Journal of The Korea Society of Computer and Information
Vol. 21 No. 4, pp. 63-71, April 2016
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.4.063

Analyzing empirical performance of correlation based feature selection with company credit rank score dataset - Emphasis on KOSPI manufacturing companies -

Youn Chang Nam*, Kun Chang Lee**

Abstract

This paper is about applying efficient data mining method which improves the score calculation and proper building performance of credit ranking score system. The main idea of this data mining technique is accomplishing such objectives by applying Correlation based Feature Selection which could also be used to verify the properness of existing rank scores quickly. This study selected 2047 manufacturing companies on KOSPI market during the period of 2009 to 2013, which have their own credit rank scores given by NICE information service agency. Regarding the relevant financial variables, total 80 variables were collected from KIS-Value and DART (Data Analysis, Retrieval and Transfer System). If correlation based feature selection could select more important variables, then required information and cost would be reduced significantly. Through analysis, this study show that the proposed correlation based feature selection method improves selection and classification process of credit rank system so that the accuracy and credibility would be increased while the cost for building system would be decreased.

▸ Keyword : credit rating system, Ordinal Logistic regression, Correlation based Feature Selection, KOSPI

I. Introduction

본 연구는 신용평가기관이 기업의 재무정보를 활용함에 있어서 신용평가에 영향을 미치는 설명변수를 통계적 방법 및 데이터마이닝 기법을 통해 확인하고, 이를 바탕으로 하는 신용평점 예측 모형을 구축하여 실증하는 것에 관한 연구이다.

신용평가는 전문적이고 객관적인 신용평가기관이 채권 및 기업의 원리금 상환 능력을 측정하고 그에 적합한 등급을 부여하는 제도이다. 이를 위해 각 신용평가기관은 현재의 기업 상태를 정확하게 진단하고 그에 따른 미래의 현금흐름을 예측해야 할 필요가 있다. 이 때, 신용평가기관은 기업의 현금흐름에 영향을 주는 다양한 재무적 요소들과 함께 해당 기업과 관련한 각종 사건 및 시장 환경의 변화와 같은 비재무적 요소 등의 변수, 그리고 산업 환경 과 경영환경, 사업경쟁력 등 다양한 평가지표들을 종합하여 신용등급을 부여하게 된다. 즉, 개별기업마다 독자적인 신용평가시스템(Credit Rating System)을 구축하고 기업의 재무제표 정보 및 역사적 부도 경험 관계를 통계적인 방법론을 활용한 계량 모형에 반영하기도 하며 비재무적 위험 요인을 전문가적인 판단을 비계량 모형에 적용하는 것이다. 따라서 신용평가기관은 신용평가제도를 통하여 자금을 제공하

• First Author: Youn Chang Nam, Corresponding Author: Kun Chang Lee
*Youn Chang Nam(nyc0512@nate.com), Dept. of Global Business Administration, Sungkyunkwan University
**Kun Chang Lee (kunchanglee@gmail.com), SKK Business School/SAIHST, Sungkyunkwan University
• Received: 2015. 11. 30, Revised: 2016. 03. 22, Accepted: 2016. 04. 16.
• This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A3A2038108).

는 투자자와 자금을 필요로 하는 기업을 연결시켜주는 매개역할을 한다. 신용등급을 기반으로 정보 불균형 문제를 해소하여 투자자의 정보획득 비용과 시간을 절약하게하고 불확실성을 감소시켜 투자의사결정을 보조하는 것이다. 이는 곧 증권 시장의 활성화를 이끌어내므로 신용평가기관의 정확한 신용분석과 평가등급 부여를 바탕으로 하는 정보의 신뢰성의 증진은 자본시장의 발전을 위한 선결과제라고 할 수 있다.

본 연구에서는 기업 재무 환경의 변화를 고려하여 최신 자료들을 사용함과 더불어 순서형 로지스틱 회귀분석과 데이터마이닝을 통한 특성선택 기법을 접목하여 신용평가에 영향을 미치는 변수를 식별하고자 한다. 또한 도출된 설명변수들이 기업의 신용등급에 관한 실질적인 예측능력을 지니는 지를 분류 모형에 적용하여 표본들을 신용등급에 맞게 분류하고 그 실증적 예측 능력을 비교하고자 한다.

II. Preliminaries

1. Credit Evaluation

우리나라의 신용평가 회사들은 원리금 상환능력정도에 따라 AAA~D까지 20개의 신용등급을 제공하고 있다. AAA~BBB까지가 원리금의 상환능력이 인정되는 투자적격등급이며, BB~D는 환경변화에 따라 원리금의 상환능력의 변동이 있는 투기적 등급으로 분류된다. 신용평가회사가 회사채 신용등급을 평가하는 요인으로, 기업의 상환능력에 영향력이 있는 기업 내부 및 외부의 위험요소, 재무 및 비재무 요소 등을 포괄적으로 고려하고 있다. 이때, 재무자료는 다른 어떠한 자료보다도 기초적으로 중요하기 때문에 신용등급에 많은 영향을 미친다. 따라서 신용평가에 대한 선행연구들은 대부분 재무자료에 기초하고 있다.[1,2] 다음 Table 1은 본 연구에서 다루고 있는 NICE평가정보의 신용평점 등급 체계에 관한 표이다.

