2023-2 데이터마이닝 팀 프로젝트 보고서

12조
2019122044 김대한
2020122037 김채성
2021122082 백예주
2018122001 전세연

주제: 서울시 상권 분석 서비스 데이터를 활용한 상권별 상권변화지표 예측

[목차]

[Contributions]

김대한 - 결측치 분석, 결측치 처리, 회귀분석 및 Randomforest 모델링, 보고서 작성, 발표
김채성 - 결측치 분석, 파생변수 생성, EDA 실행 및 분석, Decision Tree 및 Random Forest 모델링, 결과 시각화 및 분석, 보고서 작성, 코드 통합 및 정리, PPT 작성
백예주 -데이터 수집, 데이터 전처리, EDA 실행 및 분석, 보고서 작성, 보고서 내용정리 및 형식 통일
전세연 - 데이터 수집, 데이터 전처리, EDA 실행 및 분석, 보고서 작성, 코드 통합 및 정리, 결론 도출, 발표

1. 서론

자영업자들이 사업체 개업을 위한 입지를 선정할 때에는 다양한 인구학적, 지리적 요인을 고려하게 된다. 서울시 내에서는 같은 구나 동 안에 속한 상권이라도 그 위치나 주변 환경에 따라 상권의 특성이 크게 달라지기도 한다. 전통시장이나 상가와 같이 기존 업체들이 장기간 생존하는 상권이 있는가 하면, 소위 '핫플'이라고 말하는 지역에서는 신규 업체들이 끊임없이 들어오고 영업 지속 기간이 짧기도 하다. 상권을 해석할 때에는 물론 인구, 면적, 소득, 소비 등 요인을 종합적으로 고려해야 하지만, 간단한 지표를 사용한다면 사업체의 유입과 존속 여부를 직관적인 기준으로 판단할 수 있을 것이다. 서울시 '상권변화지표'는 각 상권 내의 생존/폐업 사업체의 총 영업 기간을 서울시 전체 평균과 비교한 지표이다. 이를 통해 영업기간을 비교하여 상권의 안정성과 정체 정도를 추정할 수 있다.

상권변화지표는 영업과 폐업 업체의 평균 영업 기간에 따라 H 또는 L로 구분되며, 이를 조합하여 HH, HL, LH, LL 4가지 카테고리로 구성된다.
각 카테고리별 상권의 특성을 다음과 같이 설명된다.

HH	정체	상권 내 사업체의 교체주기가 느리고 오랜 기간 영업한 사업체들이 상업공간을 구성
HL	상권축소	기존 사업체가 입지를 굳건히 하고 있어 신규 사업체의 진입이 어려운 상업공간
LH	상권확장	오래된 사업체가 새로운 사업체로 교체된 상업공간
LL	다이나믹	새롭게 형성된 상권이거나, 상업 공간 내 사업체 교체가 빠른 상권

상권에 새롭게 진입하고자 하는 자영업자나, 기존에 머물던 상권에서 이동을 고려하는 자영업자라면 이 지표를 참고해 본인이 고려하는 상권에서 가질 수 있는 입지를 파악할 수 있다. 4가지 카테고리로 구성되는 이 지표는, 결국 서울 대비 평균 영업 기간을 나타내기 때문에 연속형 변수로도 표현이 가능하다. 카테고리가 아닌 연속형 변수로 나타내면, 그 숫자에 따라 각 상권의 영업 기간이 얼마나 길고 짧은지 알 수 있어 세밀한 분석과 비교가 가능하다.

이 보고서에서는 2019년 4분기부터 2022년 3분기 까지의 상권 데이터를 활용해 상권변화지표를 연속형 변수로 예측하는 모델을 학습시키고, 2022년 4분기의 상권변화지표를 예측하여 모델을 검증하는 것을 목표로 한다. 또한 변수 중요도를 점검하여 어떤 요인이 상권변화지표, 즉 상권의 영업 기간에 큰 영향을 미치는지 알아볼 것이다. 주어진 정보로 상권변화지표를 예측할 수 있다면 사업자는 본인에게 적합한 상권과 위치를 찾을 때 도움을 받을 수 있을 뿐만 아니라 사업체 진입 여부 결정을 위한 상권 분석 시 중점을 두어야 할 요소를 알고 의사결정을 할 수 있을 것이다.

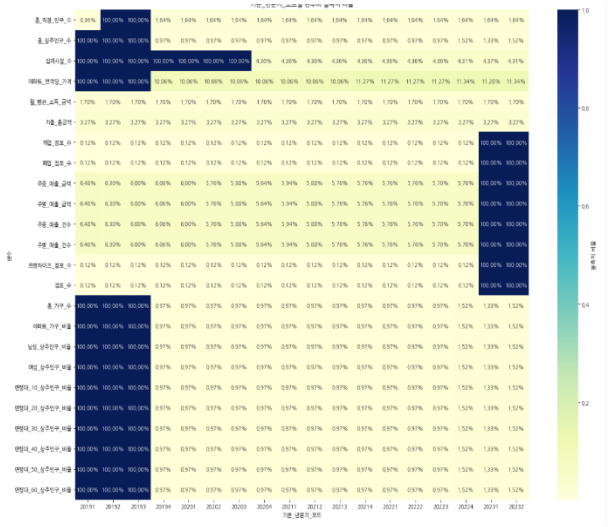
2. 데이터 설명

- 원데이터

'서울시 상권분석서비스' 카테고리에는 총 10개의 데이터가 등록 되어있다. 그 중에서 '영역' 데이터는 자치구가 아닌 상권을 기준으로 분석이 진행되었기 때문에 이를 제외한 9개의 데이터를 활용하고자 하였다. 또한 '영역' 데이터는 해당 상권의 좌표 값을 포함한 데이터로 개별 상권의 특징을 모두 조사하기가 어렵고, 단순히 이를 분석함으로써 의미 있는 결론을 도출하기는 어려울 것이라고 판단했다.

