빅데이터분석및응용

[과제 1] 인공신경망을 활용한 부도예측 모형 구축

변현정

1. Data & Variables

1) 입력변수와 출력변수

입력변수	x1(안전성-고정자산구성비율)
	x2(안전성-고정장기적합률)
	x6(안전성-순부채/총자산)
	x7(안전성-순자산배율)
	x13(안전성-유형순자산/총자산)
	x14(안전성-이익잉여금구성비율 1),
	x15(안전성-재고자산/유동자산)
	x27(수익성-이자부담률)
	x30(수익성-세후순손익추세)
	x32(유동성-당좌비율)
	x34(유동성-단기유동성비율)
	x36(활동성-매입채무회전기간 3)
	x39(활동성-경영자본회전율 1)
	x40(활동성-영업자산회전율 2)
	x43(생산성-총자본투자효율 2)
	x44,(현금흐름-경상수지비율 3)
	x45(현금흐름-경상현금흐름/총차입금)
	x47(현금흐름-현금이자보상배율 3)
	x48(현금흐름-DSCR2)
출력변수	Output(0,1)

입력변수로는 19 개의 변수가 선택되었다.이는 점이연상관계수, 다중공산성진단, 단계적 선택법을 차례대로 시행하였다. 종속변수와의 상관성, 독립변수들 간의 다중공산성, 변수의 중요한 정도를 평가하여 선택하였다. 출력변수로는 0 과 1 로 구성된 이항변수인 output 변수를 선택하였다. 입력변수로 안전성, 수익성, 유동성, 활동성, 생산성, 현금흐름 등 고르게 선택되었다.

2) 변수 선정 기준

독립변수를 선정하기 위해서 아래와 같이 세 가지 기법을 사용하였다.

-점이연상관계수 (point biserial correlation coefficient)

-다중공산성진단

-단계적선택법(stepwise selection)

첫 번째로 독립변수들 각각이 종속변수를 얼마나 잘 설명하고 있는 가를 알아보기 위하여 상관분석을 실시하였다. 독립변수는 연속 변수이고, 종속변수는 이항변수이기 때문에 상관분석 중 점이연상관계수를 구하였고, 그 상관관계가 확률이 유의미한 지 알아보았다. 결과로는 각각의 독립변수들은 종속변수에 대하여 모두 확률적으로 유의미하게 나타났다. 아래와 같이 별표 하나에서 두개의 표시를 보였고, 이는 모든 독립 변수들이 종속변수를 잘 설명하고 있다고 볼 수 있다. 그러나 어떤 독립변수를 제외하여야 하는지 상관관계분석을 통해서 알 수 없었다.

<독립변수들의 상관 분석 결과>

		output																	
output	Pearson Correlation	1	x9	Pearson Correlation	-.436**	x19	Pearson Correlation	.235**	x29	Pearson Correlation	-.361**	x39	Pearson Correlation	.078**
	Sig. (2-tailed)			Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	2766		N	2766		N	2766		N	2766		N	2766
x1	Pearson Correlation	.093**	x10	Pearson Correlation	-.385**	x20	Pearson Correlation	.385**	x30	Pearson Correlation	-.229**	x40	Pearson Correlation	.164**
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	2766		N	2766		N	2766		N	2766		N	2766
x2	Pearson Correlation	-.139**	x11	Pearson Correlation	-.392**	x21	Pearson Correlation	.360**	x31	Pearson Correlation	.180**	x41	Pearson Correlation	.073**
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	2766		N	2766		N	2766		N	2766		N	2766
x3	Pearson Correlation	.083**	x12	Pearson Correlation	-.286**	x22	Pearson Correlation	.153**	x32	Pearson Correlation	.235**	x42	Pearson Correlation	.057**
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.003
	N	2766		N	2766		N	2766		N	2766		N	2766
x4	Pearson Correlation	-.079**	x13	Pearson Correlation	.540**	x23	Pearson Correlation	.362**	x33	Pearson Correlation	.437**	x43	Pearson Correlation	.273**
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	2766		N	2766		N	2766		N	2766		N	2766
x5	Pearson Correlation	.043*	x14	Pearson Correlation	.672**	x24	Pearson Correlation	.339**	x34	Pearson Correlation	.487**	x44	Pearson Correlation	.359**
	Sig. (2-tailed)	.025		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	2766		N	2766		N	2766		N	2766		N	2766
x6	Pearson Correlation	-.605**	x15	Pearson Correlation	-.279**	x25	Pearson Correlation	.396**	x35	Pearson Correlation	.167**	x45	Pearson Correlation	.416**
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	2766		N	2766		N	2766		N	2766		N	2766
x7	Pearson Correlation	.501**	x16	Pearson Correlation	-.414**	x26	Pearson Correlation	.428**	x36	Pearson Correlation	-.248**	x46	Pearson Correlation	.473**
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	2766		N	2766		N	2766		N	2766		N	2766
x8	Pearson Correlation	-.454**	x17	Pearson Correlation	-.517**	x27	Pearson Correlation	-.531**	x37	Pearson Correlation	-.046*	x47	Pearson Correlation	.429**
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.016		Sig. (2-tailed)	.000
	N	2766		N	2766		N	2766		N	2766		N	2766
	Pearson Correlation	-.436**	x18	Pearson Correlation	.202**	x28	Pearson Correlation	-.523**	x38	Pearson Correlation	-.280**	x48	Pearson Correlation	.383**
				Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
				N	2766		N	2766		N	2766		N	2766

