**ppt1**

안녕하세요 4조 발표를 맡게 된 응용통계학과 18학번 신예진입니다.

**ppt2**

목차는 다음과 같습니다. 간략한 조원소개와 데이터 설명을 한 후, preprocessing과 feature extraction의 과정을 소개해드리겠습니다. visualization에서 그래프를 몇 개 보여드리고, 다음주 주제인 modeling의 진행상황을 짧게 보여드린 후 발표를 마치겠습니다.

**ppt3,4**

승지님과 태환님, 민주님이 데이터 전처리와 모델리을 맡아서 해주셨고, 민회님과 일찬님, 제가 변수조사와 변수선택을 도왔습니다.

**ppt5,6**

2번 데이터는 폴란드 회사의 파산 예측과 관련된 dataset입니다. 회계원리에서 봤을 법한 개념들과 생소한 개념들이 각 기업들의 특성을 말해주는 변수를 이루고 있었습니다.

**ppt7**

데이터에 대한 좀 더 자세한 설명을 드리겠습니다. 데이터셋은 총 6855개 회사의 64개의 특성을 보여줍니다. 여기에 추가적으로 class라는 변수가 있는데, class는 categorical data로 1은 파산기업을, 0은 정상기업을 나타냅니다. 왼쪽의 그래프를 보면 알 수 있듯이, 파산한 기업은 6855개 중 361개 밖에 없어, 대부분이 정상기업임을 알 수 있었습니다. 또한 자료소개에는 결측치가 없다고 되어있었는데, NA값이나 nan값이 없었을 뿐, csv파일을 훑어보니 물음표로 표현된 결측치들이 있었습니다. 따라서 물음표의 개수를 알아봤고, 총 6132개의 결측치가 있었습니다. 오른쪽에 보이시는 것과 같이 총 15개의 변수에 결측치가 존재했고, 특히 37번 변수의 경우 총 3100개의 결측치가 존재했습니다. 따라서 preprocessing 단계에서 해결해야겠다고 생각했습니다. 또한 복잡한 변수들이 64개로 너무 많아 데이터 분석이 어려웠습니다.

ppt8

따라서, 변수들의 correlation을 통해 변수선택의 단계를 거쳤습니다.

ppt9

우선, 64개 변수 모두, 각 변수별로 corrleation이 0.7 이상인 다른 변수들을 모두 추출했습니다. 그리고 이 추출 결과를 바탕으로 변수들의 성질이 비슷하다고 판단되는 것끼리 분류해 묶었습니다.

ppt10

이후, 각 그룹 내의 변수끼리 다시 correlation을 조사했습니다. 이를 바탕으로 그룹 내의 변수의 성질들을 대표하는 하나의 변수를 선정했습니다. correlation을 살펴보는 과정에서 한눈에 보기 쉽게 왼쪽과 같은 correlation plot을 그려 판단하였습니다. 이 과정을 통해 총 37개의 변수를 제거하였습니다.

ppt12

다음으로 outlier를 제거하는 과정을 거쳤습니다. outlier의 기준은 각 변수에서 가장 작거나 큰 값이 3개로 정하였습니다. 또한 outlier의 제거는 곧 데이터의 제거이기 때문에 최대한 적은 outlier를 제거하기 위해 여러 변수들에서 공통적으로 outlier로 여겨지는 data를 조사하였습니다. 왼쪽 표와 같이, 3개 이상의 변수에서 outlier로 여겨지는 총 10개의 회사 데이터를 추출하였습니다.

ppt13

또한 각 변수의 plot을 그려봐 눈에 띄는 outlier을 제거해 주었습니다. 제거 후 오른쪽 그래프와 같이 데이터의 경향이 훨씬 잘 드러남을 볼 수 있습니다. 이렇게 두 가지 방법을 통해 아래와 같이 총 20개의 데이터를 제거해주었습니다.

ppt14

다음으로, 처음 데이터 설명에서 언급했던 결측치를 해결하기 위한 NA Imputation 작업을 해주었습니다.

ppt15

우선, 결측치가 자료의 50%에 가까운 37번 변수는 제거하였습니다.

이후, 다른 결측치들은 R mice package의 mice함수로 pmm을 사용해 채워주었습니다.

ppt16

MICE 패키지는 mulitple imputation 방식으로 결측치를 채워넣는 방식입니다. 몬테카를로 방법을 통해 채워놓은 n개의 complete dataset을 만들고, 그 결과를 합쳐 최종 결과물을 내놓는 방식입니다.

predictive mean matching은 baesian linear regression 방식으로 아래와 같은 5개의 단계가 반복되며 결측치를 채워가는 방식입니다. 더 자세한 내용은 아래 링크를 참조하시면 될 것 같습니다.

ppt17

이후 skewed된 변수에 적절한 transformation을 취해주었습니다.