활용한 데이터는 크게 상권 내의 인구 관련 데이터와 상권 내 점포 관련 데이터로 나누어 볼 수 있다. 먼저 인구 관련 데이터로는 각 변수에 대한 성별/연령대별 정보를 담고 있는 '상주인구', '직장인구', '길단위인구'와 병원, 학교 등 인구를 유입시키는 시설의 수에 관한 데이터인 '집객시설', 상권 내 인구의 소득과 분야별 소비에 관한 '소득소비', 마지막으로 아파트의 면적, 가격, 세대 수에 대한 '아파트' 데이터를 활용하였다. 상권 내 점포에 관한 데이터로는 예측의 대상이 된 '상권변화지표'와 개업/폐업/유사업종 점포 수에 관한 데이터인 '점포', 그리고 요일/시간/연령대별 '추정매출'이 있다.

현재 주어진 전체 데이터셋에서, 모든 행에 대해 결측인 데이터 구간을 제거하고 사용하기 위해, 분석하고자 하는 기간을 결측인 변수들이 존재하는 20191~20193, 20231~20232 기간을 제외한 2019년 4분기 ~ 2022년 4분기로 제한하였다(아래 그림).



- **데이터 전처리(가져온 변수, 파생변수,결측치 처리)**

채택된 데이터에 변수가 적게는 9개, 많게는 27개 포함되어 있어 전처리 과정에서 변수들을 집약적으로 포함시키는 데에 집중하였다. 상주인구, 직장인구, 그리고 길단위인구의 경우, 성별과 연령대를 기준으로 세분화된 데이터까지는 필요하지 않다고 판단하여 총 상주인구, 총 직장인구, 총 길단위인구 수만 최종 데이터에 포함했다. 마찬가지로 집객시설의 수의 경우에도 해당 상권에 인구를 밀집시키는 요인을 파악하고자 총 집객시설 수를 사용했다.
데이터를 결합할 때는 모든 데이터셋에 공통적으로 있는 변수인 '기준_년분기_코드'와 '상권_코드'를 기준으로 결합했다. 최종적으로 생성된 데이터의 단위는 분기별 상권이다.

[가져온 변수]

데이터셋	변수
공통 변수 (primary key)	기준 년분기 코드, 상권 코드, 상권 코드 명, 상권 구분 코드, 상권 구분 코드 명
서울시 상권분석서비스(점포-상권)	점포 수, 프랜차이즈 점포 수, 개업 점포 수, 폐업 점포 수
서울시 상권분석서비스(소득소비-상권)	월평균소득금액, 지출총금액
서울시 상권분석서비스(추정매출--상권)	주중매출건수, 주말매출건수, 주중매출금액, 주말매출금액
서울시 상권분석서비스(상주인구-상권)	총 상주 인구수, 성별/연령대별 상주인구수, 총 가구 수
서울시 상권분석서비스(직장인구-상권)	총 직장 인구 수
서울시 상권분석서비스(상권변화지표-상권)	상권변화지표, 운영영업개월 평균, 폐업영업개월 평균, 서울 운영 영업 개월 평균, 서울 폐업 영업 개월 평균
서울시 상권분석서비스(길단위인구-상권)	총 유동인구
서울시 상권분석서비스(집객시설-상권)	집객시설 수

[파생변수 생성]

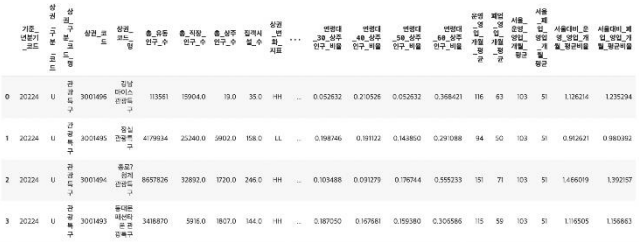
파생변수	변수 설명	원데이터에서 활용한 변수
점포 수(상권 전체 합)	기존 데이터에서 업종별로 주어진 점포수를 한 상권에 대해 모두 합침	점포 수(업종별)
남성 상주인구 비율, 여성 상주인구 비율	남성 상주인구 / 총 상주인구 수 여성 상주인구 / 총 상주인구 수	남성 상주인구 수, 여성 상주인구 수, 총 상주인구 수
연령대 10 상주인구 비율 ~ 연령대 60 상주인구 비율	각 연령대가 해당 상권의 전체 상주인구 중 차지하는 비율	연령대 10(,20,30,40,50,60이상) 상주인구 수, 총 상주인구 수
서울대비 운영 영업 개월 평균비율, 서울대비 폐업 영업 개월 평균비율 => Y변수	상권변화지표인 H,L을 결정하기 위한 지표로, 각 지표가 1 이상이면 서울시 전체보다 평균 영업 개월이 긴 것이므로 H, 1 미만이면 평균 영업 개월이 짧으므로 L로 판단하는 지표임.	운영 영업 개월 평균, 폐업 영업 개월 평균, 서울 운영 영업 개월 평균, 서울 폐업 영업 개월 평균
	'운영 영업 개월 평균 / 서울 운영 영업 개월 평균', '폐업 영업 개월 평균 / 서울 폐업 영업 개월 평균' 으로 계산	

- **결측치 분석 및 처리**

2019 4분기 ~ 2022 4분기의 상권데이터의 각 변수에 대한 결측치 분석과 처리 결과는 다음과 같다.

