2022 하계 빅데이터 분석 경진대회

상가거래 데이터를 활용한 서울시 상권 및 권리금 분석

SEOUL-GOODWILL

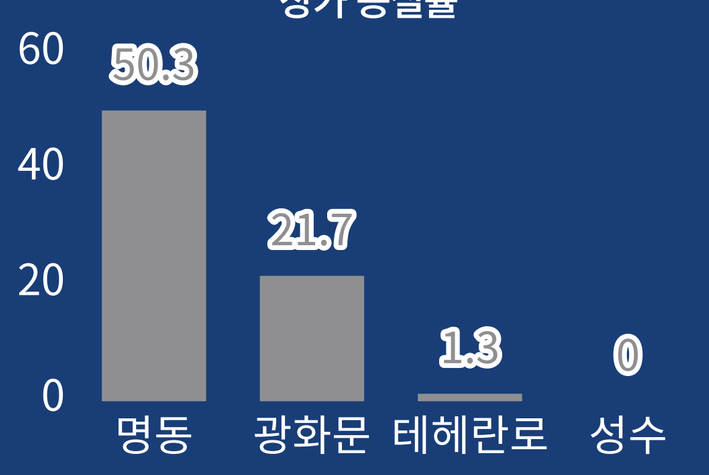
경영정보학과 5569602 김차미
통계학과 5526369 안유나

목차

문제 정의

[상권의 변화] 한국부동산원

뜨는 상권 vs 지는 상권



어려운 시기에도
뜨는 상권이 존재한다 ?

- 상권의 침체와 활성화를 결정짓는 요인은 무엇인가 ?
- 우리 상가는 뜨는 상권에 속할까 ?

과제 1 - 상권의 흥망성쇠 및 영향 요인 분석

사용 데이터

상가 거래

서울시 상가 거래 데이터 생성

jumpo 데이터 중
address == '서울시' 인
데이터 추출

상권 구분

서울시 상권 영역 정보

골목, 발달, 전통시장, 관광특구
총 4개 유형에 대한 정보를 포함한
행정동별 상권 구분 데이터

지하철

서울시 역사 마스터 정보

서울시 역사에 대한
역사 ID, 역사명, 호선명,
좌표 정보데이터

대형 점포 인허가

서울시 대규모점포 인허가 정보

대형마트, 백화점,
쇼핑센터, 복합쇼핑몰 등
다수의 사업자로부터 납품받아
판매하는 업소정보

상권 변화

서울시 행정동별 상권 변화 지표

행정동별 상업공간의 변화를
4개 등급으로 나눈 지표를
포함하는 데이터

대학교

서울시 대학 및 전문대학 DB

서울시내 대학 및 전문대학의
종류별 학교명 및 상태, 주소 등
관련 정보

과제 1 - 상권의 흥망성쇠 및 영향 요인 분석

전처리

대형점포 인허가 데이터

영업일자, 휴업일자, 폐업일자

» 휴업일자 & 폐업일자가 Null 이거나, 영업일자가 해당 연도에 존재하는 경우 **영업 중**

영업 중인 점포

» 연도별, 행정동별로 추출

	0	2018	2019	2020	2021	2022
0	신월동	0.0	0.0	0.0	0.0	0.0
1	도화동	0.0	0.0	0.0	0.0	0.0
2	면목동	1.0	1.0	1.0	1.0	1.0
3	역삼동	3.0	2.0	2.0	2.0	1.0
4	보문동7가	0.0	0.0	0.0	0.0	0.0

영업 중인 대형 점포의 수

» 연도별, 행정동별 영업 중인 대형 점포의 수 count

	0	Unnamed: 1	0.1
0	신월동	2018	0
1	신월동	2019	0
2	신월동	2020	0
3	신월동	2021	0
4	신월동	2022	0

bmart 컬럼 생성

» jumpo 데이터와 merge

address_d	contract_year	bmart
신월동	2022	0
도화동	2022	0
면목동	2022	1
역삼동	2022	1
보문동7가	2022	0
...	...	...

