Analysis of Sales in Alley Market in Seoul

서울시 골목상권 매출액 분석

- 클러스터링 방법에 따른 설명력 변화

김건우 박동재 장상현 장우빈 허은정

Analysis of Sales in Alley Market in Seoul

개요

주제	서울시 골목상권 매출액 분석
활용 데이터	년도, 상권, 업종별 매출액 (서울시 45개 생활밀집업종), 시간, 성별, 직장인구, 유동인구, 집객시설 수, 사업체 수, 개폐업률, 배후지 평균 소득, 아파트 평균 시가 (서울 열린데이터광장)
문제 정의	상권, 업종 코드 범주화 (클러스터링) 타겟 변수인 매출액 예측 (회귀 모델링)
분석 과정	분석 배경 -> 데이터 전처리 & 탐색 -> 클러스터링 -> 모델링 -> 결론
분석 환경	Python 언어 사용, Jupyter Notebook 환경에서 작업
활용 패키지	numpy, pandas 연산, 데이터 조작 statsmodels, scikit-learn 모델링 matplotlib.pyplot, seaborn 시각화

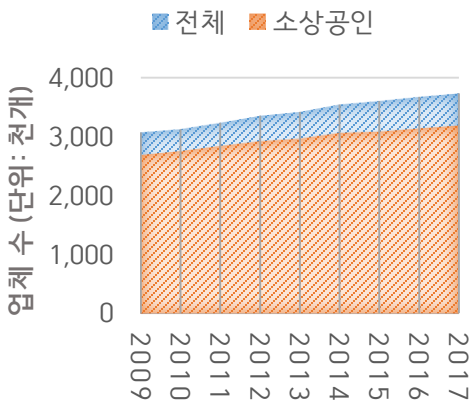
분석 배경

<환경 분석>

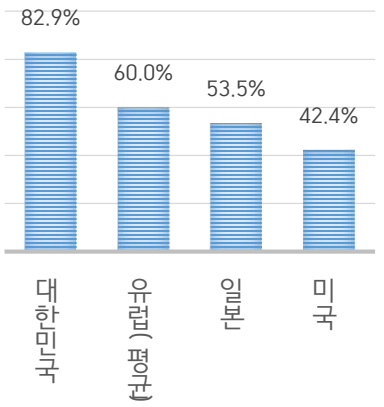
전체 사업체 대비 소상공인 비율 85.3%
(자료 1. 출처: 통계청)

자영업자, 소상공인 비중이
해외 선진국보다 훨씬 높음
(자료 2. 출처: OECD)

(자료 1)
소상공인 사업체수 현황



(자료 2)
자영업자, 소상공인 비중



"소상공인 절반 5년 내 망한다" (뉴스1 2019. 5. 30)
"소상공인, 2곳 중 1곳은 '빚'… 평균 부채 1억8100만원" (뉴시스 2019. 12. 27)

- 소상공인은 진입장벽이 낮은 골목상권 창업 중심
- 발달상권에 비해 골목상권의 낮은 생존율
- 개별 소상공인 대상 맞춤형 지원 정책 부재
- 신규 소상공인 창업 지표 부재

⇨ 골목상권 매출액 분석 필요

분석 배경

<선행 연구>

선행연구	문제 정의	결론	한계
빅데이터 분석을 통한 서울시 골목상권 분석 (2017)	업종, 지역구 기준으로 상권 정의, 대표 특성 파악(클러스터링, 회귀)	• 업종, 구별 군집 구성요소 확인 • 군집별 특징, 매출상관요인 분석	• 다중공선성 발생 • 데이터의 양적인 한계
GWR을 이용한 고객 특성별 골목상권 매출액 영향 연구 (2018)	골목상권 매출액 결정 고객 특성 분석 (OLS, GWR-지리가중회귀)	• GWR이 OLS 회귀분석보다 우수 • 골목상권별 매출액 영향 요인 식별	• 성별과 연령에 한정된 분석
서울시 골목상권 매출액에 영향을 미치는 요인에 관한 연구 (2019)	서울시 골목상권 매출액 결정 요인을 상권, 배후지, 공간구조 등으로 구분하여 규명 (회귀)	• 골목상권이 지리적 입지여건에 따라 다른 특성 • 매출상관요인 분석	• 점포 단위가 아닌 상권 단위 분석 • 업종 고려 X • 상권분석이 구체화되지 못함