변수	결측치 개수 (row)	분석 (결측 패턴 등)	처리 방법
총 상주인구 수, 총 가구 수, 성별/연령별 상주인구 비율	217	일부 분기에서만 결측, 모든 분기에서 비슷한 값 가짐	결측이 아닌 다른 분기의 값으로 채움
집객 시설 수	7247	일부 분기만 결측, 모든 분기에서 같은 값 가짐	결측이 아닌 다른 분기의 값으로 채움
월 평균 소득 금액	364	결측인 상권은 모든 분기에서 결측	중앙값으로 채움
지출 총 금액	702	결측인 상권은 모든 분기에서 결측	총 상주인구 수 등 타 변수와 관련됨을 반영해 MICE로 채움
매출금액, 매출건수 관련 변수	2147	결측 패턴 없음, 데이터 패턴 없음	행 drop
점포 수 관련 변수	26	모든 분기에서 결측 (13개 분기, 2개 상권)	매출금액, 매출건수 drop시 같이 drop됨

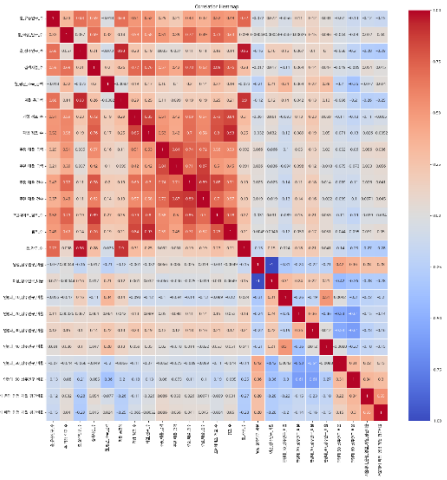
- **최종 데이터: commerce_na1.csv**



행: 19216, 열: 36

3. EDA

1) 변수들의 상관관계 분석

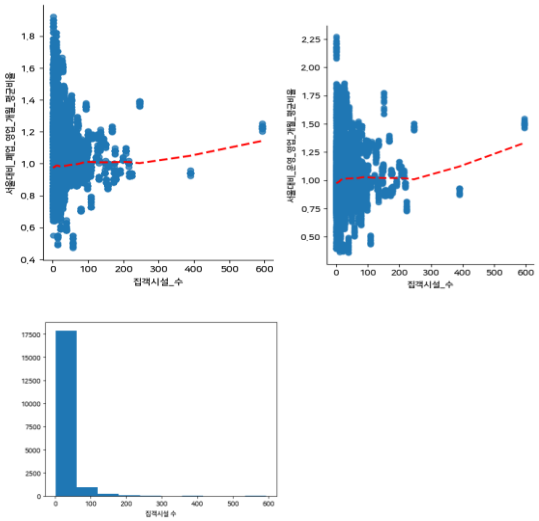


높은 상관관계를 갖는 변수들의 집합을 살펴보면,

- 지출 총금액, 총 가구 수, 그리고 상주인구 수: 해당 분석에서 활용된 지출 총금액은 상권 내 거주한 인구의 총합을 의미하므로 직장인구나 유동인구, 집객시설의 수와 같이 일시적으로 상권에 머무르는 인구가 아닌 상주인구 수와 가구 수의 영향을 받은 것을 볼 수 있다. 그리고 이 변수들 모두 서울 대비 운영 영업 평균 비율과 서울 대비 폐업 영업 개월 평균 비율과 음의 상관관계를 보이고 있다.
- 개업 점포, 폐업 점포, 프랜차이즈 점포, 총 점포수: 아주 높은 상관관계를 보이며, 개업 점포의 개수와 폐업 점포의 개수는 운영개월에도 큰 영향을 미칠 것으로 예측된다.
- 프랜차이즈 점포 수와 직장인구, 주중 매출, 집객시설: 프랜차이즈 점포 수의 경우, 평일의 전체 점포의 수보다 평일 인구 밀도의 영향을 많이 받는 것으로 보인다.

2) 초기 데이터 변수들에 대한 eda

- **데이터 시각화**

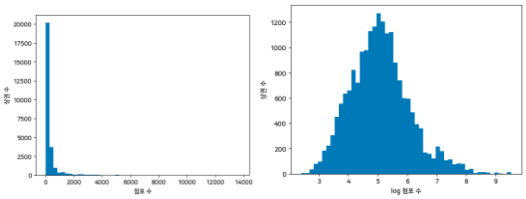


대부분의 상권에서 총 유동인구는 0~100000, 총 직장인구는 0~2000, 총 상주인구는 0~10000, 집객시설의 개수는 0~100의 범위에 분포함을 히스토그램을 통해 확인할 수 있었다. 네 변수 모두 상권에 따라 이상치를 가지는 것도 확인되며, 모두 왼쪽으로 치우친 분포를 보인다. 집객시설의 경우 400대와 600대에 분포 되어있는 이상치를 가지고 있다. 금액 관련 변수들에 대하여 정규화를 시키고 분포를 출력해보았을때 전체적으로 right-skewed인 분포를 가지는 것을 볼 수 있었다.

각 변수와 Y변수의 상관관계를 파악하고자 모든 변수에 대하여 '상권변화지표'를 기준으로 박스플롯을 그려 보았다. 그 결과, 상주인구의 분포가 육안으로 가장 뚜렷한 차이를 보여 Y변수에 유의미한 영향을 줄 것이라고 추측했고, 집객시설 수 또한 HH와 LL의 지표에서 뚜렷한 이상치를 가지므로 유의미한 변수일 것으로 예상했다. 집객시설은 운영 영업 평균 개월수 & 폐업 영업 평균 개월수와도 양의 상관관계를 가진다. 지출총금액도 마찬가지로 박스플롯의 각 범주에서 뚜렷하게 차이를 나타냈으므로 Y변수에 미치는 영향을 기대해 볼 수 있다. 반면 월평균소득금액, 주중 매출 금액 등은 박스플롯에서 큰 차이를 보이지 않아 상권변화지표에 큰 영향을 미치지 않을 것이라고 판단했다.

