

Bank Scoring Case

Predicting probability of default

4th and 2nd places on the Private Leaderboards of 2 competitions

In this material I mentioned two competitions:

<https://www.kaggle.com/c/bank-scoring-case2/overview>

<https://www.kaggle.com/c/scoring-case/overview>

Both of them were closed at the moment of predictive models creation, that's why only late submissions were available.

Model fitting was being done in this competition

<https://www.kaggle.com/c/scoring-case/overview>.

The main predictive model is XGB, it's effectiveness has been enhanced by the voting ensemble method. Initial training dataset was splitted into train and test parts with regarding to the class balance.

Firstly most fitted params were being searched by the GridSearchCV.

Secondly when needed params intervals were found, the RandomizedSearchCV was used for more precise tuning.

Since datasets that were used for public and private leaderboards score counting might have class balances that were distinct from the available data class balance, I tried to find an optimal value of the `scale_pos_weight` parameter for the Public Leaderboard dataset. Putting `scale_pos_weight = 5` (Picture 1) gave me my best public score. The next step was an improvement for the private dataset. I managed to move from the 13th place (Picture 2, Public Leaderboard 0.87115, Private leaderboard 0.86335, `scale_pos_weight = 13`) to the 7th place (Picture 3, Public Leaderboard 0.87134, Private Leaderboard 0.86394, `scale_pos_weight = 1.975`).

trial_1.csv	0.86353	0.87225	<input type="checkbox"/>
8 hours ago by Vadim			
scale_pos_weight = 5			

Picture 1

trial_1.csv	0.86335	0.87115	<input type="checkbox"/>
4 days ago by Vadim			
XGB_new_params			

Picture 2

trial_1.csv	0.86394	0.87134	<input type="checkbox"/>
6 hours ago by Vadim			
1.975			

Picture 3

I was realizing that I was doing an overfitting. Also if I was given only a public score I would probably not choose the model that gave me the 7th place on a private since I had several models with bigger public score but much lower private one. That's why I decided to increase the generalizing ability of my model in relation to the class balance. It was done with the use of voting ensemble that included the XGBoost models with previously found optimal params that differed from each other only by the scale_pos_weight param value in the range from 2 to 10 with the pace of 1. With these model I got the 6th place on a private (Picture 4, Public Leaderboard 0.87186, Private leaderboard 0.86394) and one of my biggest scores on a Public.

trial_1 (2).csv	0.86394	0.87186	<input type="checkbox"/>
4 hours ago by Vadim			
2-10			

Picture 4


The next step was that I decided to use a wider range of the scale_pos_weight param: from 1 to 13 with the pace of 0.5. I also tried to use all the train dataset for the final prediction (before that I had been fitting my model only on the train part without using the test one). That gave me not only private score increase and 4th place on the Private Leaderboard but also much bigger gain in a public score in comparison with my previous improvements (Picture 5, Public Leaderboard 0.87251, Private leaderboard 0.86413).

trial_1 (7).csv	0.86413	0.87251	<input type="checkbox"/>
6 hours ago by Vadim			
+the halves			

Picture 5

Thus this last model would be chosen also in case if I knew only my public score. Choosing this model for the final prediction would also have one important reason. This voting ensemble with different scale_pos_weight took into account as more possible datasets ratios as possible and gave good generalizing ability. That is why I decreased the model overfitting.

To enhance my belief in the generalization ability of my model I did next. This competition organizer also created a new one competition with absolutely the same dataset (<https://www.kaggle.com/c/bank-scoring-case2/overview>). But, of course, observations in Private and Public Leaderboard sets changed. I fit my model with training set that was given in new competition (old dataset was not used) and got the 2nd place both on Public and Private Leaderboards ((Picture 6, Public Leaderboard 0.86801, Private leaderboard 0.86434).



trial_1 (8).csv	0.86434	0.86801	<input type="checkbox"/>
4 hours ago by Vadim			
ensemble(voting)			

Picture 6