Table 1. KIS Credit Rating System

Rating / Score		Moody's	S&P
1st Rank	97.5 ~ 100 (extremely strong)	Aaa	AAA
2nd Rank	92.5 ~ 97.5 (Very strong)	Aa1 Aa2	AA+ AA
3rd Rank	85 ~ 92.5 (strong)	Aa3	AA−
4th Rank	75 ~ 85 (good)	A1 A2	A+ A
5th rank	65 ~ 75 (adequate)	A3	A−
6th rank	50 ~ 65 (less vulnerable)	Baa1 Baa2	BBB+ BBB
7th rank	33 ~ 50 (more vulnerable)	Baa3	BBB−
8th rank	15 ~ 33 (currently vulnerable)	Ba B Caa	B CC CCC
9th rank	5 ~ 15 (currently highly vulnerable)	Ca	C
10th rank	0 ~ 5 (extremely vulnerable)	C	D
Default			

2. Feature Selection

특성 선택은 기계학습에서 학습 알고리즘의 적용 이전에 불필요한 변수들을 배제하여 실제 분류 성능에 영향을 주는 특성 변수들만을 추출함으로써 학습 모형 자체의 정확도를 향상시키고 학습 시간 및 필요한 데이터의 양을 감소시키는 역할을 한다. 본 연구에서는 상관관계 기반에 의한 특성선택(Correlation based Feature Selection) 기법을 사용하였다.[3] 보다 세부적으로, 속성평가 방법으로는 상관관계 기반 속성 평가 방법(CfsSubsetEval)을, 속성 검색 방법으로는 최적우선탐색(Best first search)을 사용했다. 상관관계 기반에 의한 특성선택은 잘 선택된 특성 군은 상호간에는 상관관계가 높지 않지만 분류 집단(Class)과는 보다 높은 상관관계에 있다는 것에 기반을 한다. 따라서 개별 속성 값에 대한 엔트로피 및 목표클래스와 속성들 간의 피어슨 상관 계수(Pearson's correlation coefficient)를 조건부 확률로 계산하여 전체 속성들의 확률 분포도를 최대한 가깝게 표현할 수 있는 최소 개수의 속성집합을 찾는 방법이다.[4]

본 연구에는 최적우선탐색 가운데 베스트 퍼스트 작동 방식으로 후진(backward)방식을 채택하였는데 이는 후진방식이 대체로 유의미하다고 판단되는 독립 변수가 많을 경우에 예측 효율성이 증가하기 때문에 본 연구에서 확인하고자 하는 신용등급에 영향을 미치는 변수의 식별에 용이하다고 판단되기 때문이다.[5]

아래 Figure 1은 상관관계 기반에 의한 특성선택 모형을 본 연구에 맞게 수정한 연구 모형이다.

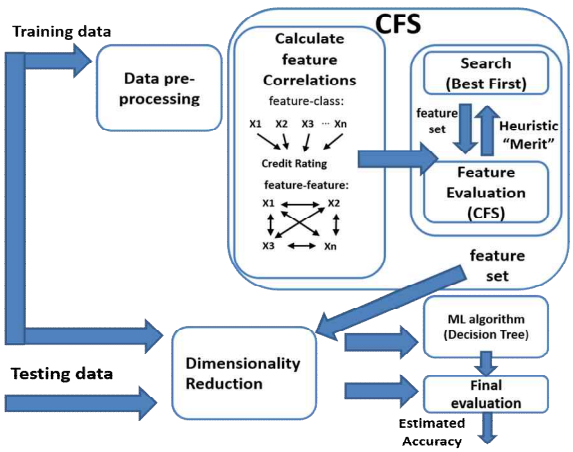

Fig. 1. Model of Correlation based Feature Selection

3. Hypotheses

본 연구는 코스피 시장을 연구의 대상으로 삼고 특성선택 기법을 활용하여 국내 기업의 신용평가에 영향을 미치는 변수들을 식별하는데 기여함과 더불어 식별된 변수들을 통해 신용평가 예측 모형을 구축함으로써 선행연구들을 확장하고자 한다.

우선적으로는 KIS-Value에서 선행 연구에서 다루어진 기업의 신용등급에 영향을 미치는 변수들을 포함하여 재무비율에 관한 정보들을 중심으로 독립변수들을 구성하였다. 이때 KIS-Value가 제공하는 가능한 한 많은 변수들을 일차적으로 선발한 후 회귀분석과 특성선택 기법을 통해 모형 구축에 필요한 변수들을 다시 선정함으로써 변수 선택의 자의성을 배제하고 설명 변수들의 적합성을 비교 검토하고자 하였다.

연구가설 1 : 신용등급 분류에 영향을 미치는 독립 변수를 선별할 때 회귀분석으로 선정된 변수들을 상관관계 기반에 의한 특성선택을 통해 다시 한 번 추출할 경우 변수의 수가 감소할 것이다.

한편, 이렇게 도출된 변수들을 바탕으로 순서형 로지스틱 회귀분석 활용하여 그 실질적 분류 예측력을 검증해보고자 한다. 이때, 앞서 제시되었듯이 특성선택 기법을 접목하였을 때 훨씬 적은 수의 변수로도 신용등급의 분류가 가능하다는 해외 선행연구들의 사례와 마찬가지로 본 연구에서도 특성선택 기법을 적용하여 구성된 데이터셋이 더 적은 변수로도 유의미한 예측력을 보일 것이라는 점을 T-test를 통해 확인하고자 한다. 즉, 각 모형의 예측값과 실제값의 차이를 비교하였을 때 유의미한 차이가 없다면 특성선택 기법이 신용등급의 분류에 효과적으로 적용될 수 있음을 의미한다.

연구가설 2 : 상관관계 기반에 의한 특성선택을 통해 선정된 변수들로 구성된 데이터셋과순서형 로지스틱 회귀분석만을 통해 선정된 변수들로 구성된 데이터셋간 예측력에 있어서 유의미한 차이를 보이지 않을 것이다

III. The Proposed Scheme and Analysis Result

1. Sample Selection

본 연구에 사용된 자료는 2009년부터 2013년까지 유가증권시장에 상장된 기업으로 한정하였다. 신용평가사의 신용평점 부여는 재무제표 공시 이후에 사후적으로 갱신되는데 기재정정 등으로 인한 수정이 이루어지는 점을 감안하여 현재 등재된 최신 자료인 2014년을 제외하고 5개년을 선정하였다.

