4차 4기 Project

|  |
| --- |
| 상가 경매 낙찰가 예측 |

**2019년 05월 22일**

**딥러닝 기반의 업종별 빅데이터 예측모델 분석 전문가**

**개 미 조**

정 호 성

이 재 희

전 혁 원

이 윤 지

김 고 은

**목 차**

**1. 프로젝트 개요**

1.1 프로젝트 기획 배경 및 목표

1.2 구성원 및 역할

1.3 프로젝트 추진 일정

**2. 프로젝트 현황**

2.1 경매 시스템 프로세스 이해

2.2 관련 논문 탐색 및 자료 조사

2.3 데이터 수집 계획

**3. 프로젝트 결과**

3.1 데이터 수집

3.2 탐색적 분석(EDA)

3.3 핵심 알고리즘(Random Forest, 선형회귀) 모델링

3.4 R shiny 구현

3.5 회차별 멘토링 결과

**4. 기대 효과**

4.1 향후 개선 사항

4.2 기대 효과

**5. 개발 후기**

**6. 강사 및 멘토 의견**

**1. 프로젝트 개요**

1.1 프로젝트 기획 배경 및 목표

* 딥러닝 기반의 빅데이터 분석 과정을 통해 배운 통계 지식을 기반으로 R program 과 Python의 데이터 분석 기술 및 크롤링 를 이용하여 데이터 수집과 가격 예측 모델 구축

1. 2018년 상반기까지 부동산 시장 활황에 따른 관심 고조
2. 주택 중심의 경매시장에서 수익형 부동산으로 변화 中
3. 경매물건 하자 유무 판단의 어려움 **-** 낙찰자가 부담해야 부분에 대한 금액 산정

[단계별 시나리오]

1) 프로토타입 모델 구축

- 분석대상: 2018년 하반기 서울시 상가 경매 물건 330개

2) 상권 분석 데이터 활용 모델 정확도 향상

3) 컨텐츠 제작

- 신건 유사물건 조회기능, 낙찰가격 대비 예상 수익률 제공

중점 사항 - 낙찰가격을 결정하는 요인 분석, 경매 프로세스 및 권리분석 이해

**권리분석 : 경매로 진행되는 부동산에 ‘인수되는 권리’ 유무 판단**

1.2 구성원 및 역할

|  |  |  |  |
| --- | --- | --- | --- |
| 이름 | 전공 | 역할 | 구현 부분 |
| 정 호 성 | 전자공학 | 팀 장 | 프로젝트 관리  data전처리,  보고서 작성 및 발표 |
| 이 재 희 | 컴퓨터공학 | 팀 원 | 경매 data 크롤링  기술 문제 해결 및 조언  R shiny 웹 구현  보고서 작성 |
| 전 혁 원 | 전자공학 | 팀 원 | 프로젝트 기획  data 크롤링 sys개발  데이터 분석 모델링 |
| 이 윤 지 | 경영정보 | 팀 원 | 공공 데이터 수집,  데이터 맵핑 및 전처리  유찰 예측 모델 구축 |
| 김 고 은 | 데이터 사이언스 | 팀 원 | 데이터 분석 모델링  모델링 시각화  보고서 작성 |

1.3 프로젝트 추진 일정

|  |  |  |  |
| --- | --- | --- | --- |
| 구분 | 기간 | 활동 | 비고 |
| 사전  기획 | 4/15(월) | 경매 프로세스 이해  전체 시나리오 논의 |  |
| 4/16(화),  4/17(수) | 관련 논문 탐색  데이터 조사 |  |
| 4/17(수)  4/18(목) | 데이터 주요항목 선정  오픈 API 정리  방향성 수립 |  |
| PJT  수행  /  완료 | 4/19(금) | 크롤링 시스템 구축  상가 실거래가 데이터 분석 |  |
| 4/25(목) | 1차 멘토링  [프로젝트 점검 및 기술자문] | 현업 멘토 참여 |
| 4/26(금) | 1차 중간 발표 | 4개 조 발표 |
| 4/29(월)~  5/3(금) | 크롤링 Data 전처리 및 기준 수립  유동인구 Data맵핑 , 상가 경매 Data 추가 항목 생성 및 전처리 |  |
| 5/9(목) | 2차 멘토링  [프로젝트 점검 및 기술자문] | 현업 멘토 참여 |
| 5/10(금)~  5/15(수) | 회귀 분석 결과 논의  유찰횟수 예측, 금액오차 개선 모델  컨텐츠 구상 및 제작 |  |
| 5/16(목)~  5/21(화) | 모델링[수치형(가격) - 선형회귀,  분류형(유찰) - Random Forest]  테스트 및 평가 |  |
| 5/23(목) | 최종 발표 (분석 완료 보고) | 최우수 한 팀 선발  멘토 평가 |