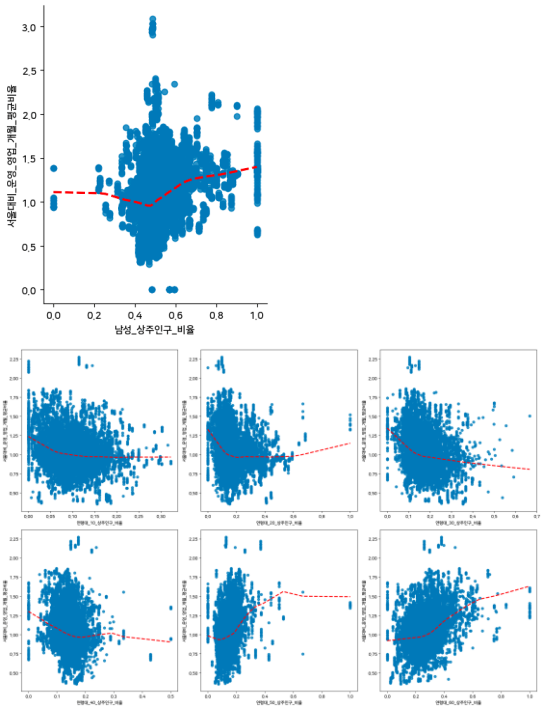
각 변수들과 Y 변수로 사용할 '서울대비_운영_영업_개월_평균비율' 과 '서울대비_ 폐업_영업_개월_평균비율'의 상관 관계를 파악하고자 scatter plot을 그려보니 위에서 내린 결론들과 비슷한 결과를 보였다.

추가로 log 변환을 통해 정규분포에 가깝게 만들 수 있는 변수들이 여럿 존재했다. 총 유동인구, 총 직장인구, 총 상주인구, 총 가구 수, 집객시설, 점포수, 매출건수, 매출금액 등의 변수들은 아래와 같이 왼쪽으로 치우친 분포를 가질뿐더러 박스플롯으로부터 상관관계를 관찰하기가 어려웠다. 이에따라 언급된 변수들에 대해 회귀분석을 시행하기 위해서는 log 변환이 필수적임을 알 수 있다.



3) 성별/연령별 상주인구 비율에 대한 추가 eda

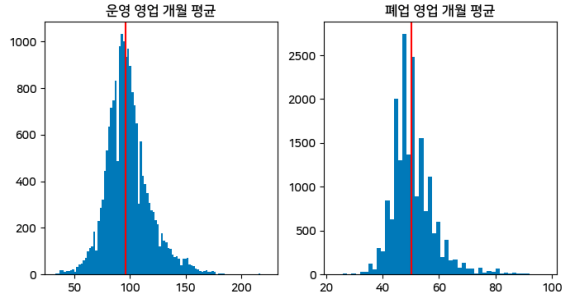
성별과 연령별 상주인구 비율은 Y변수와의 상관관계를 볼 수 있다.



남성 상주인구 비율과 서울 대비 평균 영업 기간은 양의 상관관계를 가진다. 50대, 60대 상주인구 비율은 다른 연령대의 상주인구 비율과 다르게, 상권의 평균 영업 기간과 양의 상관관계를 가진다. 영업, 폐업 기준 두 변수 모두에 대해 동일한 양상을 띤다.

4) 운영 영업 개월, 폐업 영업 개월에 대한 eda, 서울시 평균과의 비교

운영 영업 개월 평균과 폐업 영업 개월 평균은 서울시 평균값을 중심으로 전체적으로 정규분포의 형태를 보이는데, 약간 right-skewed 된 모양을 가진다.



4. 모델링

상권 변화 지표 (HH,HL,LH,LL)는 각 상권의 생존 업체 영업 개월 평균과 폐업 업체 영업 개월 평균을 서울시 평균과 비교함으로써 결정된다. 이를 단순히 H 또는 L로 표현하지 않고 숫자 지표로 표현해 서울시 평균의 영업 개월 수와 비교하기 위해,

우리는 데이터 전처리 과정에서 '서울대비_운영_영업_개월_평균비율', '서울대비_폐업_영업_개월_평균비율'이라는 새로운 변수를 생성했다. 이 두가지 변수를 y변수로 설정하고, 동일한 X변수들을 가지고 두가지 y를 예측하는 회귀문제를 설정했다.

Linear Regression, Decision Tree Regression, Random Forest Regression의 방법으로 모델링을 진행했다. 본 분석의 1차적인 목표는 회귀 모델링을 통한 두가지의 Y변수 예측이고, 2차적인 목표는 예측된 Y로 상권변화지표 4가지로 분류해 다중분류 문제로 전환하는 것이다.

본 데이터는 분기별 데이터로, 2019 4분기 ~ 2022 4분기의 데이터를 사용했다. 주어진 데이터에서 가장 최신 분기의 상권 변화 지표를 예측하는 것을 목표로 하여, 2022 3분기까지의 데이터를 train data, 2022 4분기 데이터를 test 데이터로 사용했다. train data는 17739개, test data는 1477개이다.

각 모델별 분석 방법과 결과

1) 선형 회귀

모델 선정 이유: 가구 수와 점포 수, 매출액, 인구수 등에 따라 영업개월수와 몇몇 변수 사이에 선형관계가 있음을 확인하였다. 선형회귀분석은 간단하고 해석이 쉬우며, 추정된 계수와 의미를 해석할 수 있고, 각 변수의 영향을 파악할 수 있다. 상권 변화 지표에 영향을 미치는 요인을 분석하는 데 기초적이고 강력한 방법이라고 생각하여 선택하였다.

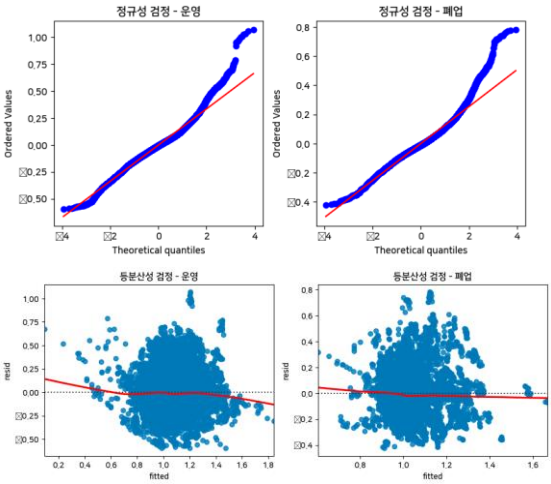
모델링 과정: 2개의 y변수에 대해 다음과 같은 절차로 모델링을 진행하였다.

