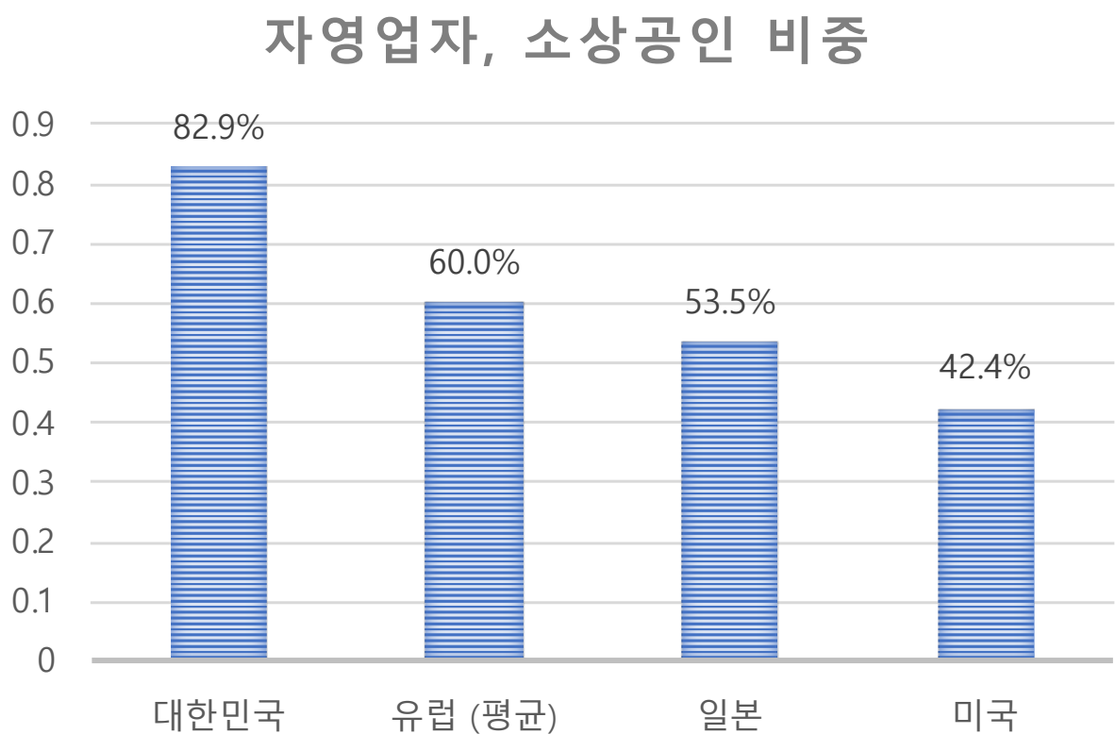
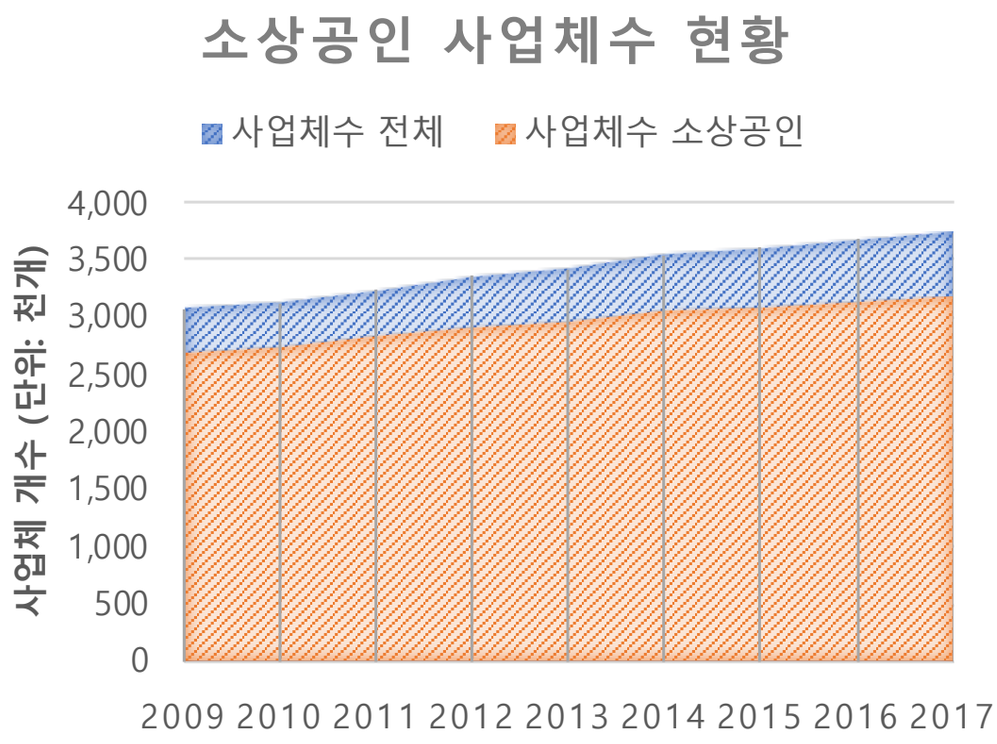
서울시 골목상권 분석 모델링

머신러닝 1조
김세진 문다영 박동재

Introduction

회귀분석 연구 요약





정부/지자체

‘소상공인 지원 정책 수립에 도움’

잠재 소상공인

‘개업 시 참고할 수 있는 지표 제시’

상권 분석에 대한 연구

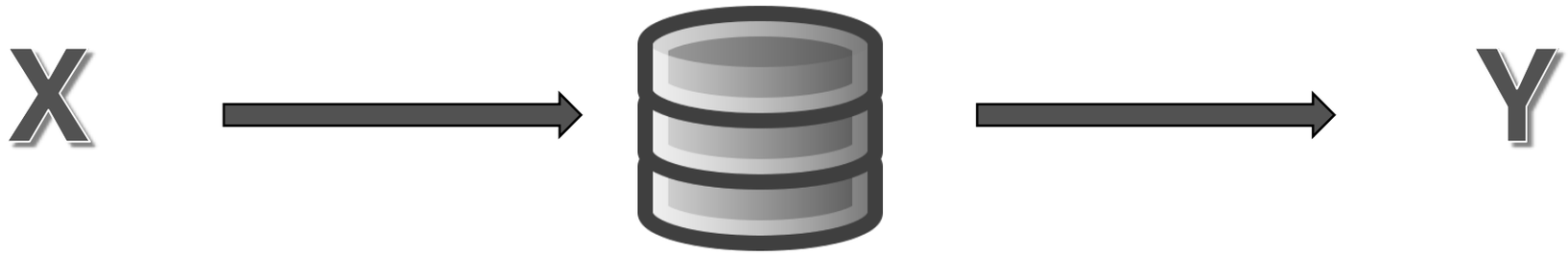
연도	제목	저자	주요 내용
2017	빅데이터 분석을 통한 서울시 골목상권 분석	오홍록 외 2	• 회귀분석(실패) • 상관관계 분석 • 군집분석
2018	골목상권 내 외식 업종 점포의 월 매출액 예측 모형에 관한 연구	임소연	• 선형모형(회귀분석 등) • 비선형 모형(랜덤포레스트 등)
2018	골목상권 매출변화에 영향을 미치는 상권 특성 연구	김지원	• 다항회귀모형
2019	서울시 골목상권 매출액에 영향을 미치는 요인에 관한 연구	김현철 외 1	• 업종 간 상관관계 분석 • 다중회귀분석