**2. 프로젝트 현황 [대제목 : 16pt / 맑은 고딕]**

2.1 경매 시스템 프로세스 이해

* YouTube ‘경매대마왕’의 “부동산 읽어주는 남자” 강의 공부

2.2 관련 논문 탐색 및 자료 조사

- 경매낙찰가 예측모델 논문 , 로보어드바이저 예측모델 자료 등

- 스피드 옥션 가입 (경매물건정보), 공공데이터포털(오픈API), 국가공간정보포털 등

2.3 데이터 수집 계획

1. 수집할 데이터 주요항목 선정 (16개)

* 실거래가, 위치 (도로명 주소), 층수 (건물 최대 층수/ 물건 층수), 평수
* 입지 (상권), 유동인구, 인접 주거시설, 교육시설(학교)
* 접근성 (근처 지하철, 버스정류장 위치)
* 행정적 조건 (그린벨트, 개발제한구역 등)
* 노후도 (건축 날짜), 전용률, 시공상태, 시설수준
* 인접 도로의 넓이, 인접 상가단지 규모
* 상가 내부 위치 (대로변 방향으로의 노출 정도)

1. 데이터관련 오픈API 목록 정리

* 자료정보(URL,API등)엑셀 파일 참조

1. 방향성 수립
   * 경매 낙찰가 예측 모델 프로토타입  만들기 ( 2주이내 )
2. 스피드 옥션 크롤링 시스템 구축

- 경매물건, 주소(KEY값), 감정가, 낙찰가 등 수치형 변수 위주 데이터

2.4 데이터 수집 및 전처리

(1). 크롤링 시스템 구축 ( BeautifulSoup + Selenium ) – speed auction page 크롤링

(지역:서울시, 수집기간 : 2015.01~2019.05 ) 상가 경매 data 3000건 수집

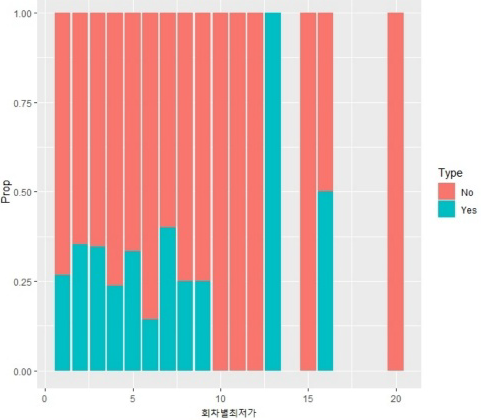
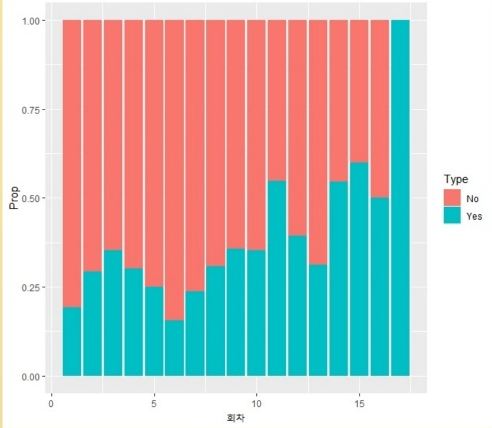
(2). 공공데이터(생활인구, 유동인구, 매출액) – 공공데이터 포탈(.csv download, 지역 – 서울시, 수집 기간 : 2018.06 ~ 2018.12) – 데이터 수집 후 상가 경매 data와 행정동 기준 및 도로명 기준으로 맵핑.

