서울특별시 역전세 위험요인 분석 프로젝트

Code States AIB 17기 Estate-3팀
권순범 김서현 김현섭 송지연

목 차

I. 서론

1. 프로젝트 배경

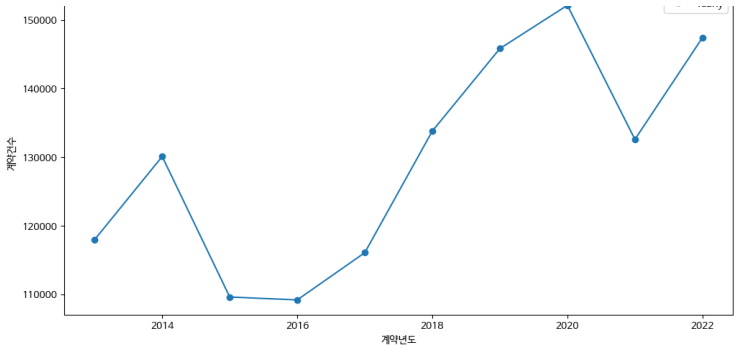
부동산 시장은 많은 사람들에게 주거 또는 투자의 기회를 제공하는 중요한 부분인데, 최근 몇 년간 역전세라고 불리는 현상이 사회 문제로 부각되고 있다. 여기서 역전세란 전세 계약 시점보다 만기에 전셋값이 떨어진 상태를 의미한다. 이로 인해 임차인은 보증금을 돌려받지 못할 수 있으며, 임대인의 부정한 행위에 대한 제재가 부족하여 피해를 보는 경우가 최근 다수 발생하면서 큰 사회적 문제로 대두되고 있다.

역전세는 주거 안정성과 임대인과 임차인 간의 신뢰 문제를 야기하여 부동산 시장의 건전한 발전에 부정적인 영향을 미치고 있다. 이러한 문제를 해결하기 위해 필요한 것은 역전세 리스크를 예측하고 이를 줄일 수 있는 방법을 제시하여 잠재적인 피해자들이 부동산 계약을 안전하게 체결할 수 있는 환경을 조성하는 것이다.

본 프로젝트는 서울특별시의 아파트와 오피스텔의 역전세에 대한 위험률을 예측하고, 이를 기반으로 잠재적인 피해자들을 돕기 위해 추진했다. 본 프로젝트는 데이터 분석과 머신 러닝 기술을 활용하여 부동산 시장과 관련된 다양한 변수를 고려하여 역전세 리스크를 평가했으며, 이를 통해 사용자들에게 각 특성을 고려한 역전세 위험 가능성에 대한 정보를 제공하여 그 피해를 최소화하고자 한다.

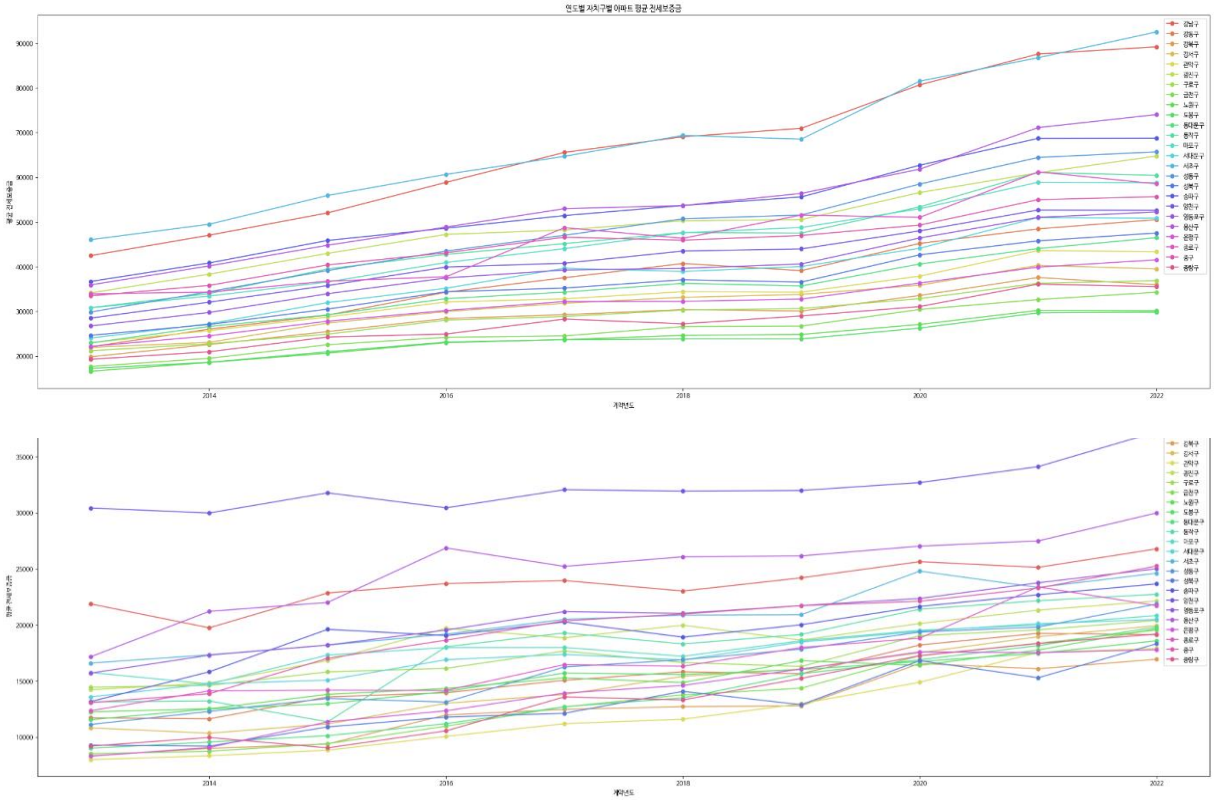
2. 현황 분석 (EDA)

현재 서울특별시 아파트와 오피스텔에 대하여 조사해본 결과를 탐색적 데이터 분석을 통해 나타내면 다음과 같다.



<그림 1> 연도별 계약건수

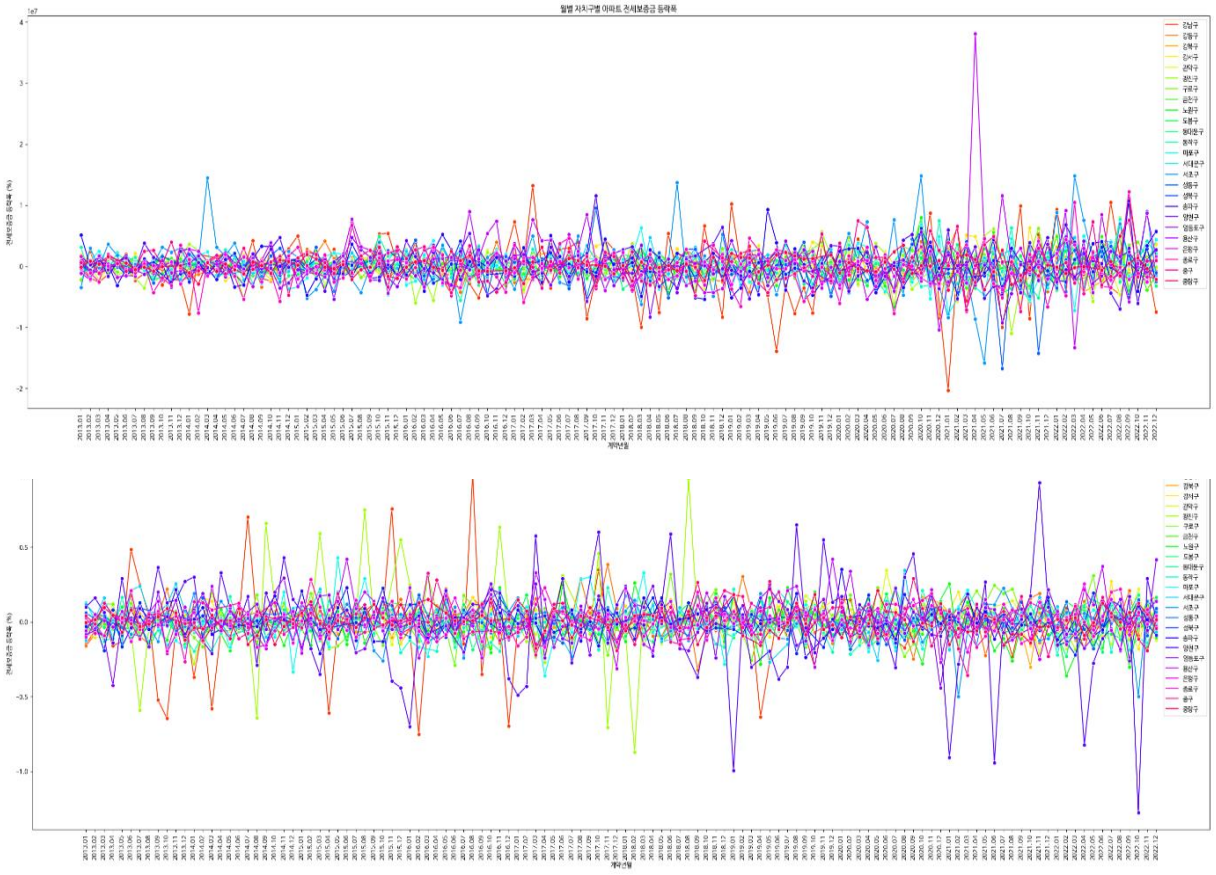
2013년부터 2022년까지의 서울특별시의 아파트/오피스텔 전체 전세 계약건수를 확인해본 결과(<그림 1>), 2014년에 증가하다가 2015년부터 2016년도까지 크게 감소했으며, 2017년부터 큰 폭으로 증가하다, 2021년 다시 하강, 2022년에 다시 증가한 것을 확인할 수 있다.



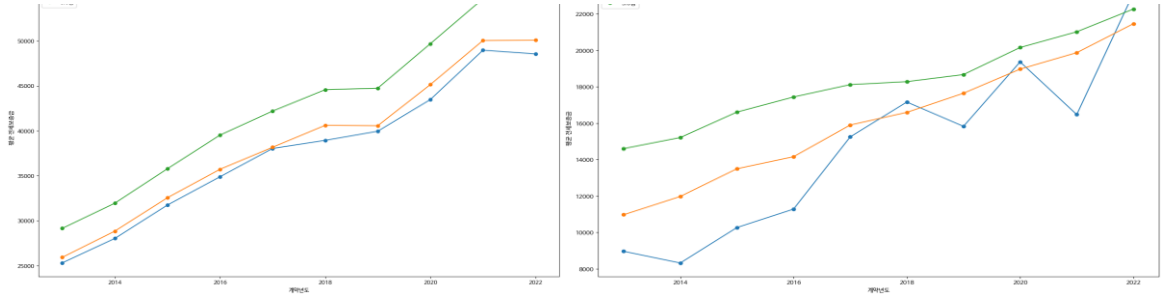
<그림 2> 연도별 자치구별 아파트/오피스텔 평균 전세보증금

연도별로 각 자치구마다 아파트 평균 전세보증금을 확인해보았을 때 (<그림 2>), 서초구와 강남구의 경우 타 자치구보다 전세보증금이 높았다는 것을 확인할 수 있으며, 도봉구와 금천구의 경우 전세보증금이 가장 낮은 것을 확인할 수 있다. 서초와 강남 다음으로 용산과 송파가 전세보증금 순위권을 이루었으며, 전 자치구 모두 증가하는 추세에 있다가 2022년부터 그 증가폭이 줄어드는 것을 확인할 수 있다. 오피스텔 전세보증금의 경우 아파트와 비교했을 때 큰 차이가 존재했는데, 송파구가 타 자치구에 비교해 훨씬 높은 전세보증금을 보이고 있으며, 그 다음으로 용산과 강남이 차지하고 있다는 것을 확인할 수 있다. 이에 반해 관악구나, 강북구, 성북구의 경우 전세보증금이 다른 자치구에 비해 낮은 것을 확인할 수 있다.

