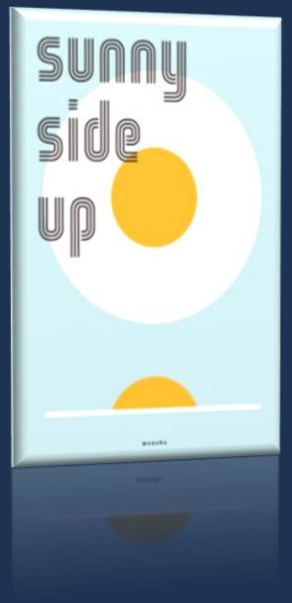
인구특성·지역특성·산업특성을 활용한

중소기업 휴·폐업 예측



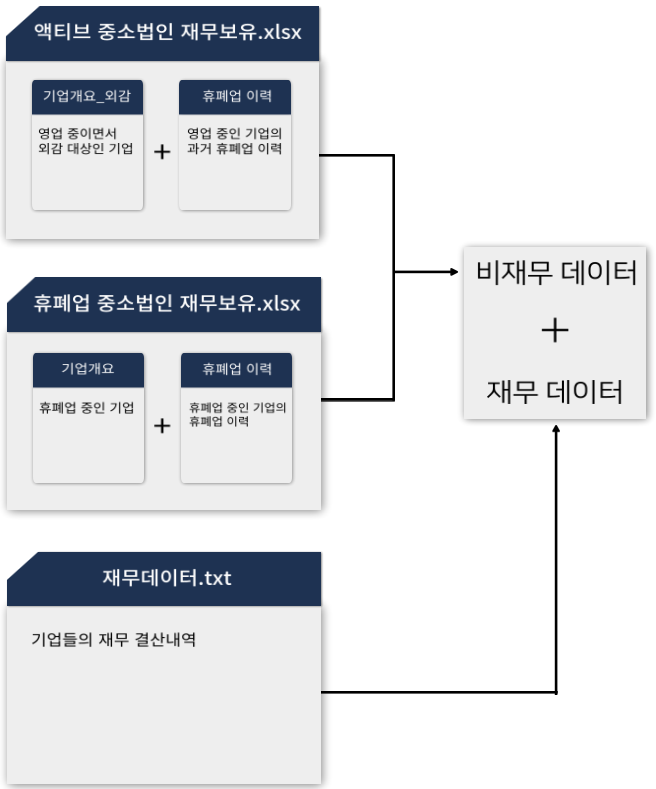
팀 장 : newdoin (신도인)

팀 원 : ollon (공한석)

팀 원 : 김바바나 (강남훈)

INTRO

데이터구조



전처리

재무데이터

- ‘NaN’ 과 같은 결측치는 0으로 보간
- ‘8.88890e+11, 1.00000e+12’와 같은 이상치는 0으로 보정
- 자본금 음수(-) & 자산총계 0인 데이터 제거
- 상위 1%이상, 하위 99% 이하 값은 1%, 99% 경계값으로 윈저라이징(winsorizing)

비재무데이터

- 휴폐업이력의 종료일자가 ‘99991231’인 데이터만 사용
- 동일한 사업자등록번호가 여러 개인 경우, 가장 결산이 많은 사업자등록번호 하나만 사용
- 고유값이 많은 변수 중 홈페이지URL, 사업내용, 대표자명 등 변수 제거

INTRO

모델링

타겟값 정의

- 휴업은 장래 영업활동 재개의사를 가지고 있지만, 신용평가기관입장에서는 주의가 필요한 상태라고 판단함
- 따라서 휴업과 폐업을 하나로 묶어 타겟값으로 정의함

데이터셋

- 전처리한 재무/비재무 데이터를 합친 데이터프레임
- 크기 : (101962, 56)

알고리즘 (LightGBM)

- 최초 LogisticRegression을 사용했으나 성능이 떨어짐
- GBM 모델 중 빠르고 과적합을 방지할 수 있는 LightGBM 알고리즘 선정

결과

모델의 평가지표 (F1-Score)

- 재현율을 중시하되, 정밀도도 고려하는 평가지표로써 모델의 성능을 가장 잘 나타낼 수 있다고 판단함
- 휴폐업 기업을 영업중이라고 판단하는 것이 큰 문제가 되는 예측 특성상 재현율이 중요함