두 번째 방법으로 독립 변수들 사이에 다중공산성이 존재하는 지 알아보기 위해 다중공산성진단을 시행하였다. 일반적으로 VIF 계수가 10 이상이고 Tolerance 가 0.1 이하면 다중공산성이 존재한다고 본다. 이를 기준으로 평가하였다. 다중공산성이 있는 변수들을 추려보니 x1, x6, x8, x9, x10, x11, x12, x17, x18, x19, x20, x21, x22, x23, x24, x25, x29, x31, x32, x33, x34, x35, x39, x41, x47 이었다.

<다중공산성 진단 결과>

Coefficients[a]

Model		Collinearity Statistics Tolerance	VIF			
1	x1	.034	29.152	x27	.093	10.797
	x2	.100	9.988	x28	.124	8.042
	x3	.168	5.941	x29	.013	76.515
	x4	.264	3.793	x30	.375	2.670
	x5	.561	1.781	x31	.035	28.346
	x6	.074	13.585	x32	.038	26.531
	x7	.423	2.364	x33	.087	11.475
	x8	.049	20.206	x34	.074	13.556
	x9	.089	11.285	x35	.019	52.668
	x10	.089	11.213	x36	.352	2.845
	x11	.063	15.774	x37	.299	3.346
	x12	.032	30.979	x38	.186	5.384
	x13	.101	9.924	x39	.033	30.084
	x14	.203	4.917	x40	.273	3.657
	x15	.191	5.245	x41	.029	33.953
	x16	.105	9.566	x42	.883	1.133
	x17	.051	19.508	x43	.396	2.523
	x18	.036	27.645	x44	.618	1.619
	x19	.018	55.699	x45	.252	3.973
	x20	.007	148.656	x46	.166	6.011
	x21	.010	96.465	x47	.061	16.332
	x22	.056	17.763	x48	.380	2.632
	x23	.012	85.196			
	x24	.008	128.862			
	x25	.065	15.406			
	x26	.148	6.766			

a. Dependent Variable: output

어떤 변수들 끼리 다중공산성이 있는 지 알아보기 위해 Condition Index 표를 살펴보았다. 그 계수가 15 이상일 때를 찾아보니 19 차원부터 49 차원까지였다. 보통 계수가 15 이상일 때를 기준으로 Variance Proportion 가 큰 나란히 존재하는 두 변수를 찾는다. 처음으로 진단했을 때 (x20, x21), (x23, x24), (x27, x28), (x31, x32), (x33, 34) 이었다.

예를 들어, (x20, x21)는 Variance Proportion 이 .87 .76 으로 나란히 존재하였고. 변수 이름을 살펴보니 총자산경상이익률과 총자산순이익률로 비슷한 변수였다. 이 두 변수 중 종속 변수와 상관계수가 작은 쪽인 x21 을 제거하였고 다른 변수 쌍도 동일한 방법으로 시행하였다.

변수 제거 후 다중공산성진단을 여러 번 반복하였고 같은 방법으로 변수를 제거하였다. 그 결과 아래의 표와 같이 VIF 계수가 10 미만인 독립 변수들을 찾았다.