계약 연월을 기준으로 전세보증금의 등락폭 (<그림 3>)을 확인한 결과, 아파트는 2021년 3월경에 용산구의 전세보증금이 급격하게 상승하고, 강동구의 경우 등락폭이 타 자치구에 비해 큰 것을 확인할 수 있다. 또한 2021년과 2022년에 서초구와 성동구, 용산구의 전세보증금이 큰 수준으로 하락했다는 것을 확인할 수 있다. 오피스텔의 경우 각 자치구마다 비교적 등락의 차이가 크게 존재했는데, 광진구, 강동구, 양천구의 등락폭이 매우 컸으며, 특정 시기와 상관없이 모든 자치구가 항상 등락폭이 컸음을 확인할 수 있다.



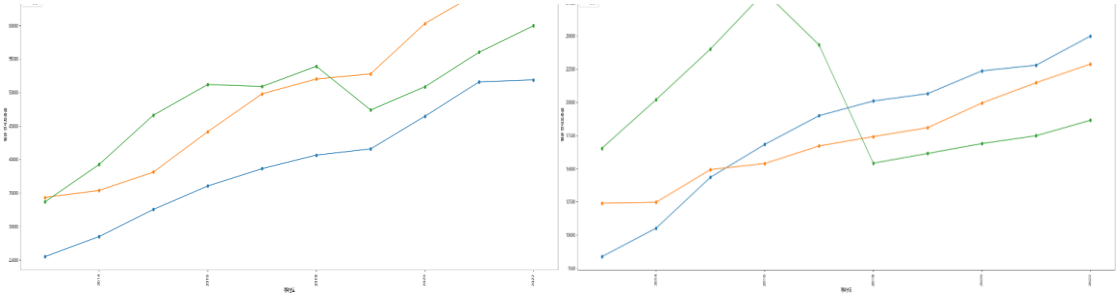
<그림 3> 월별 아파트/오피스텔 전세보증금 등락폭



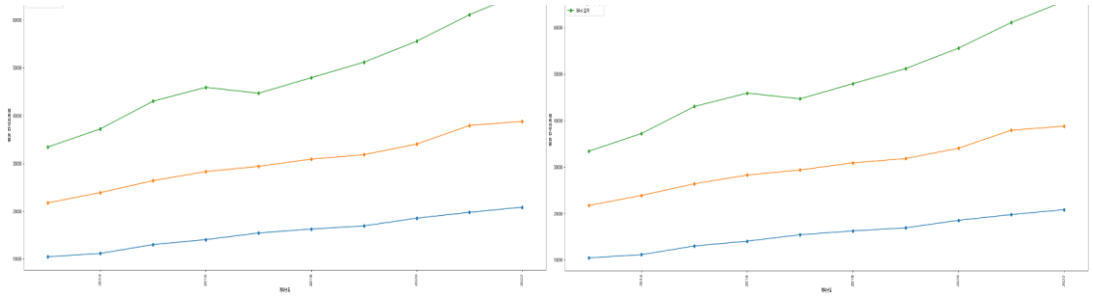
<그림 4> 연도별 아파트/ 오피스텔 층별 평균 전세보증금

(<그림 4>) 아파트와 오피스텔 모두 층이 높을수록 전세보증금이 높았으나, 오피스텔의 경우 예외적으로 1층이 고층에 비해 가격이 높아지는 시기가 잠깐 존재했다.

건물의 신축 여부별로 전세보증금을 확인해본 결과(<그림 5>), 아파트의 경우 2018년까지 전반적으로 준신축, 신축, 구축의 순으로 가격이 높았으나 그 이후로 신축, 준신축, 구축 순으로 가격이 형성되었고, 오피스텔의 경우에는 2016년도와 2017년도에는 구축이 신축보다 가격이 높아지기 시작해 구축, 신축, 준신축의 순으로 가격이 형성되었다.

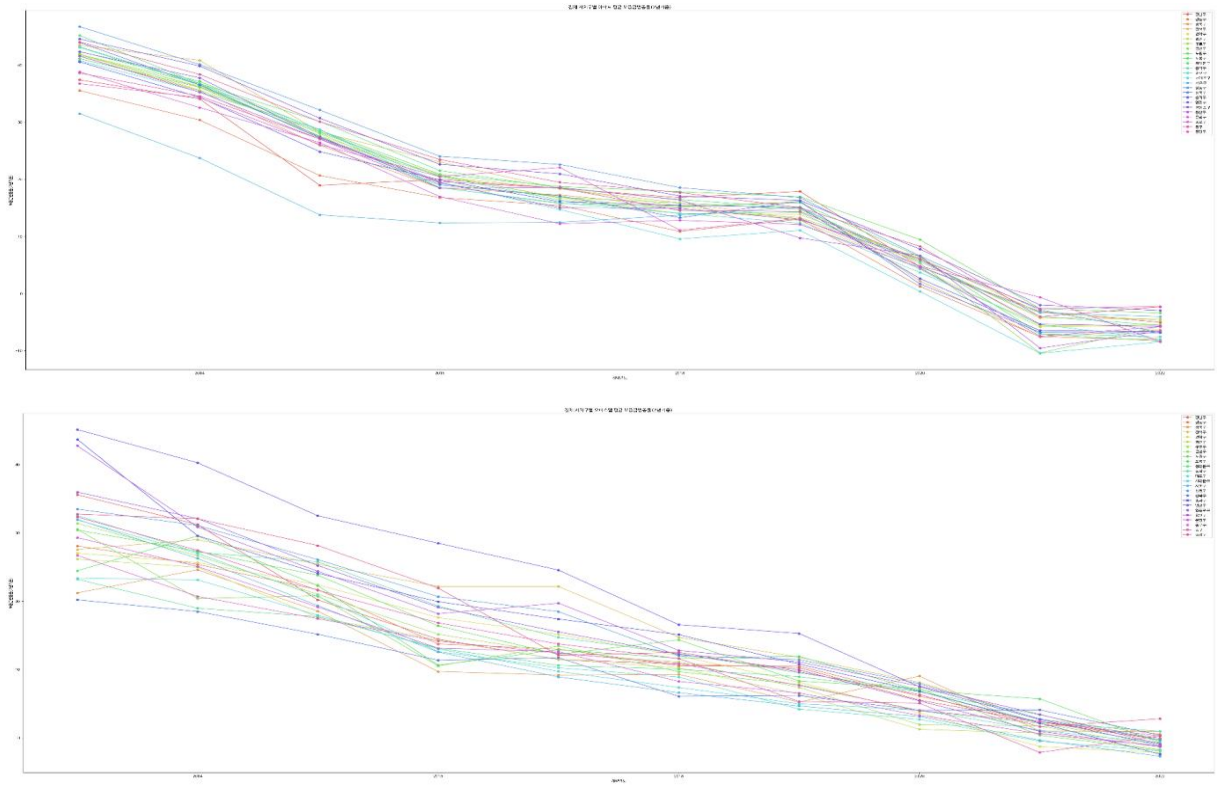


<그림 5> 연도별 아파트/오피스텔 신축여부별 평균 전세보증금



<그림 6> 연도별 아파트/오피스텔 면적구분별 평균 전세보증금

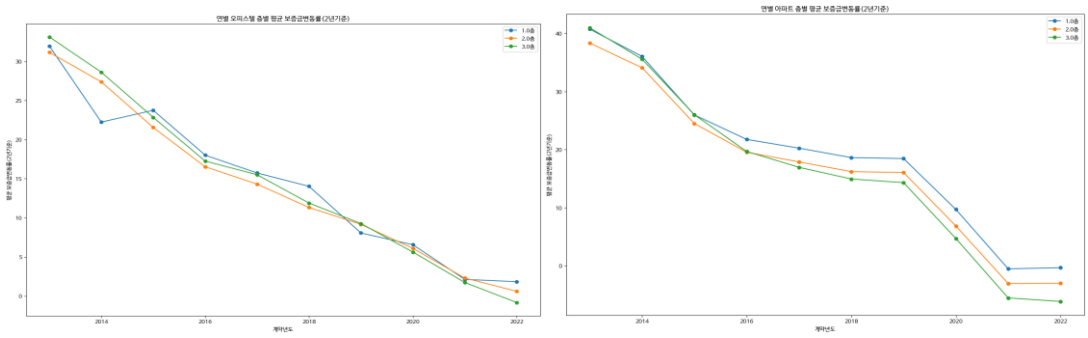
면적 별로 아파트와 오피스텔의 전세보증금을 확인해본 결과(<그림 6>) 면적이 클수록 전세보증금의 가격이 높아, 큰 특징이 없음을 확인했다.



<그림 7> 연도별 아파트/오피스텔 자치구별 보증금변동률(2년기준)

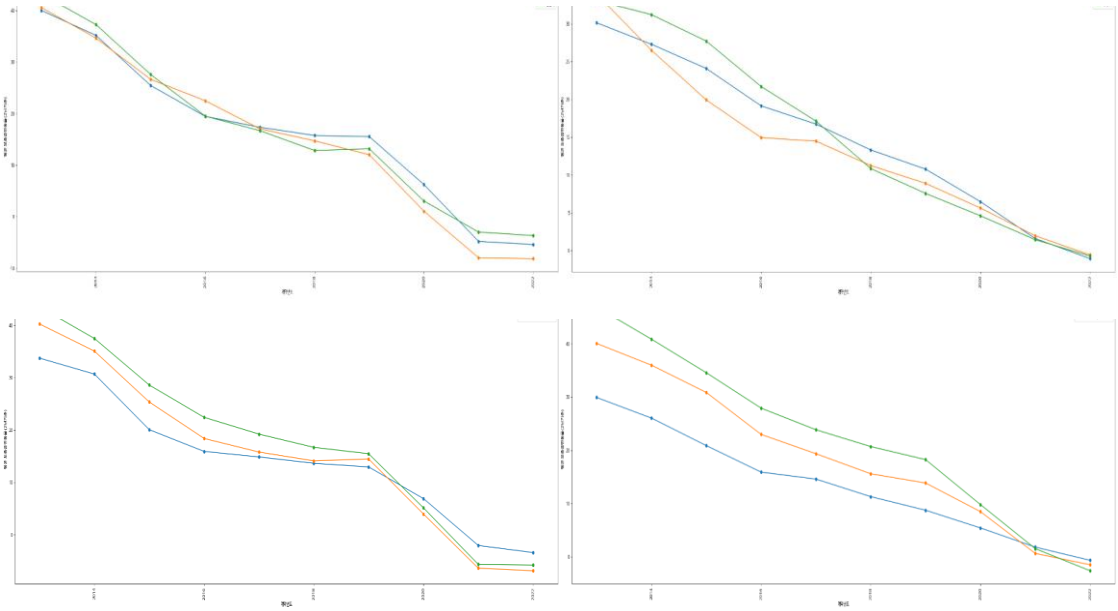
자치구별로 2년을 기준으로 한 보증금변동률을 확인했을 때(<그림 7>), 2020년까지는 아파트

와 오피스텔 모두 보증금이 계속해서 상승(보증금변동률이 0% 이상)했으나, 2021년부터는 하락하기 시작했다는 것을 확인할 수 있다.



