Data Science Contest

ПРЕДСКАЗАНИЕ ПОЛА ПОЛЬЗОВАТЕЛЯ ПО ЕГО ТРАНЗАКЦИЯМ

Результаты

30-е место в общем зачете

6-е место по первой задаче

Постановка задачи

Метрика качества auc-roc

Требуется предсказать вероятность быть мужского пола для 3000 пользователей

Публичный рейтинг оценивается на 16% данных – всего 480 человек!

Приватный рейтинг оценивается на 84% данных

Описание данных

Всего около 7 млн транзакций

Всего 15 тысяч пользователей

Для 12 тысяч пользователей известен пол

Неизвестна точная дата каждой транзакции

Данные

Customer_id – идентификатор клиента

tr_datetime – номер дня и время совершения транзакции

mcc_code - mcc-код транзакции (184 уникальных)

tr_type - тип транзакции (155 уникальных)

Amount - сумма транзакции в условных единицах

term_id – идентификатор терминала

tr_description – описание типа транзакции

mcc_description - описание mcc-кода транзакции

Данные были изменены!

Amount – цифры не похожи на траты в рублях

tr_datetime – требуется восстановить точные даты транзакций

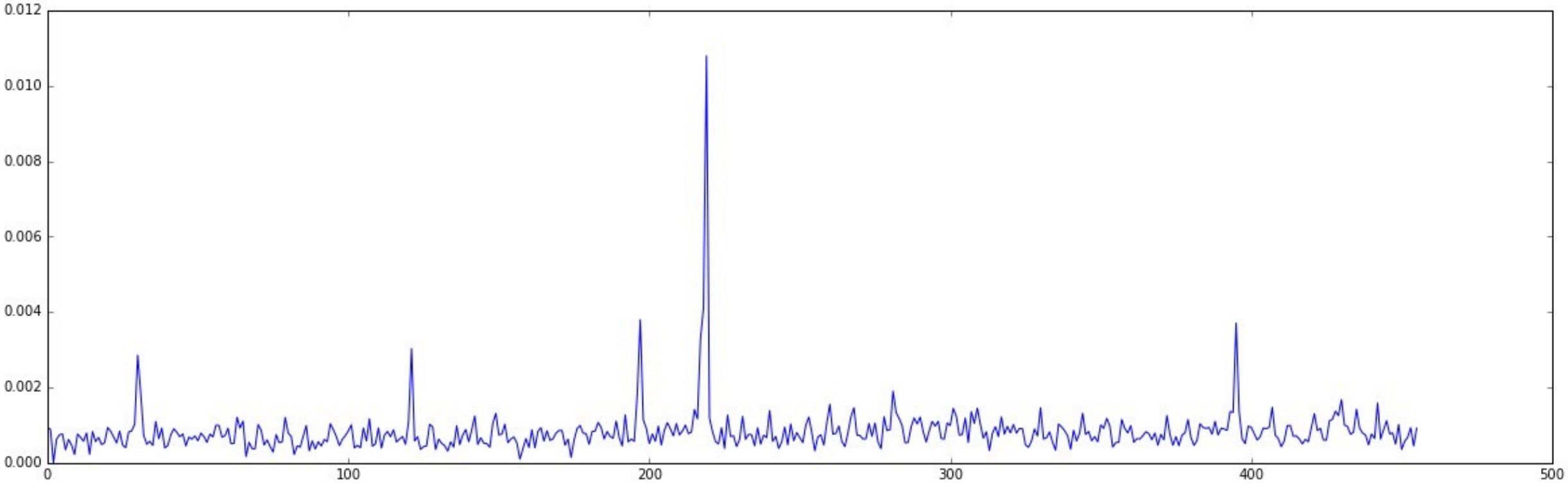
Amount до преобразования / Amount после преобразования

Amount_restored	Amount
-100	-2245.92
2500	56147.89
-2500	-56147.89
-62	-1392.47
-41	-920.83

Пример значений tr_datetime

tr_datetime
0 10:23:26
1 10:19:29
1 10:20:56
1 10:39:54
2 15:33:42

Относительное кол-во покупок по категории “Флористика”



Фичи по времени

Траты и поступления по:

- Часам
- Дням недели
- Дням месяца
- Месяцам
- Выходным и будним дням

Фичи по текстовым признакам

Нормализация каждого слова с помощью pymorphy2

Получение текста из описания всех транзакций для каждого пользователя

Векторизация текста

Нормализация по объектам

Примеры текстовых описаний mcc до нормализации

mcc_description	Tr_description
Денежные переводы	Оплата услуг банка через ВСП
Школы – бизнес и секретарей	Плата за получение наличных. Россия
Ветеренарные услуги	Перевод средств с карты на счет клиента через АТМ

Фичи по тратам и поступлениям

Средние

Медианы

Максимальные

Стандартные отклонения

Суммы

Xgboost наше все

Kaggle competitions

What society thinks I do

What my friends think I do

What I think I do

```
import xgboost as xgb
```

What I really do

Подбор параметров алгоритма

Функция, которую надо оптимизировать

```
def score(params):
    seed = int(np.random.rand()*100000)
    params['max_depth'] = int(params['max_depth'])
    lgr.info('seed = %i' % seed)
    lgr.info("Training with params : ")
    lgr.info(params)
    cv_res = xgb.cv(params, dtrain, early_stopping_rounds=100, maximize=True,
    num_boost_round=10000, nfold=5, seed = seed)
    score = cv_res['test-auc-mean'].max()
    lgr.info("Score = %f" % score)
    lgr.info('best rounds = %i' % cv_res[cv_res['test-auc-mean'] == cv_res['test-auc-mean'].max()].index[0])
    return {'loss': -score, 'status': STATUS_OK}
```