a. 모든 변수를 사용한 Default model
   i) 낮은 R-squared 값과 잔차 문제 존재
   ii) 금액과 인구수 변수의 숫자가 커서 문제 발생
   iii) 로그변환(인구수, 지출/매출금액, 소득금액, 가구수, 점포수)
b. 잔차분석을 통해 변수간 관계성을 추가한 모델
   i) 연령대와 상권 구분별 잔차 확인 후 변수 추가
c. Feature selection을 통해 변수를 최종 선택한 모델

사용한 변수:

운영_영업개월 평균비율	폐업_영업개월 평균비율
'상권_구분_코드_명_전통시장', '총_상주인구_수', '연령대_60_상주인구_비율', '주중_매출_금액', '주말_매출_건수', '개업_점포_수', '집객시설_수', '프랜차이즈_점포_수', '연령대_40_상주인구_비율', '연령대_30_상주인구_비율', '점포_수', '지출_총금액', '연령대_10_상주인구_비율', '총_직장_인구_수', '주말_매출_금액', '여성_상주인구_비율', '총_유동인구_수', '상권_구분_코드_명_관광특구', '상권_구분_코드_명_발달상권', '총_가구_수', '연령대_30_상주인구_비율:상권_구분_코드_명_전통시장', '주중_매출_건수', '폐업_점포_수'	'총_상주인구_수', '연령대_60_상주인구_비율', '주중_매출_금액', '총_직장_인구_수', '주중_매출_건수', '여성_상주인구_비율', '월_평균_소득_금액', '연령대_20_상주인구_비율', '프랜차이즈_점포_수', '집객시설_수', '상권_구분_코드_명_관광특구', '연령대_30_상주인구_비율:상권_구분_코드_명_전통시장', '연령대_50_상주인구_비율', '개업_점포_수', '점포_수', '남성_상주인구_비율', '폐업_점포_수'

최종 회귀모델 c에 대해 정규성 검정과 등분산성 검정을 수행했다.



Shapiro test 결과 p-value가 0.05보다 작아 정규성이 부합한다고 할 수 있다. 등분산성은 잘 만족하진 못하나 Default model(모델 a)의 잔차에 비해 개선되었다.

선형회귀 대한 MAPE값은 아래와 같다.

MAPE	영업		폐업	
모델	train	test	train	test
(c)	0.1284	0.1297	0.0939	0.0955

2) 회귀 나무

모델 선정 이유: EDA로 변수들과 Y변수의 관계를 살펴본 결과 선형적인 관계를 갖지 않는 변수가 다수임을 알게 되었다. 선형회귀와 다르게, Decision Tree는 비모수적 모형이기 때문에 비선형성 또한 고려가 가능하다. 또한 변수의 중요도를 분석하기 쉬우며 좋은 설명력을 가지기 때문에, 상권 변화 지표에 영향을 미치는 요인을 분석하기 용이할 것이라고 생각하여 선택했다.

모델링 과정: 2개의 y변수에 대해, 다음과 같은 3 차례의 회귀나무 모델링을 진행했다. Baseline인 Default 모델 a에서 성능을 향상시키기 위해 모델 b와 c과 같이 변형시켰다.

a. 전체 변수를 사용한 Default 모델 (max_depth = 3)
b. 전체 변수를 사용하고, GridSearchCV를 통해 얻은 최적 하이퍼파라미터 조합을 활용한 모델
c. EDA 및 모델 (a), (b)로 도출된 변수 중요도를 바탕으로 선택한 변수만을 활용한 모델

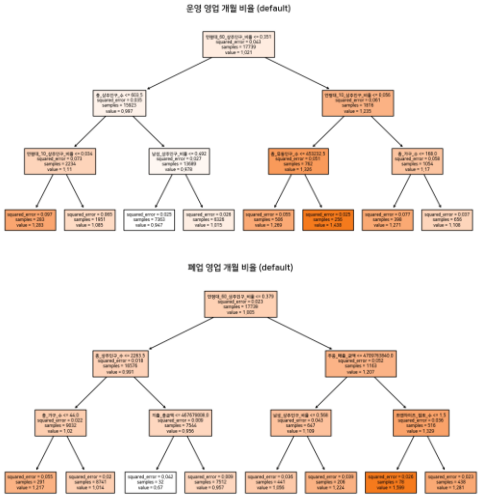
사용한 변수:

a,b,c 모두에서 사용된 변수 (EDA 및 변수 중요도를 바탕으로 선택된 변수)	총 유동인구 수, 총 직장 인구 수, 총 상주인구 수, 집객시설 수,지출 총금액, 개업 점포 수, 폐업 점포 수, 점포 수, 프랜차이즈 점포 수, 주중 매출 건수, 총 가구 수, 남성 상주인구 비율, 연령대 10 상주인구 비율, 연령대 20 상주인구 비율, 연령대 30 상주인구 비율, 연령대 40 상주인구 비율, 연령대 50 상주인구 비율, 연령대 60 상주인구 비율
a, b에서는 사용됐지만 c에서는 제거된 변수	월 평균 소득 금액, 주말 매출 금액, 주중 매출 금액, 주말 매출 건수(EDA 결과 큰 상관관계가 보이지 않아 제거) 여성 상주인구 비율 (남성 상주인구 비율과 완벽한 반비례 관계이므로 제거)

a) 전체 변수를 사용한 Default 모델 (max_depth = 3)

Default model에서는 원래 하이퍼파라미터를 따로 설정하지 않지만, 과적합 및 시간이 오래 걸리는 문제로 인해 max_depth만 3으로 제한하였다.