<다중공산성이 높은 변수 제거 후 VIF 계수/ 단계적 선택법 시행 결과>

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	x1	.209	4.789
	x2	.242	4.128
	x3	.205	4.874
	x5	.592	1.689
	x6	.105	9.559
	x7	.437	2.291
	x9	.265	3.772
	x13	.125	7.984
	x14	.215	4.660
	x15	.239	4.185
	x17	.209	4.790
	x20	.225	4.438
	x27	.371	2.693
	x30	.529	1.889
	x34	.164	6.104
	x36	.459	2.180
	x37	.311	3.214
	x38	.211	4.740
	x39	.291	3.433
	x40	.284	3.527
	x42	.943	1.060
	x43	.452	2.214
	x44	.675	1.482
	x45	.263	3.801
	x46	.240	4.171
	x47	.270	3.700
	x48	.454	2.201
	x32	.193	5.188

Model Summary[ab]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.672[a]	.451	.451	.370
2	.708[b]	.502	.501	.353
3	.748[c]	.560	.560	.332
4	.771[d]	.594	.593	.319
5	.783[e]	.613	.612	.312
6	.792[f]	.627	.626	.306
7	.796[g]	.633	.632	.303
8	.799[h]	.638	.637	.301
9	.802[i]	.643	.641	.299
10	.804[j]	.647	.645	.298
11	.808[k]	.653	.652	.295
12	.811[l]	.657	.656	.293
13	.812[m]	.660	.658	.292
14	.814[n]	.663	.661	.291
15	.816[o]	.666	.664	.290
16	.817[p]	.668	.666	.289
17	.818[q]	.670	.668	.288
18	.820[r]	.672	.670	.287
19	.820[s]	.673	.670	.287
20	.820[t]	.672	.670	.287
21	.820[u]	.673	.671	.287
22	.821[v]	.674	.672	.287
23	.821[w]	.675	.672	.286
24	.822[x]	.675	.673	.286
25	.822[y]	.676	.674	.286
26	.823[z]	.677	.674	.285
27	.823[aa]	.678	.675	.285

세번째 방법으로는 단계적선택법을 시행하여 중요한 변수를 포함하고 설명력이 떨어지는 변수를 탈락시켰고, 위의 오른쪽 표와 같은 결과를 얻었다. 이를 살펴보면 11 번 모형은 Adjusted R Square 의 값이 0.652 이었고 그 이후에 그 값이 조금씩 증가하였다. 20 번 모형에서 x3 독립변수가 탈락되었다. 그 이후 Adjusted R Square 값의 증가가 둔해졌다. 따라서 21 번 모형은 충분히 독립변수들의에 의한 설명력이 높다고 판단하였고 최종적으로 Adjusted R Square 이 0.671 인 19 개의 변수들로 구성 된 21 번 모형을 선택하였다. 선택된 독립변수들은 x14, x27, x1, x6, x13, x39, x43, x30, x44, x47, x48, x36, x15, x32, x7, x34, x40, x2, x45 이다.

2. Model Development

1) Training, Validation, Test set 비율

Training set: 50%
Validation set: 30%
Test set: 20%

2766 개의 데이터에서 Training set 은 50%로 정하였고, 과적합(overfitting)을 방지하기 위하여 Validation set 은 30%로 마지막으로 최종 모델을 평가하기 위해 Test set 20%를 사용하였다.

2) Network Architecture

Hidden layer: 1 층
Input node : 19 개
Hidden node: 10 개
Output node: 1 개

은닉층은 1 층, 은닉 노드 10 개, 출력 노드 1 개, 입력 노드 19 개로 총 노드 수가 37 개(19*2+1)를 넘지 안는 선에서 정하였다.

3) Learning Algorithm

Weight update rule: Random seed
Transformation function: Sigmoid function
Stopping rules: 15

Weight update rule 은 Set Random Seed 모드로 노드가 시행될 때 마다 매번 같은 결과가 나오도록 하기 위해 random 으로 설정하였다. Transformation function(Activation function)은 Sigmoid function 을 사용하였다. Stopping rules 는 default 의 값 15 로 두었다.

