**거시환경 변화에 따른 기업 부도확률 예측**

**데이터분석언어 2조**

**(조장)김상곤 2024712716**

**(팀원)김진영 2024712875**

**(팀원)김유진 0000000000**

**(팀원)이설화 2024712723**

**목차**

1. 배경 및 목표2
2. 가설2
3. 데이터 수집3
4. 데이터 정의4
5. 데이터 가공10
6. 데이터 분석12
7. 결론24

**1. 배경 및 목표**

2000년대 후반을 기점으로 한국경제는 저성장 국면에 진입하였고 단순 값싼 노동으로 경쟁력을 유지하던 시대 대신, 고부가 가치를 추구하며 국가 경쟁력을 확보하려는 노력에 있다. 하지만 내수시장 보단 수출기반 경제를 이룩하면서 해외 경제에 대한 의존성, 북한과 휴전상태에 있는 국가 리스크 프리미엄, 70~80년대 주 먹거리였던 저(抵) 부가가치 석유화학과 중공업을 계속 유지하면서 한국 경제는 큰 벽에 부딪히고 있다.

특히 현 상황은 한국 경제의 민감도를 더욱 확대시키고 있는데 자국 우선주의를 강조하는 트럼프노믹스의 재등장, 반도체 경쟁력 상실에 따른 무역수지 악화, 인건비/원자재 비용 인상에 따른 인플레이션과 글로벌 공급체인망 분열 등 악영향을 끼치는 요인들이 우리사회에 크게 작용하고 있다. 이에 따라 개인의 삶과 직결된 한국경제에 대한 재평가가 필요하며 이를 위해 그 초석이 될 수 있는 기업 부도율을 예측하여 내년의 한국경제를 들여다보고자 한다. 결국 대기업 수출주도로 성장한 한국 경제는 기업의 건전성이 곧 국가 경쟁력의 시발점이고 낙수효과로 그 현금흐름이 개인에게까지 이어지는 구조를 가지고 있기 때문이다.

**2. 가설**

우리 팀은 한국의 경제상황이 다른 선진국에 비해 특수성을 띄고 있다고 보았다. 예를 들어 상대적으로 낮은 인구수는 내수시장 보단 은행 신용장을 통한 수출 의존도를 높였고 주택 및 상업용 부동산 개발을 통해 내수시장을 유지하고 있기에 기업 부도율을 예측할 때 금융비용 및 차입금 의존도가 중요 변수가 될 수 있다고 생각했다(양의 상관관계). 반대로 10여년간 경쟁성장률 대비 침체되고 있는 코스피 지수(지속적으로 2,300~2,700수준 유지) 증감률은 상대적으로 설명력이 낮다고 판단되었지만 오히려 그 안에 투자하고 있는 특정 플레이어에 주목하였다. 한국 주식시장은 국내 개인보다는 외국인 및 기관 투자자가 방향성을 결정하고 있기에 해당 카운터파티의 순 매수량 및 매수금액에 집중해 보기로 하였다. 외국인은 보통 헤지펀드로, 한국 기관은 보통 국민연금 등이 큰 손으로 작용하고 있기에 투자되는 기업의 실제 내부상황을 개인 투자자보다 더 잘 알고 있었고 기업가치 평가를 통해 투기가 아닌 실질투자가 이뤄지는 구조라서 기업 부도율 파악에 용이한 변수로 작용할 수 있다고 추론되었다(양의 상관관계). 이런 기업 자체의 미시적 변수 외에도 경제 전체에 영향을 미치는 환율, 장단기 금리 또한 주요 변수로 작용할 것으로 생각되었다. 예를 들어 환율의 증가는 곧 기업수출가격 증대에 따른 부도율 감소(음의 상관관계), 금리 상승은 자금조달 비용 증가 및 환율 하락(가격 경쟁력 감소)에 따른 건정성 악화(양의 상관관계), 장단기 금리 역전은 현재의 경제상황이 미래보다 나쁘다는 것을 암시하므로 음의 상관관계가 있다고 가설을 세울 수 있었다. 위와 같이 각 변수들로 상관관계 추론은 가능하지만, 단순 이론적 추상만으로 타 선진국과 다른 경제적 특수성을 가진 한국에서는 어떤 변수가 기업의 부도율과 가장 밀접한 연관성을 띄는지 파악할 수 없으므로 한국은행 경제통계시스템의 변수를 최대한 활용하여 가장 높은 연관성을 지닌 지표 4개를 찾아 변수모형을 만들기로 하였다.