- 서울대비 운영 영업 개월 평균비율, 서울대비 폐업 영업 개월 평균비율에 대한 의사결정나무

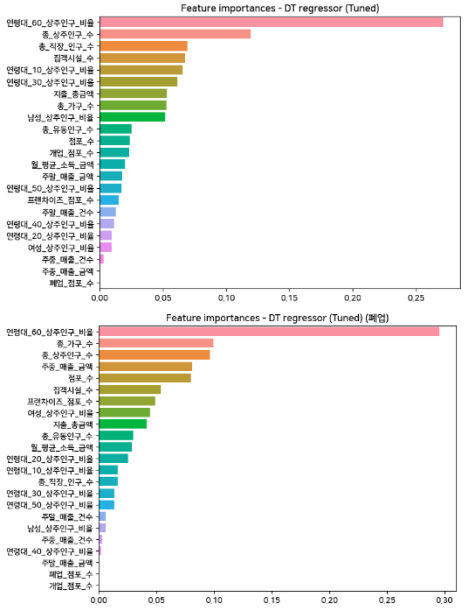


Default 하이퍼파라미터 하에 생성된 트리모델의 구조를 보면, 운영 점포의 경우에는 60대 이상 상주인구 비율, 10대 상주인구 비율, 총 상주인구 수, 남성 상주인구 수, 총 유동인구 수, 총 가구 수가, 폐업 점포의 경우 60대 이상 상주인구 비율, 총 상주인구 수, 주중 매출 금액, 총 가구 수, 지출 총금액, 남성 상주인구 비율, 프랜차이즈 점포수가 노드에 활용되었다. 두 경우 모두 공통적으로 인구와 관련된 변수들이 상권내 점포의 평균 영업 기간과 관련이 크다고 해석할 수 있다.

b) GridSearch를 통한 hyperparameter 튜닝

Cross validation Grid Search를 통해 ccp_alpha, min_impurity_decrease, min_samples_split, max_depth에 대한 튜닝을 진행했다. 튜닝 결과, max_depth는 설정한 범위 중 가장 큰 값이 최적의 값으로 나왔고, 다른 하이퍼파라미터는 default 값이 최적값으로 도출되었다. max_depth만 default보다 큰 7로 변경 후 동일하게 모델을 fitting시켰다.

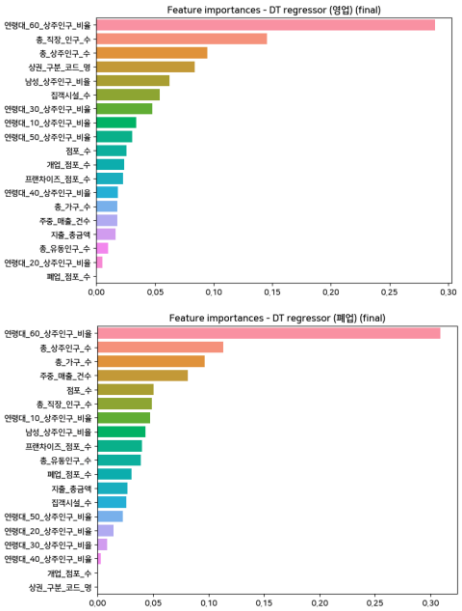
모델(b)의 변수 중요도 결과는 다음과 같다. 각각 영업 점포, 폐업 점포의 서울 대비 영업 기간에 대한 모델이다.



변수 중요도를 살펴보면, 여전히 60대 상주인구비율이 중요한 변수임을 알 수 있다. 인구 관련 변수와 더불어, 집객시설 수, 점포 수, 지출 총금액 등의 변수도 큰 영향력을 가진다.

c) 변수 선택 후 실행한 모델

(2)와 동일하게 하이퍼파라미터를 max_depth=7로 설정하고, 표에 기재한대로 변수를 삭제한 후 모델을 재실행했다.



회귀나무로 예측했을 때 MAPE값은 다음과 같다.

MAPE	영업		폐업	
모델	train	test	train	test
(1)	0.1410	0.1452	0.0954	0.0989
(2)	0.1126	**0.1429**	0.0819	0.0975
(3)	0.1151	0.1478	0.0818	**0.0959**

3) 랜덤포레스트

예측력 향상을 목적으로, 랜덤포레스트 모델을 활용했다. 선형회귀와 의사결정나무를 통해 변수의 중요도와 영향력을 파악할 수 있다면, 랜덤포레스트는 배깅의 한 종류로 의사결정나무를 결합해 예측 성능을 향상시킬 수 있다.

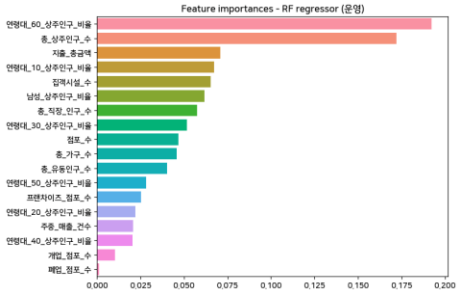
Cross Validation Gridsearch를 통해 최적의 ccp_alpha, min_samples_splits, max_depth를 찾아보았다. 의사결정나무의 하이퍼파라미터 결과와 유사하게, max_depth는 설정한 범위 중 최대의 값이 최적의 하이퍼파라미터로 선정되었고, 다른 하이퍼파라미터는 default 값이 선정되었다.

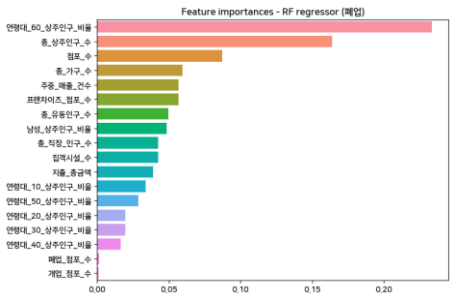
따라서 max_depth=8로 설정하여 RandomForestRegressor를 두가지 Y변수에 대해 학습시켰다.

이에 대한 MAPE는 다음과 같다.

MAPE	영업		폐업	
모델	train	test	train	test
RF	0.0910	**0.1153**	0.0731	**0.0852**

랜덤포레스트로 도출한 변수중요도는 아래와 같다. 의사결정나무와 동일하게, 가장 중요한 변수는 연령대 60 상주인구 비율이고 총 상주인구 수가 두 변수에서 모두 중요한 것으로 도출되었다. 그 외의 변수들은 두 Y변수에서 서로 다른 중요도를 보였다.