Precision = 0.8445

Recall = 0.6063

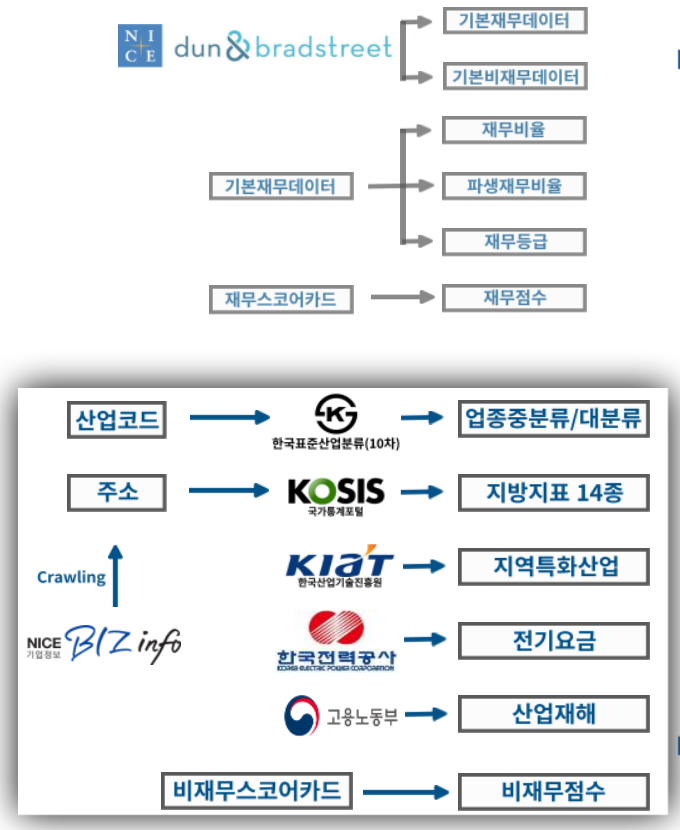
$$F1\ Score = \frac{2 \times precision \times recall}{precision + recall}$$

$$= 0.7058$$

INTRO

Challenge Flow

① 데이터 수집



[데이터 준비 및 수집]

② 변수 선정

통계 검정

T-Test

Chi-Square Test

변수 선정

[자료형에 맞는 통계 검정 실시]

③ 데이터셋 구축

모델링

[학습 및 예측 데이터셋]

④ 학습 및 예측

휴폐업 예측

Logistic Regression
Naive Bayses
Random Forest
LightGBM
XGBoost

[머신러닝 모델링]

CONTENTS

고려사항

1. 데이터는 누구에게나 공개 되어있고 수집이 용이 해야한다.
2. 외부인으로서 개별기업의 내부정보를 얻기 힘들기 때문에 범주를 확장하여 분석한다.
3. 따라서 기업이 속한 지역 또는 인구 특성 분석을 위해 주소, 산업특성을 분석하기 위해 업종분류가 필요하다.

주소

수집 과정·전처리

- 병렬크롤러로나이스비즈인포에서주소데이터수집
- 수집된주소를 '시·도'와 '시·군·구'로 분리

NICE 기업정보 BIZ info

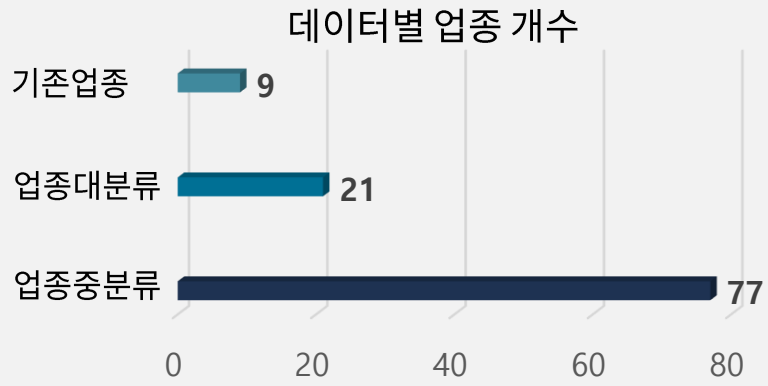
기업명	시·도	시·군·구
한일가스산업	서울특별시	구로구
도영운수	인천광역시	연수구
코팅코리아	전라북도	김제시
…	…	…

업종 분류

수집 과정·전처리

- 나이스비즈인포에서산업코드1의결측치를크롤링하여보간
- KSIC 10차를기준으로업종중분류·업종대분류생성

한국표준산업분류(10차)



인구특성

선 정 이 유

가설

- 지역별 인구구조에 따라 기업경영환경이 달라져 휴폐업에 영향을 미칠 것이라 판단함
- 이를 반영할 수 있는 지표로 고령인구비율, 외국인비율 등을 선정함

관련자료

"고령화로 인건비 부담이 증가하고 생산성이 하락하여 상당수의 중소기업이 경영악화에 직면할 것으로 예상"

-인구고령화가 기업에 미치는 영향(KDB미래전략연구소)

수집 & 분석

수집

KOSIS 국가통계포털

- 통계청의 e-지방지표에서 시도별 인구특성에 해당하는 데이터
- 수집데이터 목록 : 5개

(인구수, 1인가구비율, 고령인구비율, 외국인비율, 가구별소득)

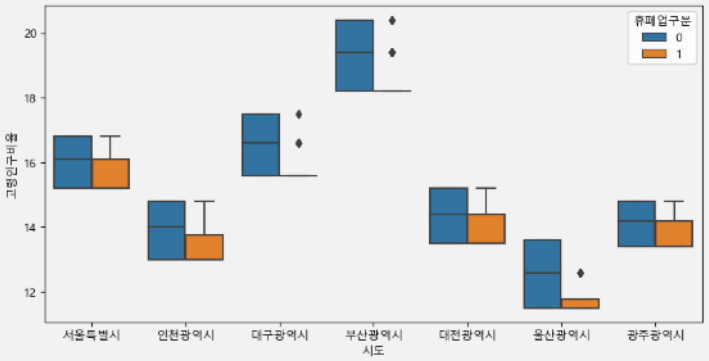
분석

- 지역별로 각 변수에 대해 영업중인 기업과 휴폐업 기업의 분포차이를 확인함

결 과

분석결과

영업중인 기업과 휴폐업 기업의
고령인구비율 분포 차이



- 영업중인 기업과 휴폐업 기업의 고령인구비율 분포차이가 확인됨

지역특성

선 정 이 유

가설

- 지역내경기나지역의발달정도에따라 휴폐업률이상이할것이라고판단함
- 이를반영할수있는지표로광공업생산지수, 재정자립도,경제활동참가율등선정

관련자료

"이들변수들이클수록경기가좋은것을 의미하는데,자영업체의폐업률을낮추는작용을 하는것으로분석되었다"