3. Result & Analysis

1) 통계 기법, 인공 신경망 기법, 의사결정나무 기법 비교

회귀분석모델의 경우 위의 Model Summary 표에서 보는 것과 같이 R^2 가 0.673 으로 충분히 컸고 모델의 설명력이 좋다고 할 수 있다. 아래는 인공 신경망 기법을 통해서 분류한 결과이다. Training set 에서 정확도가 94.32%, 오차율이 5.673%였고, Test set 에서 정확도가 94.32%로 Training set 과 둘 다 90%이상 매우 좋은 정확성을 보였고, 그 수준도 비슷하여 모형의 일반화에 문제가 없었다. 두 모델을 비교하였을 때, 회귀분석의 설명력도 충분히 컸고 인공신경망의 경우 분류 정확도가 94%이상이었다. 두 기법은 둘 다 그 성과가 우수하여 어느 모델이 탁월하게 좋다고 할 수는 없었다.

의사결정나무의분석 결과 Training 의 정확도가 95.93%으로 세 기법 중 가장 높았고 Test set 에서 91.05%을 보였다. 그러나 의사결정나무 트레이닝 셋의 정확도가 가장 높았다 하더라도 테스트 셋과의 정확도 차이가 세 가지 기법 중 가장 컸으므로 과적합의 문제가 예상되어 좋은 기법이라 할 수 없었다.

<인공신경망분석 결과>

'Partition'	1_Training		2_Testing	
Correct	2,062	94.37%	548	94.32%
Wrong	123	5.63%	33	5.68%
Total	2,185		581	

<의사결정나무분석 결과>

'Partition'	1_Training		2_Testing	
Correct	2,096	95.93%	529	91.05%
Wrong	89	4.07%	52	8.95%
Total	2,185		581	

회귀분석모델의 경우 ANOVA 통계 검증을 시행한 결과 아래 표와 같이 유의 확률이 0 로 통계적으로 유의한 수준이었음을 확인할 수 있었다. 그러나 인공신경망의 경우 통계 검증을 어떻게 하는 지 지식이 없어서 시행하지 못했다.

<회귀분석모델 통계 검증>

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
21	Regression	465.509	19	24.500	297.704	.000[v]
	Residual	225.991	2746	.082		
	Total	691.500	2765			

2) 세 모형을 예측용 기업에 대해 부실 여부 예측

<예측용 데이터 회귀분석모델 적용 결과>

	PRE_1	PRE_2		PRE_1	PRE_2
3226	.	.60016	3257	.	.08045
3227	.	-.08772	3258	.	.44207
3228	.	.11490	3259	.	.36904
3229	.	.00719	3260	.	.06137
3230	.	.26562	3261	.	.10412
3231	.	.01306	3262	.	-.05742
3232	.	.24188	3263	.	-.07242
3233	.	-.02140	3264	.	.08588
3234	.	.01960	3265	.	-.17904
3235	.	.17016	3266	.	-.44804
3236	.	.12863	3267	.	-.14581
3237	.	.44511	3268	.	.47897
3238	.	-.23251	3269	.	.15243
3239	.	-.37681	3270	.	.24557
3240	.	.01130	3271	.	.22386
3241	.	-.18680	3272	.	.27621
3242	.	.02844	3273	.	.44856
3243	.	.19130	3274	.	.20105
3244	.	.64099	3275	.	.30915
3245	.	.24162	3276	.	.54680

위의 표와 같이 회귀분석모형을 통해 데이터를 예측해 본 결과 부도기업 371 개, 건실 기업 329 개로 나왔다. SPSS statistic 을 사용하여 output 값을 예측하였지만 카테고리 형태가 아닌 연속형 실수로 보여졌기 때문에 점이연상관분석 시 0.5 를 기준으로 함을 고려하여 0.5 보다 크면 1, 작으면 0 으로

분류하였다. 실제데이터의 부도/건실 기업이 반반 이었음을 고려할 때, 회기분석모델의 정확도는 97%, 오류율은 3% 라 할 수 있다.

