공공데이터를 활용한 소상공인 매출등급 예측 모델 제작

우리4LEE

TEAM







이승용



이기찬



이혜준

index

주제 및 개요	1
기본 계획	2
데이터	3
모델링	4
기대효과	5
Q&A	6

주제 및 개요

주제 선정 배경	1-1
현황 및 시장 분석	1-2
타겟 (대상)	1-3
모델 개발 목표	1-4



지정주제 설명

빅데이터 플랫폼 활용 분야 부동산 플랫폼 지정주제



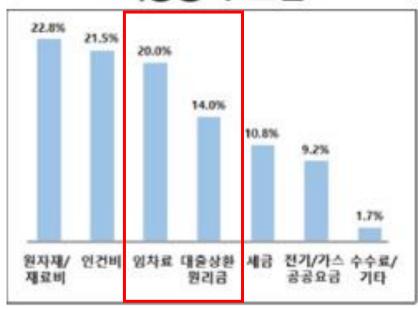
>>> 상업용 부동산 가치 창출을 위한 소상공인 매출등급 예측모형 제작 및 활용방안 제시

선정 배경

- 2023년 초 코로나 사태 종료 이후에도 고물가, 고금리, 고환율 등으로 인해 소상공인 위기감은 계속 증가하여 상업용 부동산 시장이 침체되는 결과 초래
- 소상공인 업종 중 진입장벽이 낮아 창업/폐업 비율이 상대적으로 높은 요식업(음료,음식점업)의 정확한 매출 진단을 통해 폐업 예방 및 상권 활성화 방안 모색 등 상업용 부동산의 가치 창출 도모

'자영업자, 예비 창업자 관점'

〈비용증가 요인〉



출처: https://www.fki.or.kr/main/news/statement_detail.do?bbs_id=00034797&category=ST

Live 중기 지금 중기에선

빚으로 버티는 중소기업·소상공인 늘어나는데

경기가 잘아나지 않는게 눈세 …매울 졸달수 있는 정부장책 잘잘

기사입력 2023-08-29 00:00

김범규기자(bgk7@junggi.co.kr) | 다른기사보기 >

우리 경제가 고물가·고금리·고환율이라는 삼고(三高)에서 벗어나지 못하고 계속 침체기를 겪고 있다. 여기에 아직도 가시지 않은 코로나19의 여파 등 대내외 여건으로 인해 중소기업과 소상공 인은 사업 운영에 차질을 빚을 수밖에 없다. 이에 자금은 부족하고, 구인난은 가속화하고 있는 등 위기 상황에 직면한 중기·소상공인을 위해 각계에서 다양한 지원책을 마련하고 있지만, 현장 에서는 더 실질적인 대책이 필요하다고 목소리를 높인다.

소상공인연합회(KFME) 한 관계자는 중기이코노미와의 통화에서 "현장에서 실질적으로 필요로 하는 것은 매출을 끌어올릴 방안을 마련해주는 것"이라며, "정부 차원에서 좀 더 근본적인 해결책을 위해 노력해야 할 때"라고 말했다.

중소기업·소상공인 대출 부실 가능성…연체율 우려 수준

돌석 : nttps://www.junggi.co.kr/article/articleview.ntmi/no=30955

'임대인, 부동산 투자자 관점'