회귀분석 연구 요약

선행 회귀분석 연구의 모델들보다, 더 나은 모델을 도출

- 예측력 / 오차 모델의 성능
- 더 많은 데이터 정보를 포함 가능

분석을 위해 사용한 변수들



범주형 데이터 (총 2개 변수)

- 상권 코드 (1007종)
- 업종 코드 (45종)

수치형 데이터 (총 17개 변수)

- 여성의 매출액 비율
- 20-30대 매출액 비율
- 시간대별 매출액 비율
- 여성 직장인구 비율
- 유동인구
- 사업체수
- 개업 점포수
- 폐업 점포수
- 지역 월별 평균소득
- 아파트 평균시가

113256개

- 월별 매출액

선형 회귀 프로젝트 에서의 문제점

- 1007개나 되는 상권 코드 때문에 데이터의 차원이 과도하게 높아짐
- 고차원의 데이터는 선형 회귀분석 모델링 시,
  많은 메모리를 소비하고 연산 처리에 부담을 줌
- 고차원 데이터는 선형 회귀분석 모델의 성능 저하를 초래함.

회귀분석 이후 문제 정의

- **‘비선형 회귀모형은 선형회귀 모형에 비해,
매출액 예측 성능이 우수하다.’**
  → 서울시 골목상권 매출액 회귀분석 모델별 성능 비교

- **‘데이터에 주어진 지리적 라벨링보다 클러스터링을 통한 라벨링이
데이터 간 공통된 특성을 더 잘 대표한다.’**
  → 서울시 골목상권 군집화 분석

- **‘클러스터링을 이용한 범주형 데이터의 차원축소는
회귀분석 성능 향상에 도움을 준다.’**
  → 서울시 골목상권 클러스터링 후, 회귀분석

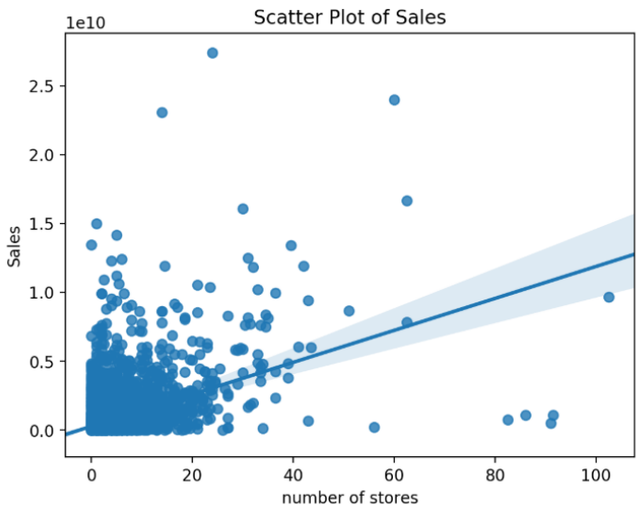
Experiment 1

‘비선형 회귀모형은 선형 회귀모형에 비해, 매출액 예측 성능이 우수하다.’

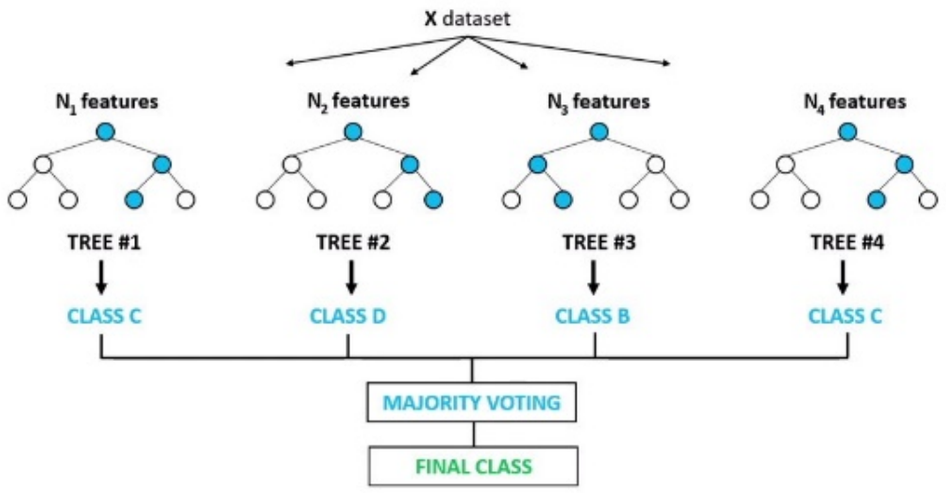
실험1:

‘비선형 회귀모형은 선형 회귀모형에 비해, 매출액 예측 성능이 우수하다.’