-국내자영업의폐업률결정요인분석(한국은행)

수집 & 분석

수집

KOSIS 국가통계포털

- 통계청의 e-지방지표에서 시도별 지역특성에 해당하는데이터
- 수집데이터목록 : 9개

(광공업생산지수,재정자립도,경제활동참가율, 실업률, 고용률, 취업자증감률,특허출원개수,도시면적,전기사용량)

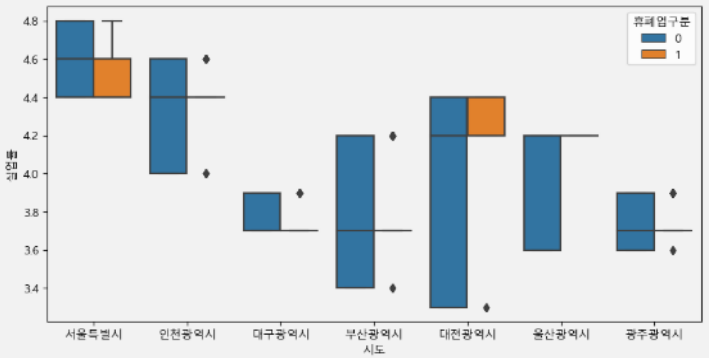
분석

- 지역별로각변수에대해영업중인기업과 휴폐업기업의분포차이를확인함

결 과

분석결과

영업중인 기업과 휴폐업 기업의
실업률 분포 차이



- 영업중인 기업과 휴폐업 기업의 실업률 분포 차이가확인됨

산업특성 - 산업재해 발생여부

선 정 이 유

가설

- 산업재해발생시 사업주는 1년 이상의 징역 및 10억원 이하의 벌금에 처해짐
- 위와 같은 처벌로 인해 경영상 어려움을 겪을 것이라고 판단함

관련자료

"기업에서 발생한 산업재해는 생산차질, 기업 이미지 하락, 노사관계 악화, 노동력 상실 등을 발생시켜 경영성과에 영향을 미칠 수 있다"

-안전보건공단 산업안전보건연구원

수집 & 분석

수집

고용노동부

- 고용노동부에서 공표한 연도별 산업재해 발생 현황자료

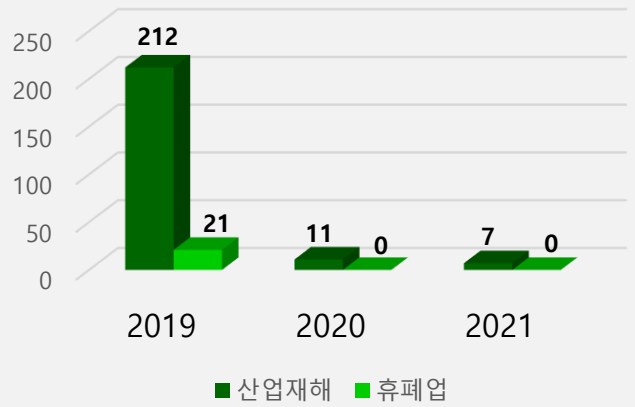
분석

- PDF파일을 XLSX으로 변환한 뒤 산업재해 발생 기업 명단을 클렌징함
- 각 연도별로 기업명과 사업장 주소를 기준으로 산업재해 발생여부를 판단함

결 과

분석결과

연도별 산업재해 발생기업 및 휴폐업 현황



- 산업재해가 발생한 230개 기업 중 휴폐업 기업은 21개로 9.1%에 해당함

(산업재해 미발생기업의 휴폐업률 5.1%)

산업특성 - 지역특화산업 해당여부

선 정 이 유

가설

- 지역특화산업에해당하는기업의경우 기술개발과사업화를위한혜택과연구 개발비를지원받음
- 따라서해당하는기업은경제적이점을 가지고기업활성화에유리함

관련자료

"비수도권의 시도별 육성중인 지역특화산업과 관련있는 기업에 기술개발 및 사업화 자금 1조 4000억원 투자"

-제21차비경제중앙대책본부회의中

수집 & 분석

수집

KIAT 한국산업기술진흥원

- 한국기술진흥원(KIAT)에서 공표한연도별 지역특화산업보고서

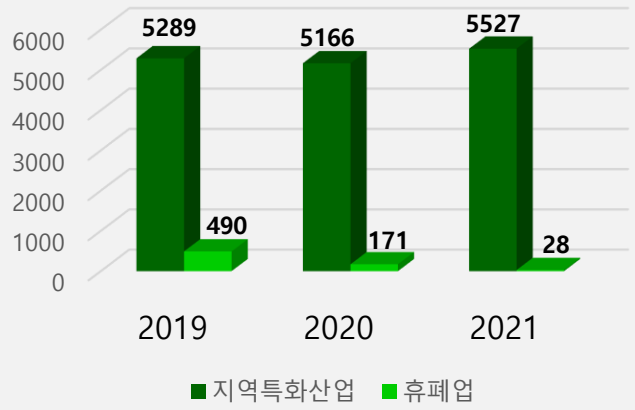
분석

- 수도권의 경우, 특화산업이지정되어있지 않기때문에많이분포한산업군을 특화산업으로간주함
- 특화산업에관련된 산업코드를가진 기업은1, 나머지는0으로지정하여파생변수 생성