<목표 설정>

1) 선행 연구 한계 극복, 발전된 모델 제안

→ **지리 변수 + 업종 변수 추가**

→ **군집화(Clustering)를 통해 예측력 강화**

2) 5년(2015~2019) 간 축적된 데이터셋 활용

10개 테이블, 1144개 컬럼, 약 39만 건의 데이터

→ **기존 연구의 데이터 양적 한계 극복**

3) **더 높은 R-squared 값을 갖는 모델 도출**

데이터 전처리 & 탐색

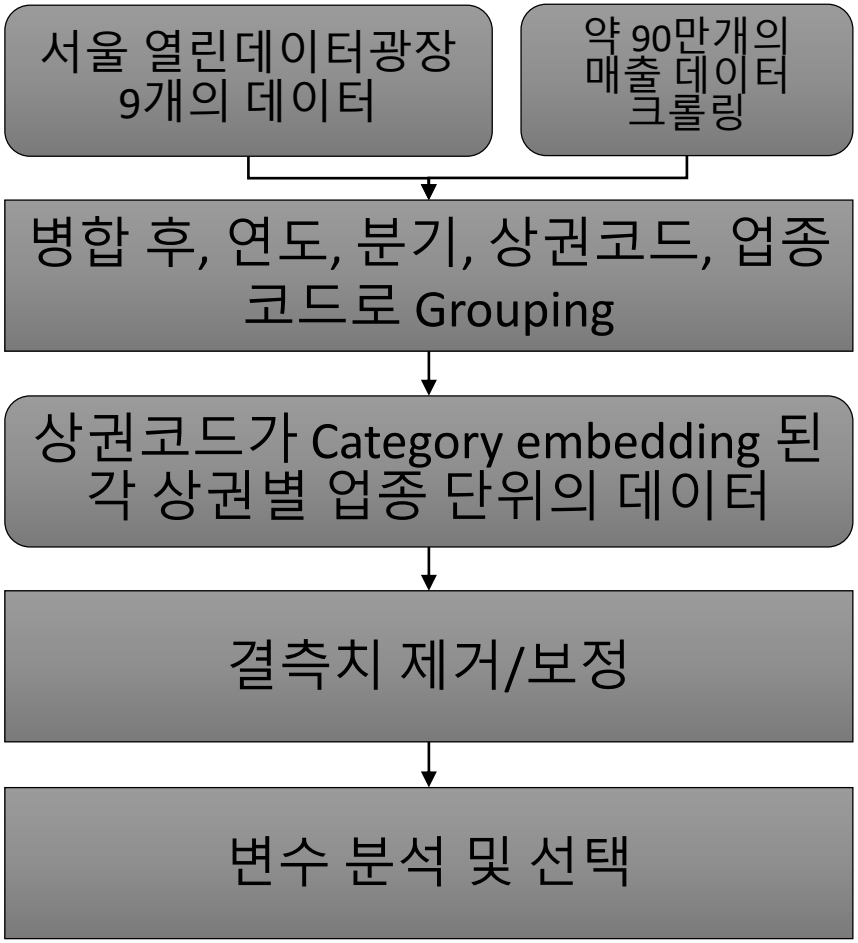
골목상권의 사전적 정의

골목점포 정의

① 생활밀접업종을 포함한 점포
② 발달상권에 포함되지 않는 점포
③ 배후지가 주거밀집 지역에 포함되는 점포
④ 전통시장에 포함되지 않는 점포
⑤ 길에 위치한 점포

골목상권 정의

① 일정 점포 수 이상의 상권
② 골목점포의 밀집도가 높은 상권



데이터 전처리 & 탐색

클러스터링

1 상권코드(1007개) 더미 변수 포함
2 시군구코드(25개) 더미 변수 포함
3 업종코드 제외한 수치형 변수에 k-means 알고리즘 적용한(3개 군집) 더미 변수 포함
4 t-sne 활용한 차원 축소,

