
Computer
Science
Center

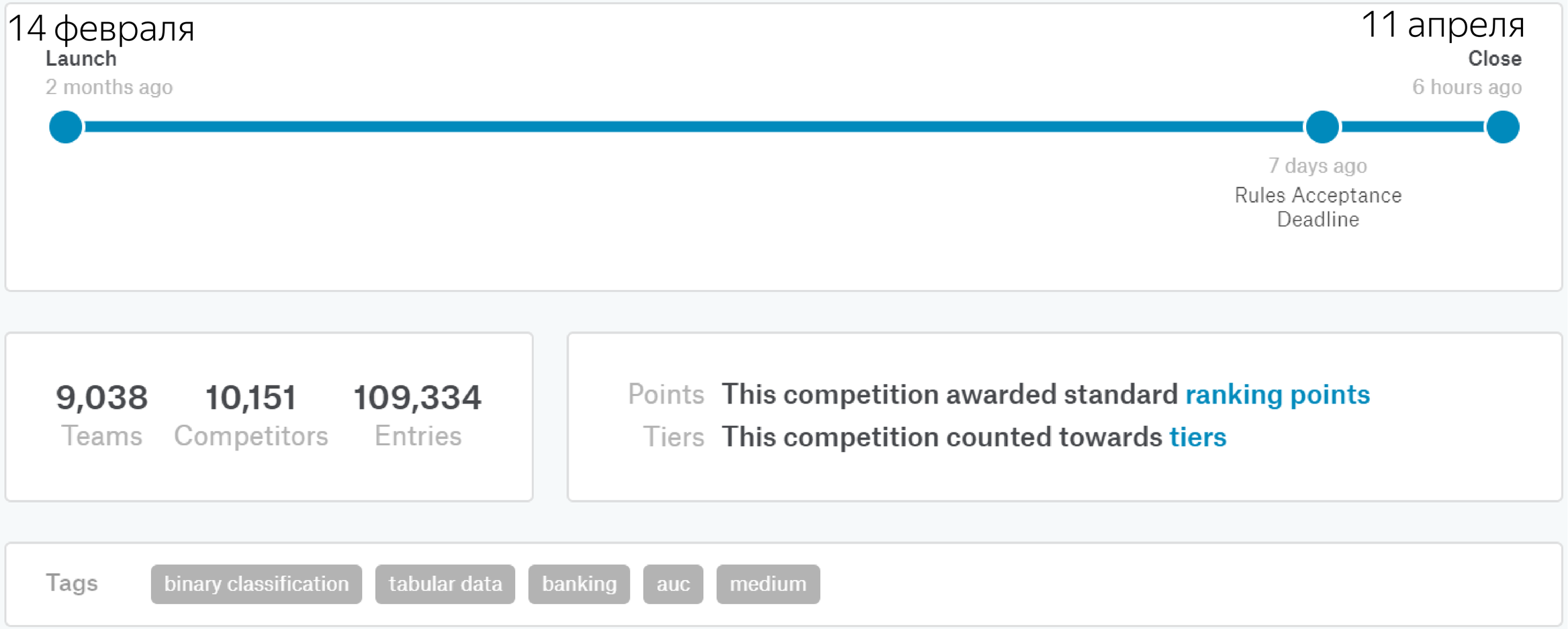
Santander-2019 at Kaggle или в поисках «магии»

Александр Авдюшенко

15 апреля 2019

Задача

Бинарная классификация – предсказать, сделает ли клиент покупку в будущем.
Анонимизированные и полусинтетические данные.



https://www.kaggle.com/c/santander-customer-transaction-prediction

Dataset

Количество строк
200K – training_set
200K* – test_set

Data
var_0, … ,var_68, ….

Обучение 20-30 мин
Применение 5 мин

* 100K/50K/50K fake/public/private

	ID_code	target	var_0	var_1	var_2	var_3	var_4	var_5	var_6	var_7	var_8	var_9	v
0	train_0	0	8.9255	-6.7863	11.9081	5.0930	11.4607	-9.2834	5.1187	18.6266	-4.9200	5.7470	2
1	train_1	0	11.5006	-4.1473	13.8588	5.3890	12.3622	7.0433	5.6208	16.5338	3.1468	8.0851	-
2	train_2	0	8.6093	-2.7457	12.0805	7.8928	10.5825	-9.0837	6.9427	14.6155	-4.9193	5.9525	-
3	train_3	0	11.0604	-2.1518	8.9522	7.1957	12.5846	-1.8361	5.8428	14.9250	-5.8609	8.2450	2
4	train_4	0	9.8369	-1.4834	12.8746	6.6375	12.2772	2.4486	5.9405	19.2514	6.2654	7.6784	-