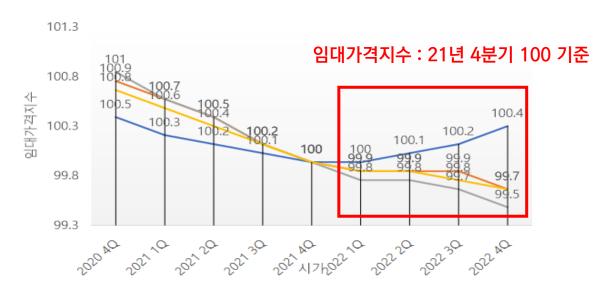
22년 연초 대비 공실률 추세



소규모 상가의 공실률 증가

분기별 임대가격지수 추세

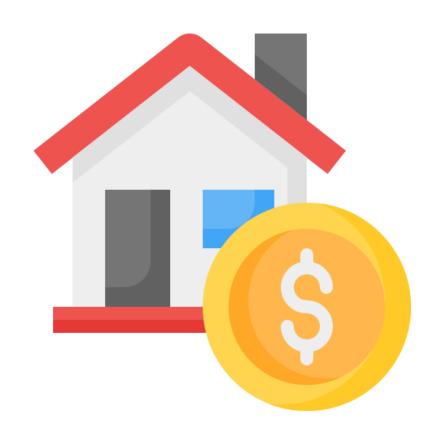
_오피스—중대형 상가—소규모 상가—집합



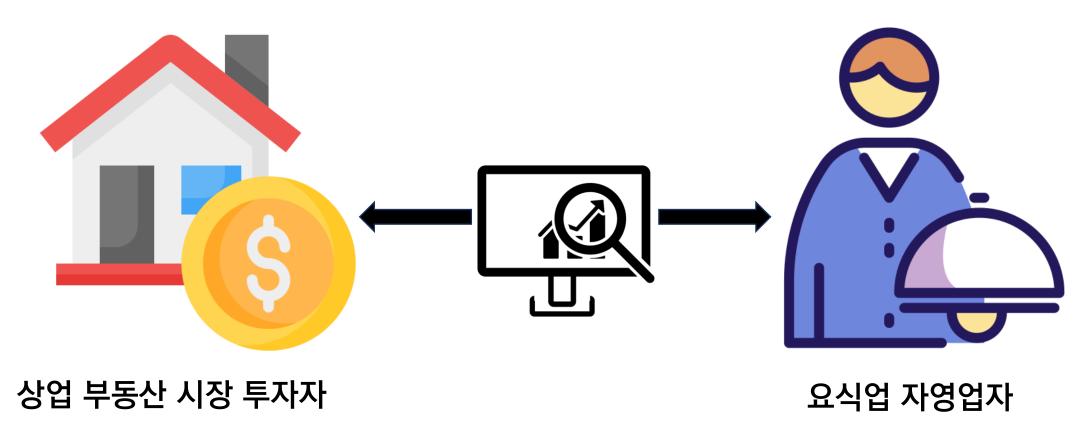
중대형, 소형 상가 분기별 임대 가격지수 감소

출처: https://www.reb.or.kr/reb/main.do

주제선정 배경 - 현황 및 시장 분석- <mark>타겟(대상)</mark> - 모델 개발 목표



상업용 부동산 투자에 관심있는 개인 및 기업



- 정확한 비즈니스 의사결정

- 상권현황 및 지원 아이디어 제시

개발 계획

개발 환경	2-1
연구 방법	2-2



👬 + a b | e a u

QGIS

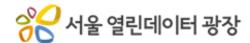




	사용 라이브러리	
∂ python [™]	pandas	NumPy
ver. 3.10.11	ver. 1.5.3	ver. 1.23.5
matpletlib	seaborn	
ver. 3.7.1	ver. 0.12.2	ver. 3.1.0
learn	LightGBM	XGBoost
ver. 1.2.2	ver. 4.1.0	ver. 2.0.0

개발 환경 - 연구 방법













1. 데이터 수집 및 전처리



2. 피쳐엔지니어링 및 EDA



3. BaseLine 구축 및 ML 모델 구축



4. 최적화



5. 파이프라인 구축

데이터

데이터 개요	3-1
활용 데이터	3-2
데이터 전처리	3-3
피처엔지니어링	3-4



데이터 개요 – 활용 데이터 – 데이터 전처리– 피처엔지니어링

특성(Feature)

타겟 레이블(Label)

			-	
데이터 기준년월	데이터 기준년월 필지고유번호		업종코드	매출등급
DATA_CRTR_YM	LT_UNQ_NO	STDG_EMD_CD	INDUSTRY_CD	SLS_GRD
202201	1.11E+18	11110101	A03	4
202201	1.11E+18	11110101	A01	2
202201	1.11E+18	11110101	A11	3

(1156040, 5)

대회 제공 데이터 정보

1. 데이터 정보: 필지단위 소상공인 매출등급 데이터

2. 기간: 2022년 1월 ~ 12월

3. 범위: 서울특별시

4. 매출등급은 자체 구축한 상권분석 알고리즘을 통해 기 분석된 값

〈매출 등급 산정 기준〉

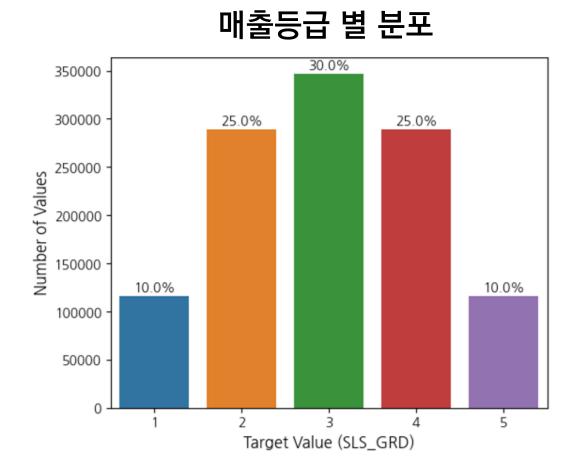
매출 등급	분류	퍼센트(백분위)
1	매우 안정	10% (백분위100~90)
2	안정	25% (백분위 90~65)
3	보통	30% (백분위 65~35)
4	취약	25% (백분위 35~10)
5	매우 취약	10% (백분위 10~0)

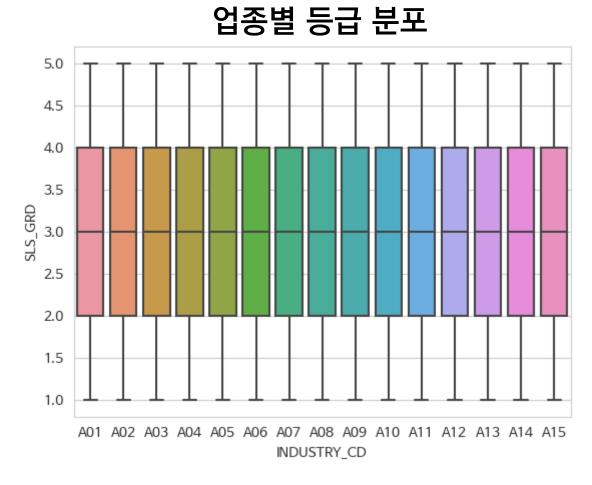
매출등급(Target Data) = 평당월임대료

데이터 개요 - 활용 데이터 - 데이터 전처리- 피처엔지니어링



데이터 개요 – 활용 데이터 – 데이터 전처리– 피처엔지니어링





데이터 개요 – 활용 데이터 – 데이터 전처리– 피처엔지니어링

최종 수집 및 활용 데이터





데이터 개요 - 활용 데이터 - 데이터 전처리- 피처엔지니어링

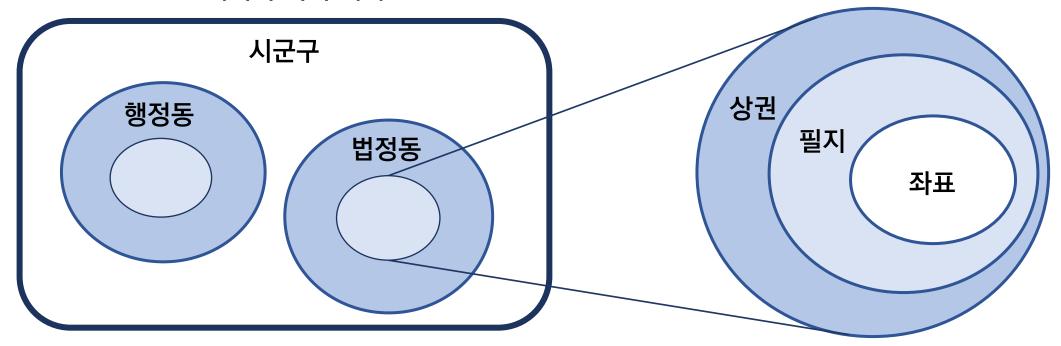


수집 데이터들의 지리적 단위가 상이한 상황



가능한 하위 단위로 통일하여 결합

데이터 위치 체계





데이터 개요 – 활용 데이터 – 데이터 전처리 – 피처 엔지니어링

좌표데이터

1	1.11E+18	서울특별시 종로구 청운동 48-0
2	1.11E+18	서울특별시 종로구 청운동 59-1
3	1.11E+18	서울특별시 종로구 청운동 59-5