Linear Regression



Random Forest

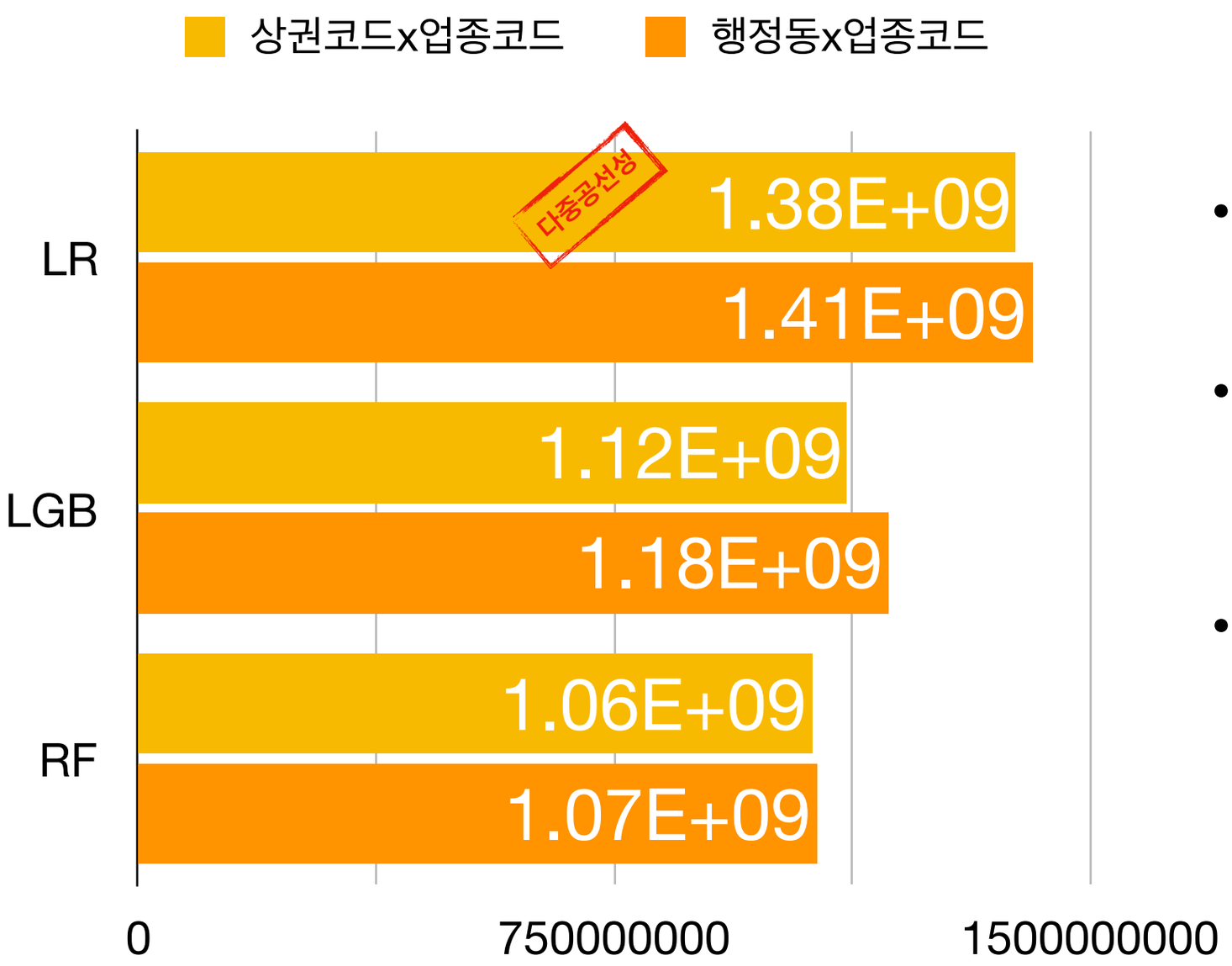


Light GBM

RMSE를 최소화시키는 Regression 모델 알아보기

실험1:

‘비선형 회귀모형은 선형 회귀모형에 비해, 매출액 예측 성능이 우수하다.’



- **비선형 회귀모형 성능이 선형 회귀모형 성능보다 우위**

- 비교대상 중, **Random forest regression이** 매출액 예측에 **제일 우수**한 회귀모형

- 비선형 회귀 모형이 같을 때, 범주형 데이터가 적을 때(행정구)보다 많을 때(상권코드 1004개)가 오차가 적은 점이 인상적.

Experiment 2

‘데이터에 주어진 지리적 라벨링보다,
클러스터링을 통한 라벨링이 데이터의 공통특성을 잘 대표한다.’

실험2:

‘데이터에 주어진 지리적 라벨링보다, 클러스터링을 통한 라벨링이 데이터의 공통특성을 잘 대표한다.’

- **양이 방대하고, 다변수 데이터 (113,265 x 17)**
  → 분석에 사용할 수 있는 클러스터링 모델이 많지 않음.

- **적절한 군집 수를 찾을 수 있는 지표 부재**
  → 유일한 지표인 Silhouette score의 한계

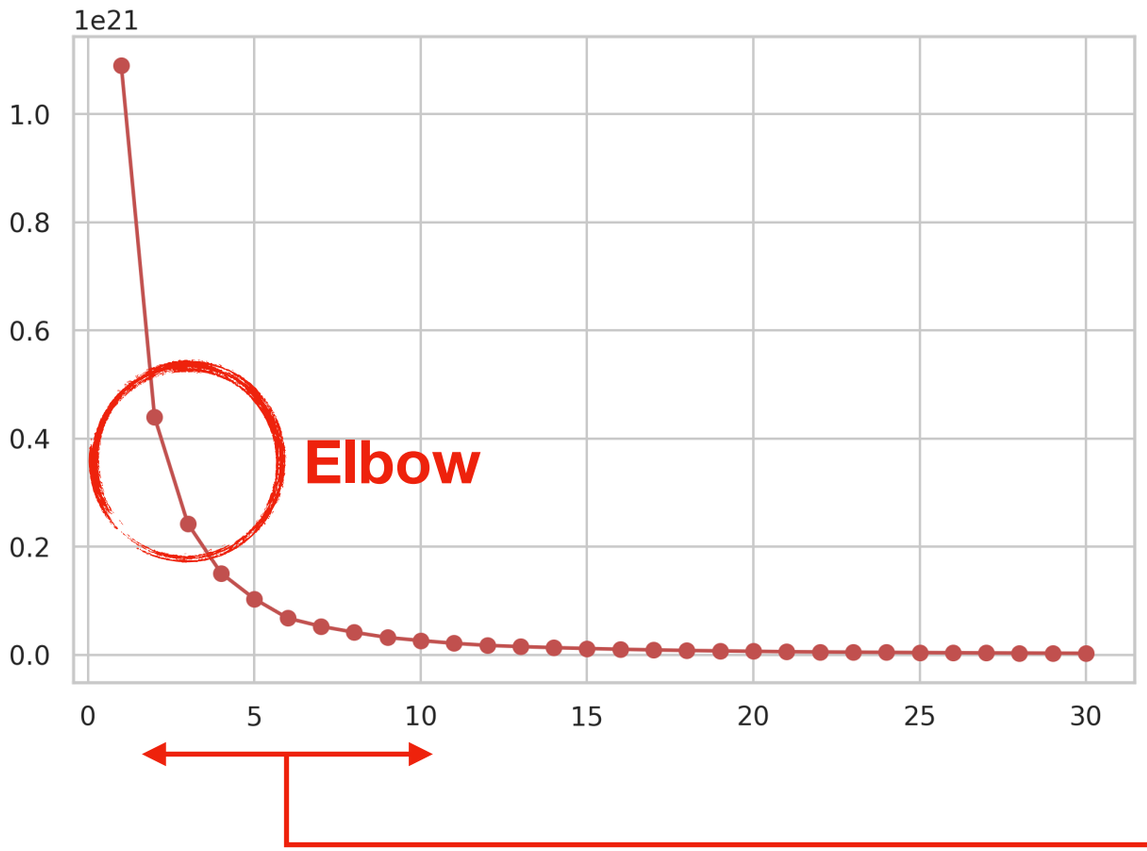
- **컴퓨터 리소스 부족 (8코어, 64GB 메모리 기준)**
  → 8코어 CPU, 64GB 메모리 기준으로 모델 실행 3시간 소요

실험2:

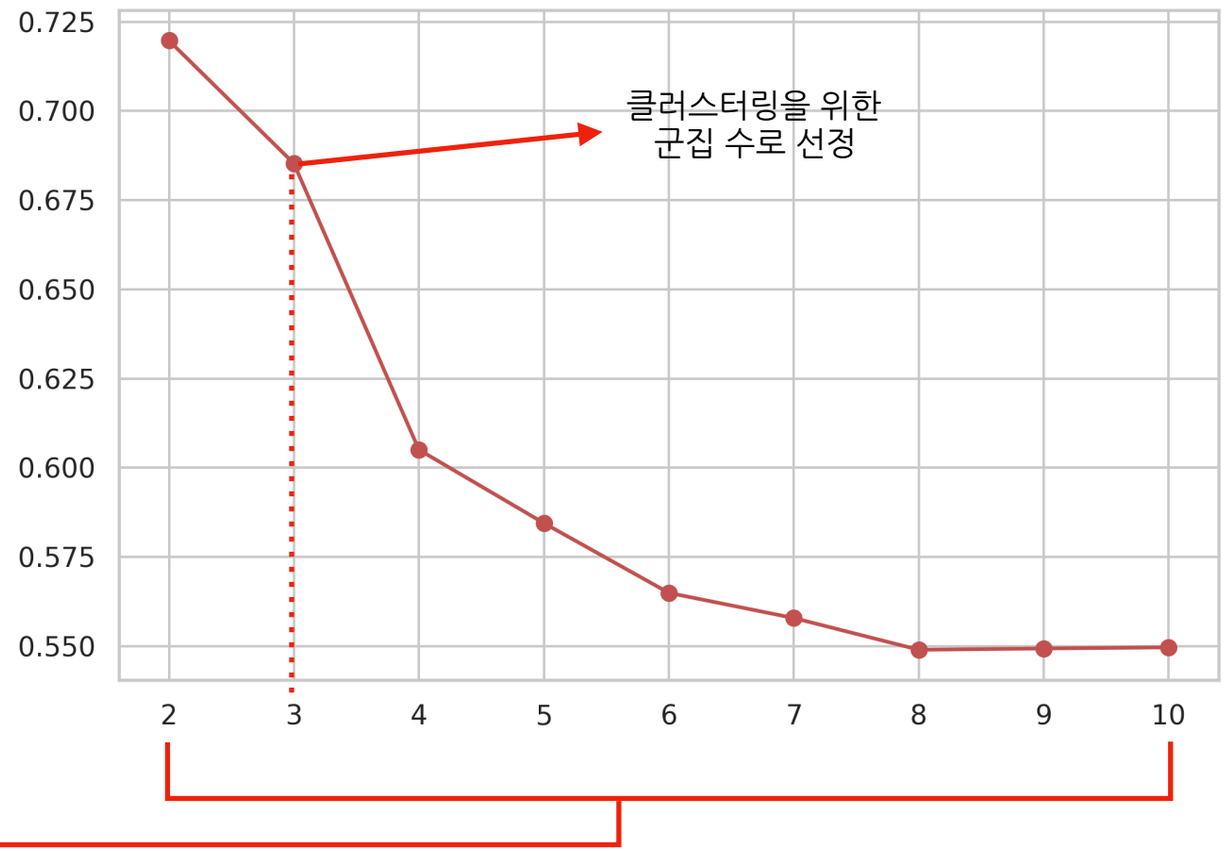
‘데이터에 주어진 지리적 라벨링보다, 클러스터링을 통한 라벨링이 데이터의 공통특성을 잘 대표한다.’

군집 수 선정 과정

Inertia



Silhouette Score

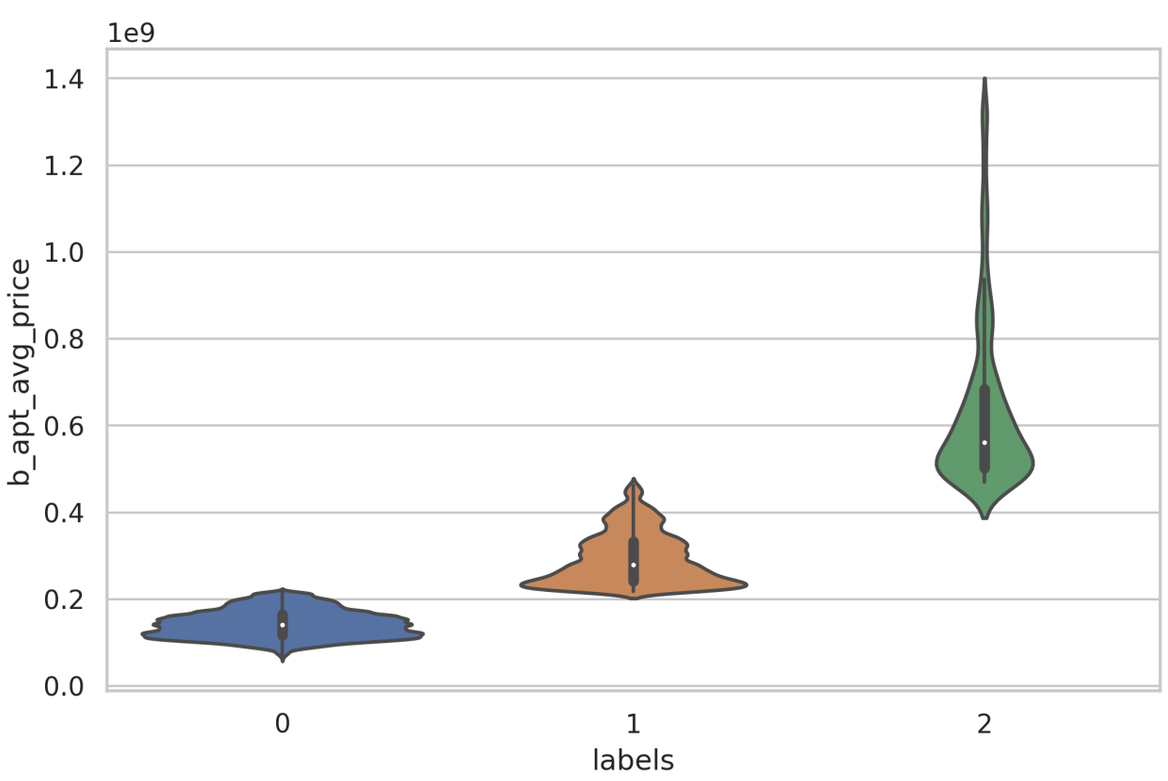


실험2:

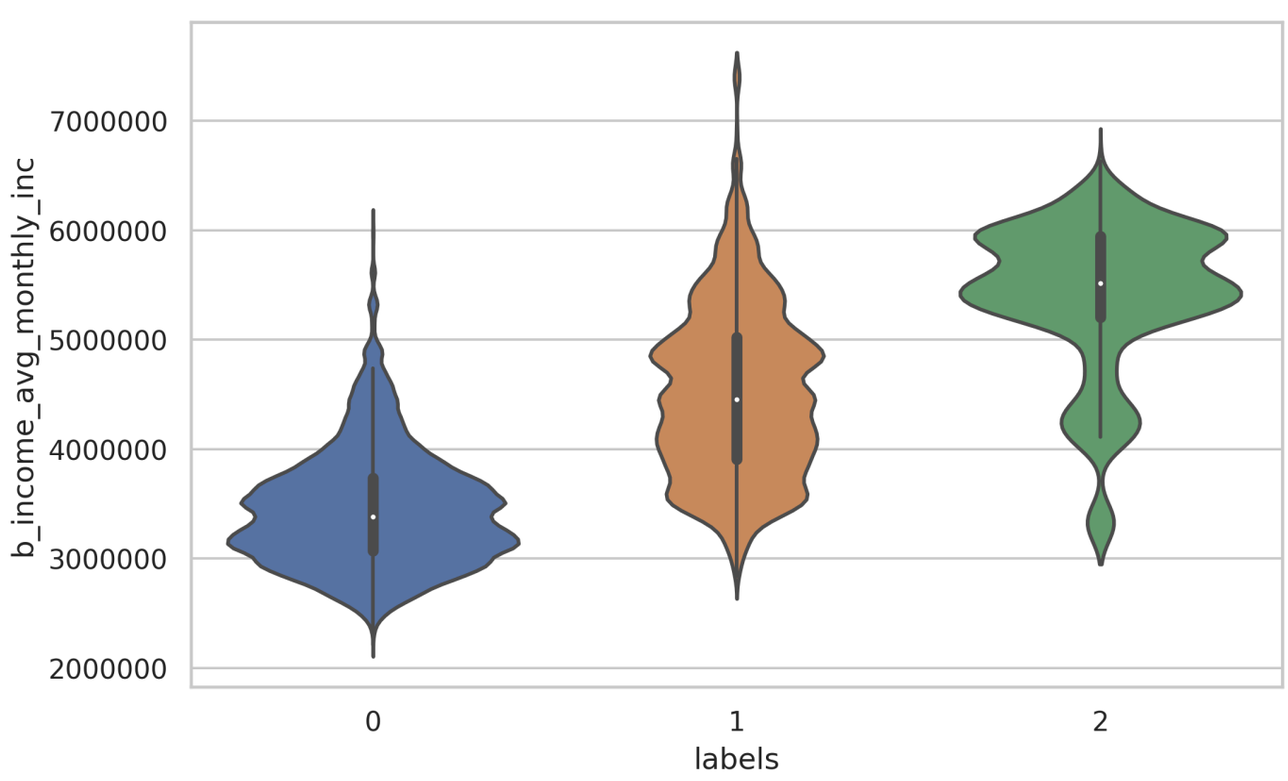
‘데이터에 주어진 지리적 라벨링보다, 클러스터링을 통한 라벨링이 데이터의 공통특성을 잘 대표한다.’