(3). 경매물건 Data 기본정보 및 추가 항목 생성 – 상권분석Data 전처리(대지권, 전용면적, `감정가, 최저가, 청구금액등, 신건여부, 경과기간), 낙찰가, 입찰수, 유찰 총 횟수, 매각기일/최저매감금액/, 회차별 하락률

**3. 프로젝트 개발 결과 [대제목 : 16pt / 맑은 고딕]**

3.2 탐색적 분석(EDA)

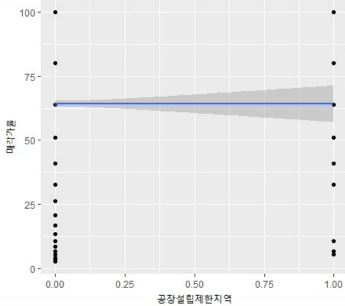
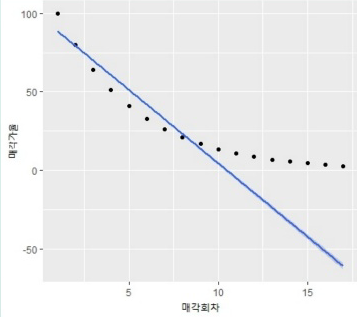
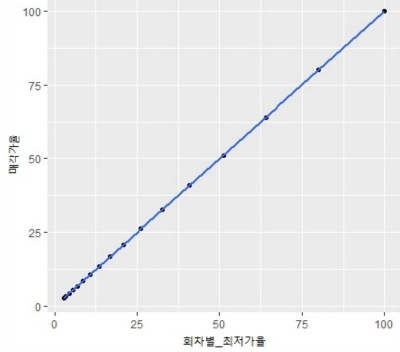
(1). 낙찰 여부에 대한 탐색적 분석



어떤변수가 낙찰여부에 영향을 미치는지 보기위해 시각화 및 분석

* 1. '회차'와 '낙찰여부' 간 선형적 관계 확인.
* 회차 - 명목형 변수로 구간별 낙찰여부 확인 - 뚜렷한 패턴 볼수 없음.
* T-test 결과 - 유의수준 0.05이하 두 변수 간 평균 차이가 있음을 증명.약간의 선형성이 존재.
* 2. 회차별 최저가 와 낙찰여부간 패턴 확인
* 최저가변수를 20개의 그룹으로 범주화 시켜 시각화
* 두 변수 간 아무런 상관관계가 없음.

(2). ‘매각가율’에 영향을 미치는 변수에 대한 분석

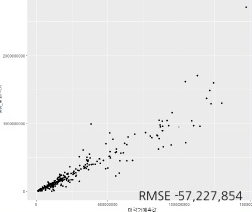
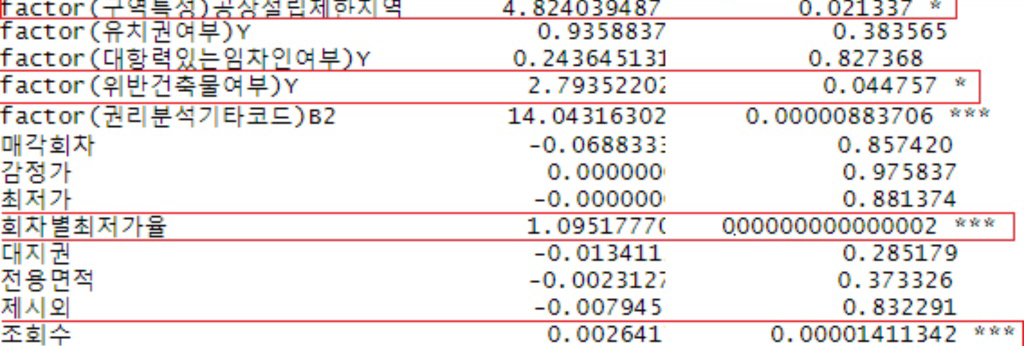


* 1. 회차별 최저가율과 매각가율에 강한 선형성 확인 - 회차별 최저가율 변수 선정
* 2. 매각회차와 매각가율의 산점도 그래프를 통해 나머지 변수들 보다 선형성을 띄는 것 확인 - 매각회차 변수 선정
* 3. 공장 설립제한지역과 매각가율의 선형성이 없지만 상관붑ㄴ석시 변수 자체의 영향력이 높은것을 확인 - 공장설립제한지역 변수 선정