결 과

분석결과

연도별 지역특화산업 해당 기업 및 휴폐업 현황



- 지역특화산업에 해당하는 15,982개 기업 중휴폐업기업은689개로4.4%에해당함

(지역특화산업 미해당 기업의 휴폐업율은 5.4%)

산업특성 - 전기요금

선 정 이 유

가설

- 산업별 전기사용비중과 전기요금 혜택이 다름
- 전기사용량이 많고 전기요금이 비싼 산업일수록 경영성과가 좋지 않을 것이라 판단함

관련자료

"산업용 전기요금 내달 오른다…기업들, 경쟁력 훼손 불보듯…한숨"

"전기요금 인상에 부담 커진 산업계…1조4천억 원 요금 증가 예상"

-서울경제/한국경제

수집 & 분석

수집

한국전력공사 KOREA ELECTRIC POWER CORPORATION

- 공공데이터포털의 한국전력공사 산업분류별 전력사용량 및 전기요금 데이터셋
- 주요전력통계 항목에서 기간을 선택하여 수집

분석

- 각 연도별로 지역, 산업을 그룹화하여 분석
- 전기요금은 매년 상승하기 때문에 연도별로 min-max 정규화를 진행하여 산업별 전기요금 차이를 반영함

결 과

분석결과

정규화 범위	휴폐업 비율
0.0 ~ 0.2	0%
0.2 ~ 0.4	4.5%
0.4 ~ 0.6	4.7%
0.6 ~ 0.8	5.0%
0.8 ~ 1.0	5.7%

- 정규화 값이 높을수록 전기요금이 비싸다는 의미임
- 전기요금이 비쌀 수록 휴폐업비율이 높아지는 분포를 보였음

비재무점수

선 정 이 유

가설

- 전문신용평가기관에서는 신용평가등급을 의사결정의참고지표로활용하고있음
- 따라서 비재무 데이터를 하나로 점수화 했을 때휴폐업예측에도움이될수있다고판단

관련자료

"신용평가등급이란 특정기업에 대한 제반 환경을 평가함으로써 일정한 기호를 이용하여 신용도를 등급화하는 제도입니다"

- 나이스신용평가 신용평가의의

수집 & 분석

수집

- 앞서 생성한인구특성,지역특성,산업특성 데이터활용

분석

- 각변수들을구간화한뒤구간별가중치산출
- 산출한가중치를로지스틱회귀모형에 학습시켜서나온회귀계수를통해구간별 점수를생성하여스코어카드 구성
- 비재무 데이터들을 대입, 총점을 계산하여 파생변수를생성

결 과

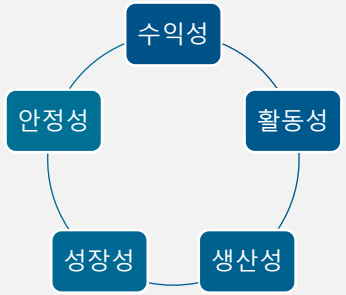
분석결과

사업자등록번호	기업명	점수
1018100340	대일건설	556.663
4028118651	신원건설	530.225
1018101126	우성개발	462.994
…	…	…

- 스코어카드기준기업별비재무점수생성
- 비재무 점수의 구간별 휴폐업률에 차이가 있음을확인

(1) 비율

- 기업의 규모차이를 고려한 재무비율을 분석에사용하고자 함



- 기업부실화 예측 관련 논문을 참고하여 부실화와 관련된 재무비율 50개를 선정함

생성된 변수
30개

(2) 파생비율

- 경영성과나 재무상태의 변화를 반영하기 위해 '재무상태변동성' 을 생성

재무상태변동성(차이) = 당기 재무비율 - 전기 재무비율
재무상태변동성(비율) = 당기 재무비율 / 전기 재무비율

- 산업내 우량기업 또는 산업평균의 재무상태와 비교하기 위해 '산업수준상태변동성'을 생성

산업수준상태변동성(top10 차이) = 기본재무비율 - 대분류별 상위 10개 기업 총자산수익률 평균
산업수준상태변동성(top10 비율) = 기본재무비율 / 대분류별 상위 10개 기업 총자산수익률 평균
산업수준상태변동성(all 차이) = 기본재무비율 - 대분류별 산업 전체 총자산수익률 평균
산업수준상태변동성(all 비율) = 기본재무비율 / 대분류별 산업 전체 총자산수익률 평균

생성된 변수
180개

(3) 등급

- 다양한 관점으로 재무데이터를 분석함과 동시에 사용하는 재무변수를 줄이고자 '재무등급'을 생성

- 재무비율 데이터를 바탕으로 상위 n % 에 해당하는 개수로 등급을 나눔

선정 기준	등급
상위 n% 이상인 컬럼 24개 이상	SSS
상위 n% 이상인 컬럼 22개 이상	SS
상위 n% 이상인 컬럼 20개 이상	S
…	…