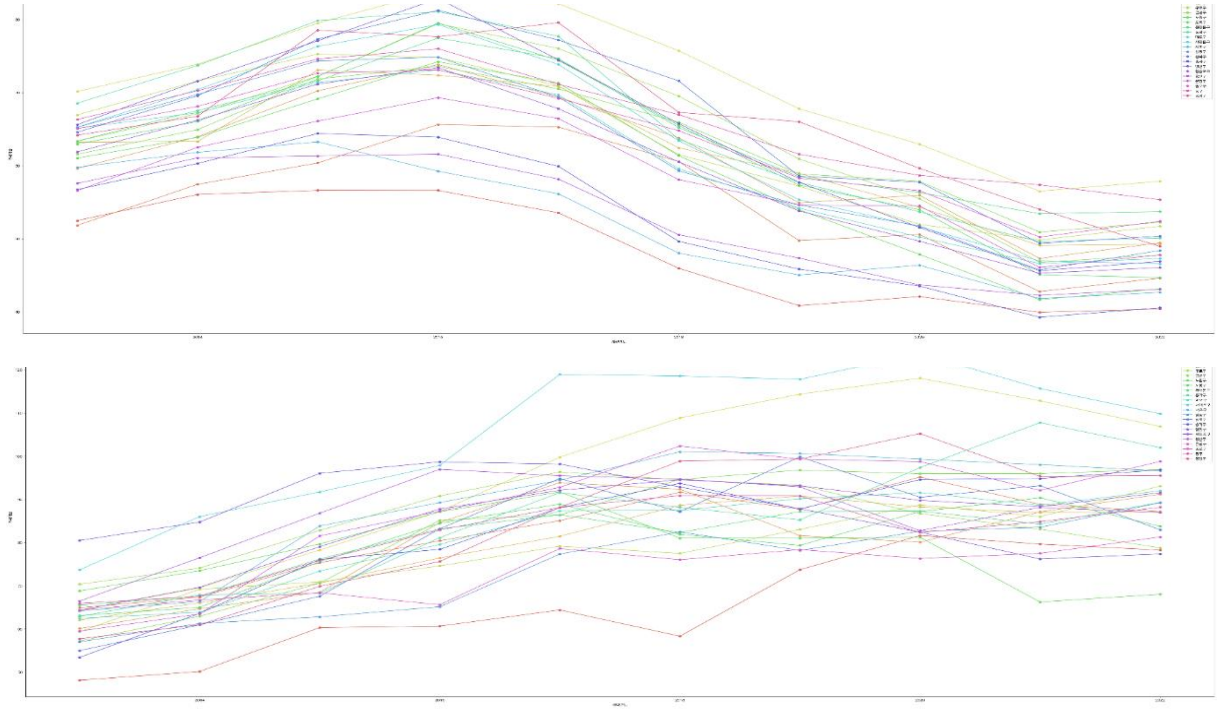
<그림 8> 연도별 아파트/오피스텔 층별 보증금변동률(2년기준)

(<그림 8>)아파트 층별 보증금변동률은 일반적으로 저층일 경우 고층일 경우보다 보증금변동률이 큰 것으로 파악되었다. 오피스텔의 층별 보증금변동률은 아파트와 비교했을 때 저층과 고층의 보증금변동률의 차이가 크지 않았으나, 전반적으로 보증금변동률이 계속해서 감소한다는 것을 확인할 수 있었다.



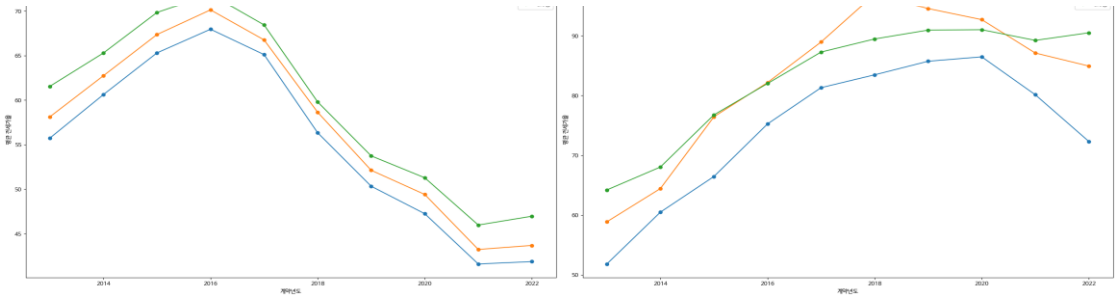
<그림 9> 연도별 아파트/오피스텔 신축여부별/면적구분별 보증금변동률(2년기준)

아파트와 오피스텔 모두 신축여부나 면적구분과 관계없이 보증금변동률이 지속적으로 감소하고 있다. (<그림 9>)



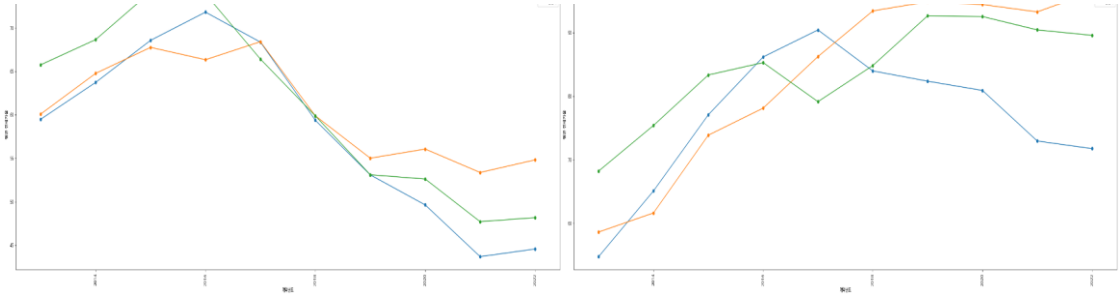
<그림 10> 연도별 아파트/오피스텔 자치구별 평균 전세가율

각 자치구마다 전세가율을 비교해보았을 때 (<그림 10>), 아파트는 관악구의 전세가율이 가장 높고, 강남구의 전세가율이 가장 낮았으며, 2016년에 전반적으로 전세가율이 최고점이고 2017년부터는 점차 하락하는 것을 확인할 수 있었다. 오피스텔의 경우 아파트와는 달리 하락하지 않고 오히려 상승하는 경우가 대부분이었다. 서대문구와 노원구가 타 자치구에 비해 전세가율이 높은 것으로 파악되었으며, 강남구의 전세가율이 2013년부터 2019년까지 가장 낮았으나 2019년 이후로 많이 상승하는 추세를 보였고, 2021년 이후에는 노원구가 가장 낮은 것으로 파악되었다.



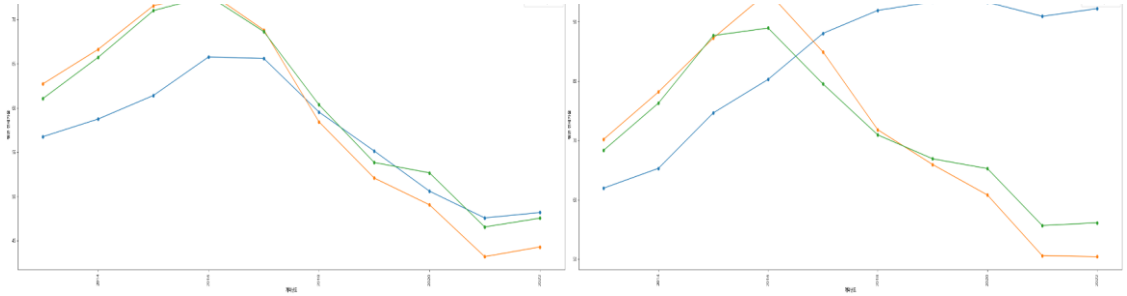
<그림 11> 연도별 아파트/오피스텔 층별 평균 전세가율

(<그림 11>) 아파트의 경우 고층 순으로 전세가율이 형성되어 있는 것을 확인 할 수 있었으며, 오피스텔의 경우 2015년도까지는 고층 순으로 전세가율이 형성되어 있었으나, 2016년부터 2층이 고층보다 전세가율이 높게 형성되기 시작했고, 2021년부터 다시 고층의 순서로 전세가율이 형성되기 시작했다. 또한 저층(1.0)의 경우 2021년부터 급격히 하락하는 것을 확인할 수 있었다.



<그림 12> 연도별 아파트/오피스텔 신축여부별 평균 전세가율

신축 여부별로 아파트와 오피스텔의 전세가율을 확인해보았을 때 (<그림 12>), 아파트는 2016년 까지 준신축의 전세가율이 가장 높았으나, 2017년부터 신축여부와 관계없이 비슷한 수준을 유지하다가 2019년 이후로 신축-준신축-구축의 순으로 전세가율이 형성되었고, 오피스텔의 경우 2015년 까지는 준신축의 전세가율이 가장 높았으나 일정기간을 걸쳐 2018년부터 신축-준신축-구축 순으로 전세가율이 형성됨을 확인할 수 있었다.



<그림 13> 연도별 아파트/오피스텔 면적구분별 평균 전세가율

면적구분별 아파트의 전세가율을 확인했을 때 (<그림 13>), 면적이 크다고 해서 전세가율이 항상 높은 편은 아니라는 것을 확인할 수 있었으며, 2013 ~ 2017년까지는 중간면적과 적은면적이 높은 전세가율을 형성한다는 것을 확인할 수 있다. 오피스텔의 경우 2017년을 기점으로 적은면적의 전세가율이 가장 높은 비중을 차지하게 된 다는 점을 확인할 수 있으며, 중간면적과 큰면적은 2017년 이후 지속적으로 하락한다는 것을 파악할 수 있었다.

II. 분석자료 및 분석방법

1. 분석 자료

이 프로젝트에서 사용한 메인 데이터는 서울 열린데이터 광장에서 제공하는 전월세가 정보로, 자치구, 법정동, 건물명, 전월세구분, 보증금, 임대료, 계약년도 등의 정보를 사용했다. 매매가를 이용해 전세가율을 계산하기 위해 서울 열린데이터 광장에서 제공하는 서울시 부동산 실거래가 정보를 사용했다. 이후 다시 언급할 데이터 전처리 과정 중 반전세의 월세 부분을 보증금으로 전환하는 과정에서 KOSIS의 지역별 전월세전환율 데이터에서 서울시 월별 전월세전환율을 사용했다. 마지막으로, KOSIS의 성별/연령별/점유형태별 1인가구(일반가구)-시군구 자료에서 서울특별시 자치구별 전세 연령별 비율을 사용했다.

아래에서 다시 언급하겠지만, 건물용도가 아파트, 오피스텔인 자료만 사용하고, 보증금이 3,400만원에 미달하거나 3년 치 월세보다 작은 표본들은 제외했다. 보증금이 3년치 월세를 초과하는 보증부월세의 경우 전월세전환율을 적용하여 전세로 변환했고, 완전 전세의 경우 환산할 필요가 없으므로 모두 사용했다. 월세의 경우 역전세 리스크가 적을 것으로 판단되어 전세 데이터만 활용하기로 결정했다. 표본기간은 2013년 1월부터 2022년 12월까지이며, 최종적으로 사용한 표본 수는 1,294,607건이다.

2. 분석 방법

이 프로젝트에서 분석 자료의 전처리는 *민병철(2021) 전세가격 변동률 분포를 활용한 역전세 위험의 측정*에서 실시한 전처리를 기준으로 수행했다. 자치구명, 법정동명, 본번, 부번이 모두 동일하지만 건물명이 다른 경우 건물명이 변경된 것이라 판단하여 가장 최근에 거래된 건물명으로 통일했다. 위 논문에서 언급한대로 기초적인 특성들이 같은 아파트들은 시장에서 같은 상품으로 취급되어 가격이 같을 것이라는 가정하에, 주소, 단지명, 건축년도, 면적, 층수가 같은 거래 건들을 같은 상품에 대한 거래로 간주했다. 주소는 시, 군, 구, 동 단위까지 사용했다. 건물용도가 아파트, 오피스텔인 자료만 사용했고, 임대면적은 1의 자리로 반올림을 진행했다. 22.01~24.01 형식으로 표기된 계약기간을 24 형식의 개월수로 변경했다. 위에서 언급했던 보증금이 3년치 월세를 초과하는 보증부월세의 경우 <식 1>을 이용하여 전세로 환산했다.

<식 1> 전세보증금 = (월세 * 계약기간 / 전월세전환율) + 월세보증금

이 때 사용한 서울시 전월세전환율 데이터에서 제공하는 주택유형은 아파트, 단독주택, 연립다세대로 오피스텔에 해당하는 전월세전환율이 존재하지 않는다. 세 가지 주택유형 중 아파트가 오피스텔과 가장 유사하다고 판단하여 아파트와 오피스텔 모두 서울시 월별 아파트 전월세 전환율을 사용했다. 보증금이 3,400만원 이하인 경우는 월세에 가깝다고 보고 월세로 판단했다. 여기까지 위 논문을 기반으로 전처리를 진행했고, 이후에는 추가로 필요한 전처리를 진행했다.