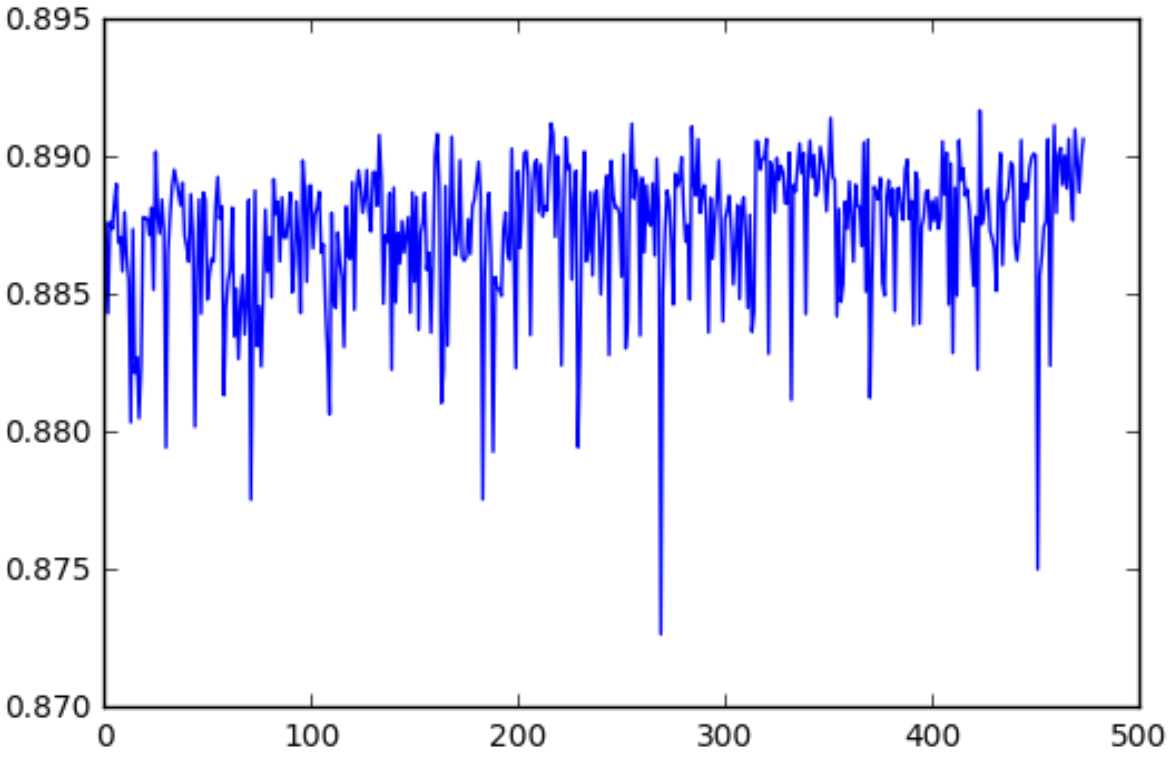
Hyperopt + Xgboost

Оптимизация функции, определенной на предыдущем слайде

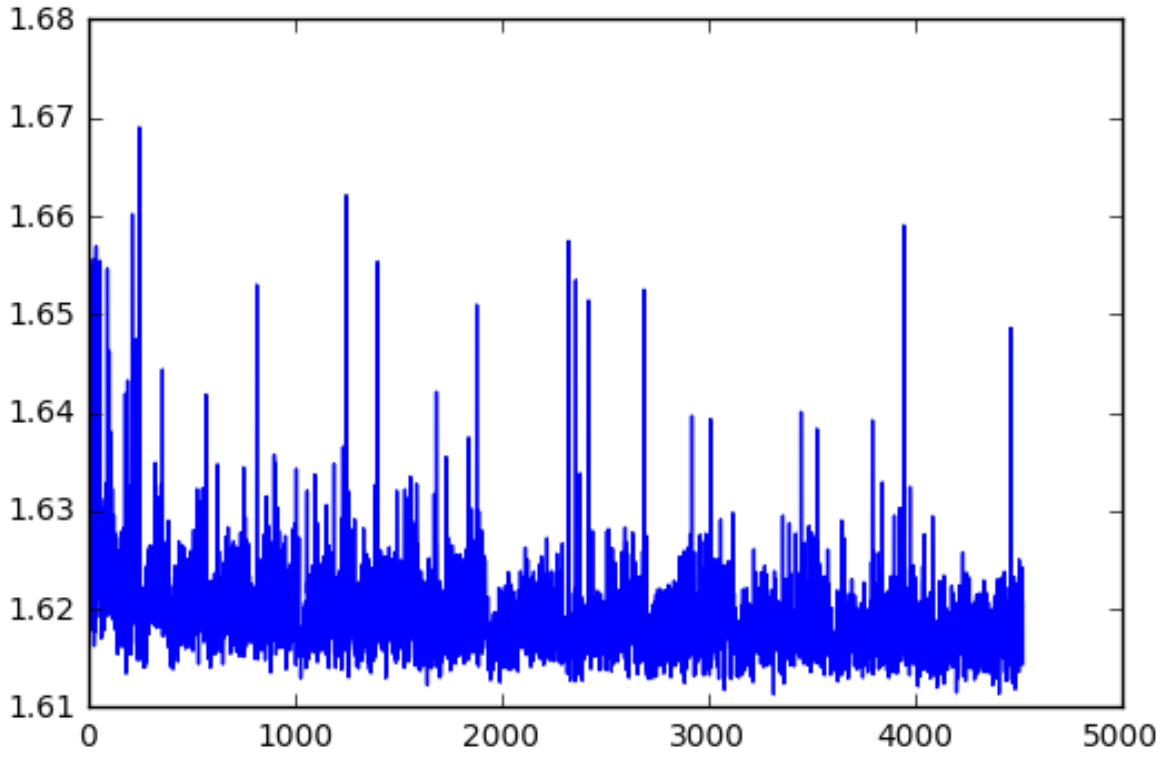
```
space = {
    'eta' : hp.quniform('eta', 0.001, 0.1, 0.001),
    'max_depth' : hp.quniform('max_depth', 3, 15, 1),
    'min_child_weight' : hp.quniform('min_child_weight', 1, 30, 1),
    'subsample' : hp.quniform('subsample', 0.5, 1, 0.05),
    'gamma' : hp.quniform('gamma', 0.1, 2, 0.05),
    'alpha': hp.quniform('alpha', 0.001, 2, 0.05),
    'lambda': hp.quniform('lambda', 0.001, 2, 0.05),
    'colsample_bytree' : hp.quniform('colsample_bytree', 0.01, 1, 0.01),
    'eval_metric': 'auc',
    'objective': 'binary:logistic',
    'booster': 'gbtree',
    'nthread' : 11,
    'silent' : 1
}

best = fmin(fn = score,
space=space,
algo=tpe.suggest,
trials=trials,
max evals=500)
```

Результаты

Первая задача



Вторая задача



Что не сработало

Нейросеть

Фичи по интервалам

Фичи по тратам перед 8-марта и 23-е февраля

Понижение размерности

Спасибо за внимание!