전체 성능 비교

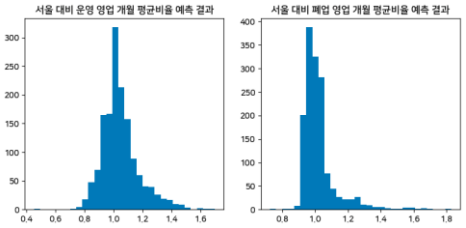
종합적인 예측 성능은 MAPE(Mean absolute percentage error)로 비교했다.

MAPE	영업			폐업		
	train	test	R2	train	test	R2
Linear Regression	0.1284	0.1298	0.328	0.0939	0.0955	0.261
Decision Tree Regression	0.1133	0.1427	0.07	0.0818	0.0955	0.226
Random Forest Regression	0.0910	**0.1153**	**0.3357**	0.0731	**0.0852**	**0.4023**

예상과 같이, 예측 오류는 Random Forest에서 가장 작았고 결정계수도 가장 커서 예측 성능이 가장 크다.

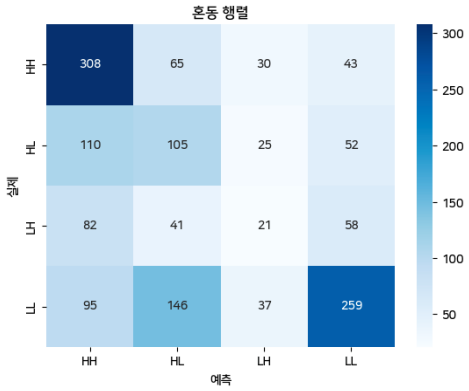
5. 결과 분석

결과적으로 모델링 예측 오류가 가장 적었던 Random Forest 모델을 사용하여 2022년 4분기의 서울대비 영업 개월 평균 비율을 예측한 결과를 아래의 히스토그램으로 나타냈다. 실제 값은 서울시내 상권들과 서울 전체 평균을 비교하는 값이므로 1을 중심으로 고르게 퍼진 분포를 가진다. 예측 결과 값의 분포는 두 y변수에 대해 모두 right-skewed 형태를 보인다. 영업 기간이 서울시 평균보다 짧은 경우에는 서울 평균과 차이가 덜 나는 0.8~1.0 사이 구간에 주로 분포하는 반면, 서울시 평균보다 긴 경우에는 1.0~1.8 사이 구간까지 넓게 분포한다. 총 1477개 상권 중, 운영중인 점포는 H가 525, L이 952개, 폐업 점포는 H가 769개, L이 708개로 예측되었다.



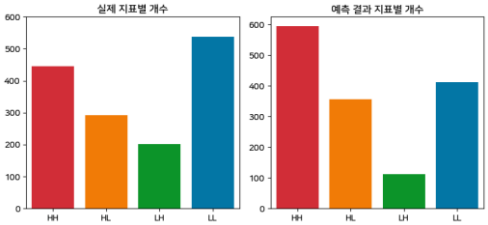
예측한 두가지 비율을 다시 상권변화지표 (HH,HL,LH,LL) 로 변환해 다중분류 accuracy를 계산해보았다. 2022년 4분기 데이터에 대해 예측한 '서울대비 운영 영업 개월 평균비율'과 '서울대비 폐업 영업 개월 평균비율'이 각각 1이 넘으면 H, 그렇지 않으면 L로 변환하여 4가지의 지표로 만든 후, 그에 대한 정확도를 평가해보았다.

가장 예측 오류가 작았던 랜덤포레스트의 예측 결과로 만든 confusion matrix와 f1 score이다.



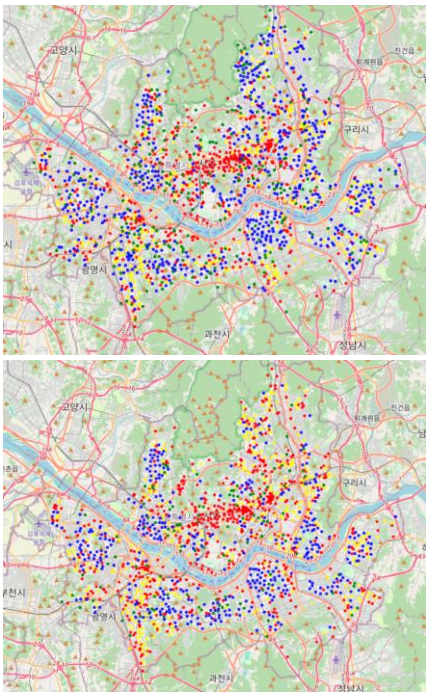
	precision	recall	f1-score	support
HH	0.52	0.69	0.59	446
HL	0.29	0.36	0.32	292
LH	0.19	0.10	0.13	202
LL	0.63	0.48	0.55	537
accuracy			0.47	1477
macro avg	0.41	0.41	0.40	1477
weighted avg	0.47	0.47	0.46	1477

HH와 LL에 해당하는 상권 수가 많기 때문에 해당 범주들의 정확도가 높게 나타나는 것을 볼 수 있다. 그러나 전체적인 f1-score는 0.4를 웃돌고, 이 모델의 예측력이 좋다고 말하기는 어렵다.



예측 결과 및 실제 지표별 개수 그래프와 혼동 행렬을 살펴보았을 때, 실제보다 HH와 HL에 해당하는 상권은 많게, LL과 LH에 해당하는 상권은 적게 예측됨을 알 수 있다. 즉, 운영중인 점포의 평균 영업 기간은 길게, 폐업한 점포의 평균 영업 기간은 짧게 예측이 된다.

2022년 4분기의 실제 지표와 예측된 지표를 서울시 지도 위에 표현해보았다.