Baseline – score CV 0.900

```
folds = StratifiedKFold(n_splits=10, shuffle=False, random_state=44000)
oof = np.zeros(len(train_df))
predictions = np.zeros(len(test_df))
feature_importance_df = pd.DataFrame()

for fold_, (trn_idx, val_idx) in enumerate(folds.split(train_df.values, target.values)):
    print("Fold {}".format(fold_))
    trn_data = lgb.Dataset(train_df.iloc[trn_idx][features], label=target.iloc[trn_idx])
    val_data = lgb.Dataset(train_df.iloc[val_idx][features], label=target.iloc[val_idx])

    num_round = 1000000
    clf = lgb.train(param, trn_data, num_round, valid_sets = [trn_data, val_data], verbose_eval
=1000, early_stopping_rounds = 3000)
    oof[val_idx] = clf.predict(train_df.iloc[val_idx][features], num_iteration=clf.best_iterati
on)

    fold_importance_df = pd.DataFrame()
    fold_importance_df["Feature"] = features
    fold_importance_df["importance"] = clf.feature_importance()
    fold_importance_df["fold"] = fold_ + 1
    feature_importance_df = pd.concat([feature_importance_df, fold_importance_df], axis=0)

    predictions += clf.predict(test_df[features], num_iteration=clf.best_iteration) / folds.n_s
plits

print("CV score: {:<8.5f}".format(roc_auc_score(target, oof)))
```

And nothing works...

Vadim Nareyko

173rd place

Options

List of experiments that haven't broken 0.9

posted in Santander Customer Transaction Prediction a month ago

92

I hope it will help you not to spend too much time in these directions. And you may focus on something new ;)

1. I have generated 100K+ features like (a+b, a*b, a-b, a/b, np.exp(a)*np.exp(b), np.exp(a)/np.exp(b). Then run separate random models to select most important features (limited it to 600). Then removed some features using permutation.
2. I have LGBM, XGBoost, Catboost, Pytorch and Keras predictions based on different feature sets. 5-folds and 10-folds
3. Different ensembling using mean, rankmean, weighted.

So far 0.9.

Surely, I have more ideas and it's in the process.

Public kernel LB 0.901 (augmentation trick)

10 submissions for **Aleksandr Avdyushenko** Sort by Most recent

All **Successful** Selected

Submission and Description	Private Score	Public Score	Use for Final Score
cb10K_pe_ue_submission.csv 15 hours ago by Aleksandr Avdyushenko cb10K pe ue	0.89757	0.89773	☐
lgb40K_pe_ue_me_submission.csv 2 days ago by Aleksandr Avdyushenko lgb40K pe ue me	0.90014	0.90200	☐
lgb40K_pe_ue_submission.csv 6 days ago by Aleksandr Avdyushenko lgb40K pe ue	0.90324	0.90484	✓
lgb_best_submission.csv 8 days ago by Aleksandr Avdyushenko lgb best from team	0.89951	0.90051	☐
lgb_submission.csv 10 days ago by Aleksandr Avdyushenko lgb aug from public	0.89990	0.90115	✓
submit_blend09.csv 14 days ago by Aleksandr Avdyushenko 09_lgb + 01_lr	0.87820	0.88133	☐

silver

bronze (!)

gold (28) >= 0.92283
silver (451) >= 0.89996
bronze (903) >= 0.89989

My first silver

Зашло
- lgbm tuning
- kts
- attempt to merge into team
- count encoding («magic»)

How to win a Data Science Competition
https://www.coursera.org/learn/competitive-data-science

Не зашло
- target mean encoding
- catboost

В целом снова очень позитивные впечатления!
Тратил примерно 5-10 ч времени в неделю.
Куча навыков, ещё и медаль.

TOP solutions

«Magic» actions	Places
remove fake from test, concat train and test, count encoding	#2, #4, #5, #21
unpivot all vars(so we have 200k x 200 = 4m train samples)	#2, #4
standard scaling	#2, #21
count round encoding	#2
convert prediction(200k x 200) into odds. We used (9 * p / 1 - p)	#2
augmentation	#5, #21