층수는 저층의 가격이 상대적으로 낮다는 점을 고려하여 지하층/ 1층/ 2층/ 3층 이상으로 나누었고 지하층은 -1, 2층 초과는 3으로 변환했다. 3층 이상일 때 층수 별 가격 차이, 조망 등은 무시되었는데, 이는 본 프로젝트의 한계점이다. 월세의 경우 전세보다 역전세 리스크가 적을 것으로 판단되

어 위 논문 기반 전처리 조건을 만족하는 월세 데이터를 전세로 전환한 뒤, 월세 데이터를 제거했다. 전월세구분이 전세인 데이터 중 월세가 0원이 아닌 데이터는 반전세라고 판단하여 제거했다. 계약기간이 결측치가 아닌 데이터 중 아파트, 오피스텔의 계약기간의 90.7%, 80.9%가 24개월이므로 계약기간의 결측치는 24개월로 채웠다. 건축년도를 이용해 계약년도와 건축년도의 차이가 5년 이하이면 신축, 6~10년이면 준신축, 10년 초과이면 구축으로 구분하여 신축여부를 구분했다. 그리고, 임대면적이 59㎡미만, 59㎡이상 84㎡이하, 84㎡초과로 구분하여 면적구분을 구분했다.

보증금변동률의 비교 기간은 일반적인 전세 계약 기간인 2년을 기준으로 잡았다. 보증금변동률은 같은 건물에 대한 2년 전 보증금을 기준으로 산출했다. 여기서 2년 전 보증금은 같은 건물의 매월 마지막에 거래된 보증금이다. 아래에서 언급하겠지만 보증금변동률은 역전세 위험을 판단하는 기준으로 사용했다. 따라서 보증금변동률이 결측이면 역전세 위험을 판단할 수 없어 결측치 5,718개를 삭제했다.

각 거래에 대한 전세가율을 구하기 위해 서울시 부동산 실거래가 정보 데이터를 '본번', '부번', '건물명', '층', '임대면적', '계약년월'로 그룹화해 평균 매매가를 계산한 뒤, 전월세가 데이터의 '본번', '부번', '건물명', '층', '임대면적', '계약년월'이 같은 레코드의 매매가를 저장했다. 결측치를 최소화하며 계약시기에 대한 경향을 반영하기 위해 '계약일' 대신 '계약년월'을 사용했다. 그럼에도 서울 열린데이터 광장에서 제공한 전월세가 데이터와 실거래가 데이터의 '본번', '부번', '건물명', '층', '임대면적', '계약년월'이 모두 일치하는 레코드는 많지 않았기 때문에 많은 결측치가 발생했다. 본 프로젝트의 목적은 '전세가율'과 '보증금변동률'을 이용해 역전세 위험을 판단하는 것이므로 '전세가율'이 결측이라면 역전세 위험을 판단할 수 없어, 회귀분석을 이용해 '매매가' 결측치를 처리했다. 회귀분석에서는 '자치구명', '법정동명', '건물용도', '계약년월'을 Label Encoding한 변수와 '층', '임대면적', '건축년도'를 Standard Scaling한 변수 총 7개를 사용했다. 회귀분석을 진행하기 전에, 다중공선성을 확인해 독립변수들 간의 상관관계가 없어 회귀분석을 진행했다. 회귀모델의 성능 중 결정계수와 수정된 결정계수는 0.957로 높았다. RMSE와 MAE의 값이 높았지만 2013년부터 2022년까지의 데이터를 사용하여 모델을 적합했으므로 해당기간동안 실제 매매가의 차이인 27,561(만원)을 고려하여 오차가 있을 수 있다고 판단하여 이를 감안하고 회귀모델을 사용했다.

Features	VIF
자치구명_encoded	3.222843
법정동명_encoded	3.074009
건물용도_encoded	1.373685
계약년월_encoded	2.721010
층_scaled	1.239920
임대면적_scaled	1.212308
건축년도_scaled	1.087777

RMSE	10496.392
MAE	6794.199
R^2	0.957
Adjusted R^2	0.957

<표 1> 회귀모델 다중공선성 & 성능

매매가의 결측치 처리 후에 <식 2>을 이용해 전세가율을 계산했다. 다음으로 한국도시연구소의 2022년 상반기 실거래가 분석 연구보고서 내용 중 *'전세가율이 10% 미만이거나 200% 이상인 자료*

*는 오류로 판단하여 삭제함*을 참고하여 전세가율이 10이하, 200이상인 레코드 5,159개를 삭제했다.

<식 2> 전세가율(%) = 전세가격 / 매매가격 * 100

역전세 위험을 판단하는 데 사용할 변수를 모두 생성했으므로 역전세 위험이 있는지(1) 없는지(0)에 대한 변수를 생성할 차례다. <식 3>의 조건을 만족한다면 역전세 위험이 있다(1)고 판단했다. 갭차이에 대한 계산은 <식 4>를 이용했다.

<식 3> 보증금변동률 하락폭 >= 갭차이

<식 4> 갭차이 = 100(%) - 전세가율

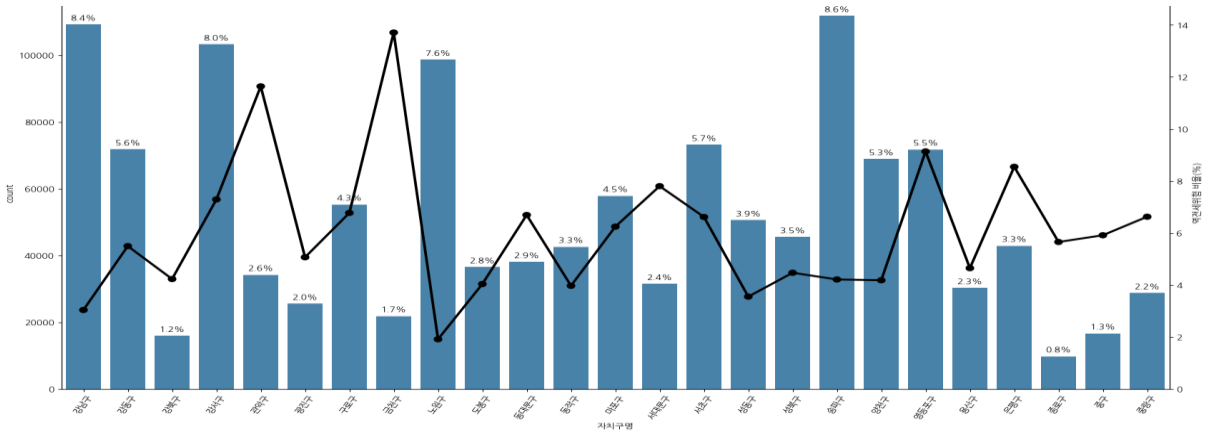
3. 가설 검증

각 요인들이 역전세 위험과 연관이 있는지 살펴보기 위해 가설 검증을 시행했다. 본 분석에서 세운 가설들은 다음과 같다.

<가설 1> 서울특별시의 자치구 간 역전세 위험은 차이가 없다.
<가설 2> 계약시기는 역전세 위험에 영향을 미치지 않는다.
<가설 3> 건물용도는 역전세 위험에 영향을 미치지 않는다.
<가설 4> 건물의 층은 역전세 위험에 영향을 미치지 않는다.
<가설 5> 건물의 건축년도는 역전세 위험에 영향을 미치지 않는다.
<가설 6> 임대면적은 역전세 위험에 영향을 미치지 않는다.

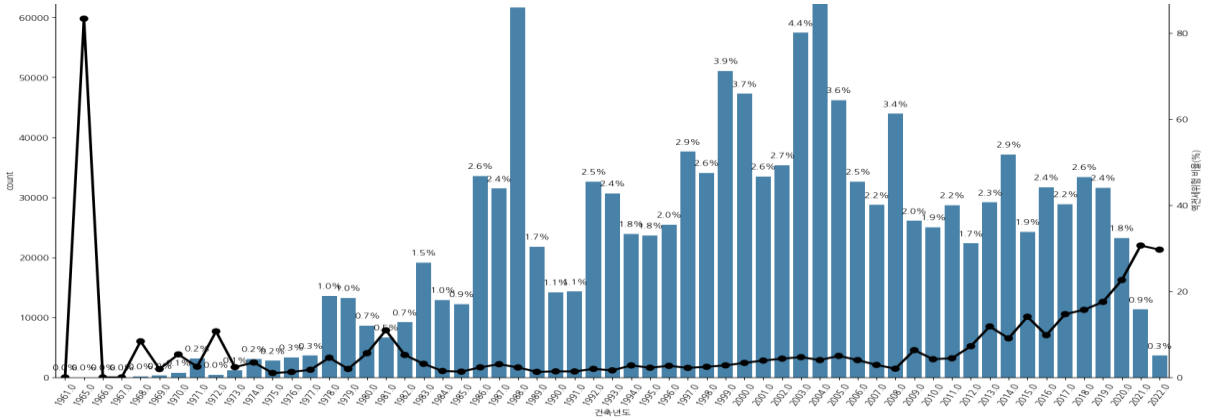
각 가설의 검증을 위해 순서대로 '자치구명', '계약년월', '건물용도', '층', '건축년도', '임대면적' 변수를 사용했으며 이 변수들과 '역전세위험' 변수의 관계를 파악했다. 가설 검증을 위해 카이제곱 검정과 로지스틱 회귀분석을 시행했다. '역전세위험' 변수는 범주형 변수이므로 비교 변수가 범주형 변수인 경우 카이제곱 검정을 실시했고, 비교 변수가 연속형 변수인 경우 카이제곱 검정과 로지스틱 회귀분석을 실시했다.

<가설 1> '서울특별시의 자치구 간 역전세 위험은 차이가 없다.'를 검증하기 위한 카이제곱 검정에서는 유의수준을 0.05로 가정했을 때, 검정 통계량이 14351이고 유의확률이 0.0 으로 자치구에 따라 역전세 위험은 차이가 존재한다는 것을 알 수 있다. 아래 <그림 14>를 확인해보아도 각 자치구별 역전세 위험의 비율에 차이가 존재하는 것을 알 수 있다.



<그림 14> 자치구명 분포와 역전세 위험 비율

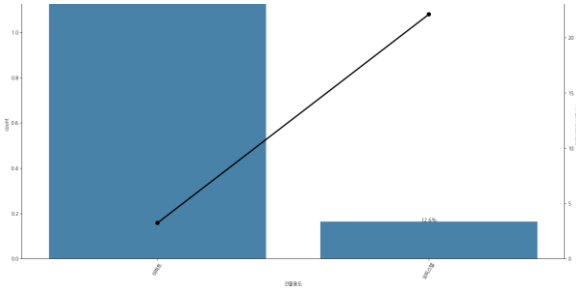
<가설 2> '건물의 건축년도는 역전세 위험에 영향을 미치지 않는다.'에 대한 검증 결과는 카이제곱 검정 통계량이 71655이고 유의확률이 0.0으로 건축년도가 역전세 위험에 영향을 미칠 수 있다는 것을 확인했다. <그림 15>를 확인했을 때, 건축년도가 1983년부터 2008년인 데이터는 역전세 위험 비율이 비슷한 정도를 보이지만, 이 기간 전후를 봤을 때 등락이 존재하여 건축년도에 대한 역전세 위험 비율에 차이가 존재함을 알 수 있다.



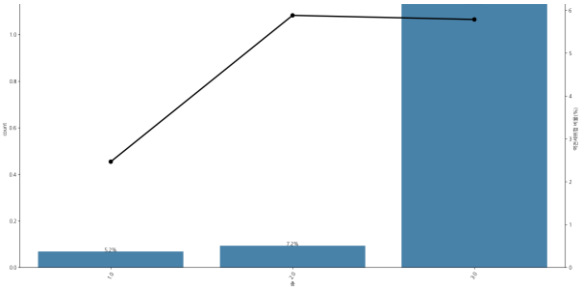
<그림 15> 건축년도 분포와 역전세 위험 비율

<가설 3> '건물의 임대면적은 역전세 위험에 영향을 미치지 않는다.'에 대한 검증 진행했다. 임대면적은 다른 요소들과 다르게 유일하게 연속형 변수이므로 카이제곱 검정과 로지스틱 회귀분석을 추가로 진행했다. 검정 결과는 카이제곱 검정 통계량이 122841이고 유의확률이 0.0이며, 로지스틱 회귀분석을 진행한 결과에서도 유의확률이 0.0으로 임대면적이 역전세 위험에 영향을 미칠 수 있다는 것을 확인했다.