**3. 데이터 수집**

부도율 예측에 영향을 주는 독립변수 x는 한국은행 경제통계시스템(ECOS)의 거시경제지표변수 중 2. 가설에 따라 우리 팀이 관련성이 높다고 생각하는 변수들을 선택하여 수집하였고, 부도율의 종속변수 y는 NICE 신용평가에서 제공하는 연간 부도율을 활용하였다.

데이터 수집의 시점은 2007년 4월 서브프라임 모기지 사태로 시작된 세계 규모의 경제위기를 고려하여, 2008년부터 2023년까지의 데이터를 수집하였다. 또한, 각 독립변수 x는 최대 1년까지 영향을 미친다고 가정하였고 시나리오의 입력 편의를 고려하여 연단위로 데이터를 수집하였다.

한편, 부도율 예측을 위한 독립변수 x를 설정함에 있어, 기존 한국은행 경제통계시스템(ECOS)의 데이터를 활용하여, 심리지수지표 데이터의 5개의 항목(BSI 종합지표, 경제인식조사, 경제전망조사, 소비지출전망조사, 가계저축 및 부채조사)은 각 지표 항목별로의 가중치를 두어 변수를 생성하였다.

**4. 데이터 정의**

**4.1 독립변수 x**

- 총 165개 (raw data 내 var\_1의 구분 변수였던 연도변수 제외)

- 데이터 수집 범위 : 연간 (2008 ~ 2023)