생성된 변수
1개

(4) 점수

- 개인신용점수 산출에 사용되는 기법을 활용하여 재무데이터를 점수화하고자 함

- 비재무점수를 산출한 방법과 동일한 방법으로 '재무점수'를 생성

사업자등록번호	기업명	점수
1018100340	대일건설	475.348
4028118651	신원건설	497.049
1018101126	우성개발	468.530
…	…	…

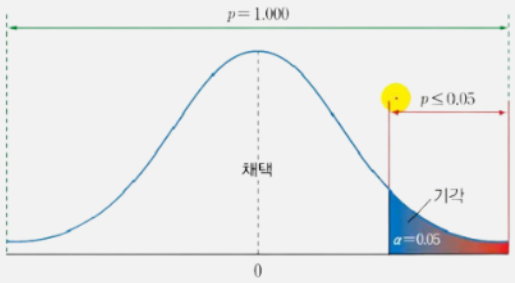
생성된 변수
1개

통계검정을 통한 사용변수 선정

STEP 1. 변수 목록

- 제공받은 데이터 변수 - 56개
- 인구특성 변수 - 4개
- 지역특성 변수 - 10개
- 산업특성 변수 - 3개
- 재무데이터 파생변수 - 212개

STEP 2. 통계 검정



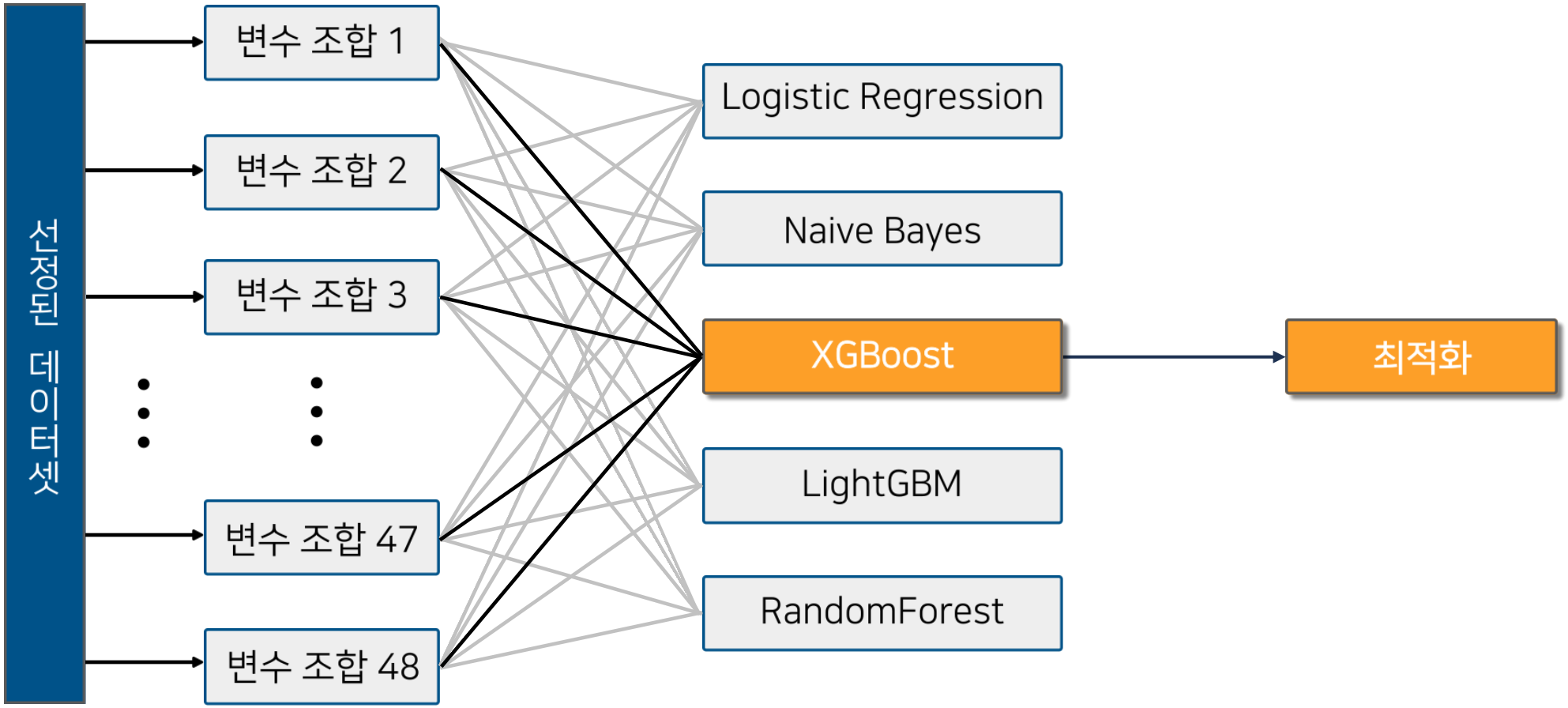
- 범주형 데이터