전처리

상권 구분 데이터

상권 구분 코드명 추출

» 행정동-법정동 맵핑 정보 활용하여
법정동에 맞는 상권 구분 코드명 추출

법정동별 상권 구분 코드명 count

» 법정동별 groupby를 통해 각 법정동에 존재하는 상권 구분 코드명 count

법정동	상권_구분_코드_명	
가락동	골목상권	5
	발달상권	5
가리봉동	골목상권	2
	발달상권	1
	전통시장	1
...	...	...

상권 구분 컬럼 생성

» pivot_table 활용하여
각 법정동에 어떠한 상권이 몇 개씩 있는지 count 후 merge

address_d	contract_year	bmart	골목상권	관광특구	발달상권	전통시장
신월동	2022	0	16.0	0.0	0.0	7.0
도화동	2022	0	7.0	0.0	2.0	2.0
면목동	2022	1	16.0	0.0	1.0	4.0
역삼동	2022	1	9.0	0.0	7.0	2.0
보문동7가	2022	0	1.0	0.0	0.0	0.0

과제 1 - 상권의 흥망성쇠 및 영향 요인 분석

전처리

상권 변화 지표란 ?

상권의 변화를 생존한 사업체의 평균 영업 기간과
폐업한 사업체의 평균 영업 기간을 기준으로 4개 등급으로 나눈 지표

상권 변화 데이터

binomial data로 변환

» 1 -> 경쟁력 있는 상권 : 신규 창업 우위, 기존 업체 우위
0 -> 주의 상권 : 창업 진출입 시 세심한 주의 필요

법정동별 상권 변화 지표 count

» 각 법정동 지역별 상권 변화 지표 맵핑 후
법정동별 상권 변화 지표 count

법정동	상권_변화_지표	
가락동	0	74
	1	22
가리봉동	0	2
	1	30

sg_change 컬럼 생성

» 각 법정동별 0과 1의
상권 변화 지표 개수 비교 후,
우세한 지표를
sg_change 컬럼으로 생성

상권_변화_지표	0	1	sg_change
법정동			
가락동	74.0	22.0	0.0
가리봉동	2.0	30.0	1.0
가산동	32.0	0.0	0.0
가양동	77.0	51.0	0.0

sg_change 컬럼 생성

» 앞서 생성한 데이터셋과
merge

전통시장	sg_change
7.0	0.0
2.0	0.0
4.0	0.0
2.0	0.0

과제 1 - 상권의 흥망성쇠 및 영향 요인 분석

전처리

지하철 데이터

역지오코딩

» 카카오 API 활용하여
지하철 역사의 위도, 경도를 주소로 변환

	역사_ID	경도	위도	ADDRESS	CODE
0	9996	37.560927	127.193877	경기도 하남시 망월동	4145010900
1	9995	37.557490	127.175930	서울특별시 강동구 강일동	1174011000

서울시 데이터 추출 및 행정동별 지하철역 count

» 서울시 데이터 추출 후,
각 행정동별 지하철역 개수 count

subway 컬럼 생성

» 앞서 생성한 데이터셋과 merge

전통시장	subway	sg_change
7.0	0.0	0.0
2.0	4.0	0.0

대학교 데이터

행정구별 대학교 개수 count

» 대학교가
존재하지 않는 법정동이 훨씬
많기 때문에, 행정구별
대학교 개수 count

행정구	학교명
강남구	1
강북구	1
강서구	2
관악구	1

학교명 컬럼 생성

» 앞서 생성한 데이터셋과 merge

관광특구	발달상권	전통시장	subway	sg_change	학교명
0.0	0.0	7.0	0.0	0.0	0.0
0.0	2.0	2.0	4.0	0.0	3.0
0.0	1.0	4.0	3.0	0.0	1.0