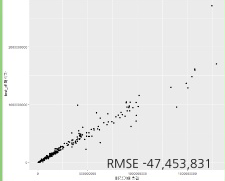
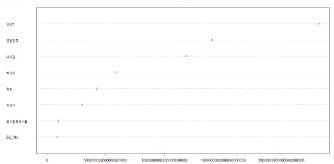
3.3 핵심 알고리즘(Random Forest, 선형회귀) 모델링

(1). 선형회귀

* 1. 선형회귀에서 P-value와 회귀계수 확인
* 2. 더미변수로 유의미한 항목들 확인(조회수, 회저별 최저가율, 권리분석 코드,)
* ※ RMSE- 57,227,854 (원)값 확인

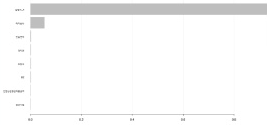


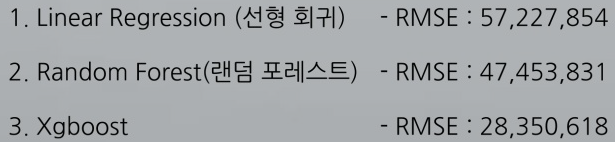
(2). Random Forest

* Random Forest에서 유의미한 항목들 확인(감정가,전용면적,대지권)
* ※ RMSE - 47,453,831 (원)값 확인
* 

(3). Xgboost

* Xgboost에서 유의미한 항목들 확인(감정가,매각회차)
* ※ RMSE - 28,350,618 (원)값 확인





-> R-MSE값을 지표로 성능이 가장 좋은 Xgboost 최종 모델 채택

3.4 R shiny 웹 구현

(1). Data set의 Traing data set으로 구현된 선형회귀 및 Random Forest을 통해 모델 구축

(2). 예측 모델Test set와 선택된 경매 물건과 Traing data유사 물건을

(3). 입지 분석 - 인접 상권명, 업종별, 월 매출 기준 Best업종, 유동인구 대비 비중이 큰 인구등

경매 크롤링 Data의 ‘물건현황 항목’을 텍스트 마이닝을 통해 위치/ 교통/ 편의시설 분류를 시각화하여 사용자에게 시각화하여 도출

(4). 예측 결과 – 사용자가 원하는 경매 물건을 선택 -> Random Forest로 예측된 유찰 횟수, 선형회귀로 예측된 최저가를 간단히 예측분석 가능

3.5 회차별 멘토링 결과

|  |  |
| --- | --- |
| 회차 | 내용 |
| 1회차 | 1. 새로운 주제 선정 – 프로젝트 주제 “주가 예측 모델 계획”에 대한 어려움과 data를 수집하기 어려움이 있어 새로운 주제 모색 - 상가 경매 낙찰가 예측 새로운 주제 모색 |
| 2회차 | **공통의견: 논문 모형 빠르게 구현해볼 것**   1. 윤효근멘토님: 시스템 아키텍쳐 및 최종 결과물에 대한 조언  * 논문관련 : 헤도닉가격모형(HPM) 이해 필요 * 시스템 구조 및 DB설계 : 물건명세/기일현황 DB분리                                          모델(산출식) / DB정의 / 컨텐츠(시각화) 업무 분담 의견제시   * **상권코드 맵핑시, 지번 -> 도로명변환 API 활용할 것** * 물건현황/계획 Text데이터 실용성低 -> 지도 GPS 활용 필요 * **결과물: 신건 입력시 기존DB에서 유사물건 조회 정보 제공, 가격 결정요인 제시**   **예상낙찰가격(1,2,3차, 입찰자수에 따라)제안, 지역별 매각가율 시각화 등**   * 향후계획 모색: 사업화 및 특허 출원  1. 박용진멘토님: 분석목표 및 고객정의 고민 필요(단순히 정확도만 높일것일지..)  * 결과물에 대한 활용방안 및 컨텐츠 고려 * 모델링시 주가예측 모델 참고 추천 * **차후 일정 : 5/3일까지 모델링 구현, 남은 2주간 컨텐츠(시각화) 제작 및 테스트**  1. 허주용멘토님: 참고사례 추천, 스토리텔링  * [데이터베이스진흥원 우수사례](http://www.dbguide.net/knowledge.db?cmd=specialist_list&boardConfigUid=104) 참고 * 상가 전체에 대한 예측력이 나쁘다면, 세그먼트 분류후 예측력이 좋은것을 사용할것   ( 10개 유형의 정확도 50~60% 보다는  Binary유형 정확도 80%가 낫다)   * **고객정의:  (소상공인 보호) 임차인을 위한 상가 경매 예측** |
| 3회차 | 1. 배당종결(매각완료)건에 대한 선형회귀 분석 결과 R-Squard: 0.98 (매우 우수함)               2. 향후방향                    1) 상권분석 데이터 활용 정확도 향상 확인  **2) 유찰여부 예측 모형 필요**  - 최종회차의 최저가를 변수로 사용해야함   3. 더미변수  권리분석코드 중 B2:공유물분할경매 만 유의하게 나옴        - 유치권, 대항력있는임차인여부 영향력 없음 확인 |