- 데이터 수집 기관 : 한국은행 ECOS

|  |  |  |
| --- | --- | --- |
| **변수명** | **항목명** | **비고** |
| var\_1 | (환율지표) 원/달러 |  |
| var\_2 | (환율지표) 원/100엔 |  |
| var\_3 | (통화지표) Lf평잔 |  |
| var\_4 | (통화지표) Lf평잔증가율 (%) |  |
| var\_5 | (통화지표) M2평잔 |  |
| var\_6 | (통화지표) M2평잔증가율(%) |  |
| var\_7 | (통화지표) M1평잔 |  |
| var\_8 | (통화지표) M1평잔증가율(%) |  |
| var\_9 | (통화지표) 본원통화평잔 |  |
| var\_10 | (통화지표) 본원통화평잔증가율(%) |  |
| var\_11 | (금리지표) 국고채3년(평균) |  |
| var\_12 | (금리지표) 국고채5년(평균) |  |
| var\_13 | (금리지표) 국고채10년(평균) |  |
| var\_14 | (금리지표) 회사채3년(AA-, 평균) |  |
| var\_15 | (금리지표) CD 91물(평균) |  |
| var\_16 | (금리지표) 콜금리(1일물,평균) |  |
| var\_17 | (금리지표) 기준금리 |  |
| var\_18 | (GDP) 국내총생산(명목GDP) |  |
| var\_19 | (GDP) 경제성장률(실질GDP성장률) |  |
| var\_20 | (기업자금조달 은행대출 지표) 기업대출 |  |
| var\_21 | (기업자금조달 은행대출 지표) 대기업 |  |
| var\_22 | (기업자금조달 은행대출 지표) 중소기업 |  |
| var\_23 | (기업자금조달 직접금융 지표) 직접금융 |  |
| var\_24 | (기업자금조달 직접금융 지표) 대기업 |  |
| var\_25 | (기업자금조달 직접금융 지표) 중소기업 |  |
| var\_26 | (기업자금조달 직접금융 지표) 금융채 |  |
| var\_27 | (기업자금조달 직접금융 지표) ABS |  |
| var\_28 | (수출입물가지수 지표) 수출물가 |  |
| var\_29 | (수출입물가지수 지표) 농림수산품 |  |
| var\_30 | (수출입물가지수 지표) 공업제품 |  |
| var\_31 | (수출입물가지수 지표) 수입물가 |  |
| var\_32 | (수출입물가지수 지표) 원자재 |  |
| var\_33 | (수출입물가지수 지표) 중간재 |  |
| var\_34 | (수출입물가지수 지표) 최종재 |  |
| var\_35 | (국민계정 지표)  1인당 국내총생산(명목, 원화표시) |  |
| var\_36 | (국민계정 지표)  1인당 국민총소득(명목, 원화표시) |  |
| var\_37 | (물가지표)  생산자물가지수(총지수 / 기본분류) |  |
| var\_38 | (물가지표) 농림수산품 |  |
| var\_39 | (물가지표) 광산품 |  |
| var\_40 | (물가지표) 공산품 |  |
| var\_41 | (물가지표) 전력,가스,수도및폐기물 |  |
| var\_42 | (물가지표) 서비스 |  |
| var\_43 | (물가지표) 국내공급물가지수(총지수) |  |
| var\_44 | (물가지표) 원재료 |  |
| var\_45 | (물가지표) 중간재 |  |
| var\_46 | (물가지표) 최종재 |  |
| var\_47 | (물가지표) 생산재(원재료+중간재) |  |
| var\_48 | (물가지표) 소비자물가지수(총지수) |  |
| var\_49 | (물가지표) 수출물가지수(총지수 / 기본분류 / 원화기준) |  |
| var\_50 | (물가지표) 농림수산품 |  |
| var\_51 | (물가지표) 공산품 |  |
| var\_52 | (물가지표) 수입물가지수(총지수 / 기본분류 / 원화기준) |  |
| var\_53 | (물가지표) 농림수산품 |  |
| var\_54 | (물가지표) 광산품 |  |
| var\_55 | (물가지표) 공산품 |  |
| var\_56 | (통화, 유동성 지표) 본원통화(평잔,계절조정계열) |  |
| var\_57 | (통화, 유동성 지표) 현금통화 |  |
| var\_58 | (통화, 유동성 지표) M2(평잔, 계절조정계열) |  |
| var\_59 | (통화, 유동성 지표) 중앙은행 |  |
| var\_60 | (통화, 유동성 지표) 예금은행 |  |
| var\_61 | (통화, 유동성 지표)  2년이상 장기금융상품 등 1) |  |
| var\_62 | (통화, 유동성 지표)  생명보험계약 준비금 등 2) |  |
| var\_63 | (통화, 유동성 지표) 중앙은행 |  |
| var\_64 | (통화, 유동성 지표) 예금은행 |  |
| var\_65 | (통화, 유동성 지표) 종합금융회사 1) |  |
| var\_66 | (통화, 유동성 지표) 생명보험회사 2) |  |
| var\_67 | (통화, 유동성 지표) Lf(금융기관 유동성) |  |
| var\_68 | (통화, 유동성 지표) 국채, 지방채 |  |
| var\_69 | (통화, 유동성 지표) 회사채, 기업어음 2) |  |
| var\_70 | (통화, 유동성 지표) 수신합계 |  |
| var\_71 | (통화, 유동성 지표) 요구불예금 |  |
| var\_72 | (통화, 유동성 지표) 가계 |  |
| var\_73 | (통화, 유동성 지표) 기업 |  |
| var\_74 | (통화, 유동성 지표)  비은행금융기관(평잔) 1) |  |
| var\_75 | (통화, 유동성 지표) 금융자금 |  |
| var\_76 | (통화, 유동성 지표) 분할상환 |  |
| var\_77 | (통화, 유동성 지표) 주택자금 |  |
| var\_78 | (통화, 유동성 지표) 중소기업자금 |  |
| var\_79 | (통화, 유동성 지표) 축산자금 |  |
| var\_80 | (통화, 유동성 지표) 종합금융회사 |  |
| var\_81 | (통화, 유동성 지표) 상호저축은행 |  |
| var\_82 | (통화, 유동성 지표) 상호금융 |  |
| var\_83 | (통화, 유동성 지표) 가계대출 |  |
| var\_84 | (통화, 유동성 지표) 예금은행 |  |
| var\_85 | (통화, 유동성 지표) 비은행예금취급기관 |  |
| var\_86 | (통화, 유동성 지표) 보험회사 |  |
| var\_87 | (통화, 유동성 지표)  주택담보대출-예금취급기관 |  |
| var\_88 | (통화, 유동성 지표) 주택담보대출-예금은행 |  |