또한 1) 한국표준산업분류 상에서 금융업을 제외한 제조업 3) 결산월이 12월인 기업 3) 감사의견이 적정인 기업 등의 조건을 모두 만족하는 기업으로 대상을 국한하였다. 1)은 회계원칙의 적용이나 재무제표의 보고형태, 계정과목 등에서 일관성을 도모하고자 제외하였고 2)는 신용평가 시점의 동질성을 유지하기 위한 조건이고 3)은 기업의 존속가능성의 부족으로 인한 극단적인 영향을 배제하기 위한 설정이다. 자본잠식 및 재무적 불안정성으로 인해 해석에 편의(biased)를 유발할 우려가 있기 때문이다. 이때, KIS-Value에서 결측치로 제공되는 기업의 재무정보라 하더라도 동사의 KIS-LINE 및 한국상장사협의회에서 제공하는 TS2000에서 재무비율 산출식에 해당하는 개별정보들을 추가로 수집하여 직접 결측치를 보완하였다. 마지막으로 신용평점이 없는 기업의 표본을 제외하여 2009년부터 2013년까지의 5개년에 해당하는 전체 연구대상 기간동안 총 2,084개의 기업별-연도별 횡단표본을 얻었다.

상세한 내역은 다음 Table 2와 같다. 또한 아래 Table 3은 표본 기업들의 연도별 신용평점 분포와 평균 평점을 나타낸 표이다.

Table 2. Data Set of Sample Companies

Data Set	'09	'10	'11	'12	'13	Total
KOSPI Listed Frims	**819**	**842**	**896**	**918**	**922**	**4397**
Non-Manufacturing Firms	(388)	(410)	(457)	(473)	(479)	(2207)
KOPSI Manufacturing Frims	**431**	**432**	**439**	**445**	**443**	**2190**
Fiscal Year-End (Non-Dec)	(18)	(17)	(17)	(17)	(16)	(85)
No Record of Credit Rating	0	(2)	0	(3)	0	(5)
Non-unqualified Opinion	(6)	0	(1)	(3)	(6)	(16)
Lack of financial information	(7)	(11)	(6)	(6)	(8)	(38)
Sample Companies	**400**	**402**	**415**	**416**	**413**	**2046**

Table 3. Number of companies scored in Credit rating system over '09 ~ '13

Credit Rating	'09	'10	'11	'12	'13	Total
1st Rank	3	2	1	1	3	10
2nd Rank	35	38	38	45	41	197
3rd Rank	47	50	42	44	52	235
4th Rank	66	68	60	65	67	326
5th rank	78	87	78	75	76	394
6th rank	72	67	83	68	63	353
7th rank	54	50	53	52	54	263
8th rank	27	31	34	35	39	166
9th rank	16	7	13	24	15	75
10th rank	2	2	13	7	3	27
Total	**400**	**402**	**415**	**416**	**413**	**2046**
Avg. rating	**5.145**	**5.0199**	**5.352**	**5.279**	**5.138**	**5.182**

2. Variables

본 연구에서는 많은 선행연구에서와 같이 신용평점을 신용평가의 대용 변수로 제시하여 사용했다.[6,7,8] 이는 신용평점이 위임이나 강행성에 근거한 평가가 아니기 때문에 기업의 압박과 유혹에서 보다 자유롭고 객관적인 평가를 담당할 수

있다는 장점이 있기 때문이다.[9]

변수 선정 과정에서는 선행연구를 바탕으로 수집 가능한 최대한의 독립변수들을 포함시킴으로써 변수 선정의 자의성을 배제하고자 하였다. 특히 선행연구들의 경우에는 종속변수로 선정된 신용등급 및 신용평점에 대하여 분석에 사용된 독립변수들 이 외에도 영향을 미칠 수 있는 변수들을 생략하여 생기는 문제에 대한 지적에 꾸준히 제기되어 왔다[10]. 이에 본 연구는 선행연구들의 결과를 참고하되, 재무비율 등 다양한 변수를 망라하여 이러한 선행연구의 문제점을 보완하고자 하였다. 그 결과 기업 분석 정보, 기초 재무 정보, 성장성에 관한 지표, 손익의 관계 비율, 자산 자본의 관계 비율, 자산 자본의 회전율, 생산성에 관한 지표의 영역에서 총 80개의 변수들을 일차적으로 선정하였다. 분석을 위하여 각각의 변수에는 X1, X2, X3,… X79, X80까지의 코드가 순차적으로 부여되었다. 분석에 사용된 변수의 목록은 아래 Table 4와 같다.