**4. 기대 효과**

4.1 향후 개선 사항

**(1) 유찰여부 예측 모형 필요**

신건 물건의 예측 결과에서 해당 물건이 몇 회차에 낙찰 될 것인지 예측 결과값을 나타내주고,

회차에 따른 가격 예측 모델 구현하여 서비스 제공

**(2) 여러가지 모델링 적용**

더 좋은 성능의 모델링을 선별하기 위한 여러 모델링 적용 후 탐색적 분석 및 앙상블 기법을 통해 성능 개선

**(3) 최적화된 프로그램 사용**

R shiny의 특성상 구동 속도가 느린 단점을 갖고 있어 data의 양이 많아 지게되면 사용자의 요구 대기 시간이 길어지게 될 것이며, 그런 불편함을 해결해주기 위해 조금 더 속도가 빠른 프로그램으로 예측 모델 서비스 구축

4.2 기대 효과

재테크 방식에서 대표적인 방법 중 하나인 부동산 투자.

부동산 투자 법엔 여러방법이 있지만, 부동산 경매의 시스템중 하나인 유찰 시스템을 통해 보다 합리적인 가격으로 투자 가능

하지만 경매는 진입 장벽이 높고, 전문 지식이 필요하며, 특히 권리분석에 대한 부분에 대하여 통찰력 있는 접근성이 필요

권리 분석등 데이터의 포함된 내용을 추가 금액으로 설정하는 등 가격에 영향을 주는 여러 변수들을 종합하여 간단히 가격 예측을 할 수 있어 합리적인 가격을 예측 할수 있다.