| var\_89 | (통화, 유동성 지표) 예금은행 |  |
| var\_90 | (통화, 유동성 지표) 비은행예금취급기관 |  |
| var\_91 | (통화, 유동성 지표) 주택담보대출-예금은행 |  |
| var\_92 | (통화, 유동성 지표)  주택담보대출-비은행예금취급기관 |  |
| var\_93 | (통화, 유동성 지표) 식료품 및 음료 |  |
| var\_94 | (통화, 유동성 지표)  화학제품 및 의료용 제품 |  |
| var\_95 | (통화, 유동성 지표) 금속가공제품 |  |
| var\_96 | (통화, 유동성 지표)  전자부품, 컴퓨터, 영상, 음향 및 통신장비 |  |
| var\_97 | (통화, 유동성 지표) 자동차 및 트레일러 |  |
| var\_98 | (통화, 유동성 지표) 건설업 |  |
| var\_99 | (통화, 유동성 지표) 서비스업 |  |
| var\_100 | (통화, 유동성 지표) 요구불예금 |  |
| var\_101 | (통화, 유동성 지표) 저축성예금 |  |
| var\_102 | (통화, 유동성 지표) 기업대출\_은행전체 1) |  |
| var\_103 | (통화, 유동성 지표) 가계대출\_은행전체 1) |  |
| var\_104 | (통화, 유동성 지표)  신용카드대출 2)\_은행전체 1) |  |
| var\_105 | (금리 지표) 한국은행 기준금리 |  |
| var\_106 | (금리 지표) KORIBOR(3개월) |  |
| var\_107 | (금리 지표) 국고채(3년) |  |
| var\_108 | (금리 지표) 국고채(10년) |  |
| var\_109 | (금리 지표) 통안증권(1년) |  |
| var\_110 | (금리 지표) 회사채(3년, AA-) |  |
| var\_111 | (금리 지표) 정기예금 |  |
| var\_112 | (금리 지표) 정기적금 |  |
| var\_113 | (금리 지표) 표지어음(91-120일) |  |
| var\_114 | (금리 지표) 금융채 3) |  |
| var\_115 | (금리 지표) 대출평균 1) |  |
| var\_116 | (금리 지표) 기업대출 |  |
| var\_117 | (금리 지표) 대기업대출 |  |
| var\_118 | (금리 지표) 중소기업대출 |  |
| var\_119 | (금리 지표) 가계대출 |  |
| var\_120 | (금리 지표) 주택담보대출 |  |
| var\_121 | (금리 지표) 예·적금담보대출 |  |
| var\_122 | (주가 지표) KOSPI\_종목수 |  |
| var\_123 | (주가 지표) KOSPI\_시가총액 |  |
| var\_124 | (주가 지표) KOSPI\_거래량 |  |
| var\_125 | (주가 지표) KOSPI\_거래대금 |  |
| var\_126 | (주가 지표) KOSPI\_거래량 일평균 |  |
| var\_127 | (주가 지표) KOSPI\_거래대금 일평균 |  |
| var\_128 | (주가 지표) KOSPI\_종가 |  |
| var\_129 | (주가 지표) KOSDAQ\_종목수 |  |
| var\_130 | (주가 지표) KOSDAQ\_시가총액 |  |
| var\_131 | (주가 지표) KOSDAQ\_거래량 |  |
| var\_132 | (주가 지표) KOSDAQ\_거래대금 |  |
| var\_133 | (주가 지표) KOSDAQ\_거래량 일평균 |  |
| var\_134 | (주가 지표) KOSDAQ\_거래대금 일평균 |  |
| var\_135 | (주가 지표) KOSDAQ\_종가 |  |
| var\_136 | (주가 지표) 기관투자자(순매수) |  |
| var\_137 | (주가 지표) 개인(순매수) |  |
| var\_138 | (주가 지표) 외국인(순매수) 3) |  |
| var\_139 | (주가 지표) 투자자 예탁금 |  |
| var\_140 | (주가 지표) RP |  |
| var\_141 | (주가 지표) 신용융자 잔고 2) |  |
| var\_142 | (주가 지표) 계약금액 1) |  |
| var\_143 | (주가 지표) 거래량(총 옵션거래) |  |
| var\_144 | (주가 지표) 거래대금(총 옵션거래) |  |
| var\_145 | (주가 지표) 거래량(CALL 옵션) |  |
| var\_146 | (주가 지표) 거래대금(CALL 옵션) |  |
| var\_147 | (주가 지표) 거래량(PUT 옵션) |  |
| var\_148 | (주가 지표) 거래대금(PUT 옵션) |  |
| var\_149 | (심리지수 지표) BSI(종합지표) | 업황실적, 매출실적, 자금실적, 인력사정실적을 각각 0.25, 0.3, 0.35, 0.1의 가중치를 두어 종합지표 변수로 생성 |
| var\_150 | (심리지수 지표) 경제인식조사 | 현재생활형편CSI, 현재경기판단CSI 각각 0.3D의 가중치로 두어 종합지표 변수로 생성 |
| var\_151 | (심리지수 지표) 경제전망조사 | 생활형편전망CSI, 향후경기전망CSI, 취업기회전망CSI, 금리수준전망CSI각각 0.4의 가중치를 두어 종합지표 변수로 생성 |
| var\_152 | (심리지수 지표) 소비지출전망조사 | 가계수입전망CSI, 소비지출전망CSI 각각 0.15의 가중치를 두어 종합지표 변수로 생성 |
| var\_153 | (심리지수 지표) 가계저축 및 부채조사 | 현재가계저축CSI, 가계저축전망CSI, 현재가계부채CSI 각각 0.4의 가중치를 두어 종합지표 변수로 생성 |
| var\_154 | (기업경영분석지표) 부채비율 | 총부채/자기자본 |
| var\_155 | (기업경영분석지표) 자기자본비율 | 자기자본/총자산 |
| var\_156 | (기업경영분석지표) 차입금의존도 | 총차입금/총자산 |
| var\_157 | (기업경영분석지표) 유동비율 | 유동자산/유동부채 |
| var\_158 | (기업경영분석지표) 비유동비율 | 비유동자산/유동부채 |
| var\_159 | (기업경영분석지표) 매출액영업이익률 | 영업이익/매출액 |
| var\_160 | (기업경영분석지표) 매출액세전순이익률 | 세전순이익/매출액 |
| var\_161 | (기업경영분석지표) 이자보상비율 | 영업이익/이자비용 |
| var\_162 | (기업경영분석지표) 금융비용대매출액 | 금융비용/매출액 |
| var\_163 | (기업경영분석지표) 차입금평균이자율 |  |
| var\_164 | (기업경영분석지표) 인건비대매출액 | 인건비/매출액 |
| var\_165 | (기업경영분석지표) 인건비대영업총비용 | 인건비/영업총비율 |