Table 4. Contents of Variables and Definitions

Code	Variable	Definition
KIS	KIS Credit Rating	Dependent Variable
<Price Valuation>		
X1	Big4	Big4 Accounting Firms (Yes/No)
X2	Yr. End % of Foreigner (Common+ Preferred)	The share % of foreigner in the fiscal year ending trading day. Data from KOSCOM
X3	Avg. End Market cap. (Common+ Preferred)	Average price of avg. common stock price and avg. preferred stock price during the fiscal year. Data from KOSCOM
X4	Yr. End Market cap. (Common+ Preferred)	Average of market cap of common stock and preferred stock during the fiscal year. Data from KOSCOM
X5	Avg. Price	Average of closing price during the fiscal year. Data from KOSCOM
X6	EBIT (bil)	operating earnings during the fiscal year including net interest expenses
X7	EBITDA (bil)	operating earnings during the fiscal year + Depreciation + Amortization
X8	Beta	covariance of the return of an asset and the return of the benchmark divided by the variance of the market portfolio
<Financial Statements>		
X9	Total Asset (bil)	Total Current Assets+ Housing Rent Assets+ Total-Non-Current Assets
X10	Total Liabilities (bil)	Total Current Liabilities+ Total-Non-Current Liabilities+ Total Deferred Liabilities
X11	Total Stockholder's Equity(bil)	Capital Stocks+ Capital Surplus+ Retained Earnings+ Capital Adjustment+ Other Comprehensive Income/Loss Accumulation
X12	Sales(Net) (bil)	Gross Sales-Sales Allowance & Return-Sales Incentives-Sales Discount-Estimated Sales for Returning Goods -Speicifc Purchase Cost - Sales Adjustment
X13	Cost of Sales (bil)	Cost of Merchandise Sold+ Cost of Finished Goods Sold+ Cost of Merchandise & Finished Goods Sold+ COS-other+ Loss on Valuation of Inventories-Estimated Cost of Sales for Returning Goods-Purchase Discount-COS adjustment
X14	Gross Profit (bil)	Sales(NET)-Cost of Sales
X15	Selling & General Admin. Expenses (bil)	Personal Expenses+ General Administrative Expenses+ Selling Expenses+ Other
X16	Cash Flows from Operating Activities (bil)	Net Income(Loss)+ Addition of Expenses Not Involving Cash Outflows-Deduction of Income Not Involving Cash Outflows+ Changes in Asset&Liabilities Resulting from Operating Activities
X17	Cash Flows from Investing Activities (bil)	Cash Inflows from Investing Act.-Cash Outflows from Investing Act.
X18	Cash Flows from Financing Activities (bil)	Cash Inflows from Financing Act.-Cash Outflows from Financing Act.
<Ratio of Growth>		
X19	Total Assets Growth	(Total Assets(n)/Total Assets(n-1))*100-100
X20	Tangible Assets Growth	(Tangible Assets(n)/Tangible Assets(n-1))*100-100
X21	Current Assets Growth	(Current Assets(n)/Current Assets (n-1))*100-100
X22	Inventories Growth	(Inventories(n)/Inventories(n-1))*100-100
X23	Shareholder's Equity Growth	(Shareholder's Equity(n)/Shareholder's Equity(n-1))*100-100
X24	Net Sales Growth	(Net Sales(n)/Net Sales(n-1))*100-100
X25	No. of Employee Growth	(No. of Employee(n)/No. of Employee(n-1))*100-100
<Ratios of Profitability>2000		
X26	Operating Income to Total Assets	(Operating Income/((Total Assets(n)+ Total Assets(n-1)/2))*100
X27	Income Before Income Tax Expense to Total Assets	(EBIT/((Total Assets(n)+ Total Assets(n-1))/2))*100
X28	Net Income to Total Assets	(Net Income/((Total Assets(n)+ Total Assets(n-1))/2))*100
X29	Net Income Before Financial Expenses to Avg.Total Assets	((Net Income+ Financial Expenses)/((Total Assets(n)+ Total Assets(n-1))/2))*100
X30	Operating Income to Operating Capital	Operating Income/((Total Assets(n)-Construction In-progress(n)-Total Investment Assets(n)-Total Other Non-current Assets(n)-Total Deferred Assets(n)-Organization Costs(n)-Development Cost(n)+ Total Assets(n-1)-Construction In-progress(n-1)-Total Investment Assets(n-1)-Total Other Non-current Assets(n-1)-Total Deferred Assets(n-1)-Organization Costs(n-1)-Development Cost(n-1))/2)*100
X31	Income Before Income Tax Expense to Equity	(Income Before Income Tax Expense/((Total Stockholder's Equity(n)+ Total Stockholder's Equity(n-1))/2))*100
X32	Net Income to Shareholder's Equity	(Net Income/((Total Shareholder's Equity(n)+ Total Shareholder's Equity(n-1))/2))*100
X33	Income Before Income Tax Expense to Capital Stock	(Income Before Income Tax Expense/(((Capital Stock(n)+ Capital Stock(n-1)/2))*100
X34	Net Income to Capital Stock	(Net Income/(((Capital Stock(n)+ Capital Stock(n-1)/2))*100
X35	Income Before Income Tax Expense to Net Sales	(Income Before Income Tax Expense/Net Sales)*100
X36	Net Income to Net Sales	(Net Income/Net Sales*100
X37	Gross Profit to Net Sales	(Gross Profit/Net Sales)*100
X38	Operating Income to Net Sales	(Operating Income/Net Sales)*100
X39	Total Expense to Total Revenue	(Cost of Sales+ Cost of Merchandise & Finished Goods Sold+ Cos-Other)/(Net Sales+ Non-Operating Income)*100
X40	COGS to Net Sales	(Cost of Sales/Net Sales)*100
X41	Depreciation Ratio	(Depreciation)/(Tangible Assets-Construction In-progress-Land+ (Depreciation))*100
X42	EBIT/Net Sales	(Income Before Income Tax Expense+ Financial Expenses)/Net Slaes*100
X43	EBITDA/Net Sales	(EBITDA/Net Sales)*100
<Leverage(or Safety) Ratio>		
X44	Equity to Total Assets	Total Shareholder's Equity/Total Assets*100
X45	Current Ratio	Current Assets/Non-Current Liabilities*100
X46	Quick Ratio	Current Assets/Current Liabilities*100
X47	Cash Ratio	Cash&Cash Equivalents/Current Liabilities*100
X48	Non Current Assets Ratio	(Non Current Assets-Deferred Assets+ Leased Assets+ Housing Rent