**5. 개발 후기.**

|  |
| --- |
| 팀 사진 첨부 |

|  |  |
| --- | --- |
| **성명** | **후기** |
| 정 호 성 | 프로젝트 기획의 방향성이 매우 뚜렷한 장점이 있었지만, 도메인 지식이 부족한 상태에서 프로젝트를 시작하여 데이터를 분석하지 못해 전처리와 새로운 변수 설정에 있어 힘든점이 있었다.  수업을 통해 배운 전처리 기술을 토대로 전처리 기술을 접목할 때, 데이터의 자료 구조에 대해 간과하여 여러 오류들을 직면하였고 자료 구조와 알고리즘의 공부는 꾸준히 해야한다는 점을 깨닫게 되었습니다.  또한 프로젝트 막바지에 보고서 및 발표 자료를 만들다 보니 그 동안 진행하는 과정 중 주간 보고서 및 자료들의 중요성에 대해 다시 한번 생각하게 되는 계기가 되었습니다. 또한 일정 부분에서 계획 했던 것보다 시간을 많이 소요하여 급하게 끝낸거 같은 아쉬움이 많이 남게 되었던 것 같습니다. |
| 이재희 | 기획단계에서 빠르게 주제를 전환하고 방향을 잡았기에 순조롭게 진행할 수 있었습니다.  다만 처음하는 분석 프로젝트이다보니 일정관리에서 조금 미숙한 점이 있었습니다.  이번 경험을 통해서 추후 프로젝트를 진행할 때에는 일정관리에 있어서 더 업무를 세분화하고 체계적으로 계획할 수 있을것으로 생각합니다.  개인 역할에서는 최종 결과물 시현을 위해 웹/앱 개발이 필요한 상황에서 가장 빨리 구축할 수 있는 방법을 고민했었습니다.  R언어를 배워보며 느꼈던 시각화 강점을 살려서 Shiny패키지를 찾아 3주동안 목표한 웹 결과물을 만들어내어 만족스럽습니다.  경매라는 생소한 주제임에도 팀원들 다 같이 공감대부터 형성하고 새로운 도메인지식을 쌓고 의견을 나누며  일관된 방향성을 유지할 수 있었기에 좌충우돌하지 않고 시작부터 끝까지 프로젝트를 즐겁게 마무리 할 수 있었습니다. |
| 전혁원 | 프로젝트 기획부터, 개발, 발표까지 팀원 모두가 각자 자신이 잘 할수 있는 부분에서 최선을 다 해주어서 프로젝트 기간내내 순조롭게 진행되었던것 같습니다. 부족한 기획임에도 불구하고 믿고 열심히 해주신 팀원들에게 감사합니다.  다만 아쉬웠던점은 처음 해보는 데이터 분석이다보니 데이터 수집이 오래걸리고 오류가 발생하면 수정하고 다시수집하는 과정이 길어진다는점을 감안하지 못하고 일정을 계획하였고, 결과물을 꼼꼼하게 확인하지 못해 데이터 수집을 몇번이고 다시하여 전체 일정이 지연된점이 아쉬웠습니다.  또한 개인적으로 R과 파이썬 중에 파이썬을 좋아했는데, 이번 프로젝트를 통해 R과 파이썬의 장단점을 몸으로 체감하고 R도 능숙하게 다룰수 있게 되어야겠다는 생각이 들었습니다. |
| 이윤지 | 프로젝트를 진행하며 가장 먼저는 도메인에 대한 중요성을 알았다. 도메인 지식이 있어야 데이터를 읽을 수 있고 정교하게 변수를 설정할 수 있었다.  두번째, 데이터 가공은 힌 번에 계획대로 되지 않는다는 것이다. 출처가 다른 여러 데이터를 병합하는 것이 나의 주된 일이었는데 사소한 이유로 혹은 데이터 파악이 미흡해서 많은 오류들을 접했다. 이 과정에서 가공 할 데이터를 살펴보는 안목과 데이터 병합 시에 어떤 프로세스로 해야하는지에 대해 설계하는 능력이 개발되었다.  마지막은 기술적인 부분이다. 경영정보를 전공했지만 컴퓨팅 사고나 코딩에 관해 깊이 공부하지 않아서 사소한 오류에도 처리하기가 쉽지않았는데, 반복하는 오류 혹은 고민하며 해결했던 오류들을 손에 익히며 자연스럽게 코딩 공부를 했고 이론으로 공부했던 기술지식을 실제 프로젝트에 적용하려 복습하고 고민하며 미흡했던 부분을 보완하는 시간을 가졌다. |
| 김고은 | 서로 다른 도메인지식을 가진 팀원들과 하나의 주제로 분석 프로젝트를 진행하면서 정말 많은 것들을 배워갈 수 있었습니다.  다만 제가 맡은 파트에서 여러 변수들에 대한 통계적 해석의 완성도가 높지않은 것 같아 많이 아쉬웠고, 추후 프로젝트 진행 시에는 이점을 보완하여 진행해야겠다는 생각이 들었습니다.  또 팀원(고은)의 늦은 합류에도 불구하고 모든 팀원들이 조급해하지않고 부족한 점을 함께 고민하며 격려해주셔서 짧은 시간 내에 맡은바를 완료할 수 있었던 것 같습니다. |

**6. 강사 의견**

|  |  |  |
| --- | --- | --- |
| 평 가 요 소 | 배점 | 평 |
| 아이디어 :  유사한 서비스 존재 유무 및  체계성 | /20 |  |
| 2. 개발 : 실제 구현 정도 및 배포 유무,  코드의 무결성 및 난이도, 현업적용도, 실무기술 반영정도 | /30 |  |
| 3. PJT 수행력 : 일정관리 및 역할분담, 목표 일정 달성도, 팀내 참여도 등 | /30 |  |
| 4. 준비도 : 프리젠테이션 및 프로젝트 준비 정도 | /20 |  |
| 계 | /100 | 강사 의견 필수 |