**4.2 종속변수 y**

- 총 1개

- 데이터 수집 범위 : 연간 (2008 ~ 2023)

- 데이터 수집 기관 : NICE 신용평가 연간부도율 지표 자료

|  |  |  |
| --- | --- | --- |
| **변수명** | **항목명** | **비고** |
| var\_166 | 부도율 |  |

**5. 데이터 가공**

**5.1 개발 환경 설명**

본 프로젝트는 이번 학기의 데이터분석언어 수업의 일환인만큼, google colab 환경에서의 python 언어를 활용하여 데이터 전처리, 변수분석, 데이터분석을 진행하였다.

사용한 라이브러리로는, 데이터 전처리 및 분석에 numpy와 pandas를 활용하였고, 데이터 시각화를 위해 matplotlib의 pyplot, seaborn을 활용하였다.

* 1. **데이터 전처리**

1. **데이터 값 float 형으로 변환 전처리**# 각 열에 대해 쉼표 제거 및 % 기호 처리 후 float로 변환  
   for col in df.columns:   
    df[col] = df[col].replace({',': ''}, regex=True) # 쉼표 제거  
    # % 기호 제거하고 숫자로 변환 (퍼센트 값은 실제 값으로 변환)  
    df[col] = df[col].replace({'%': ''}, regex=True).astype(float)  
    # 퍼센트 값 실제 숫자로 변환 (0.70% -> 0.0070)  
    if df[col].dtype == float:  
    df[col] = df[col] / 100

전체 데이터 값에 대하여, float형 변환을 위해, 쉼표값이 있다면 쉼표를 제거하고, %형의 단위인 경우 100으로 나눈 뒤, 해당 값들을 float으로 형변환을 수행하였다.