☐ R-squared: 0.443

☐ 다중공선성 존재
→ 모델을 명확히 설명하기 어렵다

☐ 1007개 데이터를 모두 사용하니
데이터의 차원이 과도하게 높아짐

☐ 1007개를 줄여보기로 결정

☐ 1007개의 골목상권을
지리적 특성 '시군구'로 통합

☐ 다중공선성은 제거됨

☐ R-squared: 0.297
→ 너무 낮은 값이 도출됨

☐ 다중공선성이 다시 나타남

☐ R-squared: 0.406

☐ 데이터의 설명력이 부족

☐ 다중공선성은 사라짐

☐ R-squared: 0.406

☐ 여전히 데이터의 설명력이 부족

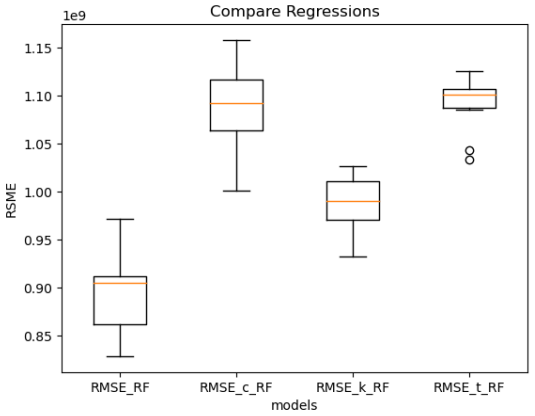
모델링

모델링

모델링

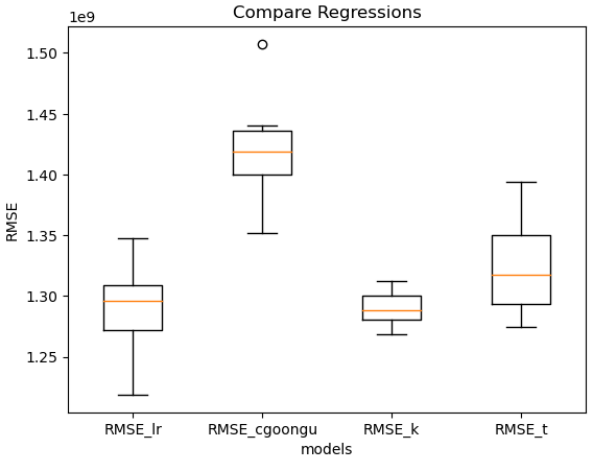
모델링

2) Random Forest



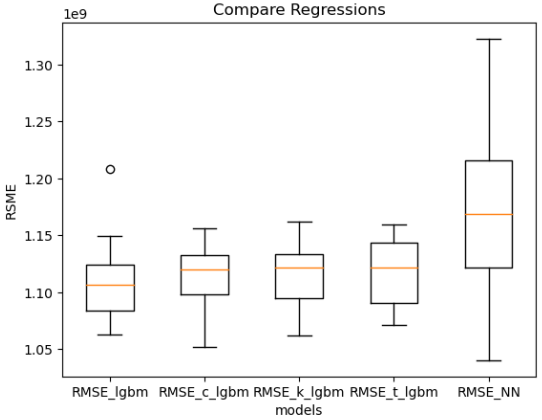
선형회귀 모델보다 낮은 오차값
약 8억~11억
가장 낮은 RMSE 값은 8억대

1) Linear regression



전반적으로 12억이 넘는 수치

3) Light GBM / Neural Network



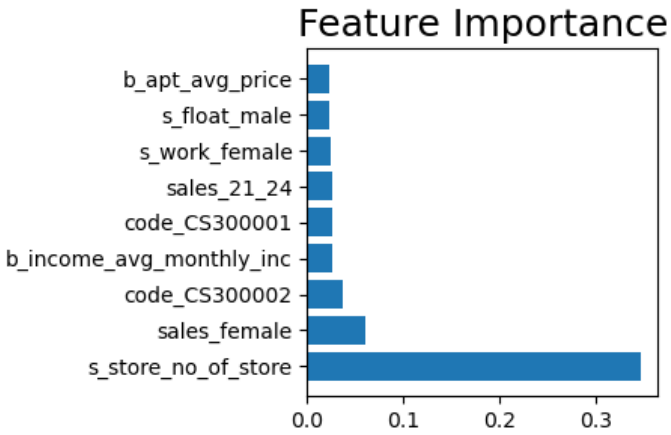
Light GBM은 11억대
Neural Network는 10억~13억대

예측률 0.7, 오차값 8억
Random Forest 분석이 가장 유의미한 결과 도출

결론

<매출액 결정 주요 변수>

1) 주변 매장 수 ⇨ 양의
2) 여성 매출액 상관관계
3) 저녁 9시 ~ 다음날 0시
4) 편의점, 슈퍼마켓 업종
5) 여성 직장인, 남성 유동 인구



<한계>

1) 클러스터링에 K-means 알고리즘만을 적용
   -> 다중공선성 제거할 추가 방법론 탐색 필요
2) 인공신경망 모델의 성능 개선 실패
   -> Hyperparameter 조정하여 개선 필요

- 끝 -

전처리 방법 / EDA 과정

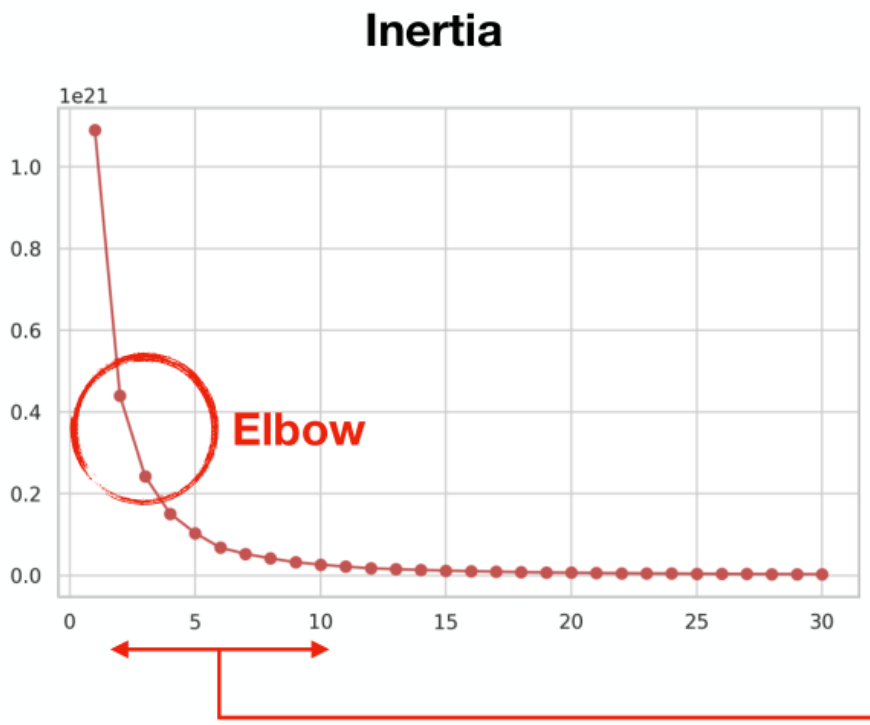
종속 변수	
Distribution	특이사항 없음
Outliers	이상치 확인
Missing Data	해당사항 없음

