Bankruptcy prediction using machine learning and an application to the

case of the COVID-19 recession1. Introduction

* 1. 부도예측모델의 중요성

(1). 부도예측을 조기에 해낸다면 투자자와 은행들, 회사의 매니저들이 부도예측 정보를 통해 시정조치를 하여 그들의 경제적 자원을 보존할 수 있다.

(2). 입안자들이 파산을 야기할 수 있는 시스템적인 문제들을 완화시킬 수 있는 시간을 제공하는데 도움이 된다.

(3). 예측은 경영자가 파산을 피할 수 있는 결정을 할 수 있도록 추가적인 시간을 제공한다.

(4). 부도예측을 통해 특정 부문의 \*\*플래그십 기업의 파산이 예상될 경우 경영자와 정책입안자들이 이를 인지하고 시정조치를 취해 해당 부문의 하락 규모와 강도를 제한할 수 있다,

\*\*플래그십(flagship)이란 조직이 소유하거나 생산하고 있는 가장 중요한 제품, 생각, 건물 등이라는 뜻인데 여기서는 해당 분야의 가장 선두에 있는 기업이라는 의미정도로 해석하면 될 것 같습니다.

(5). 투자자들이 부도예측모델을 기반으로 더 나은 결정을 내릴 수 있도록 한다.

(6). 신용평가기관들이 부도예측모델을 기반으로 해당 기업의 신용을 조사하고 평가하는데 도움을 준다.

1.2 과거 연구와의 비교

머신러닝을 이용한 부도예측모델을 만드는 과거 연구들은 대부분 상대적으로 적은 양의 회사 샘플들과 경제적지표를 사용했다. 이 연구가 과거의 다른연구들과 차별되는 점은 1970년부터 2020년까지의 약2만1천개의 데이터와 57개에 달하는 경제지표를 사용해서 연구를 했다는 점이다.

기존의 연구들에 대한 정보는 아래 Table 1에서 확인할 수 있다.

이 연구에서 부도예측모델을 만들기 위한 기법으로는 랜덤포레스트, SVM, XGBoost 3가지를 선정하였다. 3가지 기법을 활용하여 모델을 만들고 2020년 하반기 부도예측에 활용해 보고자 한다. 많은 해당 분야의 전문가들이 2008년의 경제위기에 비교하며 그때 보여진 파산율의 최고치와 2020년의 파산율이 같을 것이라 하며 파산의 쓰나미가 몰려오고 있다는 말을 한다. \*\*그러나 이 연구에서는 높아진 파산율이 2010년과 크게 다르지 않음을 발견했다.

\*\*이 부분은 파산율이 상당히 높은 것은 맞지만 가장 심각했던 2008-2009년 보다는 낮으며 조금 완화된 수치인 2010년과 비슷하다고 말하는 듯하다.

텍스트이(가) 표시된 사진

자동 생성된 설명

1.3 이 연구의 중요성

이전의 연구들은 모델에 사용되는 데이터보다 모델 자체에 초점을 맞추고 연구를 진행한 반면 이 연구는 데이터와 모델 둘 모두에 대한 균형잡힌 관점을 갖고 연구를 진행될 것이다. 데이터는 Fama and French와 같은 유명한 금융관련 논문에 사용되었으며 다양한 경제지표를 제공하는 Compustat이라는 데이터베이스의 데이터를 사용할 것이다.

\*\* 정확히 일치하진 않지만 이 연구에서 사용한 경제지표가 궁금한 경우는 아래 링크를 확인하면

대략적으로나마 확인할 수 있을 듯 하다. <https://libguides.utsa.edu/c.php?g=1009069&p=7309863>

2. Literature review

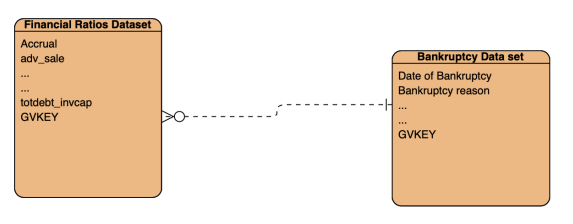
해당 부분은 Table 1의 선행 연구들에 대해 언급하는 부분이라 생략

Literarture review에 근거하면 다음의 3가지는 명확하다.

* 머신러닝 모델이 통계적모델모다 일관적으로 뛰어난 성능을 보여주고 있다는 것
* 다른 분야에 비해서 머신러닝모델을 훈련하기 위해 사용하는 데이터의 수가 상대적으로 작다는 것
* 랜덤포레스트와 XGBoost같은 앙상블 기법이 부도예측에서 다른 모델보다 좋은 성능을 보여주고 있다는 것.

3. Data and methodology

3.1 Data

수집한 데이터는 아래와 같이 두 테이블로 이루어져 있으며 GVKEY라는 기본키를 이용하여 join했다. .

3.2 Methodology

3.2.1 Data preprocessing

데이터베이스에서 21,114개의 기업에 대한 경제지표를 수집했다. 경제지표중에서 6000개이상의 기업에 대한 결측치를 보유한 경우, 즉 30%이상 결측치를 보유한 경우는 제거하였고 57개의 경제지표가 남게되었다. 57개의 경제지표의 결측값을 채우기 위해 KNN 알고리즘을 활용하였고 n\_neighbors을 3을 줬을 때 데이터의 실제분포 보존에 효과적이라는 것이 밝혀졌다. 텍스트이(가) 표시된 사진

자동 생성된 설명

3.2.2 Balancing the dataset

데이터의 90%이상이 Non-Bankrupt에 속해 있는데 연구의 목적이 Bankrupt에 속하는 기업들을 식별하는 것이기 때문에 두 데이터 셋의 크기를 맞춰주기로 결정했다. SMOTE 기법, Borderline-SVM SMOTE, ADASYN 3가지 기법을 사용했고 그 결과 non-bankrupt에 속하는 19902개의 데이터와 bankrupt에 속하는 20517개의 데이터를 가진 균형잡힌 데이터 셋이 완성되었다.