지오코딩 작업후 좌표 추가

서울특별시 종로구 청운동 48-0	126.9694096	37.5870427
서울특별시 종로구 청운동 59-1	126.9697607	37.5862078
서울특별시 종로구 청운동 59-5	126.9693531	37.5862133



데이터 개요 - 활용 데이터 - 데이터 전처리 - 피처 엔지니어링

필지단위 데이터

원본데이터 (필지) – 신규 데이터 (필지 정보 포함 | 필지 유추 가능): 필지 단위 결합



Result

	data_strd_ym		pnu	legaldong_c	d induty_cd	snp_price_scor	분기
0	202203	111101010010004	0014	1111010	D1 B01	58.61	1
1	202203	1111010100100480	0000	1111010)1 A01	73.91	1
2	202203	1111010100100500	0031	1111010)1 C05	65.64	1
	DATA_CRTR_YM	LT_UNQ	_NO	STDG_EMD_CD	INDUSTRY_CD	snp_price_scor	분기
0	DATA_CRTR_YM 202203	LT_UNQ		STDG_EMD_CD	INDUSTRY_CD B01	<pre>snp_price_scor 58.61</pre>	분기 1
0			014				
0 1 2	202203	1111010100100040	014 000	11110101	B01	58.61	1

	DATA_CRTR_YM_x	LT_UNQ_N	O STDG_EMD_CD_x	1 NDUSTRY_CD	SLS_GRD	분기	DATA_CRTR_YM_y	STDG_EMD_CD_y	snp_price_scor
0	202201	111101010010003010	0 11110101	A03	4	1	NaN	NaN	NaN
1	202201	111101010010048000	0 11110101	A01	2	1	202203.0	11110101.0	73.91
2	202201	111101010010059000	1 11110101	A11	3	1	202203.0	11110101.0	64.25

3 - 3

데이터 개요 – 활용 데이터 – 데이터 전처리 – 피처 엔지니어링

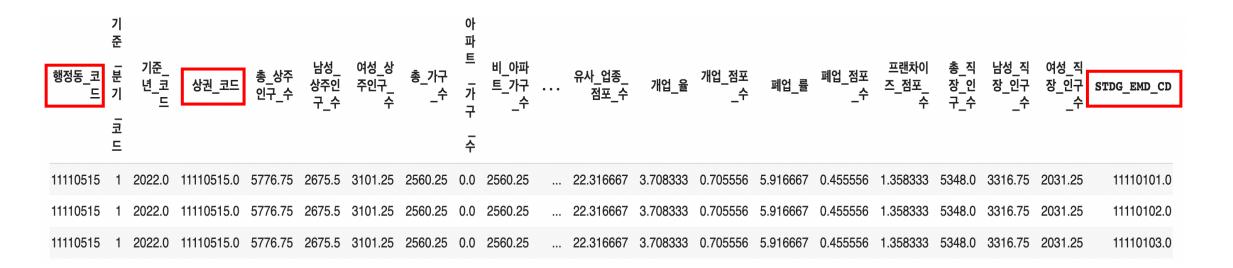
행정동, 법정동 데이터

		P 01 데이터			STEP 0 결합데이				STEP (결합데이		
	0	1	2	상권_코드	3110008	3110009	3110010		0	1	11
기준_년_코드	2.022000e+03	2.022000e+03	2.022000e+03	행정동_코드	11110515	11110550	11110560	시도코드	11	11	11
기준 분기 코드	4.000000e+00	4.00000004.00	4 0000000			_		시도명	서울특별시	서울특별시	서울특별시
					+	_		시군구코드	. 11000	11110	11110
상권_코드	2.111090e+06	2.111089e+06	2.111088e+06		0	1	2	시군구명		종로구	종로구
총_상주인구_수	6.033000e+03	1.153300e+04	1.267700e+04	행정동_코드		1.111052e+07		행정동코드	. 1100000000	1111000000	1111051500
남성_상주인구_수	2.953000e+03	5.759000e+03	6.247000e+03	기준_분기_코드		2.000000e+00		읍면동명			청운효자동
여성_상주인구_수	3 0800000	5 7740000±03	6.430000e+03	기준_년_코드	2.022000e+03	2.022000e+03	2.022000e+03		1100000000	1111000000	
				상권_코드	1.111052e+07	1.111052e+07	1.111052e+07	법정동코드		1111000000	
총_가구_수	2.683000e+03	4.395000e+03	4.928000e+03	총_상주인구_수	5.776750e+03	5.776750e+03	5.776750e+03	동리명	서울특별시	종로구	청운동
				남성_상주인구_수	2.675500e+03	2.675500e+03	2.675500e+03	생성일자	19880423	19880423	20081101
				여성_상주인구_수	3.101250e+03	3.101250e+03	3.101250e+03				
				총_가구_수	2.560250e+03	2.560250e+03	2.560250e+03				