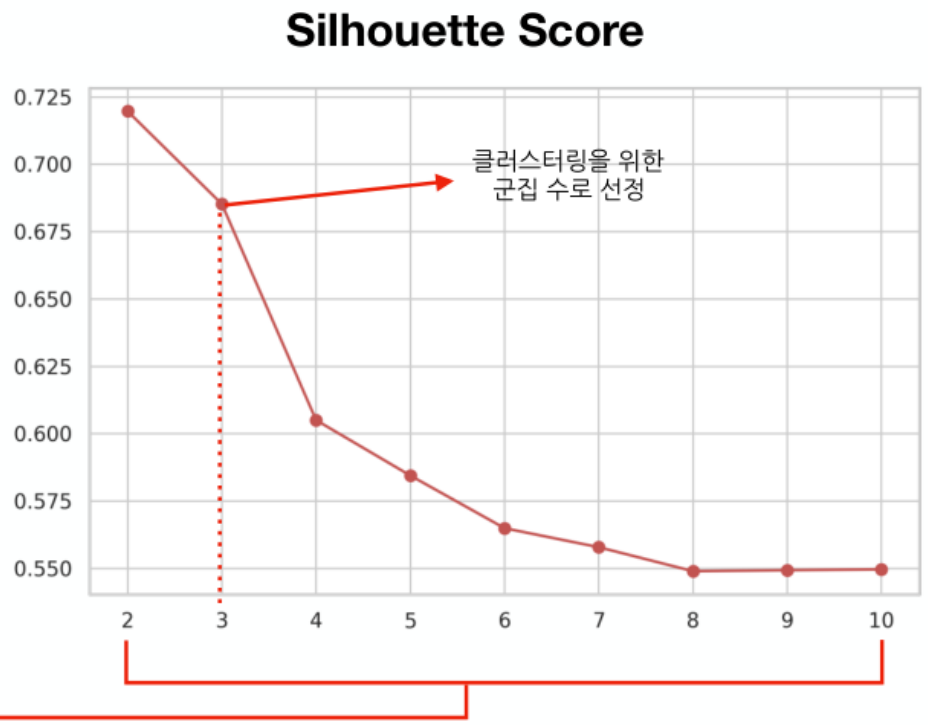
독립 변수	
Distribution	특이사항 없음
Outliers	Robust Scaler
Missing Data	0으로 대체
Correlation	수치형 독립변수엔 다중공선성을 유발하는 선형 종속이 없음

독립 변수와 종속 변수	
Scatter plot	특이사항 없음
Linear or Non-linear	특이사항 없음
Correlation	특이사항 없음

