[Paper Review] The application of machine learning in bank credit rating prediction and risk assessment

네 번째 논문 리뷰! 오늘은 최근 관심이 생긴 신용 평가와 관련된 논문을 읽어봤다. 기업을 대상으로 하는 lending 전략의 최적화를 위해 4가지 알고리즘으로 모델링하고 각각의 성능을 비교하는 내용이다. 이번 리뷰 역시 번역/요약했기 때문에 약간의 오역이나 오류가 있을 수 있다.

출처:
https://ieeexplore.ieee.org/abstract/document/9389901?casa_token=V_DSxToFAuMAAAAA:pOm4dJIRGTg7EwxBSm1wKGk49dXKggzED8e4mgLsn_0z1Iqs19JOmNORPiJVNh0DU-gu2SOIKOk

목차

Abstract

금융 산업에서 중요한 이슈인 신용 예측, 신용에 영향을 주는 여러 복잡한 특성들에 대해 성능 좋은 모델을 만들기 위해서는 효과적인 차원 축소가 필요하다. 이를 위해 본 논문에서는 머신러닝 알고리즘 중 lasso 회귀와 주성분분석(PCA) 알고리즘을 사용해 주요한 특성들을 선택한다. 또한, 지도학습 방법 중 하나인 서포트 벡터 머신 (SVM)과 랜덤 포레스트, gradient boosted classfication 알고리즘을 적용한다. 결과적으로, SVM 모델이 검증 데이터셋에 대해 86%의 best accuracy를 보였다. 더불어, 금융 생태계의 복잡성에 따른 불확실성을 측정하기 위해 Cox and Kaplan-Meier 모델을 적용해 각 특성들의 효과를 설명한다. 이를 통해, 기업의 daily revenue와 신용 등급 간 관계를 측정한다.

1. Introduction

정부와 은행은 손실과 위험을 최소화하기 위해 적절한 신용 전략을 세워야 한다. 이를 위해 기존에는 신용 등급 예측에 Jarro,R.A 등의 Markov 모델을 사용했다. 최근에는 신경망의 역전파 기술 및 서포트 벡터 머신을 기반으로 정확도를 80% 부근까지 향상시켰다. 또한, Chih-Fong Tsai 등의 연구에서는 hybrid 분류기를 사용해 가장 높은 정확도를 달성하기도 했다. 이 뿐만 아니라, 신용등급 예측을 위해 다양한 딥러닝 알고리즘이 적용되고 있는 추세이다.

더불어, 차원의 저주를 피해기 위해 feature selection이 화두가 되었다. 이와 관련해 IGDFS는 중요하지 않은 변수들을 제거함으로써 information gain을 획득하는 방식이며, Lasso는 제거되어야 할 특성들을 찾아내기 위해 각 특성들의 parameter들을 압축하는 방식이다. 또한, Cox and Kaplan-Meier 방식도 신용등급과 관련된 변수들을 결정하는데 사용된다.

본 연구에서는 lasso와 반복적인 특성 제거를 기반으로 가장 중요한 변수들을 찾

아내고 각 모델의 정확도, 민감도, 특수성을 비교한다.

2. Materials

A. Data Source

3기업으로부터 수집한 financial 등급에 대한 데이터셋은 각 기업의 시계열 소득 자료이며, 첨도, 분산, 평균, 소득이 없는 날의 수, 마이너스 소득, 매출 송장 취소 등의 정보를 담고 있다.

B. Features Selection

차원 축소를 위해 caret 패키지의 10-fold 교차 검증과 lasso-회귀에 대한 random forest 함수로 특성을 반복적으로 제거해간다. 교차 검증 끝에, 총 3개의 특성 (측도, 매출송장 취소, second sample entropy)을 최종 채택한다.

3. Methods

최종적으로 선택된 3개의 주요 변수들로 SVM, 랜덤 포레스트, gradient boosted 분류기를 학습시키기 위해 5겹 교차 검증을 수행한다. 각각의 모델은 e1071 패키지와 h2o 패키지로 구현할 수 있다.

A. Random Forest

랜덤 포레스트의 주요 목적은 여러 결정 트리들을 조합해 얻은 결과를 최종적인 예측 결과로 정하는데 있다. 이때 각각의 트리로부터 얻은 결과보다 전체 결합된 알고리즘의 정확도가 더 높게 나타난다.

B. Gradient boosted classification

부스팅은 지도학습 분류 모델 중 하나로, 성능이 낮은 학습기를 조합해 강한 학습기를 만들어내는 방식이다. 이 알고리즘은 손실함수의 기울기가 작아지는 방향으로 모델을 학습하는데, 여기서 손실함수는 모델의 unreliable한 정도를 의미한다. 따라서 이 값을 줄여 모델의 정확도를 높여야 한다.