Two Sample Chi-Squared-Test

- 연속형 데이터
Two Sample T-Test

STEP 3. 사용변수

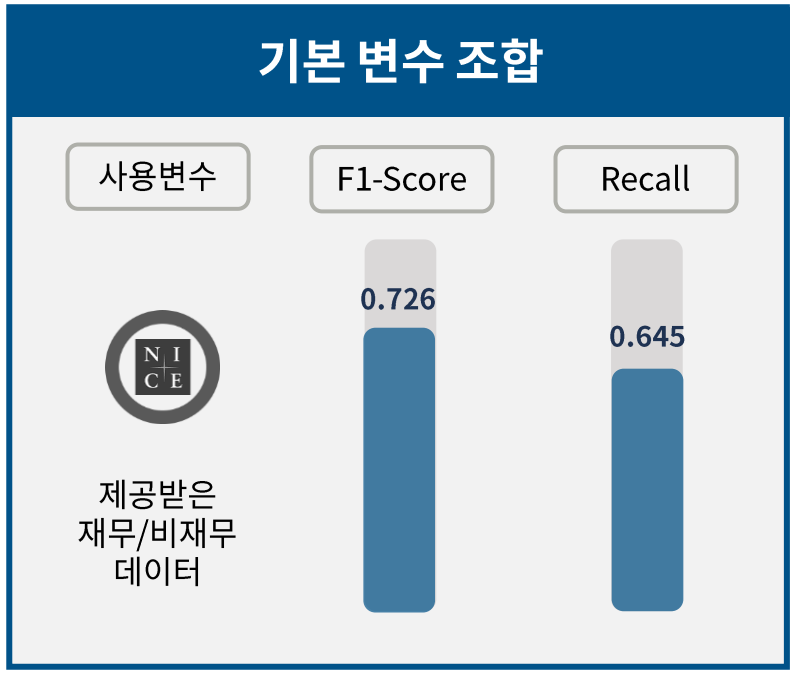
유동자산 / 매출채권 / 비유동자산 … 본지점구분	28개
1인가구비율 / 고령인구비율 / 외국인비율	3개
고용률 / 경제활동참가율 / 재정자립도 / 광공업생산지수 / 도시면적 / 가구별소득	6개
업중중분류 / 업종대분류 / 전기요금	3개
총자산영업이익률 / 유동비율 / … 재무등급 / 재무점수	12개

변수 조합

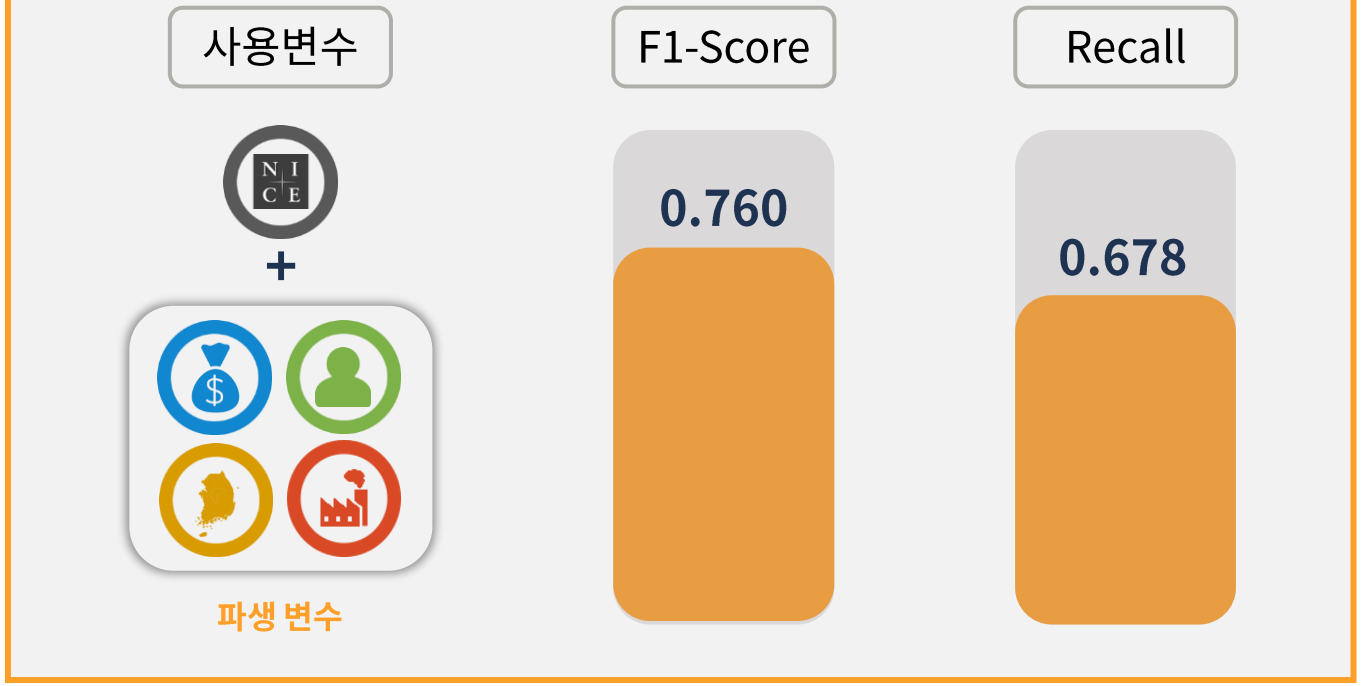

선정된 데이터셋
변수 조합 1
변수 조합 2
변수 조합 3
변수 조합 47
변수 조합 48
Logistic Regression
Naive Bayes
XGBoost
LightGBM
RandomForest
최적화