LR 분석) K-means 3개 + 업종 45개 + 수치형 데이터 17개

이너시아 밸류, 실루엣 계수를 이용한 적정 군집수 판단
N = 3



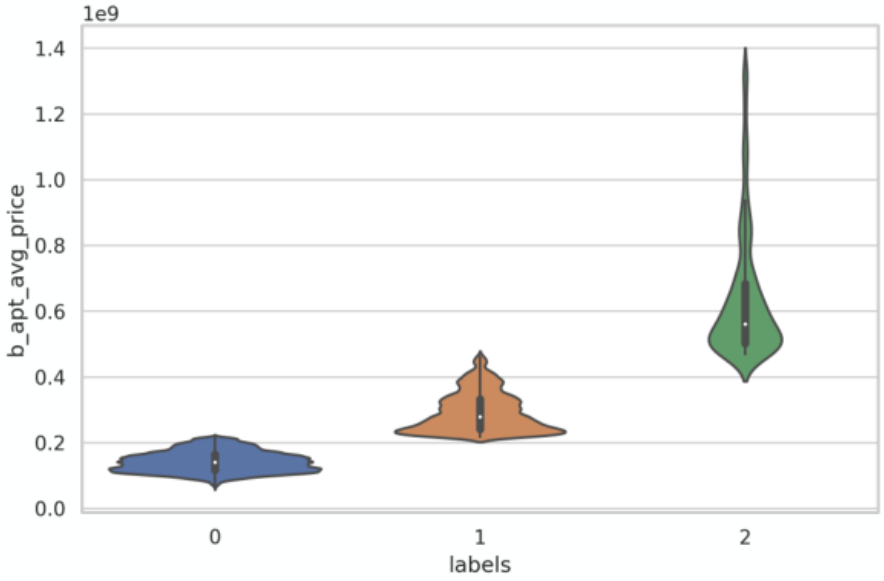


LR 분석) K-means 3개 + 업종 45개 + 수치형 데이터 17개

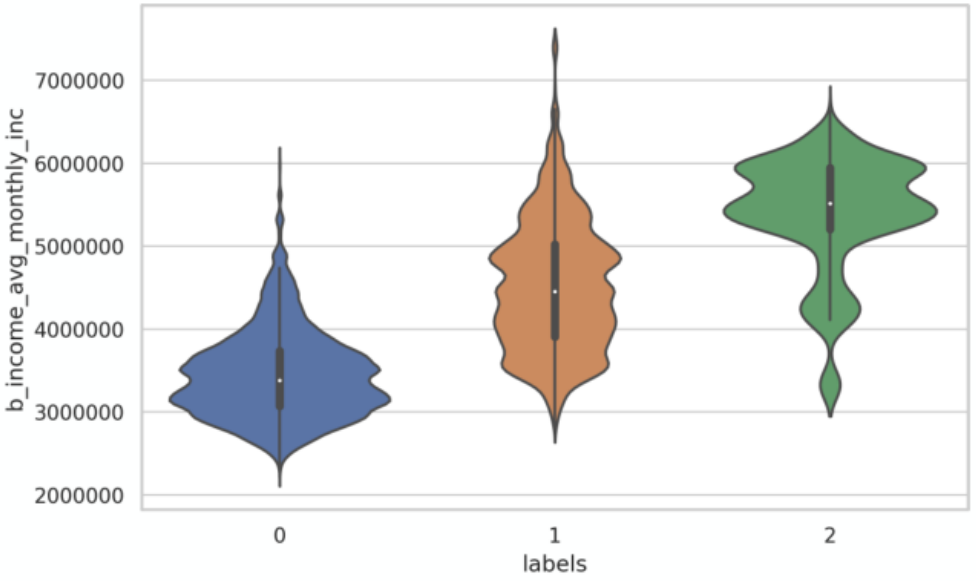
N = 3

군집 별 특성을 파악

배후지 아파트 평균시가



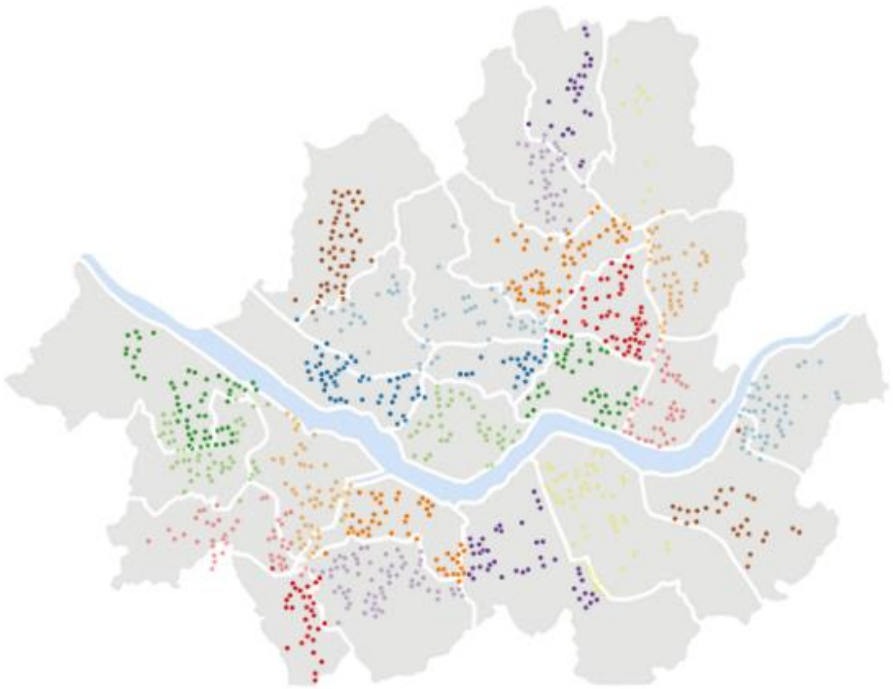
배후지 월 평균소득



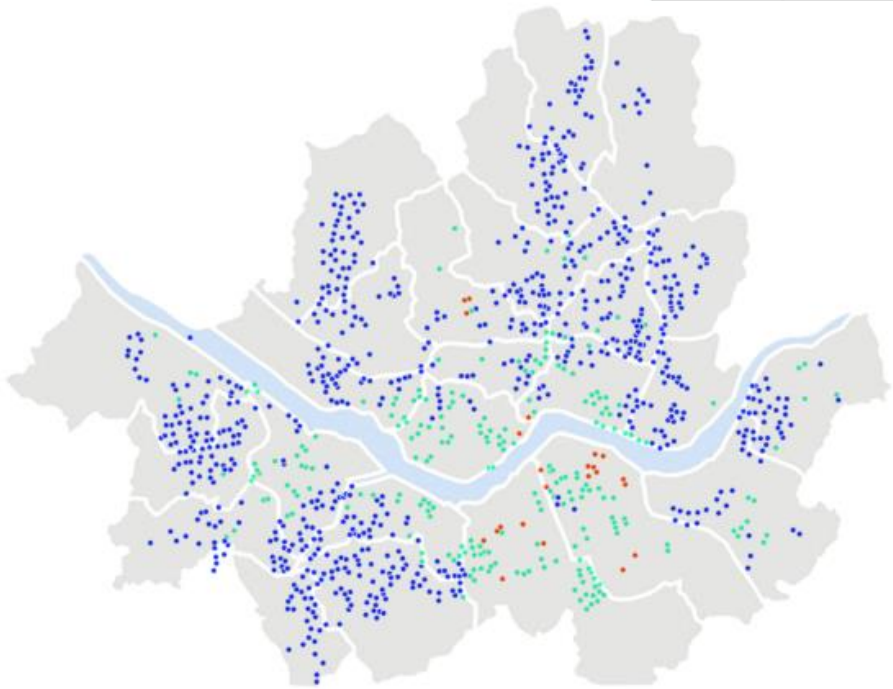
LR 분석) K-means 3개 + 업종 45개 + 수치형 데이터 17개

월 평균 소득 및 아파트 가격이 높은 지역들이
군집 2에 분포되어 있음을 확인 가능

군집	데이터 수
0	87703
1	23514
2	2048



Geographical labeling
(시군구)



Data-driven labeling
(Clustering)

LR 분석) t-sne 군집화 + 업종 + 수치형 데이터

OLS Regression Results

Dep. Variable:	sales_total	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.405
Method:	Least Squares	F-statistic:	980.9
Date:	Thu, 30 Jul 2020	Prob (F-statistic):	0.00
Time:	18:34:02	Log-Likelihood:	-2.0305e+06