N = 3

배후지 아파트 평균시가



배후지 월 평균소득



실험2:

‘데이터에 주어진 지리적 라벨링보다, 클러스터링을 통한 라벨링이 데이터의 공통특성을 잘 대표한다.’

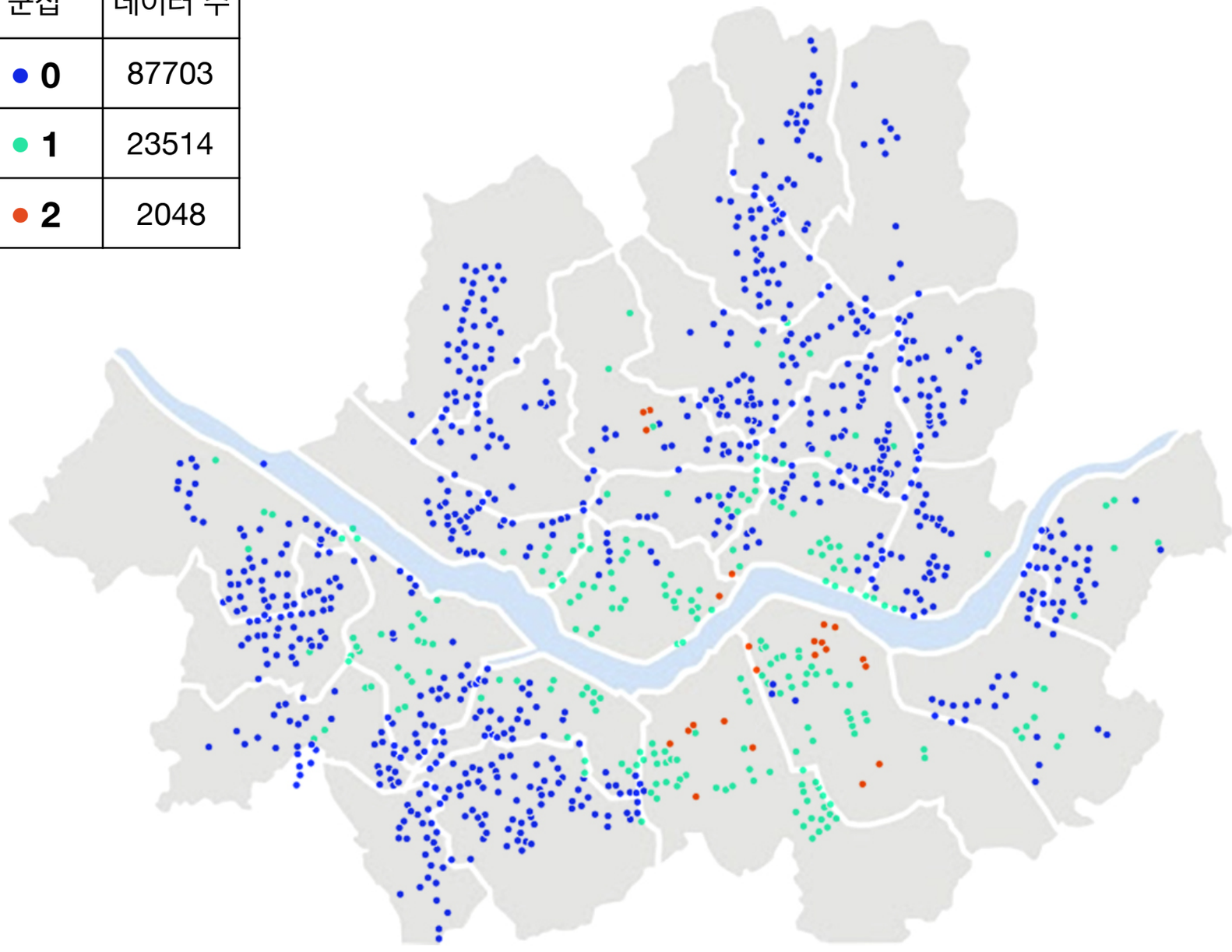
N = 3

variable	cluster	count	mean	std	min	median	max
배후지 아파트 평균시가	0	87,703	142,873,400	31,282,990	63,770,174	140,757,018	217,593,400
	1	23,514	293,001,700	59,376,220	218,269,259	279,079,598	462,629,100
	2	2,048	636,470,900	189,615,300	469,839,952	561,564,647	1,318,357,000
배후지 월평균 소득	0	87,703	3,445,343	523,879	2,213,074	3,382,338	6,080,675
	1	23,514	4,515,762	773,679	2,840,931	4,457,249	7,417,693
	2	2,048	5,390,793	682,957	3,246,230	5,516,389	6,630,464

실험2:

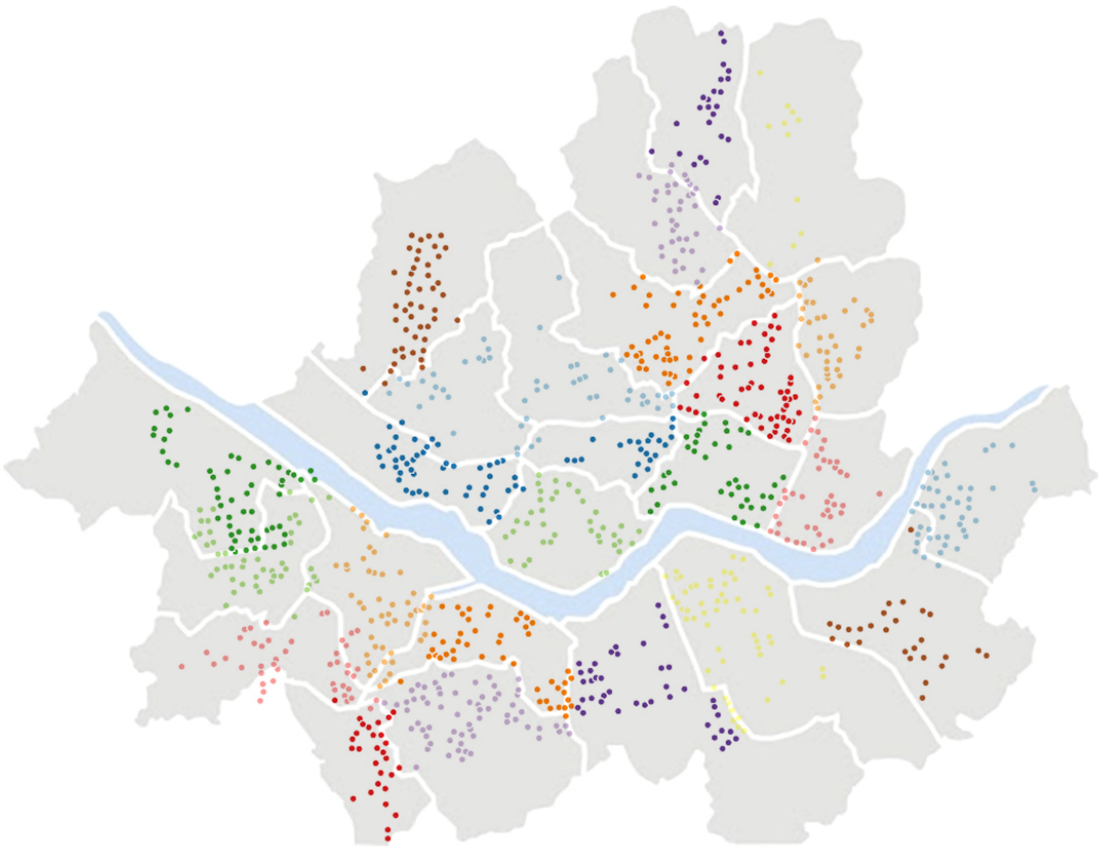
‘데이터에 주어진 지리적 라벨링보다, 클러스터링을 통한 라벨링이 데이터의 공통특성을 잘 대표한다.’

군집	데이터 수
● **0**	87703
● **1**	23514
● **2**	2048



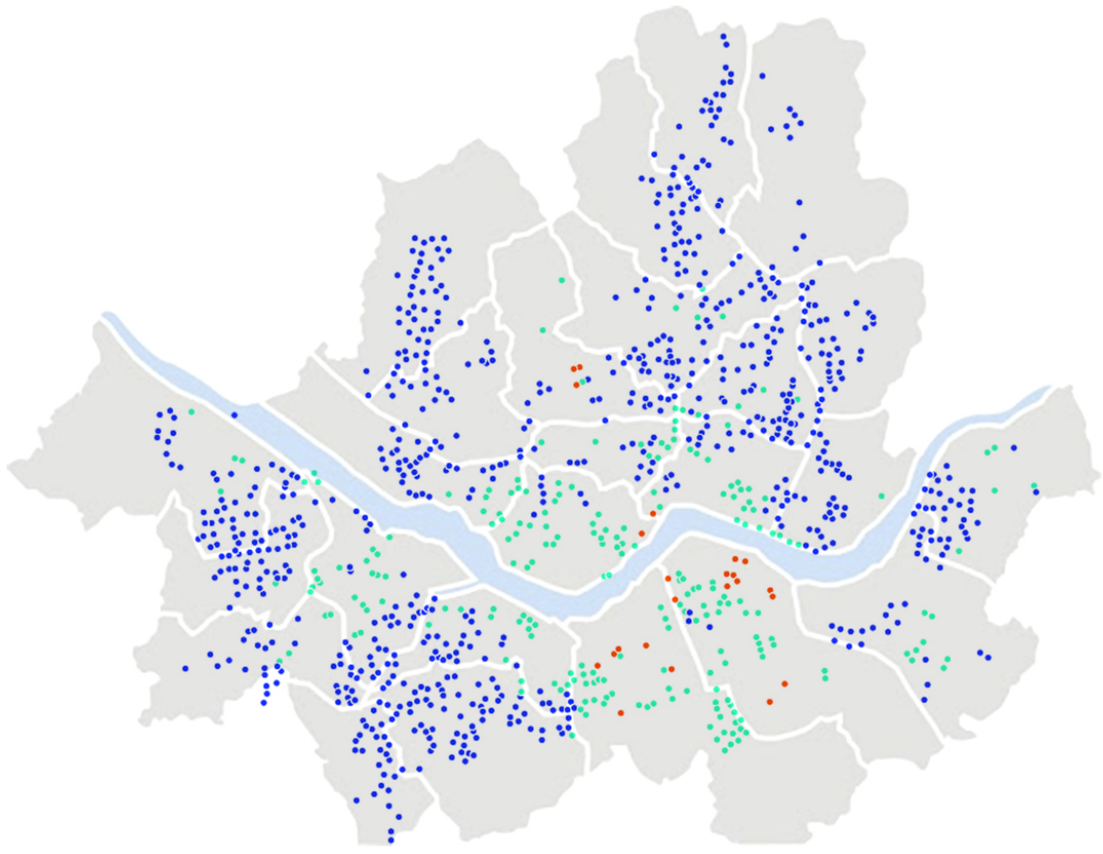
실험2:

‘데이터에 주어진 지리적 라벨링보다, 클러스터링을 통한 라벨링이 데이터의 공통특성을 잘 대표한다.’



Geographical labeling
(행정동)

〈



Data-driven labeling
(Clustering)

Experiment 3

‘클러스터링을 이용한 범주형 데이터의 차원축소는 회귀분석 성능 향상에 도움을 준다.’

“회귀모델이 다중공선성의 문제로 인한 매우 높은 설명력을 갖거나 매우 낮은 설명력을 갖는 문제가 생겨 유의미한 결과를 도출하지 못하였다…(중략)…따라서 회귀분석을 제외한 상관분석과 군집분석을 통해 상권 분석을 진행하였다.”

- 오흥록 외 2명 (2017)

“서울 열린데이터 광장 사이트에서 제공하는 데이터의 위치정보는 도로명으로 표기되어있다.
하지만 그 개수가 너무 많아 모델링에 적합하지 않아 프로파일링 정보에 포함된 행정구 정보와 매칭하였다.”

