FRE 9733 Big Data in Finance Week 10 Homework

Mengheng Xue

May 7, 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	300k
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	300k
regressors	AGE, MTM_LTV, INCENTIVE, CREDIT_SCORE, ORIG_INTRATE, UNIT_COUNT
hyper-parameters	NumTrees = 12, MaxDepth = 2

	Best MLP model setting
sample size	300k
regressors	AGE, MTM_LTV, INCENTIVE, CREDIT_SCORE, ORIG_INTRATE, UNIT_COUNT
hyper-parameters	solver = gd, hidden layers = $[6, 6, 6]$

2 Performance Evaluation

I tried to turn off the *safe_mode* and change the *sample_size* to 300k loans. Each model performances based on the settings in the previous section are shown as follows:

- In terns of average rho, all three models are better than vacuous model and the three models' performance are similar in which *avg(rho)* are all less than 0.2.
- The performance on dirty dateset for all three models outperform performance on clean dataset, which shows the robustness of our model against the noisy data in my understanding.

label	avg(distEntropy)	avg(rho)
vacuous_pred	0.18427044810088591	0.20328964917171935
rf_pred	0.2046535412664049	0.18721942246384132
lr_pred	0.20582348600876416	0.1866440399678261
mlp_pred	0.21288908801420006	0.19235926404987713
rf_pred_dirty	0.2179734988086431	0.1413131344587294
lr_pred_dirty	0.1535979503537019	0.14873216249459484
mlp_pred_dirty	0.21166103890095173	0.1407356169910324

Fig. 1: performance comparison

3 Vacuous Model

For vacuous model, I use *FLOOR(rho)* to count rho. We can see from the following figure that all loans are under 5.0.

FLOOR(rho)	rhoSum	count
0	2484992.106488333	55122762
3	5697016.074960321	1644570
4	2993079.5184511384	686506

Fig. 2: vacuous model

We can also see from the true rho histogram for the vacuous model that all loans are concentrated into three rho values. This is because in our vacuous model the only regressor is *NEXT_STATUS*. Therefore, there are only 3 next status $C \to C$, $C \to 3$, and $C \to P$. Thus, this model only gives the prediction for each loans in the test set based on the true likelihood of each status happening in the training set.

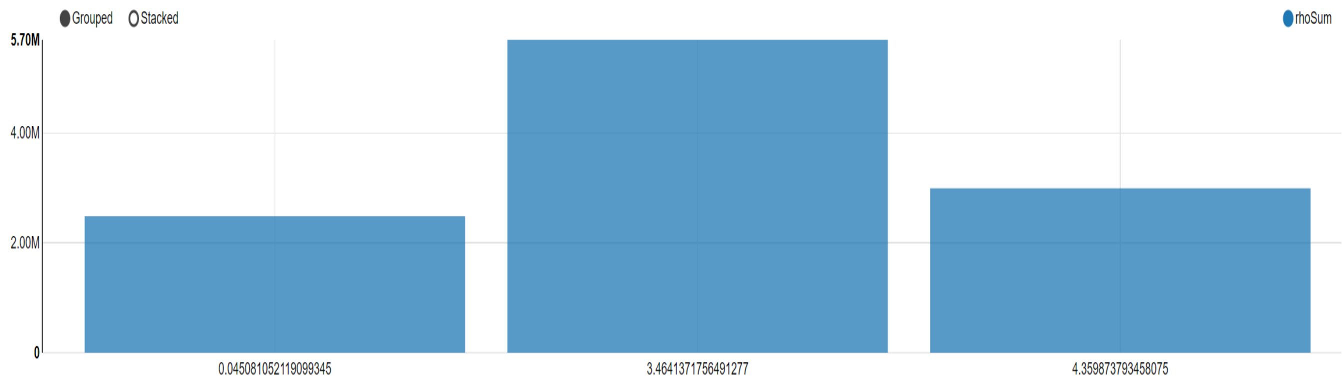

Fig. 3: vacuous model prediction

So I reduce the *limit* parameter from 100k to 5k in vacuous model, and obtain the following results

rho	rhoSum	count
0.021427980070509802	1181169.445600834	55122762
4.062845215009057	6681633.355241181	1644570
5.52145585139621	3790512.5707131787	686506

Fig. 4: vacuous model with limit 5k

We can see that on group of loan from vacuous model goes over 5.0.