<가설 4> '건물용도는 역전세 위험에 영향을 미치지 않는다.'에 대한 검증에서도 카이제곱 검정 통계량이 96113이고 유의확률이 0.0으로 건물용도가 역전세 위험에 영향을 미칠 수 있다는 것을 확인했다. <그림 16>을 보아도 아파트와 오피스텔 간 역전세 위험 비율에 차이가 크게 나는 것을 알 수 있다.



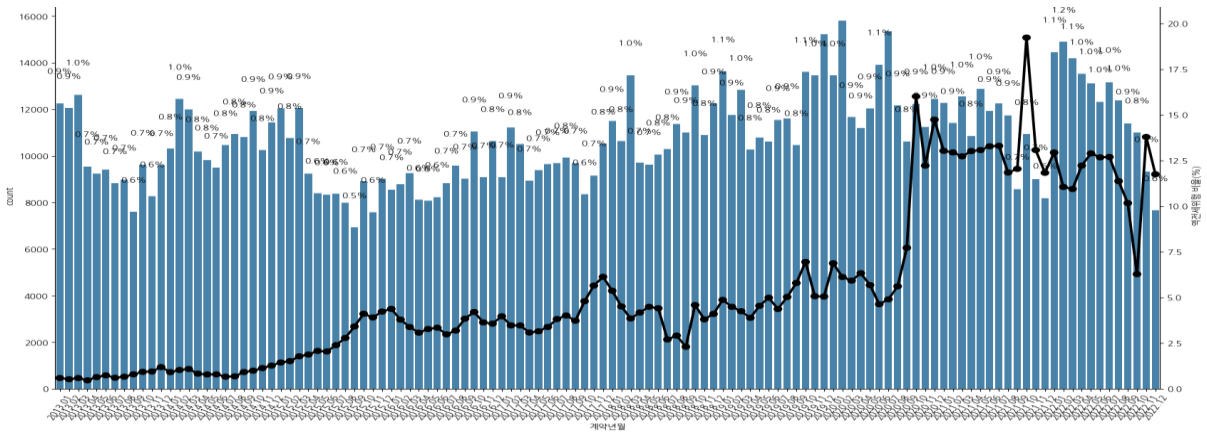
<그림 16> 건물용도 분포와 역전세 위험 비율



<그림 17> 층 분포와 역전세 위험 비율

<가설 5> '건물의 층은 역전세 위험에 영향을 미치지 않는다.'에 대한 검증에서는는 카이제곱 검정 통계량이 1336이고 유의확률이 5.56851218781213 66e-291의 아주 작은 값으로 건물의 층이 역전세 위험에 영향을 미칠 수 있다는 것을 확인했다. <그림 17>에서도 역전세 위험 비율이 2, 3층은 비슷해보이지만 이들과 1층 간의 차이가 존재하는 것을 알 수 있다.

<가설 6> '계약 시기는 역전세 위험에 영향을 미치지 않는다.'에 대한 검증에서도 카이제곱 검정 통계량이 47815이고 유의확률이 0.0으로 계약 시기가 역전세 위험에 영향을 미칠 수 있다는 것을 확인했다. <그림 18>을 보면 각 계약 연월 간에도 역전세 위험 비율에 대한 차이가 존재하지만 2020년 9월을 기준으로 더 큰 차이가 나타난 것을 확인할 수 있다.



<그림 18> 계약 연월 분포와 역전세 위험 비율

따라서 역전세 위험과의 관계가 있는지 진행했던 독립성 검정에서 '자치구명', '계약년월', '건물용도', '층', '건축년도', '임대면적' 요인들이 모두 '역전세위험' 컬럼에 영향을 미치고 있다는 것을 확인할 수 있다.

III. 분석 결과

1. 모델 선정

최종적으로 사용할 모델은 Decision Tree와 RandomForest, XGBoost를 선택했다. Decision Tree는 데이터를 분할하는 일련의 결정 규칙을 나타내는 트리 구조로 구성된 ML모델이다. 이 알고리즘은 데이터의 특징에 따라 가장 좋은 분할 규칙을 찾고, 각 분할된 영역에서 동일한 작업을 반복한다. 이러한 분할 규칙을 통해 데이터를 분류하거나 예측할 수 있다. Decision Tree의 장점은 해석력이 좋고, 데이터의 스케일링이나 정규화가 필요하지 않다는 강점을 가지고 있어 선택했다.

RandomForest는 여러 개의 결정 트리를 구성하고, 각 트리의 예측 결과를 결합하여 최종 예측을 수행하는 앙상블 기법이다. 각 결정 트리는 랜덤하게 선택된 데이터 샘플과 특징을 기반으로 학습하며, 이를 통해 다양한 결정 트리들이 서로 다른 특징을 고려하여 예측을 수행하게 된다. RandomForest는 과적합 문제를 완화시키고, 예측 성능을 향상시킬 수 있는 장점이 있어 이번 모델링 과정에서 사용했다.

XGBoost는 Gradient Boosting 알고리즘의 개선된 형태로, 여러 개의 결정 트리를 순차적으로 학습하는 앙상블 방법이다. Gradient Boosting은 이전 트리의 오차를 보완하는 방식으로 새로운 트리를 학습하며, 기울기를 계산하는 방법과 트리를 구성하는 방식에서 최적화가 이루어져 빠르고 정확한 학습이 가능하다. 또한, 과적합을 제어하기 위한 정규화 기능과 유연한 튜닝 옵션을 제공하며, XGBoost는 대부분의 데이터셋에서 뛰어난 예측 성능을 보여 해당 모델을 선택했다.

또한 RandomForest와 XGBoost의 경우 하이퍼파라미터 최적화를 위해 최대한 많은 하이퍼파라미터 조합을 시도하고, 최상의 조합을 찾아 모델의 성능 개선에 유용한 RandomizedSearch를 함께 사용했다. RandomizedSearch는 탐색 공간에서 임의의 하이퍼파라미터 조합을 선택하고, 이를 사용하여 모델을 학습하고 검증하는 과정을 반복한다. 사용자가 정의한 최대 반복 횟수 또는 시간 내에서 이 과정을 수행한 후, 가장 우수한 성능을 보인 하이퍼파라미터 조합을 반환하는 과정을 거치게 된다.

선택한 모델들을 사용하여 성능을 비교한 뒤, 최종적으로 선택된 모델을 바탕으로 순열 중요도(Permutation Importance)와 부분 의존성 플롯(Partial Dependence Plot)을 통해 모델의 주요한 특성과 각 특성이 얼마나 예측에 영향을 주는지 파악하였다.

2. 모델 적용 및 변수 선정

borough: 자치구명, floor: 층, yearmonth: 계약년월, buildinguse: 건물용도, YOC: 건축년도, NC: 신축여부, rentalarea: 임대면적, areadivision: 면적구분

features = ['borough', 'floor', 'yearmonth', 'buildinguse', 'YOC', 'NC', 'rentalarea', 'areadivision']

	Base	XGBoost	Decision Tree	Random Forest	Test-XGB
Accuracy	0.9415	0.9538	0.9453	0.9435	0.9543
Recall	0	0.3505	0.5403	0.0553	0.3483
Precision	0	0.7136	0.4314	0.7173	0.7293
f1	0	0.4701	0.4314	0.1027	0.4714
ROC-AUC	0.5	0.9291	0.4798	0.8304	0.9279

<표 2> 모델 1

features = ['borough', 'floor', 'yearmonth', 'buildinguse', 'YOC', 'rentalarea']

	Base	XGBoost	Decision Tree	Random Forest	Test-XGB
Accuracy	0.9415	0.9543	0.9468	0.9451	0.9545
Recall	0	0.359	0.555	0.0982	0.3489
Precision	0	0.7181	0.4511	0.7287	0.7319
f1	0	0.4787	0.4511	0.1731	0.4725
ROC-AUC	0.5	0.9305	0.4977	0.8745	0.9297

<표 3> 모델 2

features = ['borough', 'floor', 'yearmonth', 'buildinguse', 'NC', 'areadivision']

	Base	XGBoost	Decision Tree	Random Forest	Test-XGB
Accuracy	0.9415	0.9451	0.943	0.943	0.9449
Recall	0	0.1525	0.5374	0.0215	0.1446
Precision	0	0.6263	0.1857	0.7202	0.6245
f1	0	0.2452	0.1857	0.0417	0.2348
ROC-AUC	0.5	0.874	0.276	0.861	0.873

<표 4> 모델 3

위 세 가지 모델은 ML 진행을 위해 StandardScaler와 OrdinalEncoder를 공통적으로 진행했다. <표 2> 모델1은 모든 변수를 포함한 모델이고, <표 3> 모델2는 <표 2> 모델1에서 '신축여부'와 '면적구분'을 뺀 모델, <표 4> 모델3은 <표 2> 모델1에서 '건축년도'와 '임대면적'을 뺀 모델이다. <표 3>과 <표 4>의 모델링을 진행한 이유는 '건축년도'와 '신축여부', '임대면적'과 '면적구분'이 비슷한 의미를 가진다 판단했기 때문이다. 타겟인 '역전세위험'의 분포가 94:6으로 매우 불균형하기 때문에 모델 선정 기준은 Recall과 ROC-AUC 값을 중요한 성능지표로 판단했고, <표 3> 모델2가 가장 좋은 성능을 내어 해당 변수를 사용하기로 결정했다.

3. 최종 모델 선정

- **UnderSampling (NearMiss-3) 사용**

features = ['borough', 'floor', 'yearmonth', 'buildinguse', 'YOC', 'rentalarea']

	Base	XGBoost	Decision Tree	Random Forest	Test-XGB
Accuracy	0.9415	0.2679	0.7014	0.8759	0.268
Recall	0	0.9861	0.1278	0.5286	0.9849
Precision	0	0.0731	0.7051	0.2426	0.073
f1	0	0.1361	0.7051	0.3325	0.136
ROC-AUC	0.5	0.8717	0.2164	0.7737	0.8688

<표 5> 모델 4

타겟 불균형을 해결하기 위해 Undersampling기법을 이용하여 진행했다. 소수 클래스의 각 데이터 포인트마다 가장 가까운 데이터를 가져와 sampling을 하는 NearMiss Version 3를 사용했다. 전반적으로 Recall의 값이 많이 향상 되었으나 ROC-AUC의 성능 및 다른 평가지표도 같이 떨어져 기준에 맞지 않다고 판단했다.

- **가중치 Parameter 조절**

XGBoost는 scale_pos_weight, Random Forest와 Decision Tree는 class_weight 파라미터를 사용해 모델링

features = ['borough', 'floor', 'yearmonth', 'buildinguse', 'YOC', 'rentalarea']

	Base	XGBoost	Decision Tree	Random Forest	Test-XGB
Accuracy	0.9438	0.8647	0.9083	0.8759	0.8636
Recall	0	0.8197	0.3255	0.5674	0.8224
Precision	0	0.2688	0.5895	0.242	0.2675
f1	0	0.4049	0.5895	0.3393	0.4037
ROC-AUC	0.5	0.9236	0.4194	0.8216	0.9238

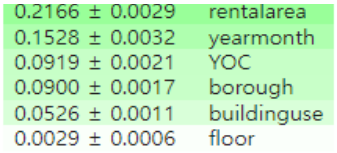
<표 6> 모델 5

Undersampling과 마찬가지로 타겟 불균형을 해결하기 위해 가중치 Parameter를 이용한 결과, Accuracy는 감소했지만 Recall이 높아졌고 F1-Score도 어느 정도 유지가 되었다. ROC-AUC도 1에 가까움으로 클래스 분류 성능이 높다고 판단하여 최종 모델로 선정했다. 선정 기준에 타겟 변수가 불균형이 존재하기 때문에 Base Model보다 Accuracy가 높아지는 것은 쉽지 않음을 감안했다.

4. 최종 모델 분석 결과

- **순열 중요도((Permutation Importances))**

최종적으로 선정된 모델의 특성 중요도를 파악하기 위하여 순열 중요도(permutation importance)를 파악해보았다. 순열 중요도는 머신 러닝 모델에서 특성(feature)의 중요도를 평가하는 방법 중 하나이며, 특성 중요도는 모델이 예측을 수행하는 데 어떤 특성이 가장 큰 영향을 미치는지를 측정하는 방법이다. 순열 중요도를 파악할 때 ROC-AUC를 평가 척도로 사용했으며, 반복 횟수는 5회로 고정했다.