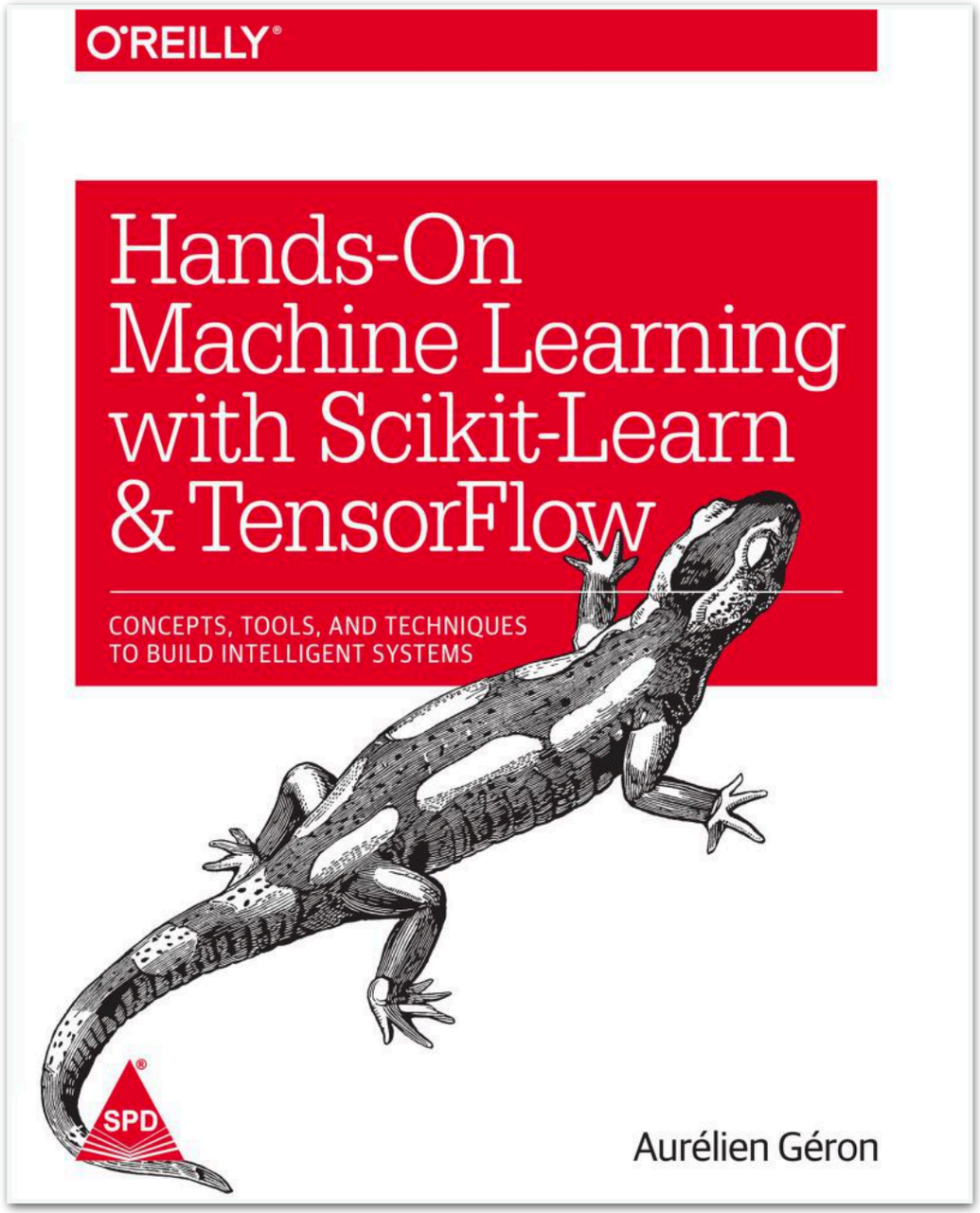
- 임소연 (2018)

“In many of these domains categorical features are common and often of high cardinality.
Using one-hot encoding in such circumstances lead to very high dimensional vector representations, causing memory and computability concerns for machine learning models.”

- Serger (2018)

“In real life, availability of correctly labeled data and handling of categorical data are often acknowledged as two major challenges in pattern analysis. Thus, clustering techniques are employed on unlabeled data to group them according to homogeneity.”

- Sarkar (2019)



"Clustering can be an efficient approach to dimensionality reduction."
- Aurélien Géron

실험3:

‘클러스터링을 이용한 범주형 데이터의 차원축소는 회귀분석 성능 향상에 도움을 준다.’

기존 연구

Independent Variables X		
Categorical variables		Numerical variables
상권코드 (district)	업종코드(code)	총 17개의 수치형 변수
1007 종	45 종	113265개

One-hot Encoding

Regressor

Dependent Variable Y
총 매출액 (sales_total)

제안하는 방법

Independent Variables X		
Categorical variables		Numerical variables
상권코드 (district)	업종코드(code)	총 17개의 수치형 변수
1007 종	45 종	113265개

Clustering

- K-mean
- T-SNE
- Hierarchical Clustering
- ⋮

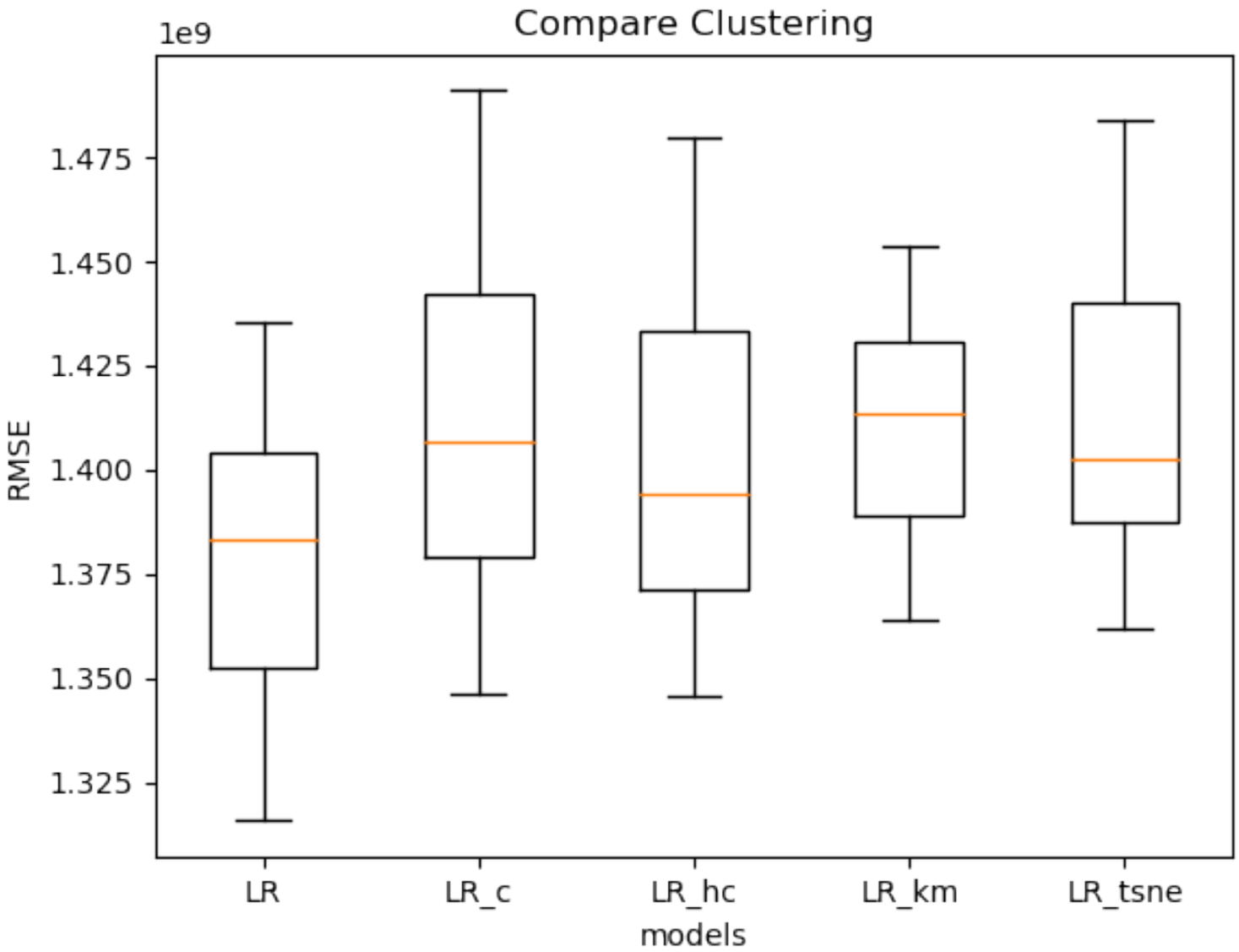
One-hot Encoding

Regressor

Dependent Variable Y
총 매출액 (sales_total)