데이터 개요 - 활용 데이터 - 데이터 전처리 - 피처 엔지니어링

행정동, 법정동 데이터

Result/최종데이터



원본데이터 (법정동코드) - 신규 데이터 (행정동, 법정동): 법정동 단위 결합

3 - 3

데이터 개요 - 활용 데이터 - 데이터 전처리 - 피처 엔지니어링

시군구 데이터

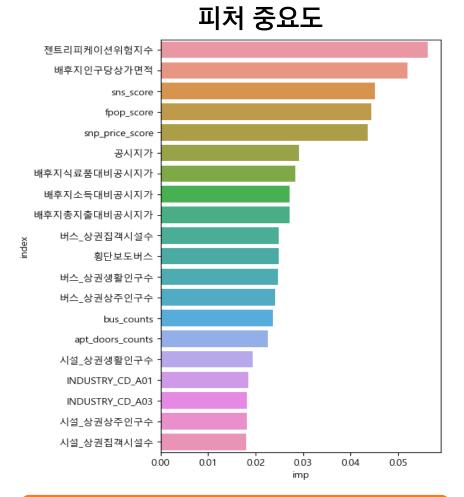
	DATA_CRTR_YM	시군구명 이	산화질소농도(ppm) 오	² 존농도(ppm)	일산화탄소농도	(ppm)	아황산가	스(ppm) 「	미세먼지 (#g / ㎡)	초미세먼지(#g/㎡)	시군구코드			
(202212	강남구	0.028	0.016		0.5		0.003	36.0	17.0	11680			
1	202212	강남구	0.028	0.016		0.5		0.003	36.0	17.0	11680			
	DATA_CRTR_YM	LT_	_UNQ_NO STDG_EMD_C	CD INDUSTRY_C	D SLS_GRD	시군구코 드	시군 구명	이산화질소능 (pp		일산화탄소농도 (ppm)	아황산가스 (ppm)	미세먼지 (<i>坤</i> ʃ/㎡)	초미세먼지 (#3/㎡)	
0	202201	111101010010	00030100 1111010)1 A0	03 4	11110	종로 구	0.0	0.017	0.7	0.004	43.0	32.0	A
1	202201	111101010010	00480000 1111010)1 A0)1 2	11110	종로 구	0.0	0.017	0.7	0.004	43.0	32.0	
2	202201	111101010010	00590001 1111010)1 A1	11 3	11110	종로 구	0.0	0.017	0.7	0.004	43.0	32.0	

원본데이터 (시군구) - 신규 데이터 (시군구 정보 포함): 시군구 단위 결합

3-3 데이터 개요 – 활용 데이터 – 데이터 전처리 – 피처 엔지니어링

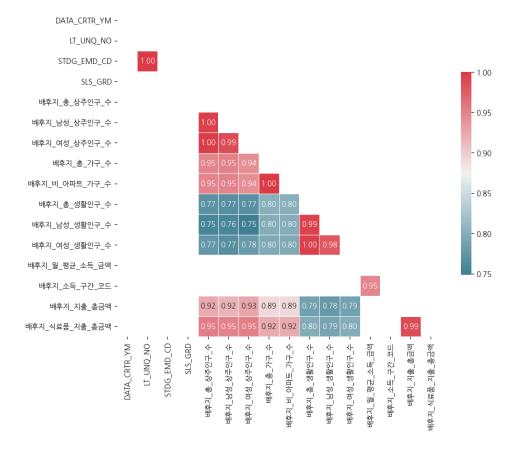


데이터 개요 - 활용 데이터 - 전처리 - 피처 엔지니어링



피처 중요도 0에 가까운 피처 제거

피처 상관관계

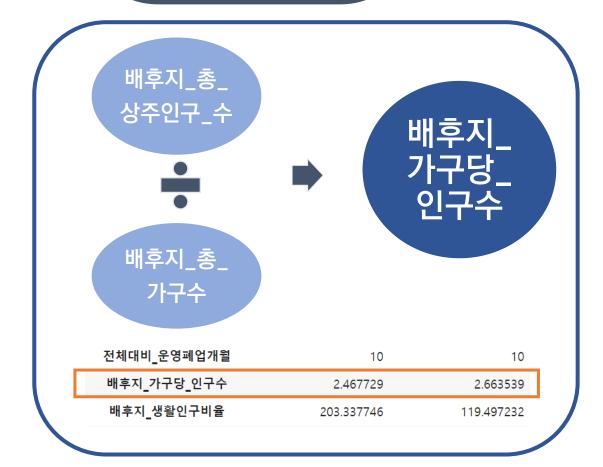


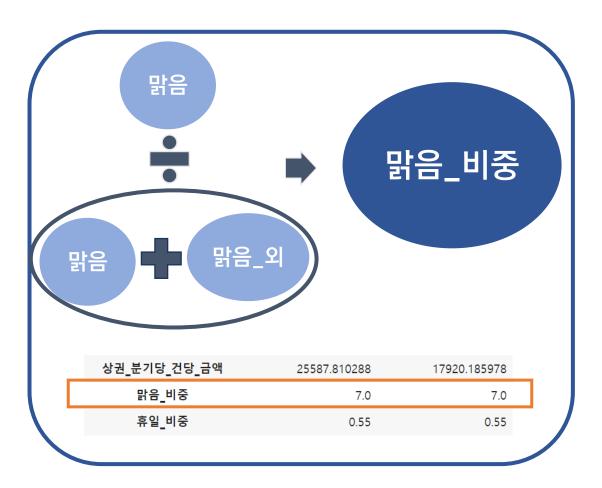
상관관계 높은 칼럼 선별

3 - 4

데이터 개요 - 활용 데이터 - 전처리 - 피처 엔지니어링

파생변수 생성 예시





피처엔지니어링 전 (1156040,88)

파생 27 삭제 25 피처엔지니어링 후 (1156040,90)