		Assets)/Total Shareholder's Equity*100
X49	Non Current Assets to Equity & LT Liabilities	(Non Current Assets-Deferred Assets)/(Total Shareholder's Equity+ Non Current Liabilities)*100
X50	Total Liabilities to Shareholder's Equity	Total Liabilities/Total Shareholder's Equity*100
X51	Current Liabilities to Shareholder's Equity	Current Liabilities/Total Shareholder's Equity*100
X52	Non-Current Liabilities to Shareholder's Equity	Non-Current Liabilities/Total Shareholder's Equity*100
X53	A/R to Trade Account Payable	(Account Receivables+ A/R Other Construction+ A/R Other_Housing Lotting_Out+ A/R Other-Operations+ A/R Other-Operations in Foreign Currency+ A/R Other-Rent)/(Trade Account Payable+ Account Payable Other_Construction+ Trade Account Payable Other Operations)*100
X54	Trade Account Payable to Inventories	(Trade Account Payable+ Account Payable Other_Construction+ Trade Account Payable Other Operations))/Inventories*100
X55	NWC to Total Assets	(Current Assets-Current Liabilities)/Total Assets*100
X56	Reserves Ratio	(Capital Surplus+ Retained Earnings+ Capital Adjustment+ Treasury Stock+ Other Comprehensive Income/Loss Accumulated Amount)/Total Shareholder's Equity*100
X57	R/E to Total Assets	(Capital Surplus+ Retained Earnings+ Capital Adjustment+ Treasury Stock+ Other Comprehensive Income/Loss Accumulated Amount)/Total Assets*100
X58	R/E to Paid-in Capital	(Capital Surplus+ Retained Earnings+ Capital Adjustment+ Treasury Stock+ Other Comprehensive Income/Loss Accumulated Amount)/Paid-in Capital*100
X59	Total C/F to Total Liabilities	Total C/F(Adjustment Net Income)/Total Liabilities*100
X60	Total C/F to Total Assets	Total C/F(Adjustment Net Income)/Total Assets*100
X61	Total C/F to Net Sales	Total C/F(Adjustment Net Income)/Net Sales*100
<Activity(or Efficiency or Asset Management Ratio>		
X62	Total Assets Turnover	Sales/((Total Assets(n)+ Total Assets(n-1))/2)
X63	Equity Turnover	Sales/((Total Shareholder's Equity(n)+ Total Shareholder's Equity(n-1))/2)
X64	Paid-in Capital Turnover	Sales/((Paid-in Capital(n)+ Paid-in Capital(n-1))/2)
X65	NWC Turnover	Sales/(((Current Assets(n)-Current Liabilities(n))+ (Current Assets(n-1)-Current Liabilities(n-1)))/2)
X66	Operating Capital Turnover	Sales/((Total Assets(n)-Total Investment Assets(n)-Total Other Non-Current Assets(n)-Total Deferred Assets(n)-Organization Costs(n)-Development Costs(n)-Construction in-progress(n))+ (Total Assets(n-1)-Total Investment Assets(n-1)-Total Other Non-Current Assets(n-1)-Total Deferred Assets(n-1)-Organization Costs(n-1)-Development Costs(n-1)-Construction in-progress(n-1)))/2)
X67	Non-Current Assets Turnover	Sales/((Non-Current Assets(n)-Deferred Assets(n)+ Non-Current Assets(n-1)-Deferred Assets(n-1))/2)
X68	Tangible Assets Turnover	Sales/((Tangible Assets(n)-Construction in-progress(n)+ 전Tangible Assets(n-1)-Construction in-progress(n-1))/2)
X69	Inventories Turnover	Sales/((Inventories(n)+ Inventories(n-1))/2)
X70	Merchandise & Finished Goods Turnover	Sales/((Merchandise(n)+ Finished Goods(n)+ Semi Finished Goods(n)+ Merchandise(n-1)+ Finished Goods(n-1)+ Semi Finished Goods(n-1))/2)
X71	Raw Materials Turnover	Sales/((Raw Materials(n)+ Raw Materials(n-1))/2)
X72	A/R Turnover	Sales/((Account Receivables(n)+ A/R other Constuction(n)+ A/R Other_Housing Lotting_Out(n)+ A/R Other-Operations(n)+ A/R Other-Operations in Foreign Currency(n)+ A/R Other-Rent(n)+ Account Receivables(n-1)+ A/R other Constuction(n-1)+ A/R Other_Housing Lotting_Out(n-1)+ A/R Other-Operations(n-1)+ A/R Other-Operations in Foreign Currency(n-1)+ A/R Other-Rent(n-1))/2)
X73	Trade Account Payable Turnover	Sales/((Trade Account Payable(n)+ Account Payable Other_Construction(n)+ Trade Account Payable Other Operations(n)+ Trade Account Payable(n-1)+ Account Payable Other_Construction(n-1)+ Trade Account Payable Other Operations(n-1))/2)
X74	Net Operating Capital Turnover	Sales/((Account Receivable(n)+ Inventories(n)-Trade Account Payable(n)+ Account Receivable(n-1)+ Inventories(n-1)-Trade Account Payable(n-1))/2)
<Productivity>		
X75	Net Sales per Employee	Sales/No.of Employees
X76	Income Before Income Tax Expense per Employee	Income Before Income Tax Expense /No. of Employees
X77	Net Income per Employee	Net Income/No.of Employees
X78	Avg.Tangible Assets, net of CIP per Employee	((Tangible Assets(n)-Construction in-progress(n)+ Tangible Assets(n-1)-Construction in-progress(n-1))/2)/No. of Employees
X79	Machinery & Equipment per Employee	((Machinery & Equipment (n)+ Machinery & Equipment (n-1))/2)/No. of Employees
X80	Total Assets per Employee	((Total Assets(n)+ Total Assets(n-1))/2)/No. of Employees

3. Methods

본 연구에서는 2009년부터 2013년까지의 총 2046개 기업의 재무 비율 등 80여가지의 독립변수들을 데이터셋으로 하여 유의미한 독립변수들을 추출한다. 따라서 순서형 로지스틱 분석을 통해 1차적으로 유의미한 변수들을 선정한다. 다음으로 특성선택 기법을 통해 그 가운데에서 변수들을 2차적으로 추출한다. 이 과정을 통해서 기업의 신용평점에 영향을 미치는 변수들을 확인한다.

다음으로 순서형 로지스틱 분석만으로 추출한 변수들로 구성된 데이터셋과 특성선택 기법을 적용하여 추출한 변수들로 구성된 데이터셋 각각이 순서형 로지스틱 분석에서 보이는 예측력의 차이를 T-test를 통해 비교한다. 개별 기업의 실제 신용평점 값과 예측 모형에서 제시된 신용평점 예측값의 차이에 대해서 각각 검증을 실시하고 이를 통하여 순서형 로지스틱 분석 및 특성선택 기법을 접목하여 추출한 변수들의 분류 예측력이 변수의 수가 적어도 여전히 유의미한 예측력을 보인다는 점을 확인하고자 한다.

4. Results

4.1 Model Comparison

우선, 순서형 로지스틱 분석을 통하여 신용평점에 영향을 미치는 독립변수들을 추출하였다. 그 결과, 앞서 제시된 총 80개의 독립변수 중에서 20개의 변수들이 통계적으로 유의미한 독립변수로 추출되었다. 순서형 로지스틱 분석의 결과 및 최

종 추출된 독립변수의 구성은 아래 Table 5와 같다.

