빅데이터 분석 기획서

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 분석 과제명 | 서울특별시 부동산 빅데이터와 AI를 이용한 실거래가 예측 분석 | | | | |
| 분석 기간 | 2020.10.26 ~ 2020.11.13 | | | | |
| 입주자(팀) | 소속기관 | KIC캠퍼스 | | 소재지 | 서울특별시 강남구 역삼동 819-10 세경빌딩3층 |
| 구분 | 성명 | 이메일 | 연락처 | 역할 |
| 팀장 | 김빛나라 | [qlcskfk02@gmail.com](mailto:qlcskfk02@gmail.com) | 010-9953-6646 | 데이터 수집 및 가공, 데이터 전처리, 데이터 분석, ppt 만들기 |
| 팀원1 | 장연주 | [duswn6193@naver.com](mailto:duswn6193@naver.com) | 010-4194-8322 | 데이터 수집 및 가공, 데이터 전처리, 데이터 분석, ppt 만들기, 점심 메뉴 정하기 |
| 팀원2 | 안솔몬 | [solmon0427@gmail.com](mailto:solmon0427@gmail.com) | 010-5530-2546 | 데이터 수집(웹 크롤링), 데이터 전처리, 데이터 분석, ppt 만들기, 발표 |
| 연구 배경 및 목적 | 주택 자산은 우리나라 국민 개인 자산 유형별 구성비의 68%를 차지할 만큼 그 비중은 매우 크다. 최근 몇 년간 주택 가격은 내 집 마련을 꿈꾸는 개인과, 부동산 시장의 안정화를 추구하는 정책당국 및 금융기관의 꾸준한 관심사로서, 중요한 경제학적 요인으로 인식되어 왔다. 하지만 부동산 시장은 수많은 이해관계가 얽혀 있고, 도시의 사회, 경제, 정치적 환경과도 매우 밀접하게 연관되어 있어 합리적인 가격 예측이 어렵다.  이에 본 연구는 주택 가격에 영향을 주는 다양한 경제적, 환경적 요인을 바탕으로 부동산 실거래가 예측 분석 모델을 개발할 것이다. 이를 통해 시민들이 부동산 시장의 흐름 변화에 발 빠르게 대응하도록 돕고, 왜곡된 가격정보로 인한 피해를 예방하며, 더 나아가 향후 정책당국과 금융기관의 부동산 정책이 나아갈 방향을 제공하고자 한다.  <데이터 > | | | | |
| 연구 및 분석 계획 | 1. 주택 가격에 영향을 주는 경제적, 환경적 요인 파악  2. 데이터 수집 및 가공  3. 데이터 분석을 통한 미래 실거래가 예측 분석 모델 개발 | | | | |
| 결론 및 한계점 | · 결론  GB모델을 통해 변수중요도를 확인한 결과 31개의 변수 중 특성중요도가 가장 높은 변수는 주택 면적이었다. 0.1이상 중요도를 가진 변수를 확인해보면 환경적인 요인인 한강과의 거리, 명문 학교의 개수, 명소의 개수 등의 변수가 각각 0.38, 0.46,0.44이었으며 아파트 내부요인으로는 동별 평균가격, 위치(경도/위도), 브랜드 이름이 0.62, 0.39, 0.36의 중요도를 가지는 것을 확인할 수 있었다. 경제적인 요인의 경우 0.01에서 0.09 사이의 중요도를 가지는 것으로 나타났다.  XGB모델로 GridSerchCV를 확인한 결과 가장 좋은 매개변수는 {'learning\_rate': 0.1, 'max\_depth': 8, 'n\_estimators': 3000}로 확인되었으며, 최종 test score는 0.98점이었다.  따라서 기존 연구의 가설대로 아파트 내부요인과 환경적 요인은 주택가격에 영향력이 큰 것으로 나타났으며, 본 연구에 사용된 변수들의 확인이 가능한 경우 다소 높은 확률로 예측이 가능함을 알 수 있었다.  · 한계점  -주택 가격에 경제적인 요인이 영향을 미치는 것으로 확인된 많은 선행연구들의 결과와는 다르게 본 연구에서는 경제적인 요인이 큰 영향을 끼치지 못했다. 경제적인 요인들에 대한 전문지식에 한계가 있어 변수연구를 통한 변수연구를 하지 못해 아쉬움이 남는다.  -기존 데이터셋에 아파트 내부요인에 대한 정보에 세대수, 방개수, 화장실개수 등 중요할 것으로 파악되는 정보가 없었다. 다양한 방법으로 데이터를 수집하려고 하였으나, 기존 정보와 정확히 일치하는 데이터는 찾지 못했으며, 일부를 제외하고 정보를 입력하는 경우 누락되는 값이 많아 반영하지 못했다.  -재건축 가능성을 반영하기 위해서는 건축년도 뿐 아니라 용적률, 단지의 크기, 안전진단 결과 등 파악해야하는 요건들이 많은데, 확인할 수 없는 정보들이 대부분이어서 변수에 반영하지 못했다.  -구별 평균 집값에 대한 정보를 구하기 위해서는 다른 변수와의 관계에 기초하여 변수의 값을 추정하는 보외법을 사용해야한다. 하지만 예상치 못한 변수로 인해 평균 집값의 변화가 큰 경우를 반영하기가 어렵다.  -생성한 모델을 통해 예측값은 구할 수 있었으나, 마감된 공모전 데이터를 이용해 분석을 해서 실제 2017년의 정답률을 확인할 수가 없었다. | | | | |