1. **Logit 변환을 통한, 데이터 스케일 맞추기**# 로짓 변환 함수 정의  
   def logit\_transform(x):  
   # 0과 1에 가까운 값을 클리핑하여 계산 오류 방지  
   x = np.clip(x, 1e-6, 1 - 1e-6)  
   return np.log(x / (1 - x))

# 데이터프레임에 로짓 변환 적용  
df\_logit = data\_renamed.apply(logit\_transform)  
변수별 스케일 차이가 큼으로 인한 데이터 분석시의 편향 및 오차 문제를 해결하기 위해, 변수 로짓 변환을 수행하였다.

**6. 데이터분석**

**6.1 변수 중요도 및 상관관계 분석**데이터 전처리가 완료된 거시경제 변수는 총 165개이다. 통상 변수가 많을수록 성능이 우수한 모형 개발이 가능하나 거시경제 변수의 특성상 다음과 같은 상호작용을 가질 수 있기 때문에 모형개발 시 불필요한 변수가 사용되어 다중 공선성으로 인해 모형이 과적합되거나 복잡해질 우려가 있다.

#### <거시경제 변수간 상호 작용 예시>

#### GDP와 실업률: 경제가 성장(GDP 증가)하면 고용이 증가하고, 실업률은 감소하는 경향

#### GDP와 소비: GDP의 증가(소득 증가)는 소비의 증가를 유발

#### 물가와 금리: 물가(인플레이션)에 따른 금리의 상승(중앙은행의 통화정책(금리, 통화 공급)이나 정부의 재정정책(지출, 세금 변화)은 여러 거시변수들에 동시다발적으로 영향을 미침)

#### 환율과 수출입: 환율 변동은 국제수지(수출입)에 영향

이에 변수 간 상관관계 분석 및 중요도에 기반하여 불필요한 변수를 제거하고 모형을 개발하여 계수 추정의 신뢰성과 해석 가능성이 향상시키는 방법으로 모형을 개발하였다.

또한 코로나로 인한 이자유예 및 유동성 제공 등 다양한 정부지원 정책을 통해, 실제 코로나기간 부도율이 매우 낮았던 점을 감안하여 코로나 기간(2020년)을 제외한 데이터 셋을 별도로 생성하여 분석을 수행하고자한다.

1. 텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

   자동 생성된 설명**변수 중요도 분석 결과**전체 변수 대상 변수 중요도는 랜덤포레스트를 활용하여 분석하였다. 모형개발은 선형회귀분석을 활용할 예정이나, 선형회귀는 변수와 종속변수 간의 선형 관계를 가정하지만, 랜덤포레스트는 비선형성과 상호작용을 포함한 다양한 관계를 탐지할 수 있기 때문에 랜덤포레스트를 통해 중요한 변수를 식별한 후 이를 선형회귀 모형에 반영하면 모형의 성능과 해석 가능성을 동시에 고려하고자 한다.   
     
   분석한 결과 중요도가 높은 변수는 TOP20변수는 다음과 같다.  
     
   **<전체 기간 분석 결과>**

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명**<코로나기간 제외 분석결과>**

1. **상관관계 분석결과**

랜덤포레스트는 변수 중요도 분석에 효과적인 도구이고, 이를 통해 선형회귀 모델 개발에 활용하면 효율적인 변수 선택이 가능하나 여전히 과적합 이슈는 존재한다.

따라서 data.corr()을 통해 독립변수 간 상관관계 분석을 추가로 수행하여 상관관계가 0.8 이상인 변수를 그룹화하여 상관관계가 높은 변수를 제거하여 모형개발에 활용 할 수 있도록 하였다.

상관관계가 0.8 이상인 그룹 중 중요도가 가장 높은 변수만을 남기는 변수 선정 작업을 진행한결과 165개의 변수 중 전체 기간의 경우 48개 변수가 제거되어 118개변수가 선정되었으며, 코로나 제외기간의 경우 116개 변수가 선정되었다. 남은 변수의 변수 중요도의 순서는 다음과 같다. 기존 전체 기준대상으로 분석한 결과 중요도가 상위였던 변수는 대부분 유지된 것으로 확인 되었다.

텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명**<전체기간>**

텍스트, 평행이(가) 표시된 사진

자동 생성된 설명**<코로나 제외기간>**

텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진

자동 생성된 설명

**6.2 최종선정변수 선정**

변수 중요도 분석과 상관관계 분석을 통해 1차 선정 시에는 독립변수간의 관계를 분석했다면 모형개발을 위해 최종 선정 변수는 종속변수인 부도율변수와의 관계를 분석하게 된다.