Table 5. Variables Selected by Ordinal Logistic regression

#	Code	Item	Estimate	Std Error	Chi-Square	Prob>Chi-Sq
Intercept		Intercept[1]	-19.9744	0.686558	846.44	<.0001
		Intercept[2]	-14.3614	0.49779	832.35	<.0001
		Intercept[3]	-12.1859	0.467642	679.02	<.0001
		Intercept[4]	-9.83282	0.435796	509.09	<.0001
		Intercept[5]	-7.20883	0.406744	314.11	<.0001
		Intercept[6]	-4.68801	0.390707	143.97	<.0001
		Intercept[7]	-2.29988	0.389594	34.85	<.0001
		Intercept[8]	0.55667	0.413303	1.81	0.178
		Intercept[9]	3.926114	0.492165	63.64	<.0001
1	X2	Yr. End % of Foreigner (Common+Preferred)	0.016806	0.003882	18.75	<.0001
2	X4	Yr. End Market cap. (Common+Preferred)	-1.30E-05	5.80E-06	4.98	0.0256
3	X5	Avg. Price	0.003211	0.000501	41.07	<.0001
4	X8	Beta	-0.6719	0.117492	32.7	<.0001
5	X19	Total Assets Growth	0.009157	0.002657	11.88	0.0006
6	X21	Current Assets Growth	-0.01206	0.001774	46.22	<.0001
7	X27	Income Before Income Tax Expense to Total Assets	0.299704	0.02775	116.65	<.0001
8	X28	Net Income to Total Assets	0.705678	0.073241	92.83	<.0001
9	X29	Net Income Before Financial Expenses to Avg.Total Assets	-0.93348	0.069871	178.49	<.0001
10	X30	Operating Income to Operating Capital	0.030454	0.007339	17.22	<.0001
11	X33	Income Before Income Tax Expense to Capital Stock	-0.00345	0.001203	8.24	0.0041
12	X34	Net Income to Capital Stock	0.004315	0.0015	8.27	0.004
13	X44	Equity to Total Assets	0.116813	0.006044	373.55	<.0001
14	X47	Cash Ratio	0.011454	0.001503	58.04	<.0001
15	X57	R/E to Total Assets	0.011147	0.00384	8.43	0.0037
16	X60	Total C/F to Total Assets	0.164204	0.012949	160.81	<.0001
17	X62	Total Assets Turnover	1.059075	0.149739	50.02	<.0001
18	X64	Paid-in Capital Turnover	-0.00818	0.002332	12.3	0.0005
19	X67	Non-Current Assets Turnover	-0.20898	0.030952	45.58	<.0001
20	X76	Income Before Income Tax Expense per Employee	-0.00261	0.000653	15.97	<.0001

위 표 5에서 해석상 유의할 점이 있다. 본 연구에 사용된 통계소프트웨어 JMP는 SAS사에서 만들어졌으며 순서형 로지스틱 회귀분석에 있어 SPSS와 알고리즘의 차이가 있다. 예를 들어, 종속변수가 Y가 순서형 로지스틱 변수로 (1, 2, …, g, …, k-1, k)이고 독립변수가 X1,…,Xj,…,Xp라고 할 때, 순서형 로지스틱 회귀분석에 사용되는 모형은 다음과 같다.

$$\text{logit}(p(Y\leqslant g))=\ln\frac{p(Y\leqslant g)}{p(Y>g)}=\beta_{0_g}+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p \quad (g=1,\cdots,k-1)$$

SPSS의 경우에는 다음과 같은 식을 사용하여 계산한다.

$$\text{logit}(p(Y\leqslant g))=\ln\frac{p(Y\leqslant g)}{p(Y>g)}=\beta_{0_g}-(\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p) \quad (g=1,\cdots,k-1)$$

따라서 양(+)의 βj값이 의미하는 바는 변수 Xj의 값이 증가하는 것이 더 낮은 범주의 Y값에 대한 오즈 값이 증가하는 것으로 해석할 수 있다.[11,12] 즉, 위의 표 5에서 X2 기말외국인보유율(보통주+우선주)의 Estimate 값이 0.016806로 양(+)의 값이므로 기말외국인보유율(보통주+우선주)의 값이 증가할 때 범주 값인 신용평점은 낮아지는 것으로 해석해야하며 이는 앞서 선행연구들에서 외국인 주식 보유율을 독립변수 및 통제변수로 사용한 것과 일치하는 결과임을 알 수 있다.[13]

한편, WEKA 소프트웨어를 이용하여, 상관관계 기반에 의한 특성 선택 기법을 적용한 결과 선택된 변수들은 아래 Table 6에 제시된 총 6개 항목이다.

Table 6. Variables Selected by Ordinal Logistic regression and Best-first Feature Selection

Code	Item	Code	Item
X27	Income Before Income Tax Expense to Total Assets	X47	Cash Ratio
X33	Income Before Income Tax Expense to Capital Stock	X57	R/E to Total Assets
X44	Equity to Total Assets	X60	Total C/F to Total Assets

Table 7. Statistics of Variables Selected by Ordinal Logistic regression and Best-first Feature Selection

#	Code	Item	Min	Max	Mean	Std Dev
7	X27	Income Before Income Tax Expense to Total Assets	-66.88	48.07	3.836	9.466
11	X33	Income Before Income Tax Expense to Capital Stock	-1163.76	7520.88	136.034	392.668
13	X44	Equity to Total Assets	2.33	95.51	56.269	18.339
14	X47	Cash Ratio	0	556.35	24.858	40.895
15	X57	R/E to Total Assets	-122.17	109.97	47.122	23.837
16	X60	Total C/F to Total Assets	-24.3	47.42	6.529	6.054

이처럼 순서형 로지스틱 회귀분석과 상관관계 기반에 의한 특성선택을 통해 추출된 설명변수들을 바탕으로 기계학습을 위한 데이터셋을 구성하였다.

위 Table 5와 Table 6, Table7에서 알 수 있듯이 순서형 로지스틱 회귀분석을 통해 식별된 설명변수들 가운데 상관관계 기반에 의한 특성선택을 통해 변수를 다시 선정한 결과 그 수가 20개에서 6개로 줄어 대폭 감소함을 알 수 있다. 따라서 연구가설 1은 성립한다.

4.2 Comparing the Forecasting Performance of two Models

한편 이렇게 선정된 변수들이 그 실질적 예측 능력에 있어서 차이를 보이는가를 검증하기 위하여 최적 우선탐색을 통해 선정된 변수들로 데이터셋을 구성하고 순서형 로지스틱 분석을 통해 그 예측능력을 확인하였다. 그 결과는 다음 Table 8와 Table 9에 오차행렬로 제시하였다.