\*\*만약 bankrupt에 속하는 기업의 수가 3%정도 되는 데이터를 가지고 훈련시켰을 때 전부 non-bankrupt에 속한다고 모델이 판단해도 97%의 정확도가 나올 수 있기에 데이터를 업샘플링 해주는 것이라고 한다. 위의 언급한 3가지 업샘플링 기법은 아래 링크를 통해 쉽게 이해할 수 있을 듯 하다. <https://wyatt37.tistory.com/10>

3.2.3 Fitting training data into models

훈련데이터와 테스트데이터를 7:3비율로 나누고 랜덤포레스트, SVM, XGBoost 3가지 모델을 훈련시켰다.

3.2.4 Performance analysis

모델의 성능을 비교하기 위해 Accuracy score. ROC, AUC 3가지를 사용하였다.

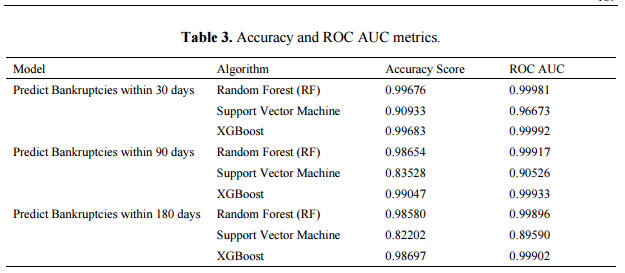
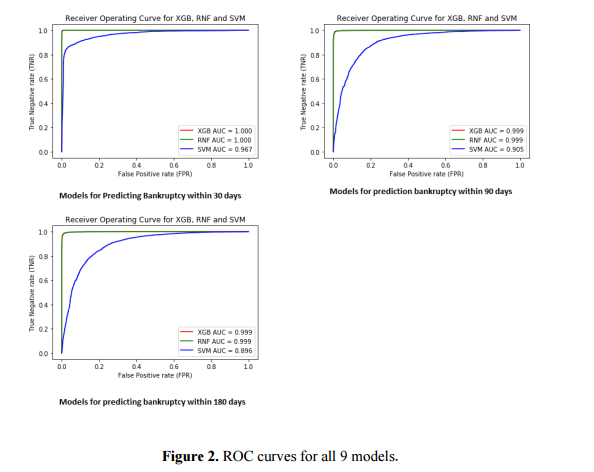
3.2.5 Predicting bankruptcies

이 연구의 목표는 다음 30일, 90일, 180일 이내에 파산할 기업들의 수를 예측하는 것이기 때문에 30일, 90일. 180일이내의 파산 수를 예측하기 위해 앞서 언급한 3가지 모델을 사용하였다. 유일한 차이점은 각 모델에 대한 train, test 데이터가 파산날짜에서 파생되었다는 것이다. 예를 들어 30일 이내의 파산 수를 예측하는 모델은 아래와 같은 방식으로 훈련되었다. 텍스트이(가) 표시된 사진

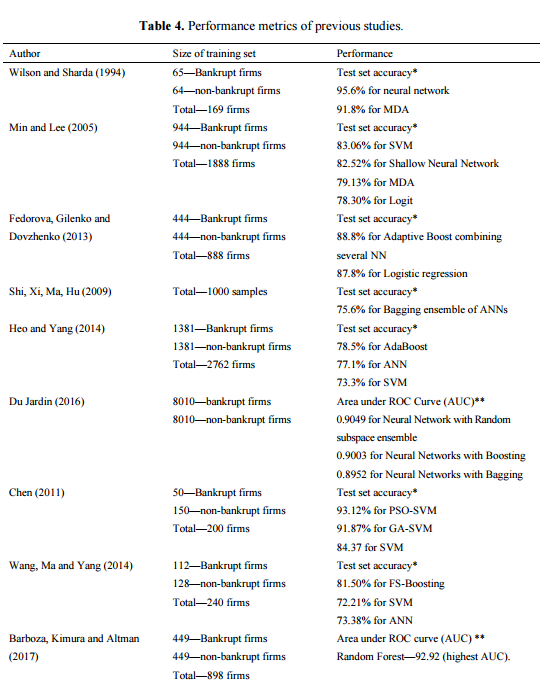
자동 생성된 설명

4. Results

앞서 언급한 랜덤포레스트, SVM, XGBoost 3가지 모델을 가지고 예측을 진행했을 때 XGBoost 모델이 다른 2가지 모델의 30일, 90일, 180일 예측에서 보인 성능보다 일관되게 높은 성능을 보였음을 알 수 있다.

4.1 Comparison of performance with previous studies



5. Finish

나머지 부분은 각 연도별 파산율에 대한 비교와 2020년 하반기에 파산으로 예측 되는 기업의 숫자를 나타낸 내용이라 생략하였다.

이 논문 리뷰를 통해 부도예측에 대한 역사? 와 데이터는 가능한 많은 양을 수집하고 훈련시킬 모델을 선정할 때 앞서 언급한 3가지 모델을 우선적으로 사용해보면 좋을 것 같다는 생각이 들었다.