FRE 9733 Big Data in Finance Week 10 Homework

Mengheng Xue

May 1, 2019

1 Best Models for LR, RF and MLP

Based on my previous assignments, my choices of best model settings are as follows

	Best LR model setting
sample size	100k
regressors	AGE, MTM_LTV, INCENTIVE, CREDIT_SCORE, ORIG_INTRATE, UNIT_COUNT
hyper-parameters	MaxIter = 200, RegParam = 0.002, ElasticNetParam = 0.008

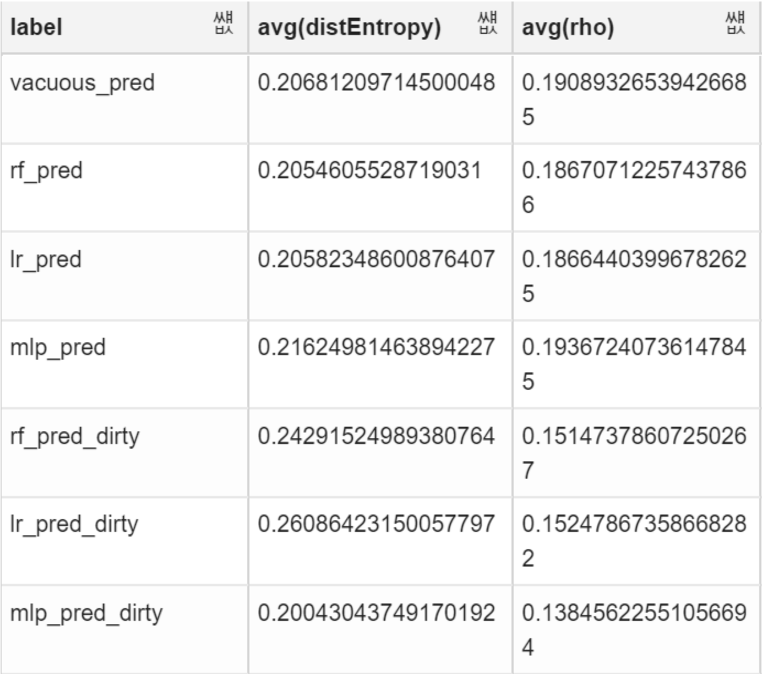
	Best RF model setting
sample size	100k
regressors	AGE, MTM_LTV, INCENTIVE, CREDIT_SCORE, ORIG_INTRATE, UNIT_COUNT
hyper-parameters	NumTrees = 12, MaxDepth = 2

	Best MLP model setting
sample size	100k
activation function	AGE, MTM_LTV, INCENTIVE, CREDIT_SCORE, ORIG_INTRATE, UNIT_COUNT
hyper-parameters	solver = gd, hidden layers = $[6, 6, 6]$

2 Performance Evaluation

The following is the performance compa rison based on the setting in previous section. We can see that:

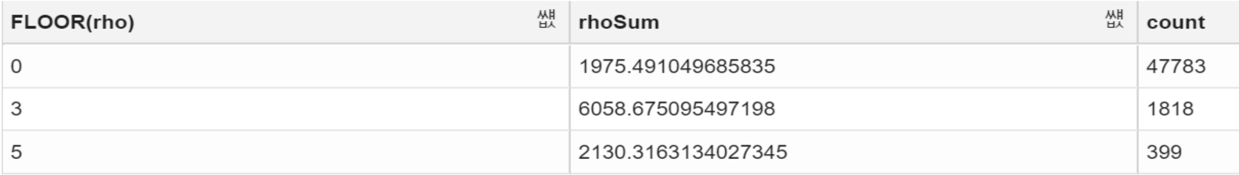
- LR and RF models beat the vacuous model on both clean and dirty dataset.
- MLP does not beat vacuous model but it obtains the best performace on dirty dataset.
- The performance on dirty dateset for all three models outperform performance on clean dataset, which is counter-intuitive.



label	avg(distEntropy)	avg(rho)
vacuous_pred	0.20681209714500048	0.19089326539426685
rf_pred	0.2054605528719031	0.18670712257437866
lr_pred	0.20582348600876407	0.18664403996782625
mlp_pred	0.21624981463894227	0.19367240736147845
rf_pred_dirty	0.24291524989380764	0.15147378607250267
lr_pred_dirty	0.26086423150057797	0.15247867358668282
mlp_pred_dirty	0.20043043749170192	0.13845622551056694

Fig. 1: performance comparison

From following figure, we could see that for vacuous model, I use FLOOR(rho), there are 399 loans which have rho above 5. Therefore, I don't need to adjust down the *limit* parameter in vacuous model.



FLOOR(rho)	rhoSum	count
0	1975.491049685835	47783
3	6058.675095497198	1818
5	2130.3163134027345	399

Fig. 2: vacuous model

Our mission in this assignment is trying to minimize the sum of rho for all loans having rho > 5.0. Here is a result summary of using my original setting:

model	round(rho,0) for rho > 5	sum(rho)
rf_clean	none	none
lr_clean	5	25.991
mlp_clean	none	none
rf_dirty	none	none
lr_dirty	5 6	67.0305 5.5905
mlp_dirty	none	none

It is very surprising that only LR model still has loans having rho > 5.0, which makes me doubt my model correctness.

Assume what I have done is correct, the only model I need to improve is LR model. Since I have already adjusted the hyper-parameters in LR model, and if I try to add or reduce some regressors in the model, it will affect other model's performance since we need to make them consistent. So except for tuning hyper-parameters or changing regressors, I don't know whether there are other changes I can do to improve the LR model.

Also, since for RF and MLP already meet our target, I am not sure whether I need to continue to modify them. Any suggestions would be appreciated!