<회귀분석을 통한 예측 결과>

회귀분석모델	
부도	371
건실	329

회기분석모델	
Correct	97.00%
Wrong	3.00%

<예측용 데이터 인공 신경망 적용 결과>

		x39	x40	x41	x42	x43	x44	x45	x46	x47	x48	field51	field52	field53	field54	$N-output	$NC-output
1	2	1.542	11.025	1.501	0.080	70.302	121.686	19.276	17.980	393.814	45.043					1	1.000
2	8	3.227	8.397	2.949	0.681	72.670	107.033	220.513	52.703	2015.000	827.273					1	0.837
3	7	2.124	7.048	1.846	-0.367	72.670	131.249	35.302	15.472	447.222	24.609					1	0.992
4	7	1.693	5.145	1.684	-0.545	51.958	120.019	42.829	21.156	779.365	787.302					1	1.000
5	6	0.679	3.225	0.679	0.068	32.654	113.784	50.832	29.263	1057.143	1057.143					1	0.991
6	1	1.125	6.527	1.098	-1.013	34.412	113.458	12.195	4.804	476.087	195.833					1	1.000
7	8	1.491	7.675	1.488	0.450	43.178	111.611	68.809	52.703	986.207	110.425					1	0.898
8	0	1.999	3.653	1.814	-0.062	30.709	123.080	181.853	20.251	3875.000	1578.571					1	0.959
9	9	1.254	2.762	1.188	0.251	37.408	112.825	55.406	22.028	766.195	54.215					1	0.977
10	8	1.798	3.919	1.668	0.158	46.251	98.946	182.474	18.995	2400.000	114.107					1	0.933
11	6	1.487	3.287	1.484	0.021	70.146	131.249	244.655	52.703	2905.556	1578.571					1	1.000
12	0	1.079	2.403	1.077	0.158	27.163	108.765	21.161	7.207	310.870	27.375					1	0.951
13	2	1.238	2.454	0.896	0.162	29.956	107.591	23.088	11.981	326.250	23.768					1	0.945
14	8	1.156	4.828	1.153	-0.171	18.928	113.345	48.452	44.250	3875.000	1578.571					1	0.999
15	4	0.897	4.465	0.834	0.062	26.111	111.754	20.946	11.243	333.333	39.441					1	1.000
16	0	1.937	7.368	1.911	0.681	55.406	100.832	67.582	44.561	1099.422	73.949					1	0.829
17	6	1.962	5.277	1.940	-0.698	61.748	98.571	27.367	14.867	735.135	24.927					1	1.000
18	0	1.220	2.104	1.177	-0.062	30.859	131.249	8.662	4.383	307.692	39.372					1	0.875
19	1	1.378	2.977	1.315	-0.145	27.382	112.361	16.639	14.559	650.000	32.517					1	0.980
20	6	1.723	5.337	1.675	0.065	45.132	124.370	48.452	28.177	769.288	234.462					1	0.976
21	0	1.701	6.636	1.681	0.260	46.249	109.013	30.896	17.696	478.462	480.000					1	0.962
22	4	2.447	4.404	2.283	-0.530	48.891	115.100	101.860	50.602	942.308	98.544					1	1.000

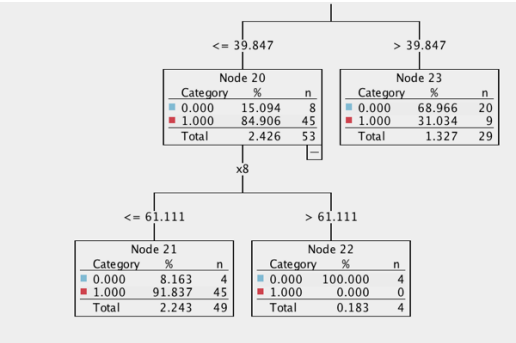
		x39	x40	x41	x42	x43	x44	x45	x46	x47	x48	field51	field52	field53	field54	$N-output
349	9	2.631	6.336	1.888	-0.526	33.152	107.818	73.446	17.980	678.356	92.632					1
350	3	1.255	6.123	1.221	-0.040	16.658	104.831	26.806	14.659	482.878	27.199					1
351	2	0.906	4.121	0.831	-0.420	5.464	94.949	5.401	7.635	381.029	93.300					0
352	7	2.314	4.695	2.312	0.083	23.750	90.357	30.248	14.589	270.650	35.273					0
353	7	1.197	3.145	1.185	-0.105	26.606	91.400	20.364	13.478	262.319	262.319					0
354	7	1.786	8.301	1.706	0.150	31.653	108.495	154.560	24.383	3153.333	180.526					0
355	2	3.227	10.205	1.990	0.681	69.783	120.947	244.655	48.566	2341.935	231.313					0
356	2	2.836	6.001	2.504	0.184	48.386	110.750	105.794	39.128	601.493	171.101					0
357	9	1.806	5.624	1.805	0.067	5.464	131.249	-10.966	-4.502	-134.906	-8.471					0
358	3	2.169	3.917	2.051	-1.013	50.916	92.717	37.613	17.709	468.116	44.298					0
359	0	1.827	3.944	1.382	-1.013	35.890	99.120	65.412	17.875	886.538	34.172					0
360	0	0.996	4.276	0.971	-0.503	35.876	80.309	48.990	21.037	449.451	33.252					0
361	1	2.115	8.308	1.793	0.073	64.087	106.058	8.817	-1.278	189.024	29.468					0
362	3	1.589	3.968	1.567	0.551	19.103	80.309	18.372	9.971	332.779	51.506					0
363	7	1.674	4.715	1.602	-0.037	21.814	90.221	17.305	8.629	341.270	22.775					0
364	6	1.587	3.156	1.267	-0.437	35.975	110.974	22.600	10.678	360.563	44.226					0
365	2	1.283	2.682	1.272	0.025	33.447	115.068	14.971	22.471	210.526	42.070					0
366	9	1.614	3.172	1.382	0.455	20.948	86.232	12.480	16.408	210.791	34.499					0
367	4	0.600	1.590	0.557	-0.424	5.464	80.309	-10.966	-7.598	-134.906	-8.471					0
368	7	2.062	4.287	1.813	0.137	37.720	87.948	-1.786	-1.373	51.111	8.233					0
369	5	0.807	2.280	0.751	-0.277	13.514	86.529	8.460	1.041	222.165	23.536					0