과제 1 - 상권의 흥망성쇠 및 영향 요인 분석

전처리

store_type 컬럼 정리

- store_type value 축소
  - 중복 되거나 너무 세분화된 store_type 통합, 리스트 만들어 store_type의 value 축소

```
data_6['store_type'].nunique()
```

```
60
```

```
data_7['store_type_n'].nunique()
```

```
10
```

contract_year 컬럼 정리

- contract_year 라벨 인코딩
  - contract_year 2018 ~ 2022를 1~5의 숫자로 인코딩

```
data_7_t['contract_year'].unique()
```

```
array([2022, 2021, 2020, 2019, 2018])
```

```
data_7_t['contract_year'].unique()
```

```
array([5, 4, 3, 2, 1])
```

과제 1 - 상권의 흥망성쇠 및 영향 요인 분석

모델구축

모델 성능 비교

- **Train, Test set 분리**
  - Train : Test = 7 : 3
- **Robust 정규화**
  - area 컬럼에 대하여 Robust 정규화 진행
- **3가지 모델 적용**
  - SGD, Logistic Regression, Gradient Boosted Trees

area
-0.347753
-0.403788
-0.370503
1.608988
-0.553962

- **모델 성능 비교**

	Accuracy	Precision	Recall	F1-Score
SGD	0.81	0.59	0.15	0.24
LR	0.80	0.48	0.11	0.18
GBT	0.94	0.95	0.70	0.81

과제 1 - 상권의 흥망성쇠 및 영향 요인 분석

모델구축

모델 성능 비교

최적화

- 3가지 모델 중 가장 평가지표 점수가 높았던 GBT에 대해 최적화 수행
- RandomizedSearchCV 사용
  cv = 10 으로 교차검증 수행
- max_depth: 4
  min_impurity_decrease: 0.0089 일 때 최적의 성능을 냄