모델링

베이스라인 구축	4-1
모델 성능 비교	4-2
모델 개선	4-3
최종 모델	4-4

베이스라인구축 - 성능비교- 모델 개선 - 최종모델

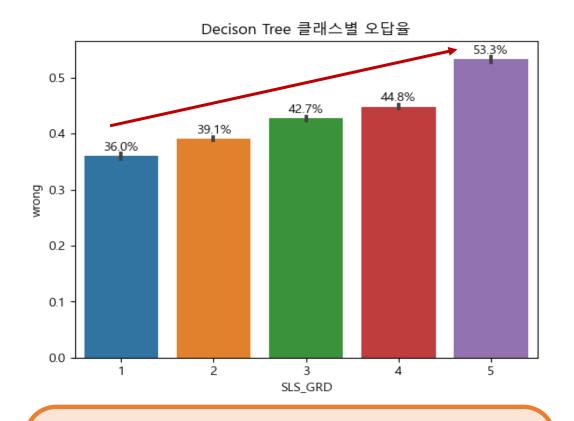
Decision Tree 개념

데이터에 있는 규칙을 학습을 통해서, 자동으로 찾아내고, 트리 기반으로 규칙을 만들어서 분류하는 알고리즘

Decision Tree 선정 이유

학습 과정에서 이상치에 대한 영향을 최소화할 수 있는 트리 모델 중 가장 베이스가 될 수 있는 DecisionTree 선택

구분	베이스라인(DT) 성능			
roc_auc	0.7302			
mse	0.8449			
accuracy	0.5729			

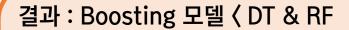


1등급 ~ 5등급 순으로 오답률이 올라가며, 5등급에 대한 학습이 잘 이루어지지 못하고 있음

베이스라인구축 - 성능비교- 모델 개선 - 최종모델

Pycaret 소개 및 선정 이유

기계학습 워크플로우를 자동화하는 오픈소스로서 빠르게 베이스라인 모델보다 개선된 모델을 찾기 위해 사용



이유: 과적합이 잘 되기 쉬운 데이터로 의심됨

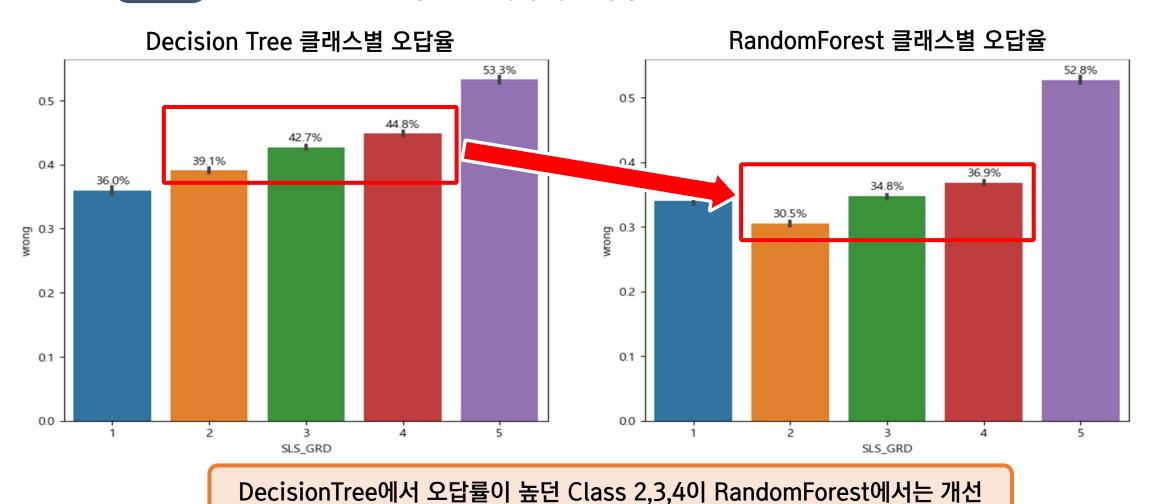
	Model	Accuracy	AUC	Recall	Prec.	F1	Карра	мсс	TT (Sec)
rf	Random Forest Classifier	0.600							
dt	Decision Tree Classifier	0.538	0.696	0.533	0.533	0.533	0.396	0.396	
xgboost	Extreme Gradient Boosting	0.445	0.733	0.461	0.455	0.440	0.252	0.255	
catboost	CatBoost Classifier	0.441	0.731	0.458	0.455	0.434	0.245	0.248	
lightgbm	Light Gradient Boosting Machine	0.413	0.706	0.432	0.426	0.408	0.207	0.210	
gbc	Gradient Boosting Classifier	0.381	0.669	0.402	0.394	0.373	0.160	0.163	2499.648
ada	Ada Boost Classifier	0.362	0.641	0.385	0.370	0.352	0.132	0.135	
knn	K Neighbors Classifier	0.358	0.633	0.359	0.361	0.353	0.151	0.152	
lr	Logistic Regression	0.355	0.639	0.380	0.375	0.343	0.119	0.122	
ridge	Ridge Classifier	0.353	0.000	0.381	0.360	0.341	0.115	0.118	
lda	Linear Discriminant Analysis	0.351	0.635	0.365	0.359	0.336	0.121	0.124	
dummy	Dummy Classifier	0.300	0.500	0.327	0.098	0.151	0.000	0.000	
svm	SVM - Linear Kernel	0.287	0.000	0.291	0.288	0.285	0.064	0.064	
nb	Naive Bayes	0.193	0.576	0.181	0.269	0.143	0.054	0.069	
qda	Quadratic Discriminant Analysis	0.191	0.584	0.170	0.288	0.128	0.056	0.072	

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT(Sec)
RandomForest Classifier	0.600	0.844	0.601	0.595	0.595	0.472	0.473	270.368