No. Observations:	90612	AIC:	4.061e+06
Df Residuals:	90548	BIC:	4.062e+06
Df Model:	63		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
s_store_no_of_store	5.378e+08	5.03e+06	106.895	0.000	5.28e+08	5.48e+08
s_store_no_of_opening	3.315e+07	4.83e+06	6.866	0.000	2.37e+07	4.26e+07
s_store_no_of_closing	6.503e+07	5.09e+06	12.774	0.000	5.51e+07	7.5e+07
s_work_female	1.863e+07	1.91e+06	9.745	0.000	1.49e+07	2.24e+07
s_float_male	2.303e+08	1.58e+07	14.603	0.000	1.99e+08	2.61e+08

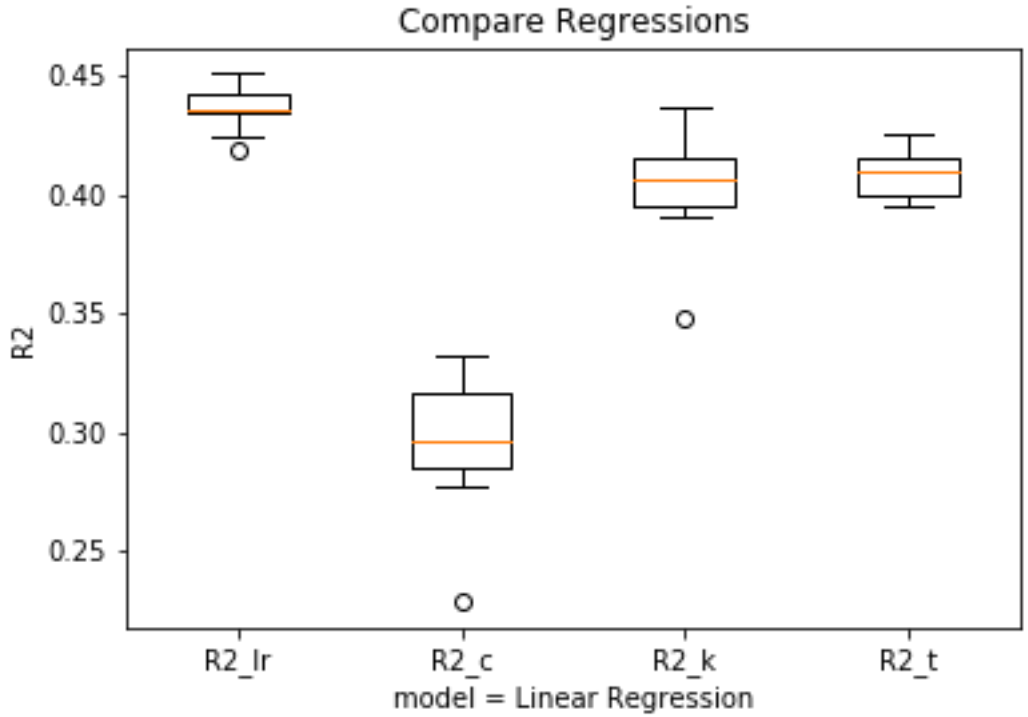
Omnibus:	117558.789	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	88574307.755
Skew:	6.739	Prob(JB):	0.00
Kurtosis:	155.573	Cond. No.	739.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- ☐ 다중공선성은 사라짐
- ☐ R-squared: 0.406
- ☐ 여전히 데이터의 설명력이 부족