0.2166 ± 0.0029	rentalarea
0.1528 ± 0.0032	yearmonth
0.0919 ± 0.0021	YOC
0.0900 ± 0.0017	borough
0.0526 ± 0.0011	buildinguse
0.0029 ± 0.0006	floor

<그림 19> 특성중요도

최종 모델을 바탕으로 특성 중요도(PI)를 확인해 본 결과 '임대면적', '계약년월', '건축년도', '자치구', '건물용도', '층' 순으로 모델의 성능에 기여하는 것을 확인했다.

- **PDP**

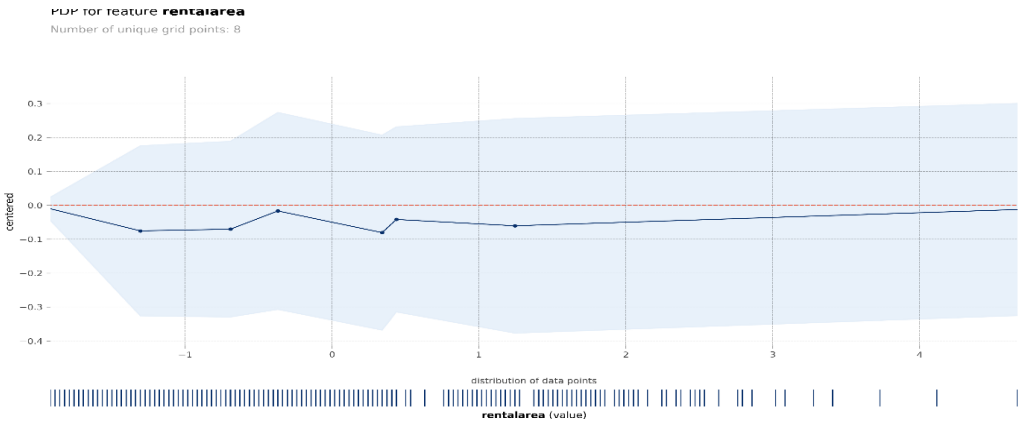
부분 의존성 플롯(Partial Dependence Plot, PDP)은 머신 러닝에서 사용되는 분석 기법이다. 이 기법은 특정한 특성(변수)이 모델의 예측 결과에 어떤 영향을 미치는지를 시각적으로 나타내는 데 사용하며, 특성 하나의 값을 변화시키면서 모델의 예측값이 어떻게 변화하는지를 그래프로 표현한다. 이를 통해 특정 특성이 모델의 예측에 얼마나 중요한지, 어떤 패턴을 가지는지를 파악할 수 있다.

PDP isolate는 부분 의존성 플롯을 개별적인 특성에 대해 생성하는 방법이다. 주어진 특성 하나에 대해 다른 모든 특성을 고정시키고, 해당 특성의 값을 변화시켜가며 예측 결과를 관찰한다. 이렇게 하나의 특성에 대해 다른 특성들과의 상호작용을 고려하지 않고 개별적으로 분석하며, 이 방법은 각 특성의 개별적인 영향력을 파악할 수 있어 해석력을 높일 수 있다.

PDP interact의 경우에는 부분 의존성 플롯을 두 개 이상의 특성에 대해 생성하는 방법이다. PDP isolate와 달리 특정한 두 개 이상의 특성의 값을 함께 변화시키면서 모델의 예측 결과를 관찰한다. 이를 통해 특성들 간의 상호작용이 모델의 예측에 어떤 영향을 미치는지를 분석할 수 있다.

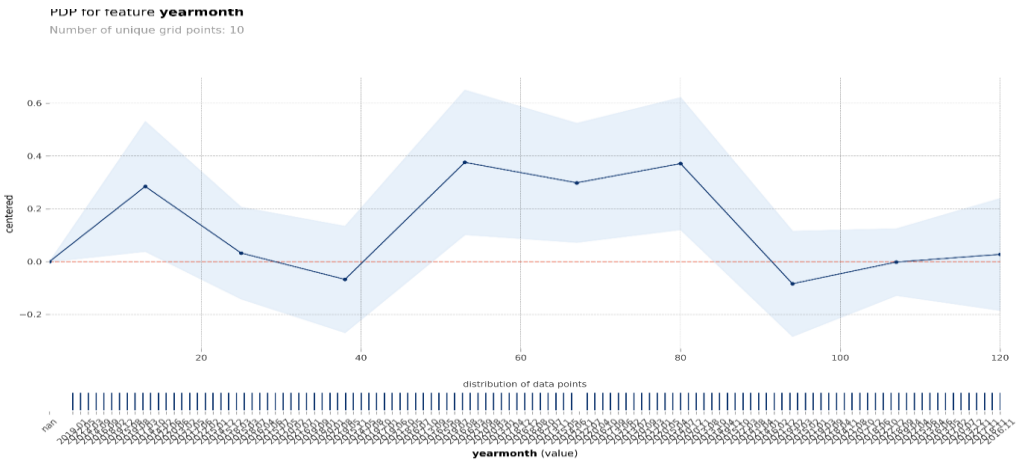
부분 의존성 플롯(PDP) isolate와 interact는 모델의 해석 가능성을 높이고, 특성 간의 상호작용을 이해하는 역할을 하여 사용했다. 이를 통해 모델의 예측을 설명하고, 각 특성의 중요성을 파악하여 해석을 실시하였다.

1) Isolate



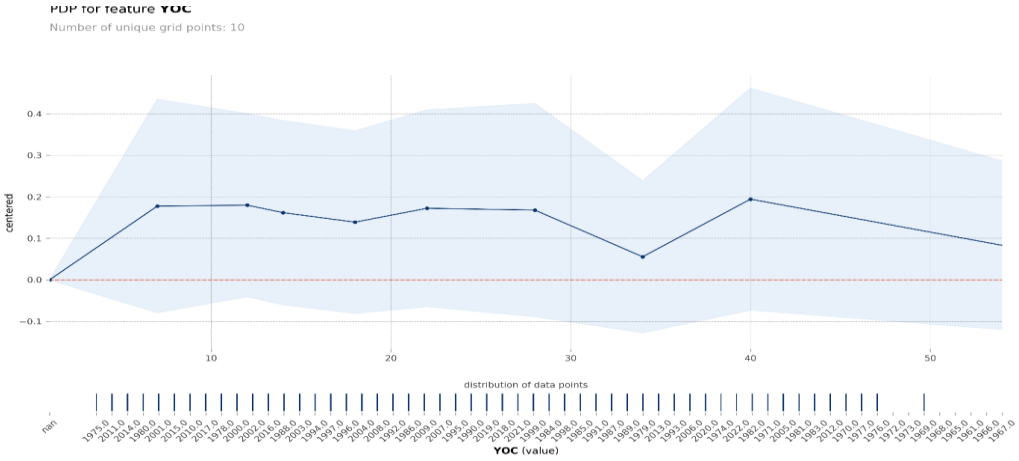
<그림 20> 임대면적 PDP Isolate

'임대면적'에 따라 플롯을 파악했을 때(<그림 20>), 모델은 평균값(0)을 기준으로 했을 때 '임대면적'이 커질수록 역전세 위험 가능성이 줄었다가 다시 증가한다는 것을 확인했으며, 평균값보다 낮아질 경우 역전세 위험이 적다는 것을 확인했다.

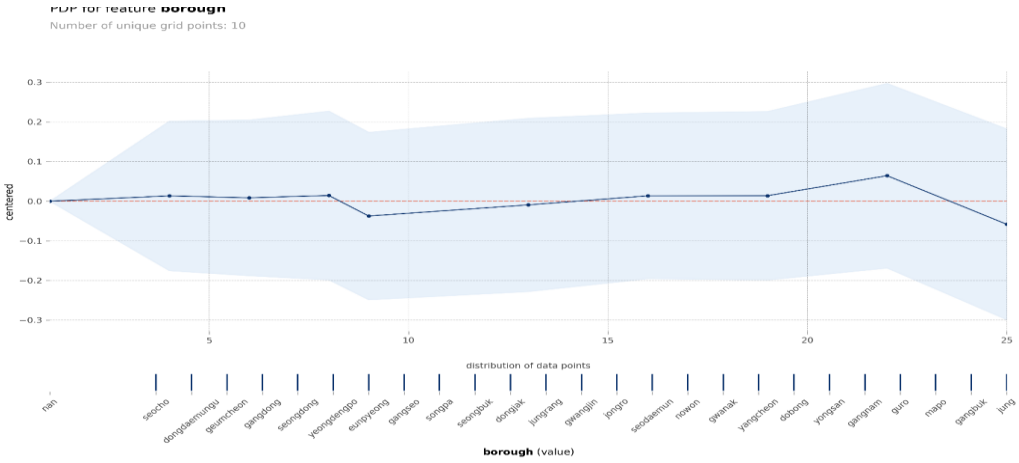


<그림 21> 계약 연월 PDP Isolate

'계약년월'을 기준으로 역전세 위험 가능성의 예측 차이를 확인했을 때(<그림 21>), 모델은 2019년 1월을 기준으로 2014년 5월과 2013년 3월은 역전세 위험 가능성이 적다고 예측했고, 2022년 6월, 2021년 9월, 2022년 7월, 2020년 12월은 역전세 위험 가능성이 높다고 예측했다.



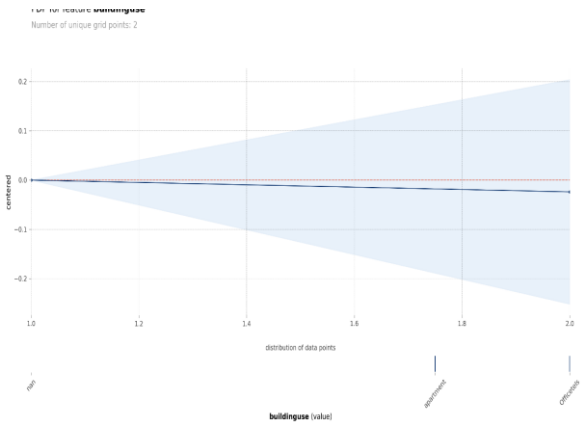
<그림 22> 건축년도 PDP Isolate



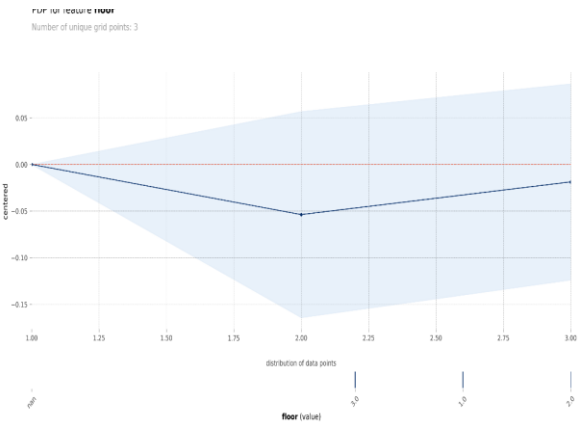
<그림 23> 자치구명 PDP Isolate

건축년도를 기준으로 역전세 위험 가능성의 예측 차이를 파악했을 때(<그림 22>), 1975년을 기준으로 2010년, 2016년, 2021년, 2020년에 건축된 것은 역전세 위험 가능성이 높다고 예측했다.

자치구별로 모델의 역전세 위험 가능성의 예측 차이를 확인했을 때(<그림 23>), 서초구를 기준으로 송파구, 강북구, 중구의 경우 역전세 위험 가능성이 낮다고 예측했고, 구로구, 강동구, 강서구, 노원구, 도봉구의 경우 역전세 위험 가능성이 높다고 예측했다.