Table 8. Confusion Matrix from OLS Model

Esitmated / Actual	1	2	3	4	5	6	7	8	9	10
1	3	6	1	0	0	0	0	0	0	0
2	2	122	60	11	1	1	0	0	0	0
3	0	52	90	83	7	3	0	0	0	0
4	0	12	56	157	93	8	0	0	0	0
5	0	0	5	68	238	80	3	0	0	0
6	0	0	0	6	82	201	57	7	0	0
7	0	0	0	0	7	88	128	38	2	0
8	0	0	0	0	0	10	64	75	16	1
9	0	0	0	0	0	0	0	37	31	7
10	0	0	0	0	0	0	0	2	12	13

Table 9. Confusion Matrix from Best-first Feature Selection Model

Esitmated / Actual	1	2	3	4	5	6	7	8	9	10
1	0	9	1	0	0	0	0	0	0	0
2	3	116	54	18	5	1	0	0	0	0
3	2	46	79	95	10	3	0	0	0	0
4	0	20	63	134	95	14	0	0	0	0
5	0	1	9	69	229	80	6	0	0	0
6	0	0	1	7	80	205	56	4	0	0
7	0	1	0	0	10	93	116	39	4	0
8	0	0	0	0	0	17	71	64	11	3
9	0	0	0	0	0	0	6	33	30	6
10	0	0	0	0	0	0	0	4	13	10

위 오차행렬을 바탕으로 두 모형간 예측능력의 차이를 알아보기 위하여 개별 기업의 실제 신용평점 값과 두 모형에서 각각 예측한 값들의 오차에 대한 T-test를 실시하였다. 그 결과는 다음 Table 10과 같다.

Table 10. T-test Result

Contents	t	Sig.
Estimated values from OLS Model vs Estimated values from Best-first Feature Selection Model	-0.02703	0.5108

Table 10에 따르면 두 방법으로 추출한 변수들로 구성된 데이터셋에 대하여 순서형 로지스틱 회귀분석 모형의 예측값에는 차이가 없다고 할 수 있다. 즉, 변수선정에 있어서 회귀분석뿐만 아니라 상관관계 기반에 의한 특성선택을 적용할 때 보다 핵심적인 변수를 추출할 수 있다는 것을 의미한다.

결과적으로, 순서형 로지스틱 회귀분석만을 통해 변수를 선정하였을 경우보다 상관관계 기반에 의한 특성선택을 적용하여 변수를 선정하였을 경우에 선정변수가 20개에서 6개로 절반 이하로 줄어들었음에도 해당 변수들로 구성된 데이터셋은 그 예측력에 있어서 유의미한 차이를 보이지 않았다. 이를 통해 상관관계 기반에 의한 특성선택을 접목하였을 때 실제 분류에 필요한 정보의 수를 감소시키면서도 모형의 예측력을 유지할 수 있다는 점을 알 수 있기에 연구가설 2는 성립한다.

IV. Concluding Remarks

본 연구는 코스피 시장에 상장된 제조업 기업들에게 부여된 신용평점을 바탕으로 다양한 재무정보 가운데 신용등급에 영향을 미치는 변수들을 순서형 로지스틱 회귀분석 및 상관관계 기반에 의한 특성선택 기법을 통해 식별하고 실제 분류에서의 예측값들에 대한 비교 검정을 통해 그 타당성을 확인하였다.

신용등급은 기업의 자금조달을 매개하는 중요한 의사결정 지원 도구일 뿐만 아니라 전문경영인의 경영 성과를 측정하는 보조 지표가 되고 시장에서 기업의 상태를 확인할 수 있는 유용한 정보를 제공한다. 따라서 신뢰성 있고 가치 있는 신용등급의 측정과 부여는 건전한 자본주의 육성과 금융시장 발전을 위한 경영과학의 중요한 연구 분야이며 사회적으로도 관심과 중요성이 증가하고 있다.

특히 데이터 마이닝 기술의 발달과 더불어 유의미한 변수의 식별과 이를 통한 다수 기업들에 대한 신속하고 정확한 신용평가가 이루어질 수 있도록 해야 한다는 필요성이 제기되었다. 이에 본 연구는 특성선택 기법을 활용한 기업 신용등급 관련 설명변수 식별 및 등급 분류 연구를 수행하였다.

그 결과, 순서형 로지스틱 회귀분석과 더불어 특성선택 기법을 활용함으로써 보다 적은 수의 설명변수의 추출이 가능하다는 점을 확인하였다. 또한 T-test 검정 분석 결과로부터 특성선택 기법을 활용하였을 경우에 추출된 변수의 개수가 20개에서 6개로 절반이하로 줄어들었음에도 그 예측력에 있어서 유의미한 차이가 없음을 확인하였다.

즉, 신용평가사의 신용등급 부여 행태에 대한 관리 감독의 필요성이 증대되고 기업들의 신용등급 정보에 대한 신뢰성 검증이 중요해지는 시점에 대규모 표본에 대하여 필요한 기업 정보들을 추출하고 신용등급에 대한 분류 모형의 정확도를 높이는 데 예측하는데 활용될 수 있다는 점에서 특성선택 기법은 큰 의의를 지닌다. 재무정보에 기반을 둔 데이터마이닝 기법만으로 신용등급 분류를 완전하게 수행해내지는 못하더라도 대규모 표본에 대하여 자료들을 선제적으로 분석하여 분류 효율성을 증진시킬 수 있을 것이기 때문이다.