4 - 2

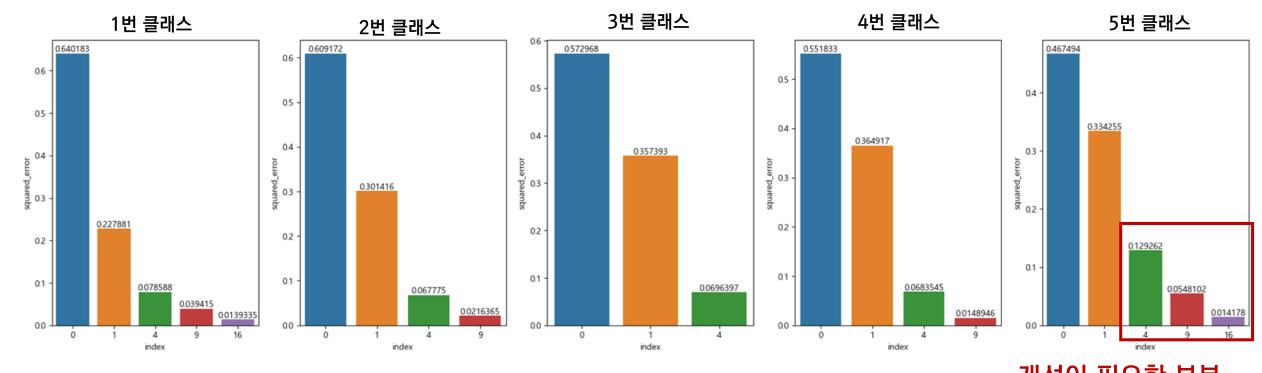
베이스라인구축 - 성능비교- 향상 시도 - 최종모델

4-2-1 RandomForest로 학습한 결과에 대한 해석



32

4-2-1 RandomForest로 학습한 결과에 대한 해석



개선이 필요한 부분 다른 클래스들에 비해 오차 제곱 분포가 너무 퍼져 있음

클래스(정답)별로 오차제곱(Squared Error)의 분포를 비교한 결과 클래스 불균형 문제가 발생하는 것으로 추정

4-3-1 하이퍼파라미터 튜닝 시도

Pycaret, GridsearchCV, Bayesian Optimization을 통해 RF, DT, LGBM, XGBoost 4개의 모델에 대해서 하이퍼파라미터 튜닝 시도

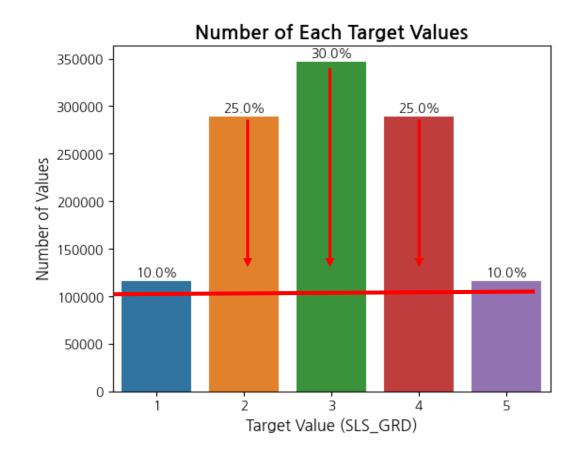
파라미터 튜닝						
Random Forest	Decision Tree	LightGBM	XGBoost			
		:	:			

결과 : 튜닝 결과 성능 향상에 영향이 없음

이유: 과적합 문제로 인해 복잡한 모델일수록 성능이 하락하는 것으로 보임



4-3-2 UnderSampling 시도

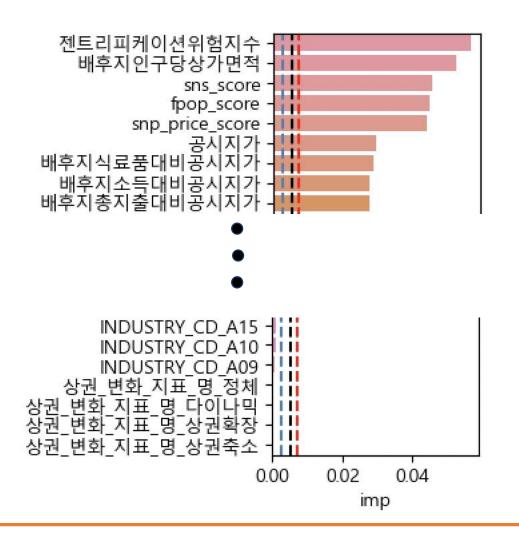


구분	Under Sampling 전	Under Sampling 후
roc_auc	0.8872	0.8446 (감소)
mse	0.6369	1.0654 <mark>(증가)</mark>
accuracy	0.6404	0.5734 (감소)

학습 결과

Random UnderSampling & Nearmiss 시도 => Undersampling 결과 성능 하락

4-3-3 피처 중요도를 통해 RF 학습 과정에서 불필요한 피처 제거 시도



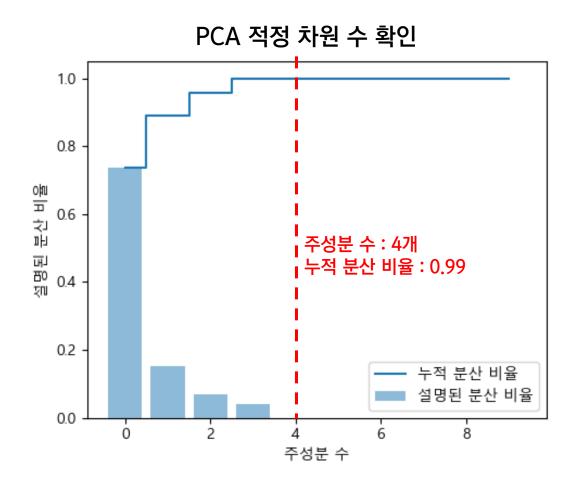
구분	roc-auc	mse			
기본모델	0.8873	0.6361			
0.0025 미만 제거	0.8889	0.6263			
0.005 미만 제거	0.8941	0.6053			
0.007 미만 제거	0.8731	0.7322			