ppt18

각 변수의 plot들이 거의 다 오른쪽으로 길게 늘어진 경향을 보여, 전체 변수에 log transformation을 취해주어습니다. 아래의 그래프는 19번 변수의 그래프입니다. 보이는 것처럼 log transformation을 거친 후 dat의 경향이 훨씬 잘 보이는 것을 확인할 수 있습니다.

ppt19

preprocessnig의 마지막 단계인 Scaling입니다.

ppt20

sclaing은 모든 data를 0과 1 사이에 위치하도록 하는 minmaxscaler를 이용하였습니다. 오른쪽의 결과를 보시면 모든 데이터들이 0과 1사이에 들어왔음을 알 수 있습니다.

ppt21

이렇게 여러단계의 preprocessing을 끝낸 후, 여러 변수를 합쳐 새로운 변수는 것이 용이한 경우가 있는지 살펴보았습니다.

ppt22

첫번째로, 1번과 17번 변수를 합치는 것이 용이하다고 판단하였습니다. 1번 변수는 net profit/total assets, 17번 변수는 total assets/total liabilities입니다. 이 두 변수를 곱하면 net profit/total liabilities를 얻을 수 있는데, 이는 ROI(Return on investment)로 한 회사의 수익성을 판단하는 지표 중 하나 입니다.

또한 오른쪽 아래의 그래프는 1번변수의 수치를 x축으로 하고 17번변수의 수치를 y축으로 하는 그래프입니다. 빨간점으로 표현된 정상기업과 파란점으로 표현된 파산기업이 잘 구분됨을 볼 수 있습니다.

ppt23

두번째로, 9번, 13번, 18번 변수를 이용해 총자산분의 감가상각이라는 새로운 변수를 만들었습니다. 감가상각 금액이 많다는 것은 자산 중에 아직 상각이 끝나지 않은 비교적 최근 자산 혹은 상각 기간이 긴 건물&토지가 많다는 정도로 생각해 볼 수 있습니다.

ppt24

또한 앞의 두 과정과 별개로 PCA를 시도해봤습니다. 앞서 preprocessing 과정에서 correltion이 높은 변수를 0.7 기준으로 한 결과 37개의 변수를 제거하였습니다. 제거한 변수 개수가 너무 많은 것은 아닌가 라는 생각이 들어, correlation이 높은 변수 3그룹을 pca로 차원축소를 해보기로 했습니다. pca 차원축소를 통해 다양한 변수를 한 차원으로 축소해 만들어진 새로운 변수는 해석이 어렵습니다. 따라서 비슷한 특성을 보이는 변수들을 1차원으로 축소하는 방법을 이용하였습니다.

ppt25

이는 3 그룹 중 한 그룹입니다. 변수들의 의미를 봤을 때, 거의 기업의 현금 유동성과 관련이 있는 변수들을 묶었습니다.

ppt26

아래와 같이 스크리 플랏을 그려보니 각 그룹의 변수들을 일차원 축소하는 것이 괜찮아 보임을 확인할 수 있었습니다.

ppt27

마지막을 visualization입니다.

ppt28, 29

preprocessing 과정을 거친 이후 남겨진 변수들이 파산기업을 구별하는 데에 어떤 정보를 주는지 알아보기 위해 boxplot을 그려보았고, 보시는 것과 같이 유의미한 차이를 확인할 수 있었습니다.

ppt30

19번 변수의 경우 매출총이익을 총 매출액으로 나눈 값에 있어 파산기업이 정상기업보다 그 분포가 더 크다는 것을 발견했습니다. 더 자세히 알아보기 위해 gross profit을 x축으로 19번 변수를 y축으로 한 그래프를 그려보니 gross profit 0을 기준으로 변화가 커진 것이 보이고, 부도기업이 gross profit을 음수값으로 더 많이 가지는 것을 확인하였습니다.

ppt31

또한 2번 변수의 feature importance가 높게 나와 boxplot을 그려보니 실제로 큰 차이가 없어 보였습니다. 더 자세히 살펴보기 위해 total asset의 크기를 나타내는 변수와 scatter plot을 그려보니 조금 경향성이 보이는 것을 확인할 수 있었습니다.

ppt32

다른 변수 간의 scatter plot도 그려보았고, 보이는 것처럼 정상기업과 부도기업의 경향성에 차이가 있는 것을 확인할 수 있었습니다.

ppt33

modeling의 진행상황입니다. 마지막에 만들어진 데이터를 바탕으로 Prototype Model(logistic regression, random forest, xgboost 사용)을 만들어봤는데 Catboost에서 AUROC, F1 score가 가장 높게 나왔습니다. 하지만 AUROC가 대략 0.69로 아직 만족할 수준이 아니라 좀 더 향상된 모델을 만들 예정입니다. 그 방법으로 PCA, FA, 파생변수 추가 등을 계획하고 있습니다.

나머지는 ppt에 적혀있는대로 설명!