XGBoost 최적화 성능비교

기본 변수 조합



기본 변수 + 파생 변수 조합



- 제공받은 재무데이터와 비재무데이터만 사용했을 때 0.726의 성능을 보임
- 인구특성, 지역특성, 산업특성을 결합했을 때의 F1-Score는 0.760으로 약 5% 향상됨
- 특정 데이터가 큰 비중을 차지하지 않고 적절히 조화를 이룰 때 성능이 가장 좋았음

요약 & 결론

요 약

- 제공받은 데이터 및 파생변수에 대해 EDA와 통계검정을 실시하여 사용할 변수를 선정함
- 최적 변수조합을 찾아 분석한 결과, 파생변수를 추가했을 때 성능이 향상됨

결 론

" 통계청, 공공데이터 포털 등에서 무료로 쉽게 수집할 수 있는 데이터도 적절히 결합되면
휴폐업 예측모델의 성능향상에 도움이 될 수 있었다는 점에서 분석의 의의가 있음 "

한계점 & 보완

한 계 점

- 인구특성, 지역특성을 나타내는 지표가 '서울특별시, 경기도' 와 같이 큰 범주로 제공되었음
- 이는 개별기업이 소재하고 있는 곳의 인구, 지역적 특성을 세부적으로 반영하지 못함

보 완

- 통계청에서는 2021년부터 '서울특별시 OO구, 경기도 OO시'와 같은 시/군/구별 지방지표를 제공함
- 따라서, 2021년 이후 데이터에 대해서 예측모델을 구축할 시 세분화된 범주를 활용하여 분석할 필요가 있음

Q & A

감사합니다

부록 | 모델링 – 변수조합별 결과

<표 11> 알고리즘별 성능

α : T-Test 재무데이터 + Chi-Squared-Test 비재무데이터
β : T-Test 재무비율, γ : T-Test 파생재무비율
δ : 재무등급, ε : 지방지표, ζ : 재무점수, η : 비재무점수

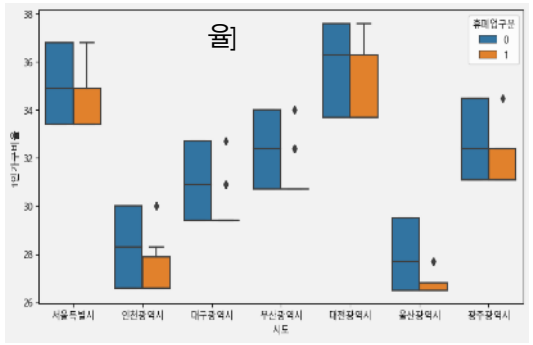
머신러닝 알고리즘	구분	사용 변수 조합	예측 성능	
			Recall	F1 Score
LightGBM	베이스라인	기본 재무/비재무데이터	0.6331	0.7264
	1st	$\alpha + \beta + \delta$	0.6344	0.7280
	2nd	$\alpha + \beta + \gamma$	0.6291	0.7263
	Worst	$\alpha + \beta + \gamma + \zeta$	0.6088	0.6960
XGBoost	베이스라인	기본 재무/비재무데이터	0.6411	0.7322
	1st	$\alpha + \beta + \gamma$	0.6464	0.7394
	2nd	$\alpha + \beta$	0.6397	0.7352
	Worst	$\alpha + \beta + \varepsilon + \eta$	0.6142	0.7042
Ramdom Forest	베이스라인	기본 재무/비재무데이터	0.5987	0.7152
	1st	$\alpha + \beta$	0.6000	0.7156
	2nd	α	0.5947	0.7127
	Worst	$\alpha + \beta + \gamma + \zeta + \eta$	0.5676	0.6864
Logistic Regression	베이스라인	기본 재무/비재무데이터	0.0040	0.0078
Naive Bayes	베이스라인	기본 재무/비재무데이터	0.9642	0.1003

<표 12> 하이퍼파라미터 최적화 및 사용 변수

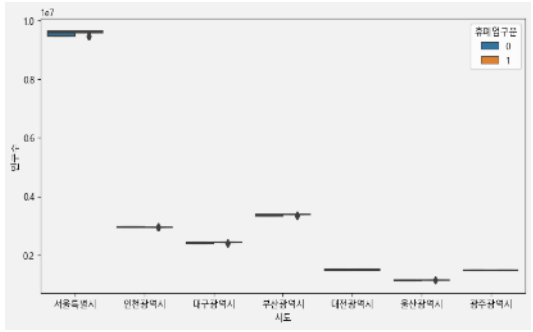
머신러닝 알고리즘	구분	사용 변수 조합	예측 성능	
			Recall	F1 Score
XGBoost	베이스라인	기본 재무/비재무데이터	0.6456	0.7269
	1st	$\alpha + \beta + \gamma + \delta + \epsilon + \zeta$	0.6777	0.7597
	2nd	$\alpha + \beta + \delta + \epsilon + \zeta + \eta$	0.6744	0.7589
	3rd	$\alpha + \beta + \varepsilon$	0.6423	0.7335