선형 회귀 모델의 문제점

- 다중공선성 문제
- 낮은 R-squared 값이 지속됨 **(0.297~0.443)**
- 각 모델을 10번씩 돌려 나온 평균 값 그래프



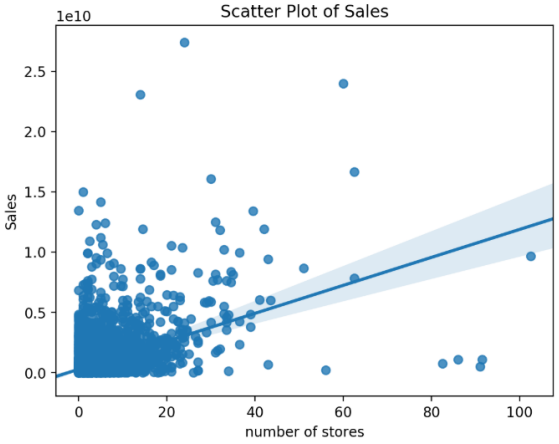
선형 회귀 모델은 분석에 적절하지 않다

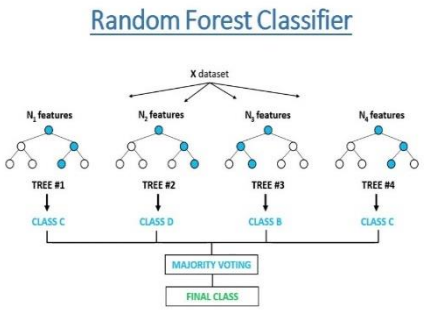
비선형 회귀 모델 비교

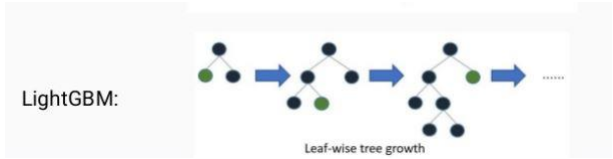
분석 내용

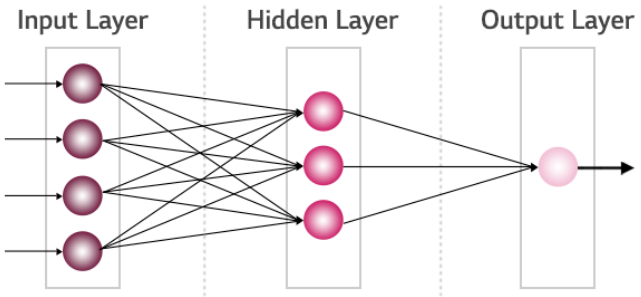
상권코드와 시군구, K-means, t-sne를 데이터에 적용해 가장 성능이 좋은 모델을 검증
R-squared 값이 가장 높고 RMSE 오차값이 가장 낮은 회귀 모델을 찾아보았음

[Linear Regression] vs [Random Forest] vs [Light GBM] vs [Neural Network]

