FRE 9733 Big Data in Finance Week 2 Homework

Mengheng Xue

02/13/2019

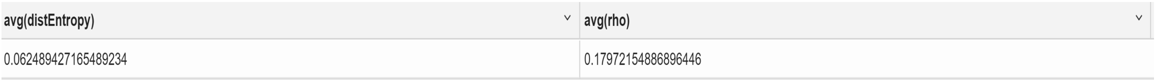
1 Random Forest (RF) vs Logistic Regression (LR)

1.1 Distribution Entropy & KL-Divergence

From Fig. 1 and Fig. 2, we could see that:

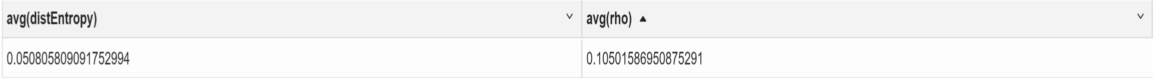
- The distribution entropy of $\{P, C, 3\}$ of LR is larger than RF. It means the uncertainty of next month states is larger when we apply LR. Also both are very small numbers since $P(C \to C) \approx 0.95$, which makes the uncertainty of next month state very small.

- The KL-Divergence of LR is also larger than RF. By the definition of KL-Divergence, it describes the "distance" between the true distribution and the predicted one. Therefore, we could say that the distribution predicted by RF is closer to the true distribution than LR. Accordingly, the prediction of RF is more accuracte.



avg(distEntropy) ˅	avg(rho) ˅
0.062489427165489234	0.17972154886896446

Fig. 1: Entropy and KL-Divergence of LR prediction of prepayment probability



avg(distEntropy) ˅	avg(rho) ▲ ˅
0.050805809091752994	0.10501586950875291

Fig. 2: Entropy and KL-Divergence of RF prediction of prepayment probability

1.2 Predicted Prepay vs Actual Prepay

As we discussed before, in Fig. 3 and Fig. 4, we can find that:

- When we group incentive feature, for RF, the predicted prepay is very close to the actual prepay probability. It means incentive is an important feature for RF model.

- However, LR does not perform very well, which is nearly uniform across the whole incentive range. It means incentive is an irreverent feature for LR model, since it does not affect our LR prediction.

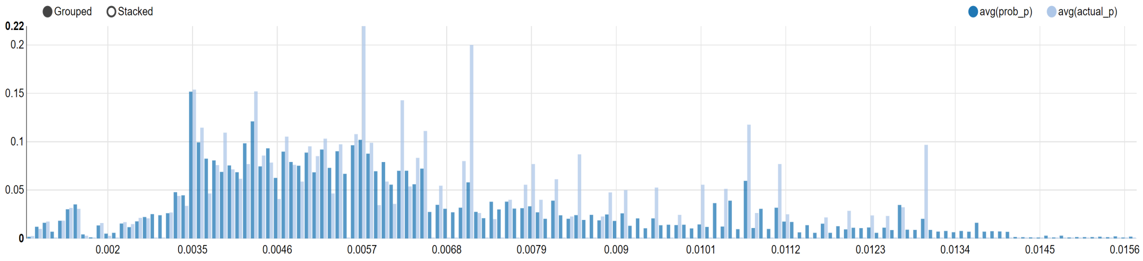


Fig. 3: predicted prepay vs actual prepay grouped by incentive using RF

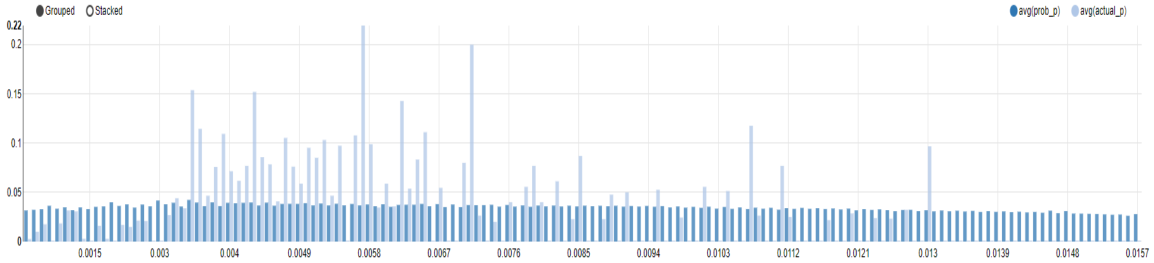


Fig. 4: predicted prepay vs actual prepay grouped by incentive using LR

1.3 Prepay Prediction vs 30 Days Delinquent Prediction using RF

From Fig. 5, we can find that age is also a well performed feature for RF model, and the prediction is very close to the actual prepay probability. However, for state 3, the prediction using RF may not as accurate as prepay prediction. We can see from Fig. 6, when age increases, RF cannot catch some extreme nonlinear trends. I think it is due to the state 3 data is more fluctuated than state P data, which makes it harder to predict.

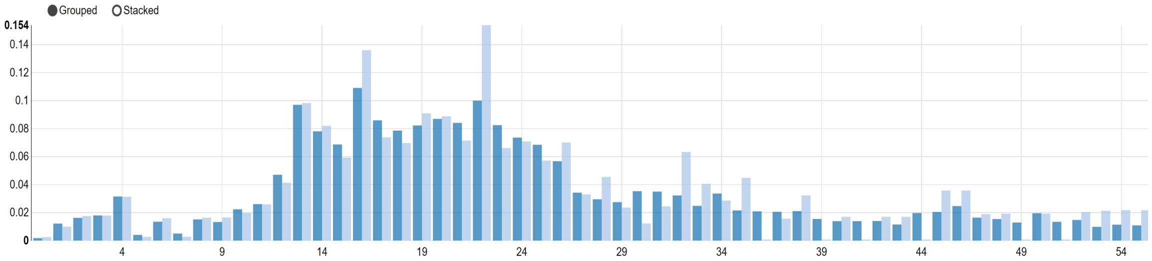


Fig. 5: 30 days delinquent probability prediction vs actual probability grouped by age using RF

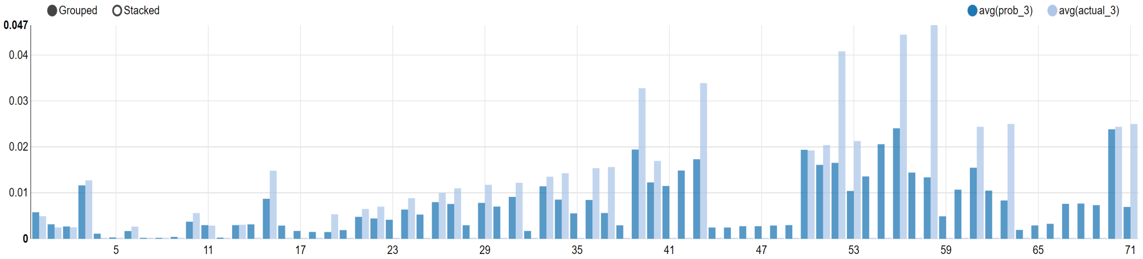


Fig. 6: 30 days delinquent probability prediction vs actual probability grouped by age using RF