결과적으로 종속변수인 기업의 부도율과 가장 상관관계가 높은 변수 4개를 선정하여 모형 개발을 진행하고자 한다. 변수의 갯수를 4개로 제한 한 사유는 너무 많은 변수가 선정될 경우 모형의 해석이 복잡해지는 것을 방지하고, 3개 이내의 적은 변수 선정 시 다양한 거시경제 변수의 영향을 제외하고 분석결과를 도출 하는 것을 방지하기 위해 적정한 변수의 갯수를 선정하였다.

이 선정된 변수들은 종속변수 기업의 부도율과 강한 선형 관계가 있는 변수들로, 모델에 중요한 특성으로 활용될 수 있으며 선정 결과는 다음과 같다.

**<전체기간>**

스크린샷, 텍스트, 직사각형, 라인이(가) 표시된 사진

자동 생성된 설명

|  |  |  |
| --- | --- | --- |
| **변수명** | **레이블명** | **비고** |
| var\_138 | 외국인(순매수) | 한국거래소 출처 투자자별 주식거래 자료 중 개인 순매수 금액 |
| var\_161 | 이자보상비율 | 한국은행 기업경영분석지표 중 전산업 대상 재무비율 자료 |
| var\_27 | 금융채 | 한국은행 금리지표 중 수신금리의 금융채 자료 |
| var\_40 | 공산품 | 한국은행 경제통계국 물가통계 부문 광산품 물가지수 |

**<코로나제외기간>**

스크린샷, 텍스트, 디스플레이, 직사각형이(가) 표시된 사진

자동 생성된 설명

|  |  |  |
| --- | --- | --- |
| **변수명** | **레이블명** | **비고** |
| var\_156 | 자기자본비율 | 한국은행 기업경영분석지표 중 전산업 대상 재무비율 자료 |
| var\_138 | 개인순매수 | 한국거래소 출처 투자자별 주식거래 자료 중 개인 순매수 금액 |
| var\_157 | 차입금의존도 | 한국은행 기업경영분석지표 중 전산업 대상 재무비율 자료 |
| var\_27 | 금융채 | 한국은행 금리지표 중 수신금리의 금융채 자료 |

* 1. **모형개발 프로세스 및 결과**

1. **모형 개발 방법 및 절차**

부도율과 거시경제변수간 Logit 회귀분석을 통해 미래 부도율을 추정하는 선형 회귀 모형 적용 하였으며, 개발 프로세스는 다음과 같다.

****

**<모형 개발 프로세스 요약>**

|  |  |  |
| --- | --- | --- |
| step 1 | 변수 변환 | 거시경제 변수의 경우 이자율에서 조단위의 통화량, 증감률까지 변수의 스케일 격차가 크므로 선형회귀모형에 적합한 로짓변환으로 변수를 변환하는 전처리 작업 수행 |
| step 2 | 중복 변수 제거 | 167개의 거시 경제 변수 중 랜덤포레스트를 활용한 변수 중요도를 기반으로 독립변수 간 상관관계가 높은 중복 변수는 제거하여 118개 변수만 남김 |
| step 3 | 변수 선정 | 118개 변수중 종속변수와 가장 상관관계가 높은 변수를 기준으로 4개 변수를 선정 |
| step 4 | 선형회귀모형 개발 | 모형결과 MSE와 R2값을 기반으로 모형적합결과 타당성 판단 |

1. **모형 개발 결과**

모형 개발 결과 다음과 같은 회귀식을 도출 할 수 있었다.

**전체 기간 개발 시 회귀식:** y = -26.86 + -0.11\*var\_138 + -7.59\*var\_161 + 1.20\*var\_27 + -1.49\*var\_40

**코로나 기간 제외 개발 시 회귀식:** y = -4.10 + -5.56\*var\_156 + -0.07\*var\_138 + 3.82\*var\_157 + -0.04\*var\_27

**<전체 기간 모형개발 결과 및 성능>**

1. **결정계수 (R²): 0.7233**

R² (결정계수)는 모델이 종속변수(y)의 변동을 얼마나 잘 설명하는지 나타내는 지표이다.. R² 값은 0과 1 사이의 값을 가지며, 1에 가까울수록 모델이 데이터를 잘 설명하는 것으로, R² = 0.7233은 모델이 데이터의 변동성을 설명하는데에 양호한 성능을 가지고 있는것으로 해석 할 수 있다.