C. SVM

서포트 벡터 머신은 지도학습 방식으로 이진 분류하기 위한 일반적인 선형 분류기이다. 이 알고리즘은 초평면 상에서 최대 마진을 갖도록 식을 세우며, 비선형적인 입력값에 대해서도 선형적인 방식과 유사하게 표현할 수 있기 때문에 수학적으로 해석하기에 용이하다.

서포트 벡터 머신은 일직선 (고차원 데이터셋에 대해서는 초평면)으로 데이터를 분류하며, 분류선으로부터 각 그룹의 점을 가능한 멀리 배치시키도록 한다. 이로써 새로운 데이터에 대해 더 좋은 분류 성능을 보일 수 있다. 분류선 (또는 초평면)은 다음과 같이 표현된다.

$$w^T x + b = 0$$

여기서 w와 b는 SVM 모델의 매개변수들로 각 평면에 따라 다른 값을 가진다.

D. Cox and Kaplan-Meier Method

Cox-회귀 모형은 비례 위험 모델이라고도 불린다. 이는 다양한 기업의 특성에 따른 기업의 생존 기간, survival outcomes을 분석한다. 이 분석의 목적은 특성값 X들과 outcome 간의 관계를 파악하는데 있으며, 다음과 같이 표현된다.

$$h(t, X) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)$$

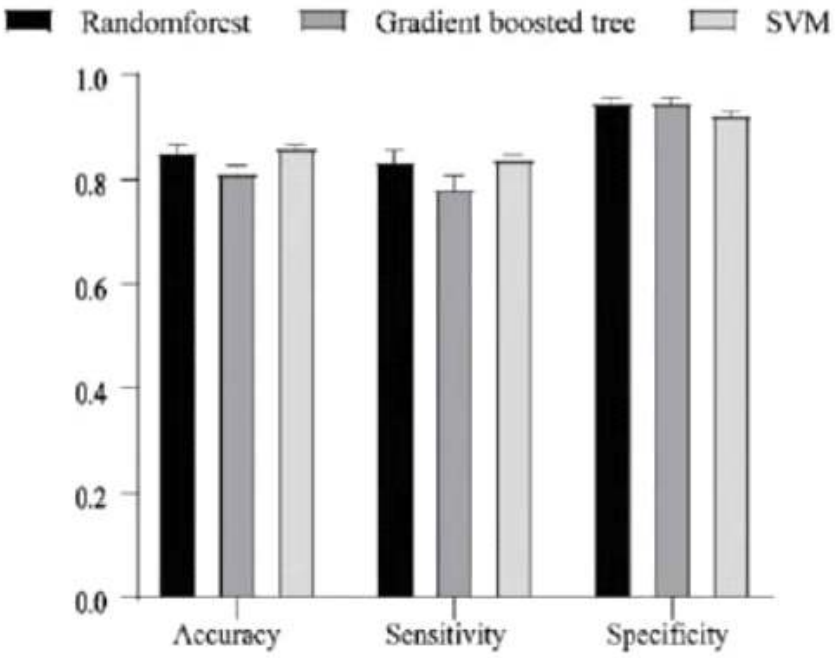
여기서 B_1, B_2, ..., B_m은 각 독립변수들의 부분 효과를 의미하며, h_0(t)는 모든 X들이 0일 때의 baseline risk를 의미한다. 본 논문에서는 이 Cox 비례 위험 회귀 모델을 활용해 기업의 위험에 대한 risk를 측정한다.

4. Conclusion

A. Model

SVM, 랜덤 포레스트, gradient boosted trees로 기업의 신용등급을 분류해보았다. 각 모델의 성능 검증에는 5겹 교차 검증을 적용했다. 결과적으로는 아래와 같이 SVM이 각각 86%, 84%로 가장 높은 정확도와 민감도를 가졌고, 랜덤포레스트와 gradient boosted tree는 95%의 더 높은 specificity를 가졌다.