향후에는 다음과 같은 한계점을 보완함과 동시에 최신 연구성과들을 반영하여 연구방향을 설정할 수 있을 것이다. 첫째, 본 연구에서는 코스피 상장 제조업 기업들만을 대상으로 하여 신용등급에 영향을 미치는 변수들을 선별하고 이를 모형화 하였다. 산업 특성이 신용 평점 및 재무 평점 그리고 재무 정보 전반에 영향을 미친다는 최신 연구들을 감안하여 업종별 가중치 등을 반영하여 모형을 확장할 수 있을 것이다.[14, 15] 둘째로 본 연구에서는 선행연구에서 주로 다루어진 재무정보 및 경영 상태에 관한 변수들 위주로 설명변수들을 구성하였다. 하지만 최근에는 기업의 사회적 책임이나 내부회계인력의 특성 등 보다 다양한 변수들과 신용등급의 상관관계에 대한 연구가 이루어지고 있기 때문에 이러한 변수들을 추가적으로 포

함하여 실증분석 할 필요가 있다. 본 연구에서 제시된 데이터 마이닝 기법을 활용할 경우 다양한 이해관계자들의 방대한 양의 비재무 정보 가운데 유의미한 변수를 빠르게 추출하고 신용 평점과의 관계를 알아내는 데 유용하게 사용될 수 있다는 점에서 향후 연구에도 도움이 될 것이다.

마지막으로 본 연구는 자료수집의 제약 및 표본의 일관성 확보를 위하여 특정 신용평가사의 KIS-Value라는 프로그램과 신용평점만을 사용하여 연구를 진행하였기에 전체 신용등급에 대하여 일반화하기에는 무리가 있다. 이러한 문제점은 여타 선행연구에서도 지적된다는 점에서 유관기관의 협조 및 최신 자료의 확보를 통해 결측치를 제거하고 추가적인 표본을 추출하여 검증한다면 보다 유의미한 결과를 얻을 수 있을 것이다.

최근 연구 성과들에 따라 국내 신용등급 고평가 현상에 대한 우려가 높아지고 있다[16]. 신뢰성 있는 신용평가를 바탕으로 하는 정보의 비대칭성의 해소를 위한 대안의 모색이 시급한 시점에 가운데 본 연구가 제안하는 데이터 마이닝 기법은 신용 평가 기법들을 재검증하고 발전시키는 데 기여할 수 있을 것이다. 나아가 본 연구에서 사용된 데이터마이닝 기법은 대규모 재무정보의 분석을 바탕으로 하고 있어 회계감사 및 부도예측 등을 비롯한 재무회계의 다양한 영역에도 접목되어 보다 실용적으로 활용될 수 있을 것이다[17].

REFERENCES

[1] Min-Seo Kim, Joon-Whan Oh, "Discretionary Accruals` Relation to Credit Score and Financial Score, Korean Accounting Journal, No.22, pp.105-131, 2013.

[2] Jong-Il Park, Eun-Sun Ki, Soo-Young Kwon, "A Review and Some New Evidence on the Effect of Book-Tax Differences on Bond Rating", Korean Accounting Review, No.39, pp.1-55, 2014.

[3] Hall, M. A, "Correlation-based feature selection for machine learning", Doctoral dissertation, The University of Waikato, 1999.

[4] Jae-hak Yu, Han-sung Lee, Young-hee Im, Myung-Sup Kim, Dai-hee Park. "Hierarchical Internet Application Traffic Classification using a Multi-class SVM", Journal of Korean Institute of Intelligent Systems, No.20, pp.7-14, 2010.

[5] Witten, I. H., and Frank, E., Data Mining: ractical machine learning toolsand techniques, Morgan Kaufmann, 2005.

[6] Dong-Young Kim, "A Study on Effects of Corporate Social Responsibility and Credit Financial Score", The Journal of Business Education, No.28, pp.123-142, 2014.

[7] Bum-Jin Park, "The Effect of Types of Venture Capitalist and Earnings Management on Credit Rating", The Journal of Small Business Innovation, No.36, pp.179-204, 2014.

[8] Bum-Jin Park, "The Effect of Types of Venture Capitalist and Earnings Management on Credit Rating", The Journal of Small Business Innovation, No.36, pp.179-204, 2014.

[9] Moon-Tae Kim, “The Effects of Entertainment Costs on Credit Evaluation in Medical Firms”, Accounting Information Review, No.33, pp.1-22, 2015.

[10] Jong-Il Park, Seong-Ho Bae, Seok-Woo Jeong, "The association between the employment of industry specialist auditors and bond rating of a client company", Korean Accounting Journal, No.22, pp.31-69, 2013

[11] Peterson, B., and Harrell Jr, F. E., "Partial proportional odds models for ordinal response variables", Applied Statistics, pp.205-217, 1990.

[12] Liu, X., "Ordinal regression analysis: Fitting the proportional odds model using Stata, SAS and SPSS", Journal of Modern Applied Statistical Methods, No.8, pp.632-645, 2009.

[13] Moon-Tae Kim, Young-Hwan Kim, "The Impacts of Foreign Ownership and Outside Directors on Bond Grading", Korean Accounting Review, No.32, pp.29-58, 2007.

[14] Sung-Yoon Ahn, “The Effect of Industrial Characteristics on the Credit Score and the Financial Score”, Korean Journal of Accounting Research, No.20, pp.55-79, 2015.

[15] Sung-Ju Choi, Sang-Won Lee, “ A Development of Hotel Bankruptcy Prediction Model on Artificial Neural Network”, Journal of the Korea Society of Computer and Information, No.19, pp.125-133, 2014.

[16] Seok-Woo Jeong, Hyun-Ah Kim, “The Implication of the Difference between Local and Foreign Credit Rating and the Effect of Analysts”, Annual Conference of Korean Accounting Association, No.1, pp.1521-1551, 2015.

[17] Rack-In Choi “A Study on An Improvement Scheme of the External Auditing System by

Enforcing K-IFRS ", Journal of the Korea Society of Computer and Information, No.19, pp.339-348, 2014.

[18] NICE homepage (http://www.niceinfo.co.kr/creditrating/bi_score_1_1.nice)

Authors

Youn Chang Nam is a college student at Department of Global Business Administration, Sungkyunkwan University, Korea. He is interested in data-mining and artificial intelligence.

Kun Chang Lee is a full professor of MIS and Health Informatics & Mining in SKK Business School and SAIHST(Samsung Advanced Institute for Health Sciences & Technology), Sungkyunkwan University. He is also in charge of Creativity Science Research Institute (CSRI) and Health Mining Research Center in SKK Business School.