(위 – 실제 지표, 아래 – 랜덤포레스트로 예측한 지표)

(red - HH, yellow - HL, green - LH, blue - LL)

예측결과를 보면, 중구와 종로구, 용산구(한강변 일대)에는 HH, 강남구와 서울서부에는 LL이 주로 밀집되어 있는 특징이 잘 반영되었다. 그러나 서울 북동쪽과 남동쪽, 강남구 북쪽 등의 지역의 실제로 LL인 상권들이 HL 등으로 잘못 예측되었다. LL은 전체 상권 중 가장 많은 수를 차지하는데, 상권의 교체가 빠른 지역으로 신규 사업체 진입 시 신중해야한다. 그런데 예측 결과 실제보다 적은 수를 LL로 예측하고, 특히 HL로 잘못 예측하는 경우가 가장 많다. HL인 상권의 특징은 기존 사업체들이 굳건하여 신규 업체의 존속이 어렵다는 것으로, 이 지표 또한 신규 사업체 진입 시 주의가 필요함을 암시한다. 실제 지표가 HL 또는 LH인 상권에서의 오분류는 주로 HH에서 발생하며 이는 정체된 상권을 의미한다.

상권 변화 지표에 관심이 많은 신규 업체의 관점에서는, 상권 확장의 의미를 갖는 LH 상권에 관심이 많을 것이다. 그러나 LH에 해당하는 상권은 실제로도 그 수가 적고, 그마저도 오분류되는 경우가 많이 발생했다. 따라서 신규 업체 진입 입지 선정을 위해 이 모델을 사용하려고 할 경우, 신규 업체에게 유리한 상권을 찾기보다는, 진입에 유의해야할 상권(특히 HL, 운영 점포의 영업 기간이 길고 폐업 점포의 영업 기간이 짧은 상권)을 알아보는 보수적인 관점으로의 활용이 더 적합할 것이다.

랜덤포레스트와 의사결정나무가 도출한 변수 중요도를 통해 상권 변화 지표에 영향을 주는 요인을 살펴보면, 공통적으로 60대 이상 상주인구 비율과 총 상주인구 수가 중요함을 알 수 있다. 60대 이상 상주인구 비율과 영업기간은 양의 상관관계를 가지고, 총 상주인구수와 영업기간은 음의 상관관계를 갖는다.

6. 결론

결론적으로, 상권별 상권변화지표 예측 시스템은 창업 전 진입하고자 하는 상권의 변화를 가늠하는 데 도움이 될 수 있으며, 더 나아가 분석을 통해 취약 상권 개선 방안을 도출해내는 데에 쓰일 수 있다. 데이터가 분기별로 업데이트 되기 때문에, 각 상권의 분기별 변화 경향성도 분석할 수 있을 것이다. 그러나 모델에 여전히 오분류가 존재하기 때문에 특정 점포 진입 시 참고하기 보다는 단기간 운영하는 홍보 목적의 팝업 스토어나 복합 매장 진입 시에 참고하기에 더욱 적합 할 것이라고 생각한다. 또한 위에서 언급한 것과 같이 신규업체가 진입할 시장을 찾기보다는 진입에 주의해야할 상권을 탐색할 때 활용하는 것이 더 적절한 모델이다.

상권변화지표에 영향을 주는 변수에는 대표적으로 연령대별 상주인구 비율과 총 상주인구 수가 있었다. 상주인구가 많을수록 상권 교체가 빨라 영업기간이 짧고, 60대 이상 인구 비율이 클수록 상권 교체가 느려 영업 기간이 길다고 볼 수 있다. 이와 같은 관계를 알고 있다면 개별 인구통계학적 요소들을 고려하여 상권 입지를 선정할 수 있을 것이다. 예를 들어, 신규 업체는 연령대가 높은 상권에 진입하기 무리일 것이라고 판단하거나, 상주인구가 많은 곳에 진입하려면 멀지 않은 시점에 점포를 옮기는 것을 염두에 둘 수 있다. 이처럼 개별 요소에 의거한 판단도 필요하겠지만 모델을 활용하면 종합적인 상권 특성을 간결하게 파악할 수 있다.

상권변화지표라 하더라도 각 상권의 개별적인 특성을 모두 반영하여 일반화할 수 있는 수치는 아니기에 목적에 따른 적절한 활용을 할 수 있어야 한다. 그리고 이러한 예측 모델은 한 상권 전체의 회전율을 빠르게 검토할 수 있다는 장점을 가지지만 특정 점포의 입지 선택에 있어서는 업종과 상권 전반의 특성을 충분히 반영하지 못한다는 점에서 한계가 있다. 가령 정체 상권(HH)의 경우 신규 업체의 진입이 쉽지 않다고 알려져 있으나, 완전히 새로운 업종이나 특정 업종에 대해서는 예외가 있을 수 있으며 그 업종만 대상으로 지표를 측정한다면 다른 등급으로 분류될 수도 있다. 뿐만 아니라 우리가 사용한 2020~2022년 데이터는 코로나라는 특수 상황이 반영되어 있는 것처럼, 사회적, 경제적 특이상황이 존재해도 모델이 완벽하게 반영하기는 힘들 것이다. 이에 대한 해결책으로 업종 단위의 데이터셋과 소득소비 데이터의 분야별 지출 데이터를 함께 활용한다면 더 세부적이고 정확한 분석이 이루어질 수 있을 것이라고 생각한다. 더욱 정확한 예측을 위해서는 다양한 데이터의 활용과 여러 변수를 고려한 분석 모델의 개발이 필요하며, 다양한 데이터 소스와의 통합을 통해 상권의 복잡한 특성을 더욱 정확하게 반영할 수 있는 방향으로 나아가야 할 것이다.

참고문헌

[1] 최은준, 천상현, & 이수기. (2021). 사업체의 생존·폐업 기간을 활용한 서울시 상업공간의 변화분석. 한국지역학회지 『지역연구』, 37(4), 3-19. https://koreascience.kr/article/JAKO202100753165594.page