2. 지난 주에는 재무 데이터에 대한 기본적인 이해와 전처리를 끝냈고, 간단히 모델을 돌려보는 작업까지 마침. 이번 주에는 feature extraction에 대해 더 이야기를 나눠보고 이를 다시 한 번 모델에 돌려보는 작업을 했다.

3. 기업의 재무상태표, 손익계산서, 현금흐름표 등 각종 항목의 연관성을 살펴보는 것이 분석의 목적 중 하나였음. 그래서 주어진 데이터에 대한 간단한 이해가 필요. 대부분이 재무비율의 항목들이었는데, 각각의 변수가 어떤 의미인지 모두 파악하긴 힘들었다. 또, 전체적으로 high correlated 되어있었고 이 부분을 고려해서 변수를 추려가는 과정을 진행함.

4. 우선 데이터가 상당이 skewed 되어있어서 이상치의 영향을 최소화 하기 위해robustscaler 방법을 사용해서 스케일링 진행함. 이후 다중공선성 관련하여 다음 두 가지 방법을 고려해 변수를 축소시켰고 결과적으로 42개의 변수가 남게 되었다.

5. 42개의 변수들에서 공통적으로 많이 등장하는 변수들을 재무상태표, 손익계산서, 현금흐름표의 각 항목별로 정리해보았다. 다음 표와 같이 정리가 되었는데, 대부분이 재무상태상태표의 항목들임을 알 수 있었다. 그래서, 기업의 파산 여부 예측 시 현재 기업의 돈이 어떻게 구성되어있는가의 상태를 나타내주는 재무상태표가 중요하게 고려됨을 알 수 있음.

6. 그래서 이렇게 정리된 변수들을 가지고 다음의 모델링 계획을 세워보았다. 처음 base model에서 전처리 이후 full data를 활용했고, 이후에는 clustering+PCA 그리고 여기에 추가적으로 correlatoin까지 고려한 방법으로 이렇게 총 세가지 시도해 봄.

7. 초기 모델에서는 unseen data를 통한 예측 결과 0.5

8. base model에서는 0.4375가 나왔다.

9. clustering과 pca방법을 사용한 경우는 다음과 같음. 우선, clustering을 총 세 그룹으로 해주었는데, 첫번째는 short term liabilities 관련 변수들, 두번째는 long term liabilities 관련 변수들 마지막으로 세번째는 그 외 변수들로 묶어주었다. 이후 PCA를 통해서 각각 3개,1개,3개의 주성분을 추출해냈다. 아래 표는 관련 논문인데, grouping 과정에서 참고하였다. 하지만 이 방법에서는 f1 score를 살펴보니 성능이 너무 좋지 않게 나와서 기업 파산 예측에 적절하지 않다고 판단함.

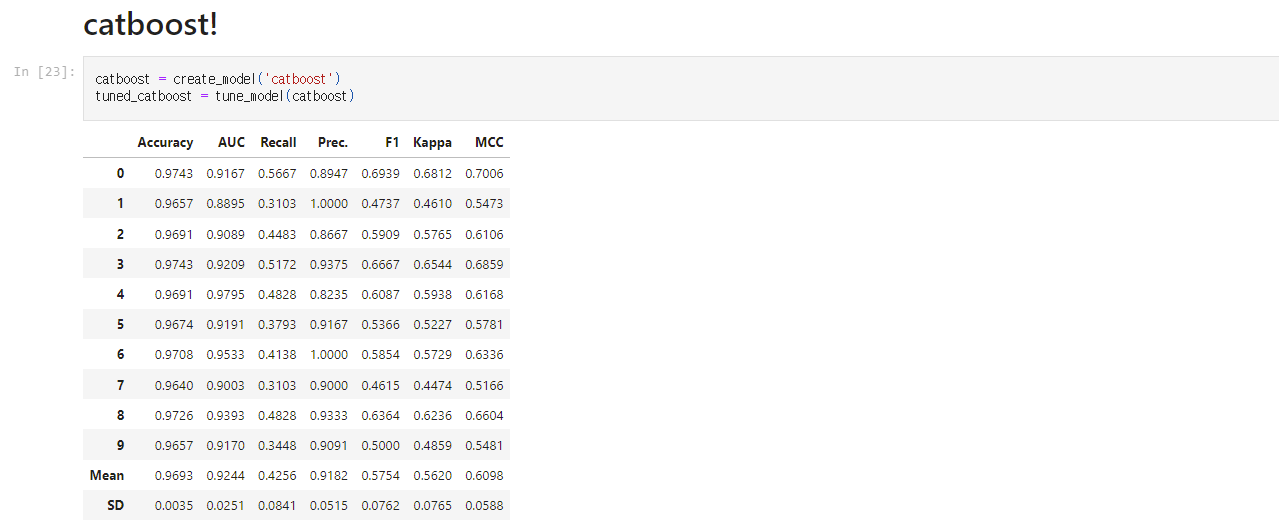
10. 마지막으로 clustering+pca+correlation 방법. 이 자료는 특정 재무비율군의 의미만 맞으면 해당 카테고리로 분류 가능하다는 것을 보여주고 있음. 사용하기로 결정한 변수들을 각각 어떤 분류로 넣을 수 있을지 조사를 해서 앞의 6가지 ratio들을 사용함.

