

Result on Tian's data

- 47,189 rows originally, 311 rows $Y = 1$
- time span: 1980-2009
- After matching 12 month bankruptcy info, 41,003 rows, 254 rows $Y = 1$
- training set: 1980-2002
- testing set: 2003-2009
- Result based on $C = 0.01$, the best out-of-sample AUC is 0.62 when log scaled $C = -1$, too little training data

Result of L1 logistic regression $C = 0.01$:

- Selected features: rsize, FAT, prc2, FMTA, $X24(\log(AT))$, $X11((LCT-CH)/AT)$

Metric	In Sample	Out Sample
Accuracy	0.55	0.11
AUC	0.77	0.55
Percentage of BK		
1	0.38	0.2
2	0.20	0.0
3	0.11	0.2
4	0.11	0.0
5	0.07	0.0
6-10	0.12	0.6

Selected features using lasso feature selection in SAS



Result on our data

- 182,300 rows originally, 311 rows $Y = 1$, 162
- time span: 1980-2015
- After matching 12 month bankruptcy info, 162,565 rows, 1,016 rows $Y = 1$
- training set: 1980-2007
- testing set: 2007-2015


Result of L1 logistic regression $C = 0.359$, c selected from 3 fold cv using AUC as metric:

- Selected features: PRICE, LCTCHAT, LTMTA, ACTLCT, INVCHINVT, SALEAT, QALCT, APSALE, LOGAT, RELCT, REAT, OIADPSALE, FAT, LOGSALE, CASHMTA, RSIZE, CASHAT, LCTLT, NIMTA, CHLCT, SIGMA, NISALE, EXCESS_RETURN, EBITDPAT, LCTSALE, WCAPAT, OIADPAT, LTAT, INVTSALE, NIAT, MBE

Metric	In Sample	Out Sample
Accuracy	0.58	0.58
AUC	0.79	0.79
Percentage of BK		
1	0.44	0.39
2	0.17	0.22

3	0.11	0.18
4	0.09	0.07
5	0.04	0.04
6-10	0.13	0.09

Selected features using lasso feature selection in SAS

- 7 selected features on tian's paper: PRICE, SIGMA, NIMTA,FAT,LTMTA,EXCESS_RETURN,LCTAT
- Our result selects five from Tian except SIGMA and LTMTA 

Result of L1 logistic regression $C = 100$, only using 7 features on tian's paper

Metric	In Sample	Out Sample
Accuracy	0.534	0.561
AUC	0.767	0.781
Percentage of BK		
1	0.43	0.32
2	0.18	0.26
3	0.11	0.17
4	0.07	0.06
5	0.05	0.03
6-10	0.16	0.11

Result of L1 logistic regression $C = 100$, only using 8 features from SAS lasso feature selection

Metric	In Sample	Out Sample
Accuracy	0.549	0.609
AUC	0.774	0.804
Percentage of BK		
1	0.43	0.38
2	0.18	0.30
3	0.12	0.14
4	0.08	0.07
5	0.04	0.05
6-10	0.15	0.07

text matched with numerical dataset:

one year before dataset

62,993 observations

228 $Y=1$

text+numerical result based on 80 tfidf, metric is AUC:

- We can not explain these 80 tfidf features due to SVD.
- use TruncatedSVD to get 80 tfidf text features, 50% variance explained

- RandomForestClassifier max_depth=1, n_estimators=150
- But for 3 types of feature input, all use the same RFC model to predict, the 8 numerical variables still works better than text+numerical.

Model	Features	In Sample	Out Sample
RandomForestClassifier	text_80_tfidf	0.76	0.55
LogisticRegression	8 lasso features	0.81	0.79
RandomForestClassifier	total	0.83	0.81
RandomForestClassifier	8 lasso features	0.81	0.85

Model	Features	In Sample	Out Sample
RandomForestClassifier	text_80_tfidf	0.76	0.55
RandomForestClassifier	8 lasso features	0.81	0.85
RandomForestClassifier	total	0.85	0.83

text result based on DAN

Kfold = 2, epoch = 5 , AUC = 0.80

Kfold = 2, epoch = 20 , AUC = 0.65