

papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning

Взял из sklearn:

- GradientBoosting
- RandomForest
- ExtraTrees
- AdaBoost

И использовал GridSearchCV:

```
param_grid = {
        'min_samples_leaf': range(1, 30),
        'max_features': np.arange(0.1, 1.1, 0.1),
    }
```

Заколдованный lightgbm_baseline

Sberbank
Artificial Intelligence

github.com/vlarine/sdsj2018_lightgbm_baseline

- Пробовал:
  - xgboost
  - lightgbm
  - catboost

- Тюнил параметры по гайдам авторов библиотек
- Пробовал RandomSearchCV и свой random search
- Пробовал категорийные фичи lightgbm и catboost
- Использовал hyperopt

Пришел к выводу, что «заколдованные» параметры – глобальный или близкий к нему оптимум в param_space LightGBM. Использовал это.

AutoML pipeline

Sberbank
Artificial Intelligence

Pipeline - набор шагов. Каждый шаг получает один или более датасетов на входе и выдает один или более датасетов на выходе.

Step (шаг) – содержит одну или более моделей. Для каждой модели генерируется тем или иным алгоритмом множество инстансов с разными параметрами.

Модель – класс-обертка (LightGbmWrapper, XGBoostWrapper, …).

Итерация – прогон всех выживших на текущий момент инстансов шага. Если на шаге возможен скоринг – расчет score и просеивание успешных моделей.

Subsampling – берем часть данных, например 4 000 строк.

Success halving – фиксируем бюджет на итерацию. На каждой итерации оставляем лучшую половину моделей, но удваиваем размер сэмпла.

Сэмплинг и train / test

Sberbank
Artificial Intelligence

- Для первого семпла пробовал брать от 500 до 40 000 строк
- Начиная с 4 000, качество обучения становилось приемлемым
- Значения до 10 000 работали достаточно быстро, еще немного поднимая качество
- Для классификации: написал свой семплер со стратификацией
- Для регрессии: np.random.choice(rows, sample_rows, replace=**False**)

- Для маленьких датасетов (по факту – только для первого) прикрутил 5-fold CV
- Для больших – разбивал семпл на train / test

optimize_dataframe() для больших файлов

Sberbank Artificial Intelligence

- downcast колонок **int**, **int32**, **int64**, **float**, **float32**, **float64**
- Преобразование колонок **object** к **category** (если уникальных значений меньше ½ числа строк)
- На данных SDSJ уменьшает потребление памяти pandas.DataFrame() в 2-6 раз
- Ускоряет обработку больших датасетов (которая упирается в размер памяти)
- Учел внутреннюю структуру pandas.DataFrame, поэтому функция работает быстро
- 50 строк кода github.com/rshekhovtsov/sdsj-2018
- Основана на статье dataquest.io/blog/pandas-big-data

Ошибки и уроки

Sberbank
Artificial Intelligence

- Стратегическая ошибка:
  - Не стоило писать automl-фреймворк общего назначения на первом соревновании
  - Потратил две недели
  - Понял, что всё успеваю как раз к Новому году
  - Прибил код гвоздями
- Ошибки новичка:
  - Не попробовал ансамбли моделей
  - Не продумал более изощренную схему валидации
- Фиксация random_state = русская рулетка. Вместо этого нужно делать CV

Что не успел сделать

Sberbank
Artificial Intelligence

- Geographic features enrichment:
  - Видел категорийные фичи с территориальными банками Сбербанка.
  - Каждый такой банк обслуживает определенный регион – федеральный округ или несколько областей. В них разное социально-экономическое положение, что прямо влияет на денежки.
  - Можно было найти социоэкономическую статистику (средние зарплаты, бюджеты) и обогатить датасеты сгенерированными по ней фичами.
- Meta-learning:
  - Сохраняем результаты экспериментов и характеристики датасетов в БД.
  - Для новых датасетов вычисляем близость к старым и используем успешно показавшие себя ранее параметры.
  - Не улучшает итоговый score, но ускоряет сходимость к нему.

Что сделал хорошо.
Что понравилось

Sberbank
Artificial Intelligence

- Ориентировался на результаты локальной валидации, а не на leaderboard

- Платформа и организация соревнований.
- Параметры lightGBM от **vlarine**: github.com/vlarine/sdsj2018_lightgbm_baseline
- Код **tyz910**: github.com/tyz910/sdsj2018
- add_holidays() от **nd7141**: github.com/contestAI/sberbank-catboost

Спасибо!