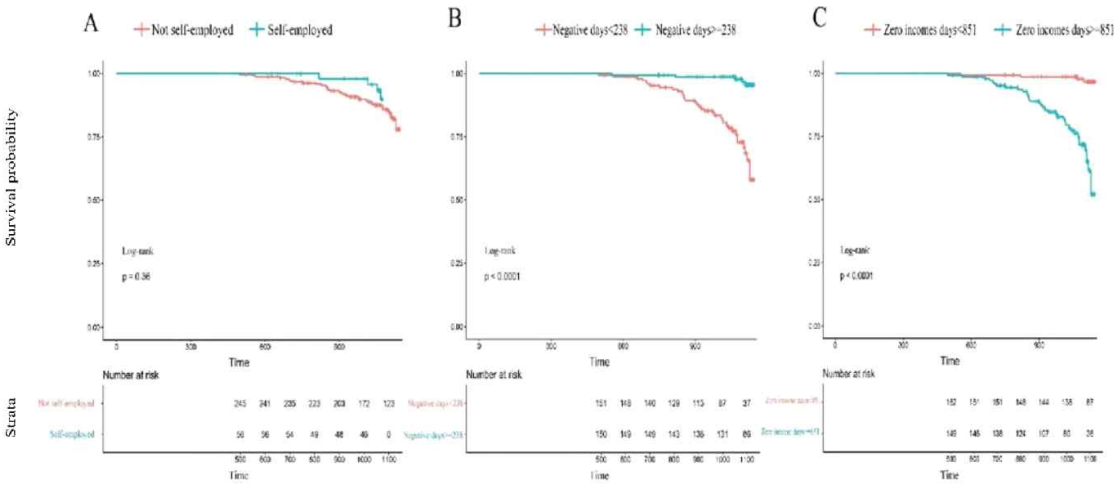
V. FIGURE AND TABLES



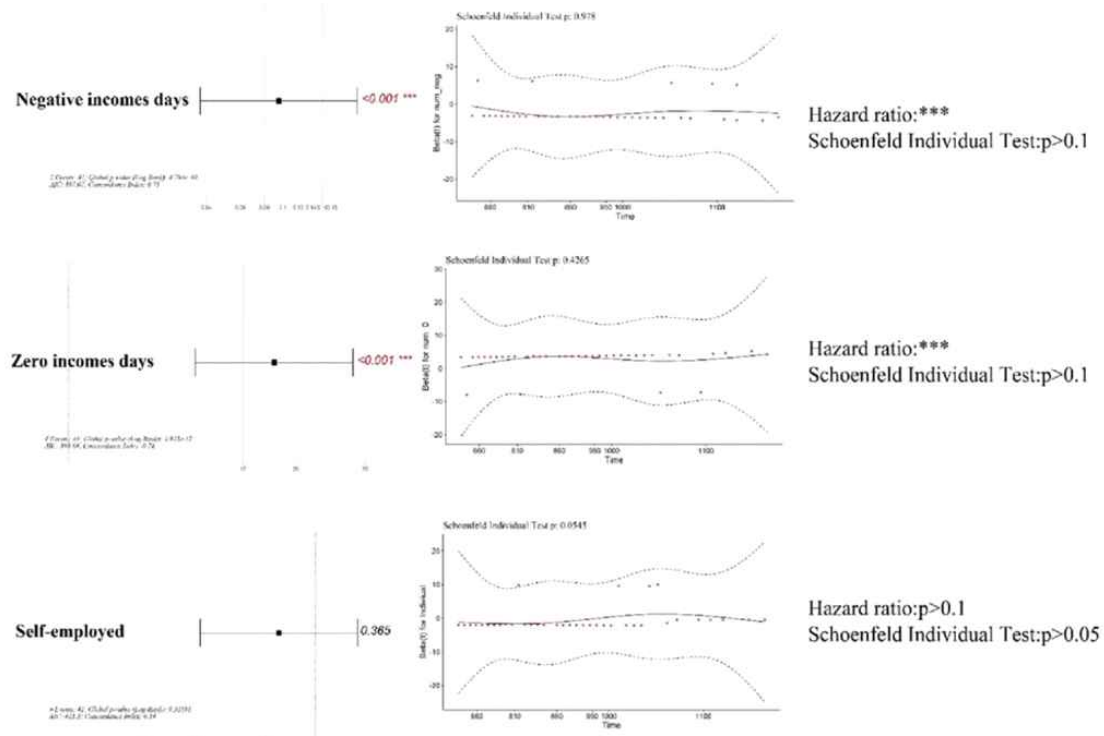
모델 학습의 결과

B. Risk assessment by survival analysis

최종적으로 선택된 3개의 특성 (self-employed, negative income days, zero income days)에 대해 Kaplan-Meier R 패키지를 통해 분석해본 결과, negative or zero income days 는 신용 등급에 유의한 영향을 미친다. (p-value < 0.0001) 반면, 기업의 self-employed 변수는 신용 등급에 그다지 유의하지 않은 결과이다. Hazard Ratio 역시 negative and zero income days 특성에 대해서는 신용 등급을 유의하게 설명한다고 reporting 하지만 (p-value < 0.01), self-employed 특성에 대해서는 그렇지 않다. 따라서, 정부나 은행은 기업의 신용 평가에 위 두 특성을 고려해볼 수 있다.



Kaplan-Meier 패키지로 분석한 변수별 유의성



Hazard Ratio로 해석한 변수별 유의성