<그림 24> 건물용도 PDP Isolate

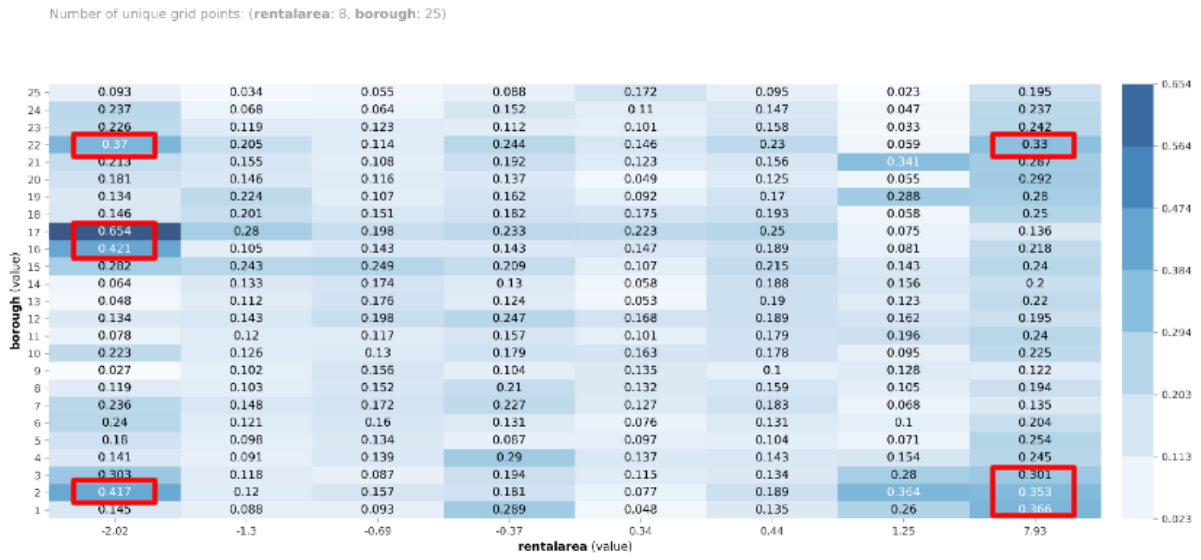


<그림 25> 층 PDP Isolate

건물 용도별 역전세 위험 예측차이를 확인했을 때(<그림 24>), 아파트보다 오피스텔이 역전세 위험 가능성이 높다고 예측했다.

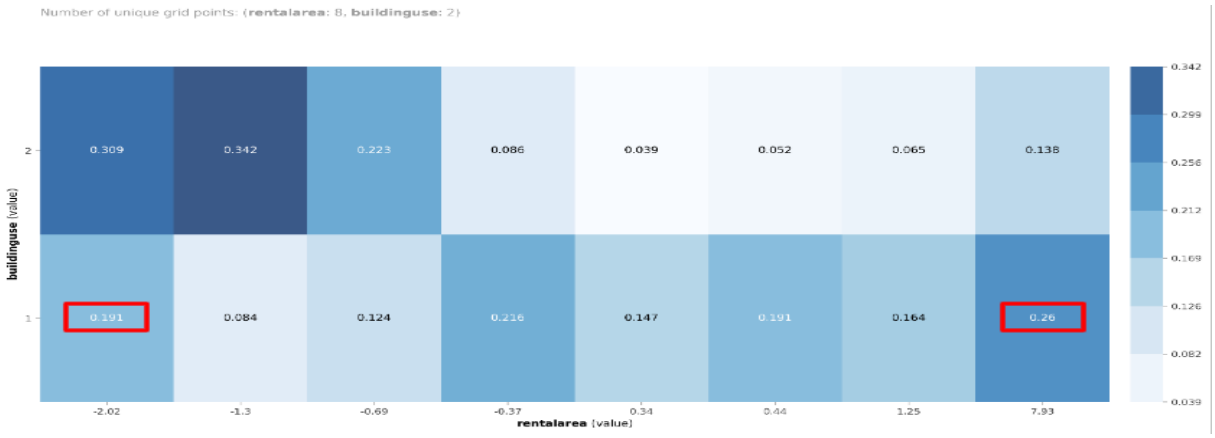
층별로 역전세 위험 가능성의 예측 차이를 확인했을 때(<그림 25>), 고층(3.0)을 기준으로 저층(1.0)은 역전세 위험 가능성이 크다고 예측했으며, 중간층(2.0)은 역전세 위험이 낮다고 예측했다.

2) Interact



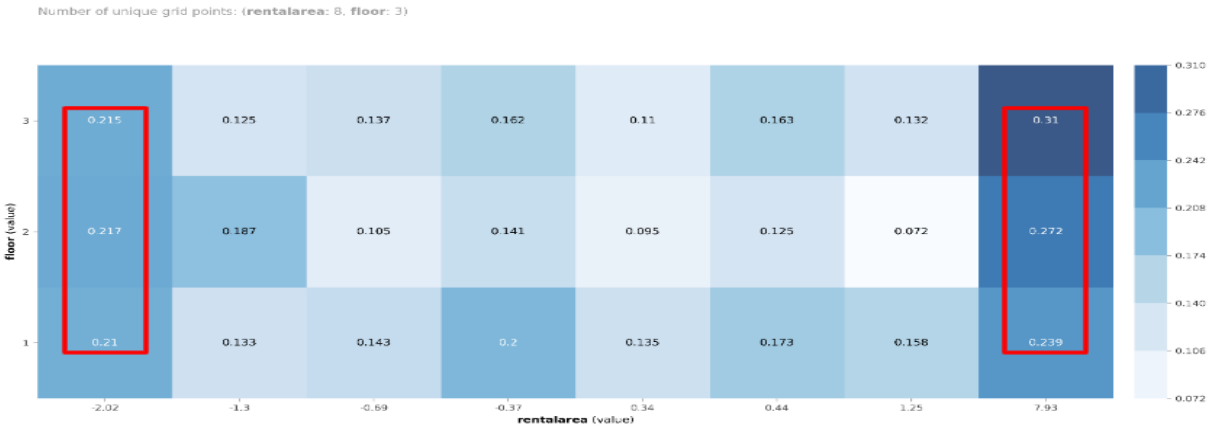
<그림 26> 임대면적 & 자치구명 PDP Interact

자치구와 임대 면적을 이용하여 역전세 위험 가능성을 예측한 결과를 확인했을 때(<그림 26>), 임대 면적이 적을 때 관악구(17), 노원구(16), 동대문구(2), 구로구(22)가 높게 나타났고, 임대 면적이 클 때 서초구(1), 동대문구(2), 구로구(22), 금천구(3)가 높게 나타난 것을 확인했다.



<그림 27> 임대면적 & 건물용도 PDP Interact

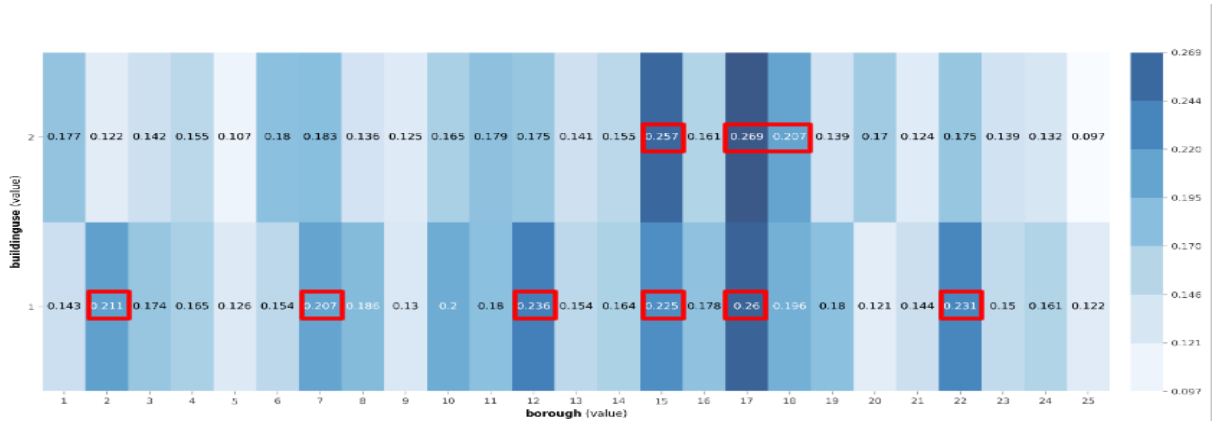
건물 용도와 임대 면적을 사용하여 예측 차이를 파악했을 때(<그림 27>), 오피스텔(2)의 경우 임대 면적이 작을 수록 역전세 위험이 크다고 예측했으며, 임대 면적이 커질수록 일반적으로 역전세 위험 가능성이 낮아진다고 예측했다. 아파트(1)의 경우 임대 면적이 작을 때도 0.191로 역전세 위험이 적지 않다고 보았으며, 임대 면적이 클 때도 0.26으로 역전세 위험이 발생할 가능성이 높다고 예측했다.



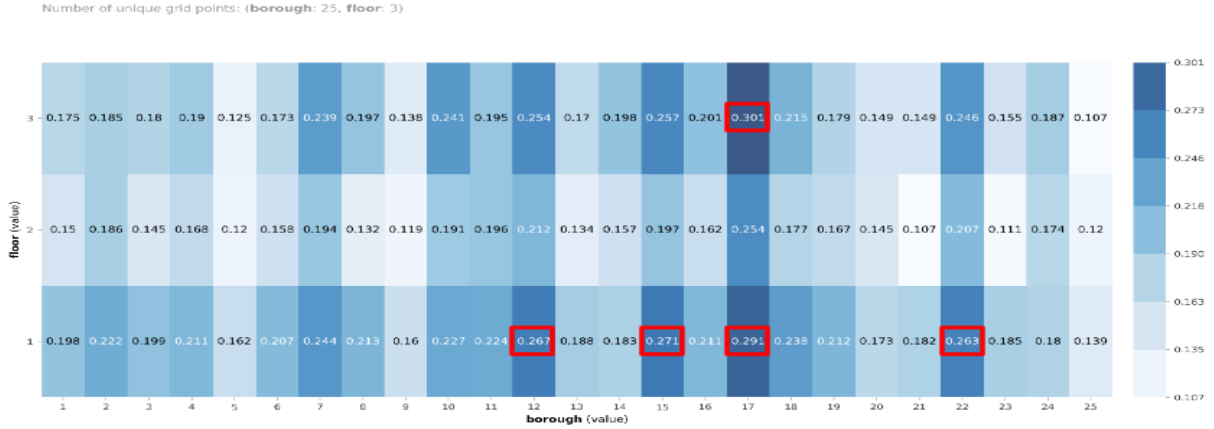
<그림 28> 임대면적 & 층 PDP Interact

임대 면적과 층을 기준으로 역전세 위험 가능성 예측 차이를 확인한 결과(<그림 28>), 임대 면적이 적을 때나 많을 때 역전세위험 가능성이 높다고 예측했고, 고층과 저층 여부와 관계없이 임대 면적별로 차이가 존재하는 것을 확인했다.

자치구와 건물용도별 역전세 위험 가능성의 예측 차이를 파악했을 때(<그림 29>), 오피스텔(2)의 경우 서대문구(15)와 관악구(17), 양천구(18)로 가장 높은 것을 확인했고, 아파트(1)의 경우 서대문구(15), 관악구(17), 구로구(22), 중랑구(12), 동대문구(2), 은평구(7)가 높은 것을 확인했다.