인공신경망으로 예측한 결과 아래 표와 같은 결과를 얻을 수 있었고 output 을 확인해 본 결과 부도가 325 개, 건실한 기업이 375 개로 나왔다. 실제 데이터에서는 부도와 건실한 기업이 반반으로 각각 350 개였지만 구축한 인공신공망의 경우 잘못 분류한 기업이 25 개이었다. 정확도는 96.5% 이고, 오류율은 3.5% 이었다.

<인공 신경망을 통한 예측 결과>

부도	325		Correct	3.57%
건실	375		Wrong	96.43%

<예측용 데이터 의사결정나무 적용 결과>



Rule 10 – estimated accuracy 86.13% [boost 99.4%]

예측용 데이터를 의사결정나무 기법으로 시행해본 결과 위의 표에서 보는 것과 같이 정확도 86.13%로 세 기법 중 가장 낮은 정확도를 보여주었고 예상했던 것처럼 과적합의 문제가 발생한 것으로 보인다.

4. Concluding Remarks

과제를 하면서 느낀 한계점은 독립변수의 수를 줄이는 과정이었다. 변수를 14 개에서 19 개까지 보는 등 여러 번 실험을 해보았다. 변수 수를 하나씩 줄일수록 오류율이 조금 줄어들었지만 매우 미미한 수준이었다. 그러나 그 작은 수준도 혹시나 개개의 회사 입장에서는 큰 타격을 줄 수 있겠다는 생각에 변수를 14 개보다 19 개로 설정하였지만 효과는 크게 없었고 분류 정확도도 크게 좋아지지 않았다.

세 가지 기법을 통해서 추린 독립변수들을 살펴보니 안정성의 변수들이 다른 변수에 비해 대거 포함되어있었다. 만약에 변수들에 대해 지식이 있다면 어떤 변수가 중요하고 어떤 변수를 제거해야하는지 통계적으로만 실행했을 때보다 더 효과적으로 변수를 선택할 수 있을 것이다.

실험을 하기 전에는 인공신경망을 통한 분석 결과가 회기분석모형보다 더 좋은 결과를 낼 것이라고 예상했지만, 회기분석모델이 0.43%정도 더 좋은 성과를 내었다. 인공 신경망의 모형도 training: 94.37%, test: 94.32%, prediction: 96.43%으로 세 가지 모두 비슷한 확률의 정확도로 과적합이 일어나지 않은 괜찮은 결과를 보여주었다. 의사결정나무의 경우 training: 95.93%, test: 91.05%, prediction: 86.13% 로 그 차이가 커서 부도 예측 데이터에서는 좋은 기법이 아닌 것으로 생각된다.

회귀분석모형이 인공신경망보다 미미하나 더 좋은 성과를 낸 것에 대한 문제점으로는 아직 인공신경망의 Architecture 부분과 Learning algorithm 을 어떻게 설정하여야 더 좋은 퍼포먼스를 내는 지에 대하여 아직 정확한 지식과 감이 없어서 발생한 것이라 사료된다. 어떤 값을 설정해야 더 좋을 지는 앞으로 더 공부해야 할 부분임을 지각하였다.