11. 이 표가 앞선 grouping 기준을 적용하여 42개의 변수들을 먼저 분류를 한 상태이다. 말한대로 이러한 6가지 분류 기준을 사용하여 clustering 완료함.

12. 그 다음 pca를 진행하였는데, 이를 위해 각 군집별로 standard scaler를 한 번 더 해줌으로써 분산의 크기를 맞춰주었다. eigenvalues 값을 확인했을 때 0.95 이상이 되는 점에서 끊겼고 총 5개의 변수가 제거됨을 볼 수 있었다. 그러나 여러 모델들의 성능 비교시에 f1 score가 너무 좋지 않게 나와서 pca 중요 가정 중 하나인 선형성에 문제가 있다고 판단! 그래서 이후 시도에서는 각 군집별로 선형성을 만족하는 변수들을 다시 추려 pca를 해보았다. 여기서는 분산의 설명력이 0.7이상일 경우 cut 했고 총 5개 변수가 제거되었다.

13. 여기서는 f1 score 기준으로 catboost가 가장 높았고, 베이스모델에서의 결과와 비교했을 때 동일 알고리즘 대비 성능이 낮음을 확인했다.

14. 그래서 튜닝을 한 이후에 다시 살펴보았는데, f1 score 기준으로 특정 fold에서 0.15 향상됨을 볼 수 있었음!



=> 중창님이 노션에 올려주신 이미지인데 f1 score가 0.15 향상되었다는 부분에 같이 보여주면 좋을 것 같습니다!

15. unseen data를 가지고 prediction 한 결과를 보았을 때 light GBM과 Catboost의 결과가 다음과 같이 나온 것을 확인할 수 있었다. 이러한 결과가 왜 나왔을까에 대해 생각을 해보면.. tree를 분할하는 방법이 light GBM은 층에 대한 제약 없이 잎에서 가지 치고 계속 치는 방식의 leaf-wise(리프 중심 트리 분할)이며 catboost는 일반적인 level-wise 알고리즘을 사용하는 방식으로 둘은 서로 tree 분할 방법이 다른 모델이었음. 또, catboost는 특히 하이퍼파라미터 튜닝에 매우 예민한 친구.

16. 마지막으로 feature importance를 살펴보았는데, 여기서 등장한 변수들이 기업의 파산 여부 예측에 있어 중요한 항목으로 작용함을 알 수 있다. 기존에 있던 34번의 변수가 빠지면서 이와 상관이 높았던 9번 변수가 등장하는 등의 변화도 볼 수 있었다. 이번 분석에서 중요했던 것이 기업 파산 여부 예측 시 어떤 요소들을 중점적으로 보아야 하느냐였는데, 이 feature importance를 참고한다면 좀 더 나은 예측 가능할 것!

17. 그래서 결론적으로 우리가 한 방법으로 기업의 파산 예측이 제대로 될까? 이를 통해 여신 여부를 결정할 수 있을까? 어느정도 가능은 하겠지만 주어진 데이터가 상당히 imbalanced 되어 있어서 변수를 줄이는 동시에 최대한의 예측력을 갖추고자 하는 시도가 있었지만 어려웠다. 그래서 이 부분이 좀 해결된 데이터를 가지고 했다면 괜찮지 않았을까 하는 바람..또 같은 이유에서 f1 score가 적절한 판단 기준인지에 대한 의문도 들었다.