부록 | 인구특성 분석 결과

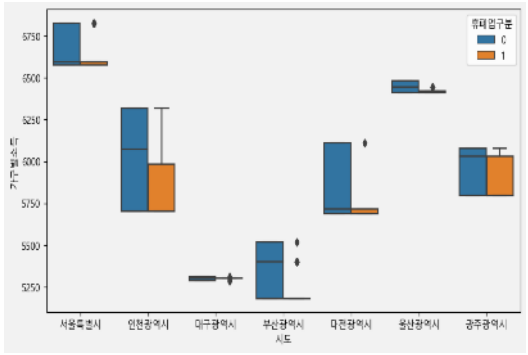
[1인가구비율]



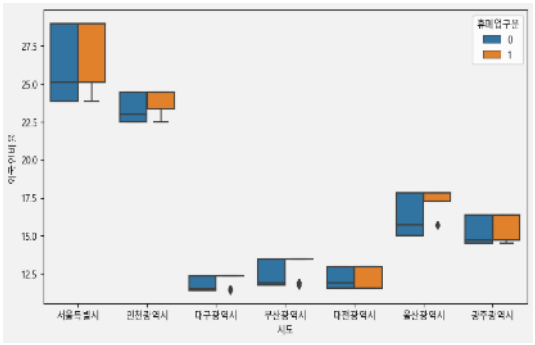
[인구수]



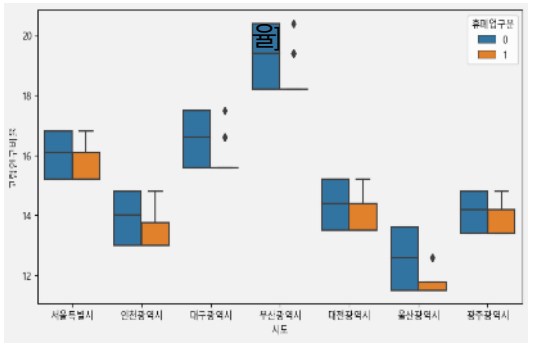
[가구별소득]



[외국인비율]

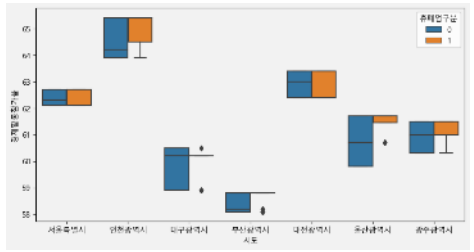


[고령인구비율]

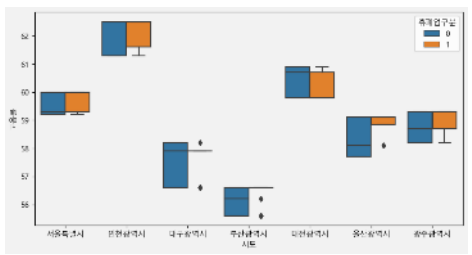


부록 | 지역특성 분석 결과

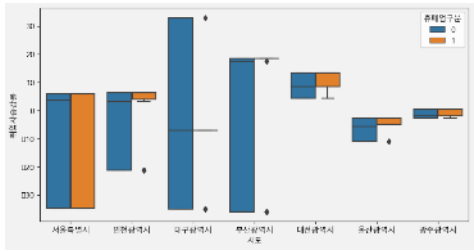
[경제활동참가율]



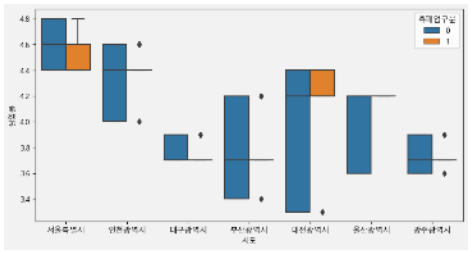
[고용률]



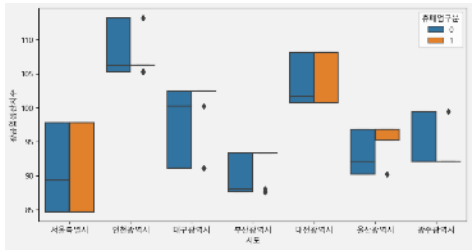
[취업자증감률]



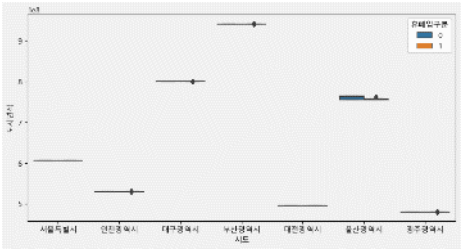
[실업률]



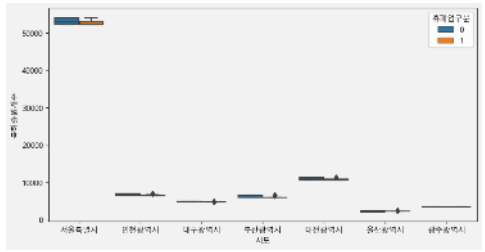
[광공업생산지수]



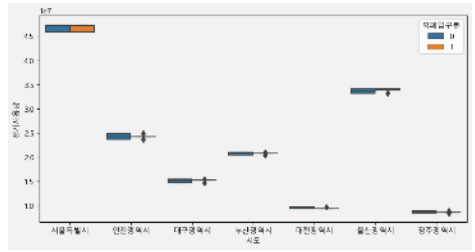
[도시면적]



[특허출원개수]



[전기사용량]



[재정자립도]