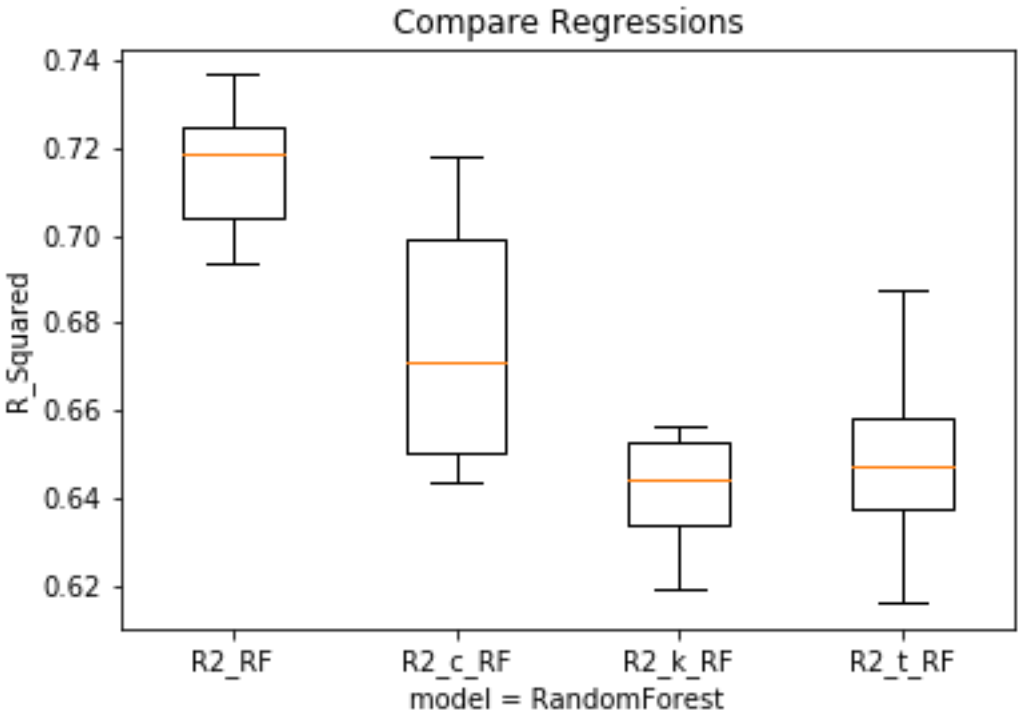








Random Forest 분석



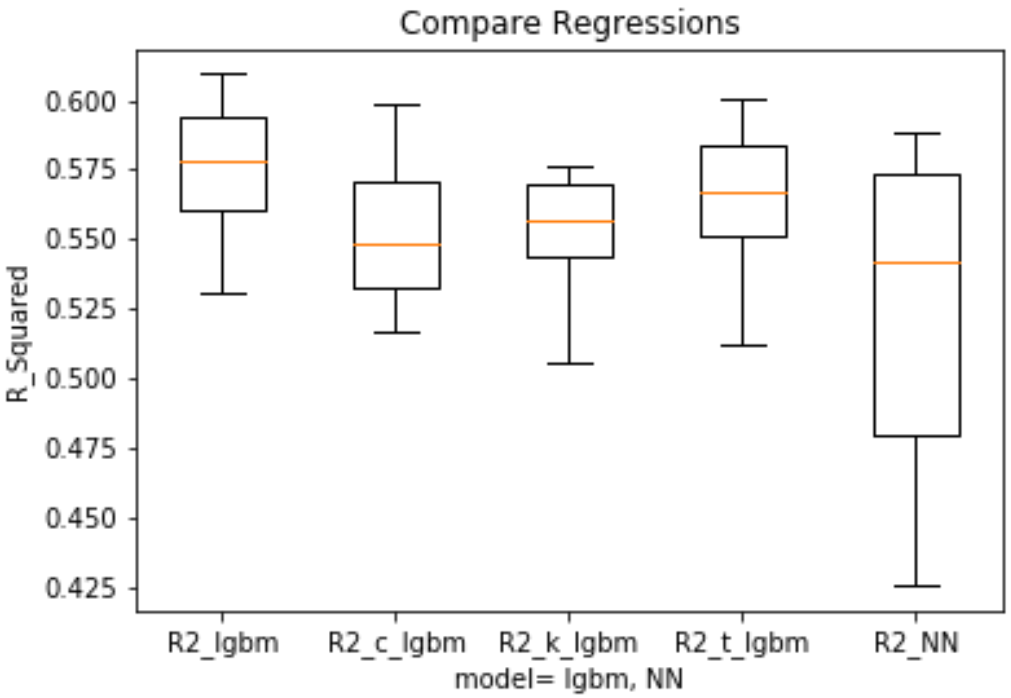
R-squared 값

최소 0.669
최대 0.728
평균 0.714

R2_RF = 1007개의 상권 데이터셋
R2_c_RF = 25개 업종 데이터셋
R2_k_RF = K-means 3개
R2_t_RF = t-sne 군집화

0.7을 넘는 매우 우수한 결과값이 도출됨

Light GBM / Neural Network 분석



LGBM/NN 모델의 R-squared 값

최소 0.515
최대 0.565
평균 0.551

R2_lgbm = 1007개 상권 데이터셋
R2_c_lgbm = 25개 업종 데이터셋
R2_k_lgbm = K-means 3개
R2_t_lgbm = t-sne 군집화
R2_NN = 1007개 + 인공신경망

선형 회귀모델보다는 결과값이 우수하나, Random Forest 보다는 낮음