중간 보고서

팀 이름 : 골목대장
1대장 - 김홍범(팀장)
2대장 - 이태영
3대장 - 하연진

CONTENTS

01 전처리

```
active['휴폐업_여부']=[0]*len(active)
closing['휴폐업_여부'] = [1]*len(closing)
df = pd.concat([active, closing], axis=0)
```

```
df['휴폐업_여부'].value_counts()

0    28620
1     6739
Name: 휴폐업_여부, dtype: int64
```

정답 데이터 라벨링

- 이진 분류(binary classification) vs 다중 분류(multilabel classification)
  -> 휴업의 데이터 라벨링이 어려움
  휴,폐업 여부를 묶은 이진분류 모델 사용
- 휴,폐업 중소법인 데이터와 액티브 중소법인 데이터 concat 진행
- 휴,폐업 기업은 1로, 액티브 기업은 0으로 라벨링
- 액티브 데이터에서 휴폐업 이력이 있는 데이터는 drop하고 순수 액티브 데이터만 선정
- Result : 35359개의 데이터셋 선정

01 전처리

Feature Selection & Engineering

- **(주) 여부’ 컬럼 추가**

CMP_PFIX_NM, CMP_NM, CMP_SFIX_NM, CMP_ENM 으로 (주)가 있는지 없는지 여부를 판단하는 컬럼으로 재가공

- **Feature drop**

공기업 구분여부 : 공기업인 기업이 거의 없으므로 삭제
개인법인구분 : 모두 법인으로 나오므로 컬럼 삭제
산업코드차수 : 10개 빼고 모두 ‘10’으로 라벨링 후 삭제
산업코드2, 산업코드3 : 결측치가 많고 산업코드1만으로 유의미한 데이터를 끌어올 수 있다 판단하여 삭제
공공기관 유형 : 값의 차이가 미세하므로 삭제

01 전처리

Feature Selection & Engineering

- **본점기업코드**

전체 데이터에서 본점기업코드의 개수를 대신 넣음
숫자가 높을수록 체인점을 많이 보유한 업종임을 판단함

- **설립일자**

결측치가 많은데 각 사업자번호별 재무제표의 첫번째 결산일로 변경함

- **상장 여부**

상장일자의 결측치가 다수 존재, 따라서 상장여부에 따른 라벨링 진행

- **홈페이지 여부**

홈페이지 url이 있는지 없는지 여부에 따라 1, 0으로 라벨링 진행

02 비재무 데이터 수집 방안

1. 가정 : 퇴사자, 입사자 수와 중소 기업의 휴,폐업이 연관성이 있다.

- **진행 방향** - 중소기업들의 특정 날짜를 지정할 수 없으므로 해당 기업의 존속기간 동안의 평균적인 퇴사자 및 입사자 수를 구함
- **데이터 수집 방법** : 타겟 기업이 중소기업이다 보니 채용 사이트인 사람인, 잡코리아 등을 크롤링 진행

2. 가정 : 외국인 노동자의 수와 관리는 곧 중소기업의 능력이고 이는 휴,폐업과 직접적으로 연관되어 있을 것이다.

- **진행 방향** - 중소기업들의 특정 날짜를 지정할수 없으므로 해당 기업의 존속기간 동안의 평균적인 퇴사자 및 입사자 수를 구함
- **데이터 수집 방법** : 타겟 기업이 중소기업이다 보니 채용 사이트인 사람인, 잡코리아 등을 크롤링 진행

02 비재무 데이터 수집 방안

3. 가정 : 기업별 산업군의 위험도와 해당 기업의 휴,폐업 여부는 연관성이 있다.

- **진행 방향**
  1. 특정 기간(나이스 디앤비 데이터 수집 기간)에 해당 산업군별 뉴스기사 크롤링
  2. 뉴스기사를 크롤링한 후에 관련 토픽을 임베딩 기법을 이용하여 정량화 시킴
  3. 해당 데이터의 가중치를 계산하여 산업군별 위험도를 측정
  4. 산업군별 위험도를 중소기업과 매칭시킴

참고 논문 : 산업별 평가를 위한 뉴스 기사 기반 산업 위험 지표 예측
(News article based industry risk index prediction for industry-specific evaluation)

02 비재무 데이터 수집 방안

4. 가정 : 중소기업의 위치적 정보(부지, 월세, 임대료, 전력) 이러한 정보는 휴,폐업과 연관이 되어 있을 것이다.

- **아이디어** : 중소기업들은 재무적 상황이 좋지 않기 때문에 부지, 월세, 임대료등에서 여러가지 문제를 겪을 수 있음 -> 해당 상황이 휴,폐업으로 연결
- **한계점** : 중소 기업과 위치적 정보와의 매칭이 현실적으로 쉽지 않음.

5. 가정 : 설립일자 ~ 최근일자 사이의 환율, 이자율, 원자재비, CPI, 등의 지표들이 기업의 휴, 폐업과 연관이 있을 것이다.

- **아이디어** : 우리나라도 과거에 비해 현재 많은 외국인 노동자분들이 있음
- 따라서 외국인 노동자의 다양성 관리가 기업들의 성과에 영향을 미침
- **참고 논문 :** 외국인노동자의 다양성관리가 기업성과에 미치는 영향 : 비재무적 성과의 매개효과 중심으로

03 향후 진행 방향

1. **가정을 바탕으로 한 비재무 데이터 수집을 우선적으로 진행**
2. **아래의 모델링 기준으로 모델 만들고 해석 진행**
3. **대안신용평가모델로서의 가치 판단 진행**

모델링(Modeling) 진행방향
이진 분류 모델 : 아직 구체적인 데이터셋이 선정되지 않았으나 모델의 성능 비교를 위해 ML/DL을 각각 진행하기로함
평가지표(Metric) : accuracy 외의 AUPRC, f1score 등의 다양한 지표 비교
결과 해석 및 XAI 이용 : 해당 모델이 예측을 진행함에 있어 특정 지표의 기여도 파악하고 이를 해석해볼 예정(SHAP OR Feature importance 사용)

TAHNK YOU