**<데이터셋>**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 번호 | 데이터셋명 | 필드 설명 | 데이터의 종류 | 출처 |
| 1 | 부동산 실거래가 데이터 | 도시, 동, 지번, 주소, 전용면적, 설립일자, 거래년월, 거래날짜, 층, 실거래가 | CSV | 데이콘 |
| 2 | 서울시 공원 정보 | 시, 구, 동, 공원이름, 공원종류, 면적, 개장년도 | CSV | 데이콘 |
| 3 | 서울특별시 아파트 정보 | 아파트명, 주소, 전체세대수, 법정동주소, 도로명주소 | CSV | 자료 수집  (서울 열린  데이터 광장) |
| 4 | 자사고 특목고 진학률 top 50 중학교 | 학교명, 소재구, 주소 | CSV | 자료 수집 |
| 5 | 서울대 진학률 top 100 고등학교 | 학교명, 소재구, 주소 | CSV | 자료 수집 |
| 6 | 관광 명소 | 상호명, 동, 구 | CSV | 자료 수집  (서울 열린  데이터 광장) |
| 7 | 상업시설 | 기간, 동, 합계, 대형마트, 백화점, 전문점, 쇼핑센터, 기타 | CSV | 자료 수집  (서울 열린  데이터 광장) |
| 8 | 서울교통공사 역사 주소 현황 | 호선, 역명, 구주소, 도로명주소 | CSV | 자료 수집  (서울교통공사 /  메트로9) |
| 9 | 한강공원 주소 | 이름, 주소, 위도, 경도 | CSV | 자료 수집  (한강사업본부) |
| 10 | 구별 청년인구 비율 | 연도, 구, 비율 | CSV | 자료 수집  (행정안전부) |
| 11 | 경기종합지수 (동행지수/선행지수) | 연도, 지수 | CSV | 자료 수집  (e-나라 지표) |
| 12 | 국내총생산 및 경제성장률 (GDP) – 연도별, 분기별 | 연도, 국내총생산, 경제성장률 | CSV | 자료 수집  (한국은행  「국민소득」) |
| 13 | 아파트매매 – 연도별, 월별 | 연도, 지역, 매매수 | CSV | 자료 수집  (국토교통부) |
| 14 | 주택담보대출금리 | 연도, 금리 | CSV | 자료 수집  (한국은행) |
| 15 | 코스피 지수 | 날짜, 종가, 오픈, 고가, 저가, 변동% | CSV | 자료 수집  (한국거래소  「증권선물」) |
| 16 | 환율 – 연도별, 월별 | 연도, 환율 | CSV | 자료 수집  (e-나라 지표) |
| 17 | 주택 구매력 지수 | 연도, 지수 | CSV | 자료수집  (한국감정원) |

**<일정관리>**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1주 (10/26~10/30)** | **2주 (11/02~11/06)** | **3주 (11/09~11/13)** | |
| **프로젝트 기획** |  |  |  | |
| **데이터 수집 및 정제** |  |  |  | |
| **분석 알고리즘 설계** |  |  |  | |
| **테스트 및 보완** |  |  | |  |
| **발표자료 및 보고서 작성** |  |  |  |  |