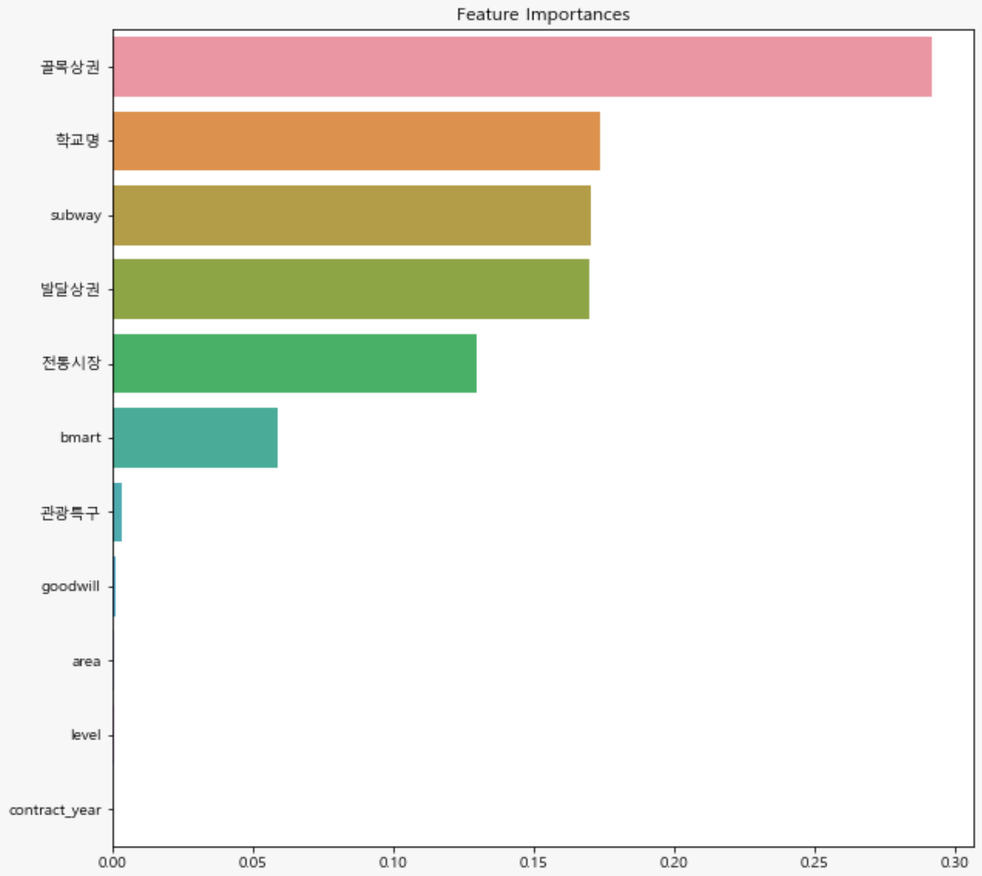
최적화 후 평가지표

	0	1	Accuracy
Precision	0.96	0.95	0.96
Recall	0.99	0.84	0.96
F1-Score	0.97	0.90	0.96

- 전체적인 Accuracy 점수가 0.96으로 오름.
  또한, 1(상권의 긍정적 변화)에 대한 Recall 값이 0.84로 오름.

상권의 침체와 활성화를 결정짓는 요인은 무엇인가?

상권의 흥망성쇠 영향 요인 파악



변수 중요도 파악

변수	변수중요도
골목상권	0.291898
학교명	0.173402
subway	0.170373
발달상권	0.169588
전통시장	0.129917

변수중요도 해석

해당 상가 지역에 **골목 상권**이 얼마나 있는지, **학교, 지하철역** 등이 얼마나 있는지 등의 **입지적 요소**가 많은 영향을 미치는 변수임을 알 수 있음

과제 1 - 인사이트 도출

우리 상가는 뜨는 상권에 속할까?

상권의 흥망성쇠 예측

- **예측값 저장**
  - 상가거래 데이터의 패턴을 학습하여 해당 **상가가 속한 상권**의 흥망성쇠 예측
- **법정동별 예측값 groupby**
  - 저장한 예측값을 test set과 merge 후, 법정동 기준으로 **groupby**
- **상권의 흥망성쇠 voting**
  - 법정동의 **예측값을 count**하여, 0과 1 중 **우세한 값**으로 상권의 흥망성쇠 최종 예측

gbt_predict dong	0	1
가락동	52.0	0.0
가리봉동	4.0	0.0
가산동	45.0	0.0

- **구 상권과 신 상권**

gbt_predict dong	0	1	상권
노고산동	10.0	0.0	0.0
동교동	21.0	0.0	0.0
성수동1가	0.0	15.0	1.0
연희동	0.0	9.0	1.0
이태원동	23.0	0.0	0.0
한남동	0.0	12.0	1.0

- 신촌
- 홍대
- 성수,뚝섬
- 연리단길
- 이태원
- 한남동

과제 1 - 인사이트 도출

상권의 침체와 활성화를 결정짓는 요인은 무엇인가?

골목상권, 대학교,
지하철 등의 개수가
중요 변수로 밝혀짐.
이는 모두 상가 자체의 요인이
아닌 상가의 입지적 요인임.

우리 상가는 뜨는 상권에 속할까?

신규 사업 진입 또는 출입 시,
세심한 주의가 필요한
상권에 속한 상가라면
주변의 골목상권, 대학교,
지하철 등의 수를 고려하여
더 좋은 상권에 속한 상가에서
사업 진출입을
고려할 필요가 있음

상가의 입지적 요인?

상가의 입지적 요인은
권리금을 산정하는데
중요한 고려 요소
-> 권리금을 둘러싼
다양한 분쟁이 존재함

문제 정의

권리금을 둘러싼 문제점

- 권리금의 구성요소
  시설권리금 + 영업권리금 + 바닥권리금
- 시설권리금과 영업권리금
  → 가시적, 분명한 산정 기준
  바닥권리금
  → 불분명한 산정 기준
- 명확한 권리금 산정 기준 필요
  → 상가의 입지 및 상권 정보 반영한 권리금 산정 모델 구축