학습 결과

피처중요도 0.005 미만의 피처를 제거하였을 때 성능 향상에 효과가 있었음

베이스라인구축 - 성능비교- 향상 시도 - 최종모델

4-3-4 PCA로 차원을 축소한 후 RandomForest 학습을 진행하는 시도



차원 축소 전/후 RandomForest 학습 결과

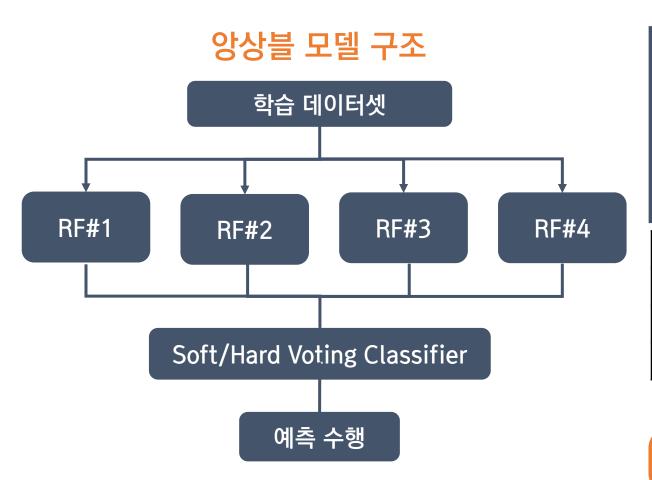
구분	차원 축소 전	차원축소(4개) 후
roc_auc	0.8871	0.6596 (감소)
mse	0.6369	1.5174 <mark>(증가)</mark>
accuracy	0.6403	0.3399 (감소)

학습 결과

차원 축소 결과 성능 하락

베이스라인구축 - 성능비교- 향상 시도 - 최종모델

4-3-5 앙상블 모델 (독립적인 RandomForest 여러 개를 모아서 soft/hard voting 시도)



앙상블 도입 이유

- 약한 분류기들의 결합을 통해 과적합에 조금 더 강건 (Robust)한 성능을 보일 것으로 생각됨
- 파라미터 기본값을 사용하는 4개의 모델을 결합하여 Voting Classifier 생성

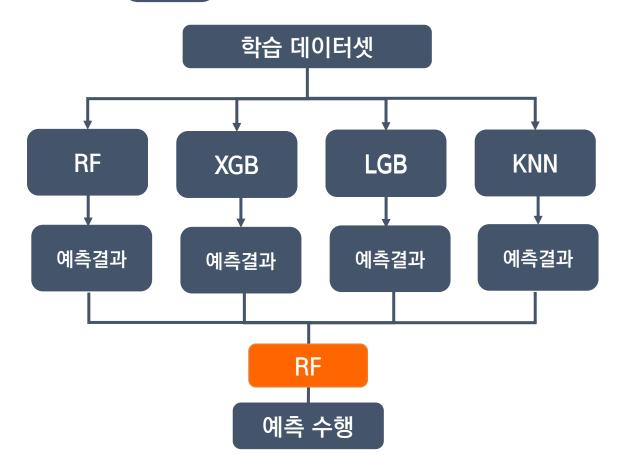
구분	앙상블 모델 성능 결과
roc_auc	0.8865
mse	0.6501
accuracy	0.6360

앙상블 결과 성능에 큰 차이가 없음

4 - 3

베이스라인구축 - 성능비교- 향상 시도 - 최종모델

4-3-6 스태킹 모델 적용



스태킹 도입 이유

여러 모델을 사용한 스태킹 모델은 다양한 특성을 반영하여 일반화 효과를 가지는 동시에 안정적인 성능을 가질것
상이한 특성을 지닌 RF, 트리모델, KNN 모델을 결합한스태킹 모델 구축

학습 결과

결론: MSE: 0.6038 / ROC-AUC: 0.7302



일반화된 성능, 안정적인 MSE/ROC-AUC 최종 모델 선택!