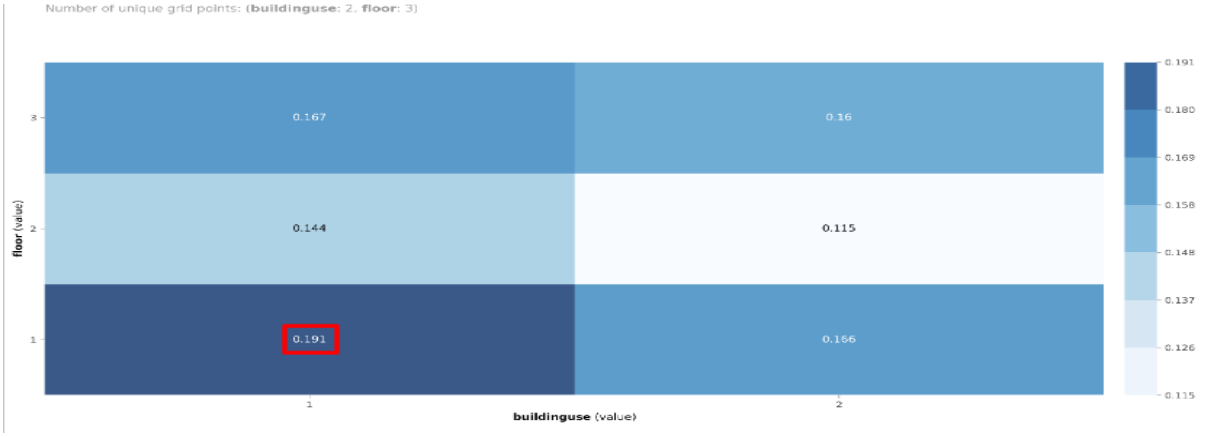
<그림 29> 자치구명 & 건물용도 PDP Interact



<그림 30> 자치구명 & 층 PDP Interact

자치구와 층을 기준으로 역전세 위험 가능성을 예측했을 때(<그림 30>), 3층 이상(1)일 때는

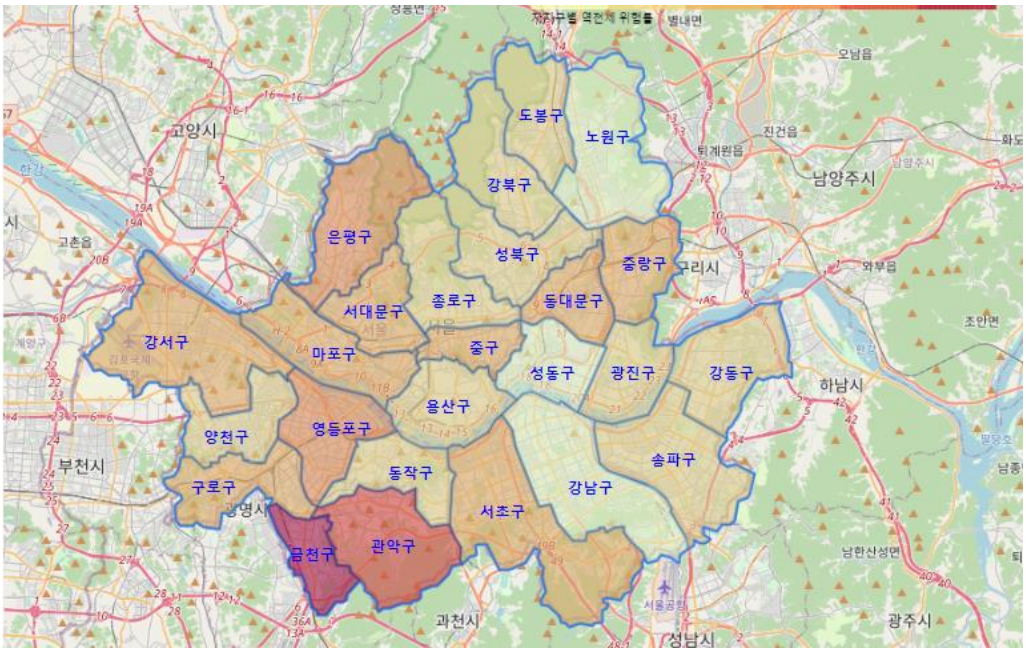
중랑구(12), 서대문구(15), 관악구(17), 구로구(22)의 역전세 위험 가능성이 높다고 예측했고, 2층(3)일 경우 관악구가 가장 높은 것으로 예측했다, 1층(2)일 경우 전반적으로 고층과 자치구 차이가 없었으며, 강남구(21)의 경우 가장 낮은 것으로 파악되었다.



<그림 31> 건물용도 & 층 PDP Interact

건물용도와 층별 역전세 위험 가능성의 차이를 파악했을 때(<그림 31>), 아파트(1)의 경우 고층(1)이 역전세 위험이 가장 높다고 예측했고, 저층(2)이 역전세 위험 가능성이 낮다고 예측했다. 오피스텔(2)의 경우에도 층별로 역전세 위험의 차이가 동일했으나, 아파트와 비교했을 때 저층에서의 역전세 위험 가능성이 타 층에 비해 많이 낮음을 확인할 수 있다.

IV. 결론



<그림 32> 자치구별 역전세 위험률

실제 데이터와 모델의 역전세 위험 가능성의 분석 결과를 종합해본 결과 '임대면적', '계약년월', '건축년도', '자치구', '건물용도', '층' 등 여러 요소들이 역전세 위험 가능성에 영향을 미친다는 것을 확인할 수 있다.

각각의 요인을 종합적으로 분석해 본 결과, '층'과 '자치구'를 기준으로 했을 때 고층일 때는 중랑구, 서대문구, 관악구, 구로구 등이 역전세 위험 가능성이 높고, 관악구의 경우 저층일 때 위험이 높은 것을 파악할 수 있다.

오피스텔의 경우 서대문구, 관악구, 양천구 등이 역전세 위험 가능성이 높은 것으로 파악되었다. 아파트의 경우에는 서대문구, 관악구, 구로구, 중랑구, 동대문구, 은평구 등이 위험한 것으로 나타났다.

임대 면적별로 역전세 위험을 파악했을 때, 임대 면적이 작을 경우 관악구, 노원구, 동대문구, 구로구 등이 위험 가능성이 높은 것을 파악할 수 있었으며, 임대 면적이 클 때 서초구, 동대문구, 구로구, 금천구 등이 역전세 위험이 높을 것을 확인할 수 있다.

전반적으로 자치구를 종합해 보았을 때 관악구나 구로구는 층, 아파트, 오피스텔 구분 없이 역전세 위험의 가능성이 모두 높은 것으로 파악되었다. 반대로 강남구나 성동구의 경우 건물용도나 임대면적의 구분 없이 역전세 위험이 낮은 것을 확인할 수 있다.

2030세대가 사회 초년생으로서 전세 계약에 익숙치 않아 역전세 위험에 취약하다고 생각하여 서울특별시 자치구별 연령별 전세 인구 데이터(2020년)를 이용하여 2030세대가 전세로 많이 거주하는

자치구를 조사해보니 관악구가 전체 연령대 중 2030세대의 비율이 0.752로 가장 높았고, 다음으로 영등포구(0.681), 강서구(0.67), 마포구(0.65), 광진구(0.635)가 높았다. 이 자치구들은 <그림 32>에서 확인할 수 있듯이 역전세 위험률도 비교적 높은 지역들에 해당하므로 해당 자치구들에서 전세 계약을 맺을 시에 더 유의할 필요가 있을 것이라 판단했다. 각 자치구별로 살펴보았을 때는 2030세대의 비율이 가장 높은 관악구는 위에서도 언급했다시피 층, 아파트, 오피스텔 구분 없이 역전세 위험의 가능성이 모두 높은 것으로 파악되었고, 두번째로 비율이 높았던 영등포구는 아파트보다는 오피스텔, 고층보다는 1층이 역전세 위험의 가능성이 높았다. 세번째로 높았던 강서구는 오피스텔보다는 아파트, 1층과 2층을 제외한 고층이 역전세 위험의 가능성이 높았다. 이와 같은 요인들과 각 자치구의 위험률을 유의하여 전세 계약을 맺어 역전세의 피해를 줄일 수 있기를 바란다.

그러나 해당 분석 이외에도 역전세를 발생시키는 요인은 상당히 많다. 임차인이 사전에 역전세를 당하지 않으려면 추가적으로 어떻게 해야 할까? 임대인에 대한 다음 2가지 정보를 확인하는 것을 추천한다. 첫 번째로 계약하려는 매물의 임대인 정보를 확인하는 것이다. 임대인에 대한 정보는 개인정보이므로 본 프로젝트에서는 다루지 못했지만, 주택도시보증공사(HUG)가 운영하는 '안심전세 앱'에서 임대인의 보증사고 이력과 세금 체납 여부, 전국 빌라·아파트·오피스텔 1252만채의 시세정보 등을 확인할 수 있다. 또, 일정 자격을 갖춘 임대인은 HUG가 발급하는 '안심 임대인 인증서'를 받을 수 있고, 2023년 12월까지 악성임대인 명단 공개 기능을 추가한다고 한다. 두 번째는 전세보증보험 가입여부를 확인하는 것이다. 전세보증보험은 HF(한국주택금융공사), HUG(주택도시보증공사), SGI(서울보증보험) 세 가지가 있고, 보증기관별 가입조건이 조금씩 상이하다. 2023년 5월부터는 모든 보증기관의 전세보증보험 가입조건이 전세가율 100%이하 주택에서 전세가율 90%이하 주택으로 하향 조정해, 기존 가입주택 중 25%가 제외대상이 되도록 강화되었다.

위의 절차를 잘 수행했음에도 불구하고 임대인이 보증금을 돌려주지 못하는 상황이 온다면 어떻게 해야할까? 전세보증금 반환보증을 신청하여 보증금의 일부를 돌려받을 수 있다. 전세계약 해지 또는 종료 후 1개월까지 정당한 사유 없이 전세보증금을 반환받지 못하거나 전세계약 기간 중 전세 목적물에 대하여 경매 또는 공매가 실시되어, 배당 후 전세보증금을 반환받지 못하였을 때 보증사고로 정의한다. 이때 전세보증금 반환보증을 신청할 수 있는 조건이 된다. 보증금액은 보증한도내에서 보증신청인이 신청한 금액이고 보증한도는 주택가격 x 담보인정비율(90%) - 선순위채권 등으로 계산된다. 갱신보증의 경우 담보인정비율이 이번년도 말까지 100%로 적용되고 그 이후에는 90%로 적용된다. 여기서 선순위채권 등은 보증신청인의 전세보증금보다 우선변제권이 인정되는 담보채권을 정의한다. 보증조건은 전세보증금과 선순위채권을 더한 금액이 '주택가격 x 담보인정비율(90%)' 이내여야 한다. 또한 등기부등본상 보증발급일 기준 주택 소유권에 대한 권리침해사항(경매신청,압류,가압류,가처분,가등기 등)이 없어야 한다. 선순위채권이 주택가액의 60% 이내여야 하고 주택의 건물과 토지(대지권)가 모두 임대인의 소유여야 한다. 이러한 보증한도와 보증금액을 잘 이해하고 있다면 전세계약을 이행할 때, 자신이 돌려받을 수 있는 금액을 생각한다면 역전세로 발생하는 피해 금액을 줄일 수 있을 것이다.

해당 분석과 모델링을 통해 자치구별 역전세 위험이 얼마나 상이한지 조사하고 어떠한 요인이

영향을 끼치는지 분석해보았다. 이 분석을 통해 조금이나마 역전세를 방지하는 방안이 모색되고, 각 지역마다 어떠한 정책을 수립할 지에 대해 도움이 되었으면 한다.

V. 참고 자료

1. 민병철(2021). 전세가격 변동률 분포를 활용한 역전세 위험의 측정, 부동산학연구, 27-2, 63-75.
2. 서울특별시(2023). 서울시 부동산 전월세가 정보, 서울 열린데이터광장, http://data.seoul.go.kr/dataList/OA-21276/S/1/datasetView.do
3. 서울특별시(2023). 서울시 부동산 실거래가 정보, 서울 열린데이터광장, https://data.seoul.go.kr/dataList/OA-21275/S/1/datasetView.do.
4. 한국도시연구소, [연구보고서] 2022년 상반기 실거래가 분석, https://kocer.re.kr/26/?q=YToxOntzOjEyOiJrZXl3b3JkX3R5cGUiO3M6MzoiYWxsIjt9&bmode=view&idx=13068129&t=board
5. 한국경제, 아파트 구축·신축 구분 왜 생겼나... "집값에 영향", 한국경제, 2018, 09.02. https://www.hankyung.com/economy/article/201809021141Y
6. 조한송, "이런 집은 피하세요" 전문가가 뽑은 역전세 위험지역 [부릿지], 머니투데이, 2022, 11.02. https://news.mt.co.kr/mtview.php?no=2022110115123462949
7. 랠리포인트, 2023년 03월 19일, https://rallypoint.tistory.com/353
8. Mighty Knowledge, 2020년 07월 17일, https://mightyknowledge.tistory.com/188
9. 동아일보, 2023년 05월 31일, https://www.donga.com/news/Economy/article/all/20230531/119549707/1
10. 머니투데이, 2023년 02월 02일, https://news.mt.co.kr/mtview.php?no=2023020212552920197