분석 목적 1

권리금 예측 모델 구축

상가 권리금을 예측하는 모델을 구축
→ 사전에 상가 권리금 예측하여 권리금 관련 사기 방지 지원

분석 목적 2

권리금 산정 지표 파악

새로 상점을 거래하고자 하는 자영업자들에게
권리금 산정에 영향을 미치는 지표 제공
→ 권리금 산정 참고 기준 설정

모델 구축

모델 성능 비교

Train, Test set 분리

Train : Test = 7 : 3

Robust 정규화

area와 goodwill에 대하여

정규화 진행

area	goodwill
1.850149	0.0
-0.234544	-0.7
-0.147322	-0.6
0.329809	0.2

3가지 모델 적용

DecisionTree, RandomForest, XGBoost

모델 성능 비교 (RMSE 기준)

모델	RMSE
DecisionTree	0.7685
RandomForest	0.7451
XGBoost	0.7373

파라미터 최적화(XGBoost)

max_depth	RMSE
3	0.7384
5	0.7373
7	0.7421

파라미터 설정

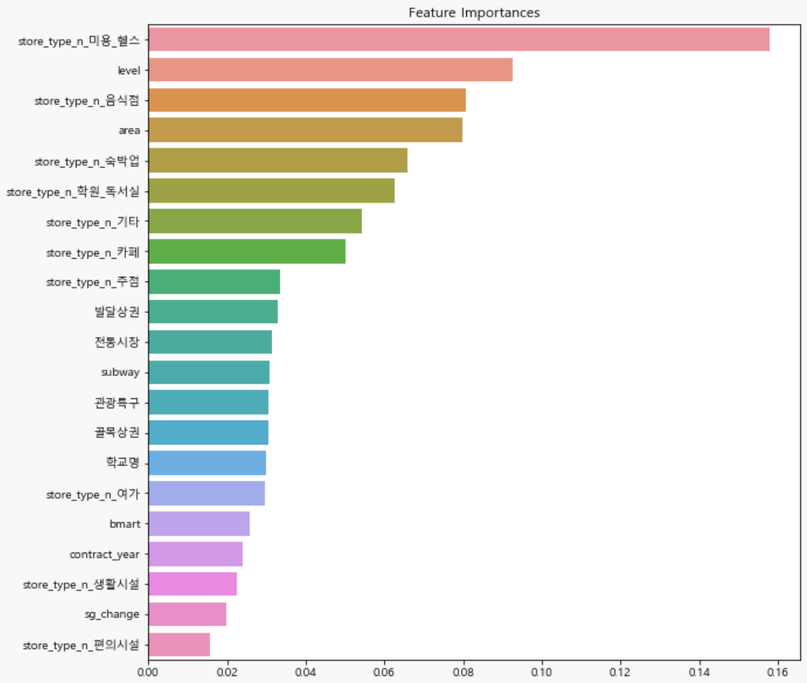
learning_rate= 0.01

n_estimators=700

XGBoost 변수중요도

권리금 산정 지표 파악

XGBoost 변수중요도



변수중요도 : 상위 6개

변수	변수중요도
store_type_n_미용_헬스	0.157927
level	0.092467
store_type_n_음식점	0.080597
area	0.079844
store_type_n_숙박업	0.065863
store_type_n_학원_독서실	0.062548