베이스라인구축 - 성능비교- 향상 시도 - 최종모델

베이스 모델(DT) 성능

구분	베이스라인(DT) 성능
roc_auc	0.7302
mse	0.8449
accuracy	0.5729

최종(스태킹) 모델 성능

구분	스태킹 모델 성능	
roc_auc	0.7302	
mse	0.6038	
accuracy	0.6382	

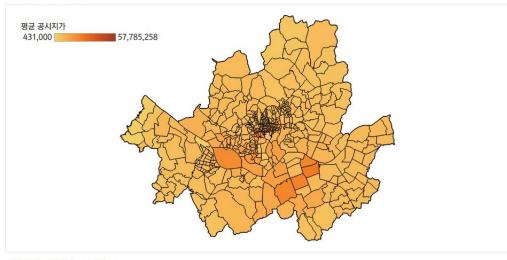
현황분석	5-1
활용방안	5-2

분석결과

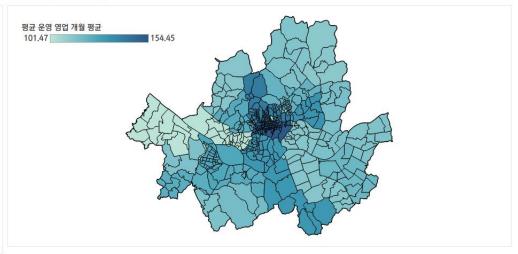
5 - 1

현황분석 – 활용방안

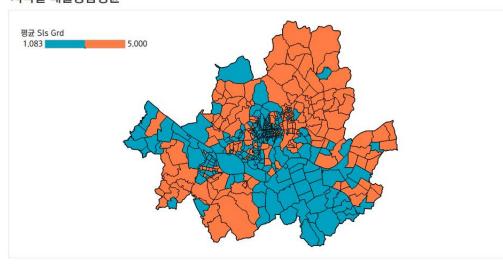




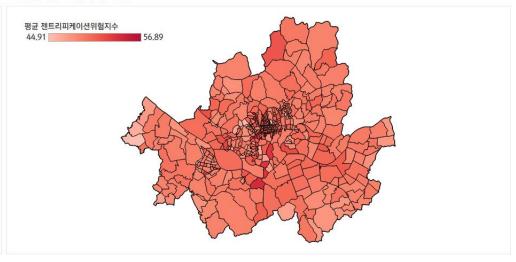
지역별 영업개월수



지역별 매출등급평균



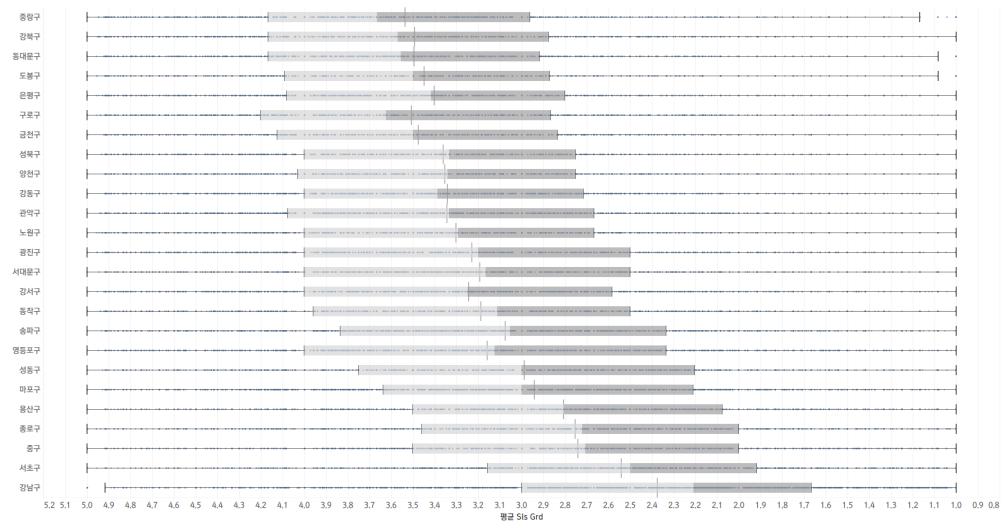
지역별 젠트리피케이션

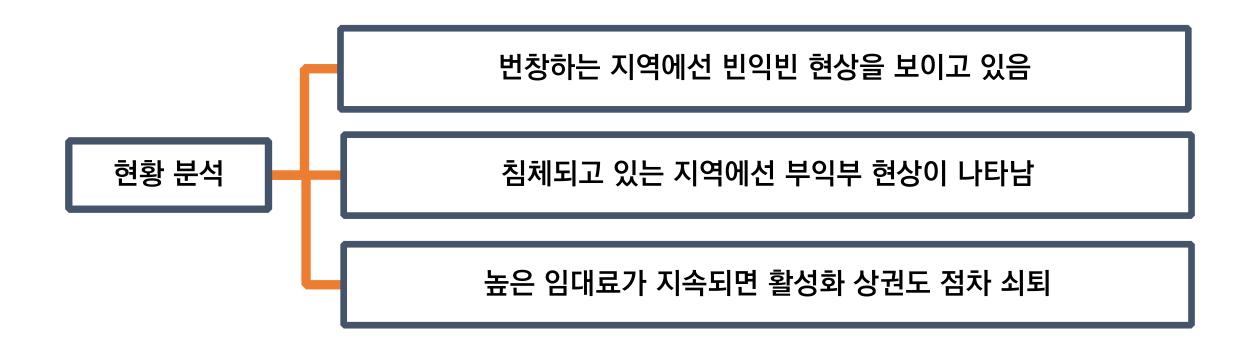




현황분석 – 활용방안

서울시 지역구별 매출등급 분포





현황분석 - 활용방안

5-2-1 어반 블록 클럽

타겟

지역 평균 매출등급. 공실률 🛊 영업평균개월 👚 : 상권 활성화 되어 있지 않고, 낮은 임대료로 진입장벽 👢



활용 케이스 - 열정도

- 6개의 매장으로 시작해 300m 골목에 50여개의 점포 생성
- 임대료 2배가량 상승
- 공시지가의 상승세: 5년만에 55% 상승
- 지역 인프라 개선
- 지역 아이덴티티 및 커뮤니티 강화

현황분석 - 활용방안

5-2-1 이반 블록 클럽

목표

공실률을 낮추고 상권 활성화를 통한 매출 등급 향상

서울'아파트상가' 굴욕…임대료 전년比 25% 급락한 지역도

롯데주류, 용산 힙플 열정도에 '클라우드 골목' 만든다

이서우 기자 | 입력 2019-06-28 16:37

활용 방안

- 지역 인프라 개선 및 안전성 강화
- 지역 아이덴티티 강조
- 커뮤니티 및 사회적 상호작용 강화
- 기업과 협력해 비즈니스 모델 개발

현황분석 - 활용방안

5-2-2 아트테리어 지원

타겟





아트테리어 사업 시행 (전)

아트테리어 사업 시행 (후)

SHAVING & CUT BARBER SHOP

활용 케이스 - 구로문화재단 아트테리어 지원

- 61% 의 참여가게가 본 사업이 지역 소상공인에게 매우 도움이 될것이라고 대답함
- 56.1%의 참여가게가 이 사업을 통해 우리가게의 매출 증가 효과가 나타날것이라고 대답함
- 79.7%의 지역주민이 결과물이 참여가게에 도움이 된다고 응답함
- 소상공인 전문업체 대비 약 70% 비용절감

5 - 2

현황분석 - 활용방안

5-2-1 아트테리어 지원

목표

평균 영업기간을 늘리고 매출 등급 향상



활용 방안

- 다양한 예술과 디자인 요소를 상권에 통합해 고객경험
- 을 향상시키고 지역매력을 높이는 것을 목표
- 주기적인 아트 이벤트 개최
- 가게 내 아트 갤러리
- 지역 아티스트와 협력

Q & A

우리4LEE