4 Sum of Rho Comparison

Our mission in this assignment is trying to minimize the sum of rho for all loans having rho > 5.0. Here are the results using first section settings.

round(rho, 0)	sum(rho)	count(1)
8	15.353058555330149	2
6	173333.47042139186	30759
5	486307.68850466213	93028
7	71.67099727674952	10

Fig. 5: $rho > 5$ for rf_pred

round(rho, 0)	sum(rho)
5	224738.02538040373
6	45217.72715222581
7	1748.97098324748
8	15.028363340112733
26	26.023101124429424
27	107.38878788485215
28	194.7600169597664
29	57.88366524473949

Fig. 6: $rho > 5$ for lr_pred

round(rho, 0)	sum(rho)

Fig. 7: $rho > 5$ for mlp_pred

round(rho, 0)	sum(rho)
5	872228.4466425541
6	623949.5359985854
7	1681.7339642164065
8	79.69776519393943

Fig. 8: $rho > 5$ for rf_dirty

round(rho, 0)	sum(rho)
5	520968.7024060456
6	153143.90305814837
7	2266.7924191301863
8	15.491371975480668
18	55.10946433232874
19	38.02460120640519

Fig. 9: $rho > 5$ for lr_dirty

round(rho, 0)	sum(rho)

Fig. 10: $rho > 5$ for mlp_dirty

Based on the above results we could say that

- MLP is the best model in terms of minimizing sum of RHO from loan having $rho > 5$, since no loan will have $rho > 5$ based on my settings.
- RF is the second best model, in which the prediction will contain some loans with $rho > 5$.
- LR performas really badly, in which the results will contain loans with rho much larger than 5.
- As expected, the prediction in clean dataset will produce less sum of rho for loans with $rho > 5$ than the dirty dataset.
- Basd on my results, I would say my MLP beats the vacuous model, whhie LR and RF need to be modified further.

5 Model Modifying

From the previous sections we could see that even with simple model such as vacuous model with only one regressor can outperform my designed complicated model in terms of reduce the sum of rho for loans with rho > 5. So instead of adding more regressors in my previous model to make it more complicated, I decided to reduce the number of regressors to make it simpler.

	Modified LR model setting
sample size	300k
regressors	AGE, UNIT_COUNT, OCCUPANCY_STATUS
hyper-parameters	MaxIter = 200, RegParam = 0.002, ElasticNetParam = 0.008

	Modified RF model setting
sample size	300k
regressors	AGE, UNIT_COUNT, OCCUPANCY_STATUS
hyper-parameters	NumTrees = 12, MaxDepth = 2

	Modified MLP model setting
sample size	300k
regressors	AGE, UNIT_COUNT, OCCUPANCY_STATUS
hyper-parameters	solver = gd, hidden layers = $[6, 6, 6]$

Following is the *AVG(rho)* performance for three modified models. We could see that when we use less regressors, the *AVG(rho)* performance does sacrifice but not too much for all three models.

label	avg(distEntropy)	avg(rho)
vacuous_pred	0.20078996530287965	0.19469654379543228
rf_pred	0.19469251281436342	0.19220886250608682
lr_pred	0.19421695465518793	0.1932240283257297
mlp_pred	0.21288908801420006	0.19235926404987713
rf_pred_dirty	0.19792408895514077	0.15486235828275488
lr_pred_dirty	0.19083685876563272	0.15346464852825425
mlp_pred_dirty	0.21166103890095173	0.1407356169910324

Fig. 11: modified model performance comparison

For RF and MLP, there will be no loans with rho > 5, and for LR, the results are as follows.

round(rho, 0)	sum(rho)
5	3684.450556229305
16	228.9984341899214

Fig. 12: $rho > 5$ for modified lr_pred

round(rho, 0)	sum(rho)
5	7516.770522592817
23	113.24601878308376

Fig. 13: $rho > 5$ for modified lr_dirty

We can conclude that

- When we decrease model complexity by reducing regressors, all three models have been improved in terms of less outlier predictions. especially for LR.
- When we try to reduce the complexity of model in order to bring down the outliers (sum(rho) for rho > 5), we also reduce the average prediction accuracy (AVG(rho)) also. So I would say there is always a tradeoff between them. You can not have one perfect model which is both accurate and robust at the same time.
- Since the tradeoff mentioned before exists, instead finding a simpler model performs better in terms of the outliers, we could try to find the best possible model in terms of all loans. And in that case, we need to find the optimal regressor selections.
- Another improvement approach for reducing prediction outliers is that instead of find a simple model by reducing regressors, we could still use the complex model but add some highly correlated or non-relevant regressors. By doing that, the overall prediction accuracy (AVG(rho)) may decrease, but it may help to reduce the prediction outliers (sum(rho) for rho > 5).