변수중요도 해석

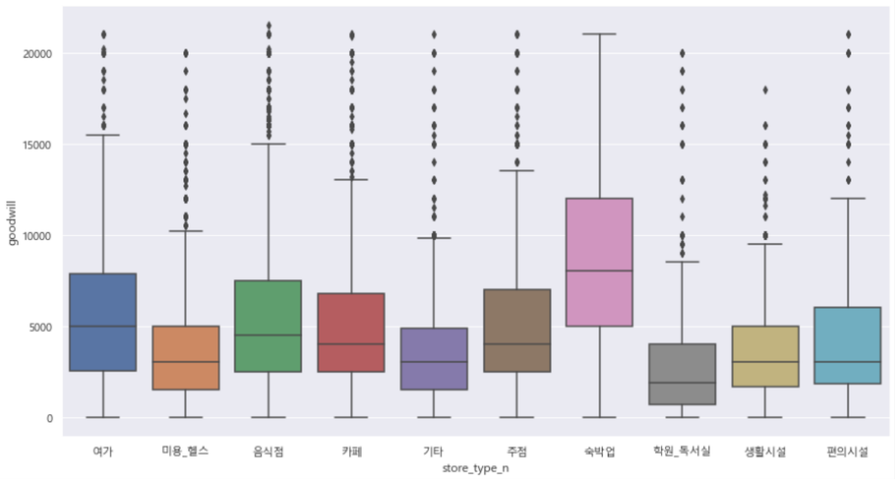
store_type_n, level, area가 예측에 가장 많은 영향을 미치는 변수

즉, 권리금 산정에 큰 영향을 미치는 변수임을 알 수 있음

권리금 산정 지표별 영향 파악

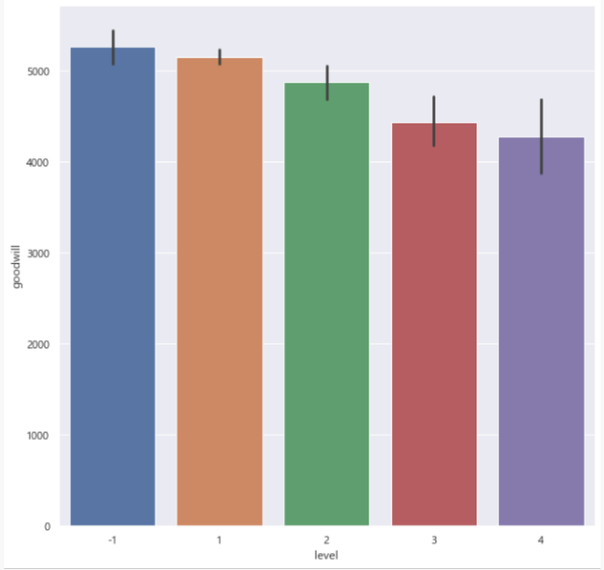
권리금 산정 지표별 영향

store_type_n에 따른 권리금



- store_type_n에 따라 권리금이 다르게 나타남
  → 숙박업, 여가, 음식점, 주점 순으로 권리금이 높게 나타남

level에 따른 권리금



- 상점의 층이 높아질수록 권리금이 낮아지는 것을 알 수 있음

권리금 가격 예측 결론

권리금 분석 예측

권리금 예측 모델 구축

1) RMSE 기준으로 XGBoost 모델 구축 및 파라미터 최적화

» 상가 권리금을 예측하는 모델을 구축함으로써 사전에 상가 권리금 예측하여 권리금 관련 사기를 방지할 수 있음

권리금 산정 지표 파악

1) 구축한 XGBoost 모델을 기준으로 변수 중요도 측정

2) 상점 유형(store_type_n), 상점 층 등의 지표가 권리금에 영향을 미치는 것을 알 수 있음

» 상점이 속해있는 업종 유형이 권리금 산정에 큰 영향을 미침

결론

상권의 흥망성쇠 및 영향요인 분석

상가가 속한 상권의 흥망성쇠를 예측해본 결과, 상가의 입지적 요소가 많은 영향을 미치는 것을 알 수 있다.

권리금 가격 예측 분석

분석을 통해 파악한 지표를 통해 자영업자들은 해당 변수를 고려하여 상가 선택이 가능하고, 해당 지표의 영향을 참고하여 권리금 거품을 파악할 수 있다.

상권의 흥망성쇠와 권리금 영향요인 비교

상가가 속한 상권이 흥망성쇠를 결정 짓는 중요 요인은 지역의 특성이며, 권리금을 결정 짓는 요소는 해당 상가가 속한 업종이다. 이렇듯 상권의 흥망성쇠와 권리금 영향요인에 차이가 있음을 알 수 있다.

감사합니다.