실험3:

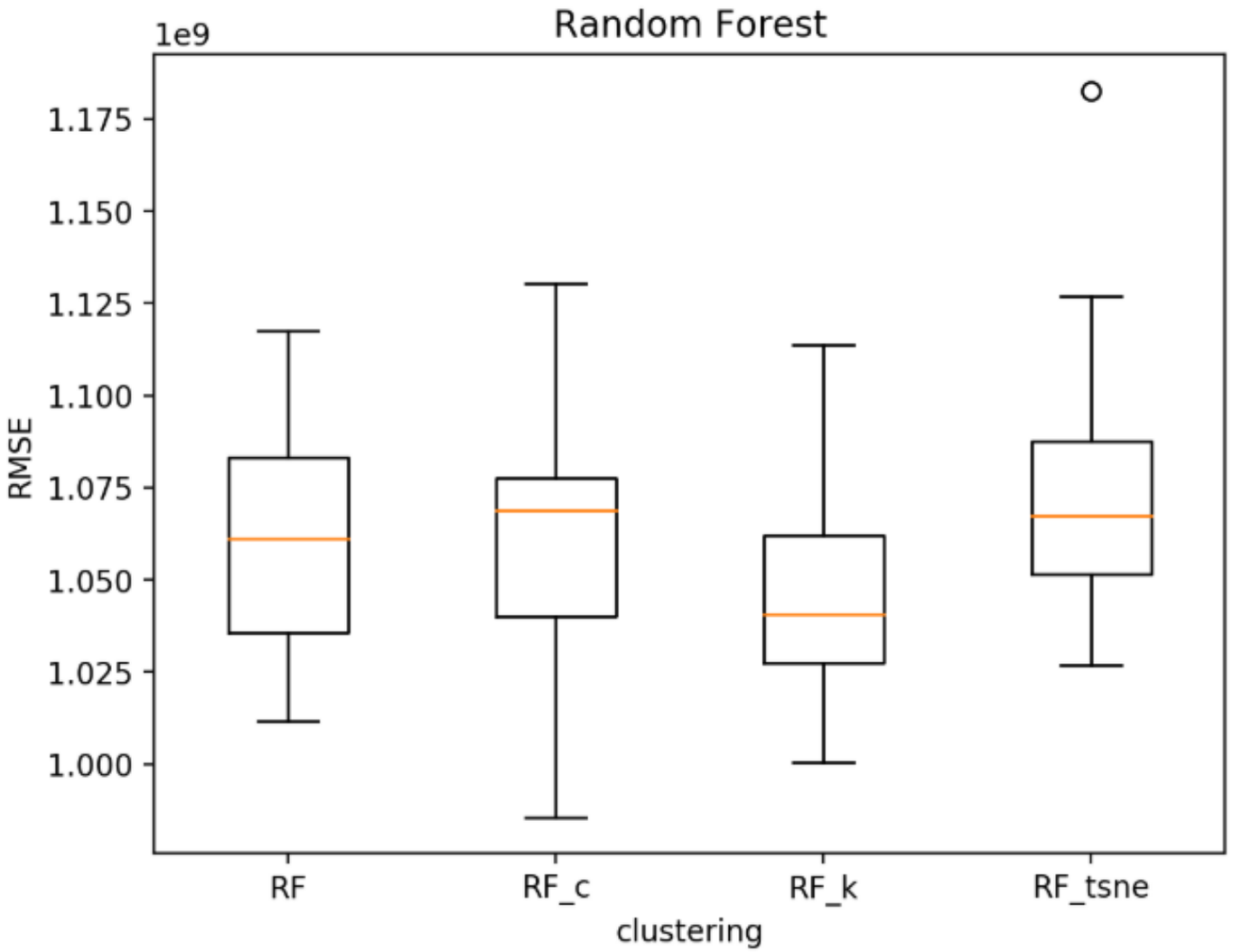
‘클러스터링을 이용한 범주형 데이터의 차원축소는 회귀분석 성능 향상에 도움을 준다.’



- **RMSE 평균이 가장 낮은 방법**
  - Hierachical Clustering
- **RMSE 분산 제일 낮은 방법**
  - K-means

실험3:

‘클러스터링을 이용한 범주형 데이터의 차원축소는 회귀분석 성능 향상에 도움을 준다.’



- **K-means 군집 라벨링 후 Random forest regression** 성능이 타 모델 대비 우위

Future Work

범주형 데이터 차원축소에 대한 연구

연도	제목	저자	주요 내용
2016	Deep Learning over Multi-field Categorical Data	Zhang 외 2	• 다차원 범주형 데이터 처리를 위한 인공신경망 기법들 제안
2017	Cat2Vec: Learning distributed representation of multi-field categorical data	Wang 외 1	• Word2Vec에서 착안한 Cat2Vec(categories to vectors)이라는 모델을 다중 범주형 데이터를 다루는 방법으로 제시
2019	고차원 범주형 자료를 위한 비지도 연관성 기반 범주형 변수 선택 방법	이창기 외 1	• 변수 선택과 변수 간의 연관성 정보를 이요하는 방법을 통해, 고차원 범주형 데이터가 가지는 '차원의 저주'를 해소하려는 시도
2019	매니폴드 학습 기반 특징값을 활용한 카테고리 분류 사례 연구	강규창 외 1	• 학습기반 특징값을 활용한 카테고리 분류 • 고차원의 입력데이터를 저차원의 특징벡터로 변환 • 오토인코드의 한 방식이라 할 수 있는 VAE를 활용하여 고차원의 입력데이터 를 저차원의 특징벡터로 변환

한계점

- 한정된 자원 때문에,
  다양한 방법의 클러스터링을 시도해보지 못 함.

- 클러스터링을 통한, 선형 회귀분석 성능 향상 정도가 아쉬움.

- 회귀분석에서 one-hot-encoding을 잘 할 수 있도록
  카테고리 재분류를 하는 것보다,
  회귀모형을 바꾸는 것이 예측 향상에 훨씬 더 큰 영향을 줌.

딥러닝을 이용한 매출액 예측

- **더 정밀한 가중치값 조정을 통한 성능 향상 기대**
  - 오차로부터 가중치값을 수정하는 back propagation으로 성능 향상 기대
- **더 많은 변수 삽입 기대**
  - 회귀분석 프로젝트 당시, 전처리 단계에서 다중공선성 방지를 위해 상관관계가 높은 변수들을 제거하여 1000여개의 변수 중 17개로 축소
- **오토인코딩 또는 CNN을 사용하는 것을 고려 중**
  - RF보다 더 복잡하고 성능이 좋은 딥러닝 모형(CNN)을 시도한다면 더 정확한 매출액 예측이 가능해질것으로 보인다