1. **평균제곱오차 (MSE): 2.5854**

MSE (Mean Squared Error)는 예측값과 실제값 간의 차이를 제곱하여 평균을 낸 값입니다. MSE 값이 작을수록 모델이 더 정확하게 예측하는 자표로ㅡ

MSE = 2.5854는 모델의 예측값과 실제값 간의 오차의 제곱 평균이 약 2.59라는 의미이다. 변수값을 로짓변환하여 종규화 진행하긴 하였으나, 이 값이 크지 않다는 점에서, 모델의 예측 오차가 비교적 작은 편에 속한다고 할 수 있을 것이다.

1. **회귀 계수 (Coefficients): [-0.1108, -7.5871, 1.1980, -1.4882]**

회귀 계수는 각 독립변수가 종속변수에 미치는 영향을 나타낸다.

선정 된 변수중 부도율에 가장큰 영향을 미치는 종속변수는 매출액세전순이익률로 매출액세전순이익률이 증가할때 부도율의 하락이 강한 영향을 받는것으로 분석되었다

|  |  |  |  |
| --- | --- | --- | --- |
| **MSE** | | 2.585352449 |  |
| **R2** | | 0.723265889 |  |
| var\_138 | 개인(순매수) | -0.11076742 | 음의 영향: 개인순매수가 증가하면 부도율이 감소 |
| var\_161 | 매출액세전순이익률 | -7.58707294 | 음의 영향: 매출액세전순이익률이 증가하면 부도율이 감소 |
| var\_27 | 금융채 | 1.19800729 | 양의 영향: 금융채가 증가하면 부도율이 증가 |
| var\_40 | 광산품 | -1.48822692 | 음의 영향: 광산품물가가 증가하면 부도율이 감소 |
| **절편** | | -26.85672566 |  |

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

**<코로나 제외 기간 모형개발 결과 및 성능>**

1. **결정계수 (R²): 0.4687**

코로나 기간 제외 시 전체 기간대비 결정계수성능은 하향하는 것으로 나타났다. 다만 R² (결정계수)는 R² = 0.4687은 모델이 데이터의 변동성을 설명하는데에 보통의 수준은 설명할 수 있다고 해석할 수 있다.

1. **평균제곱오차 (MSE): 2.9764**

MSE (Mean Squared Error) 값 역시 전체 기간 대비 다소 증가하였으나, 오차가 크지는 않다고 해석할 수 있다.

1. **회귀 계수 (Coefficients): [-5.55745137 -0.06748424 3.81882173 -0.04040352]**

선정 된 변수 중 부도율에 가장 큰 영향을 미치는 종속변수는 자기자본 비율로 자기자본의 비율이 증가할 때 부도율의 하락이 강한 영향을 받는것으로 분석되었다. 또한 동일한 금융채 변수에 있어 전체 기간 분석 시 양의 방향성을 가졌으나, 코로나 기간 제외 시 음의 방향성을 갖는 결과가 도출되었다.

|  |  |  |  |
| --- | --- | --- | --- |
| **MSE** | | 2.976416029 |  |
| **R2** | | 0.468775564 |  |
| var\_156 | **자기자본비율** | -5.55745137 | 음의영향: 자기자본비율이 증가하면 부도율이 감소 |
| var\_138 | **개인순매수** | -0.07206635 | 음의영향: 개인순매숙 증가하면 부도율이 감소 |
| var\_157 | **차입금의존도** | 3.81882173 | 양의영향: 차입금의존도가 증가하면 부도율이 증가 |
| var\_27 | **금융채** | -0.04040352 | 음의영향: 금융채가 증가하면 부도율이 감소 |
| **절편** | | -4.100980859 |  |

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

**<전체기간 실제 부도율과 예측부도율 추이>**

**<코로나 제외기간 실제 부도율과 예측부도율 추이>**

**7. 결론**

우리의 예측대로 000변수가 가장 큰 영향을 주고 있었다. 반면 000은 아니었는데 사유는**…(설화 과장님이 작성해주시는 결과 엑셀파일 보고 작성하겠습니다!)**