

## 낙찰률 예측 모형에 관한 연구

최보승<sup>a</sup>, 강현철<sup>b</sup>, 한상태<sup>1,b</sup>

“대구대학교 전산통계학과, <sup>b</sup>호서대학교 정보통계학과

### 요약

부동산 경매는 최근 새로운 부동산 투자방법 가운데 하나로 자리잡고 있다. 이는 부동산 시장의 성장과 더불어 부동산 경매 시장 또한 증가하고 있는 추세에 기인한다 할 수 있다. 본 연구는 부동산 경매에 참여하는 사람 및 기관들에게 가장 중요한 지표라 할 수 있는 낙찰률의 변화를 설명하고 예측하는 모형을 구축하고자 하였다. 월별 평균 낙찰률을 예측하기 위하여 단순한 지역별, 기간별 평균값을 보완하고 의사결정나무 분석을 이용하여 예측오차를 보정하는 방법을 제안하였고 선형회귀모형을 이용하여 개별 경매 물건별 낙찰률을 예측하기 위한 모형을 구축하였다. 구축된 모형은 전국 아파트 경매 물건에 적용하여 예측 모형을 구현하였으며 그 응용방법으로 예측결과에 대한 등급화를 함께 수행하였다.

주요용어: 부동산 경매, 낙찰률 예측, 회귀모형, 의사결정나무 분석.

### 1. 서론

한국의 부동산 시장은 IMF 위기 이후 기존의 비체계적 형태에서 벗어나 수익성과 합리성을 추구하는 선진 부동산제도를 도입하여 한 단계 성장하는 모습을 보여 주고 있으며, 이러한 시장의 변화는 1998년 6월 부동산시장의 전면적인 개방 및 선진적인 노하우를 가진 해외투자기관의 적극적인 국내 진출로 가시화되고 있는 실정이다. 최근에는 부동산펀드 등 다양한 부동산 관련 금융 상품 개발이 활발히 진행되고 있으며, 고수익 투자상품의 개발 및 부동산 투자신탁 등 전문성과 합리적인 가치평가, 수익방식에 의한 평가, 전문적인 자산관리 등 선진적인 노하우를 요청하고 있다. 최근 들어 부동산 경매 또한 중요한 부동산 투자상품으로 각광을 받고 있다.

부동산 경매시장에서 경매에 참여하고 있는 부동산 투자자들과 부동산 물건을 소유하고 있는 개인 및 기관에서 가장 관심을 가지는 분야는 경매 물건의 최종 낙찰률이라 할 수 있다. 이와 관련하여 국내 여러 기관에서는 부동산 경매와 관련된 각종 지표들을 제공하고 있다. 그러나 현재 제공되고 있는 대부분의 부동산 경매 관련 지표 및 정보들은 법원에서 제공하고 있는 경매물건에 대한 기본적인 사항에 대한 정리의 수준에 불과하며 낙찰률과 관련된 정보로는 지역별, 시기별 혹은 사용 용도별 평균을 제공하는 것에 그치고 있다.

부동산 경매와 관련된 우리나라의 연구 결과를 살펴보면 대부분의 연구가 부동산 경매에 참여한 여러 당사자들간의 법적 관계와 관련된 연구들이 주를 이루고 있으며 실제적으로 부동산 경매 자료를 이용한 실증적인 연구는 활발히 진행되고 있지 않은 형편이다. 실증적 차원에서 진행된 연구를 살펴보면 이정민과 이성근 (2007)은 부동산 경매에 있어서 낙찰률에 영향을 미칠 수 있는 여러 요인들을 조사하여 중요 요인을 찾고자 하는 연구를 진행하였고 이정민 (2008)은 상관분석을 이용하여 낙찰률과 여러 외적요인들간의 관계를 분석하였다. 그들의 연구는 실제적인 낙찰률에 대한 예측연구이기 보다는 낙찰률에 영향을 미치는 요인을 찾기위한 탐색적인 연구이다. 최보승 등 (2009)은 부동산 경매 시장의 호

<sup>1</sup> 교신저자: (336-795) 충청남도 아산시 배방면 세출리, 호서대학교 정보통계학과, 교수. E-mail: sthan@hoseo.edu

표 1: 주요 변수에 대한 기초통계

변수명	최소값	5분위수	평균	96분위수	최대값	중위수
총경매횟수	0	0	3	7	33	3
총유찰횟수	0	0	2	5	26	2
총토지전체면적	0	0	114	28,124	4,237,700,000	0
총토지실면적	0	0	5	4,145	4,181,637,152	28
총건물면적	0	0	93,825	227	733,530	3
총건물평	0	0	28	68	221,893	3
총제외면적	0	0	21	4	19,901,213	3
총제외평	0	0	6	1	6,020,117	3
총제외경매면적	0	0	21	4	19,901,213	3
총토지감정가	0	0	37,410	120,000	177,999,120	3
총건물감정가	0	0	35,626	124,481	182,266,460	3
총기계감정가	0	0	2,316	0	73,398,001	3
총감정가	0	4,656	213,877	534,523	15,342,325,000	55,000
최저경매가	0	0	89,692	268,800	4,514,100,000	27,076
낙찰금액	0	0	68,003	2,255	260,000,010	19,100
최저경매가	0	0	84,972	263,000	4,514,100,000	16,384
낙찰가	0	0	26,422	92,600	1,150,000,000	3

를 설명해 줄 수 있는 부동산 경매지수 산출을 위한 연구를 진행하였다. 낙찰률의 흐름을 모형화 하기 위하여 시차중속 회귀모형 (박유성과 김기환, 2002)을 이용하였고 구축된 모형으로부터 계산된 예측값을 지수화하여 최종 경매지수를 산출하였다.

부동산의 가격변동과 관련해서 통계청에서는 주택 및 토지에 관련된 각종 가격지수를 국가통계포털(www.kostat.kr)을 통하여 제공하고 있다. 그리고 국민은행 및 민간 부동산 정보업체에서도 자체적인 부동산 관련 가격지수를 제공하고 있다. 이들은 표본으로 선정된 주택을 대상으로 하여 등록된 공인중개사로부터 자료를 제공받아 라스파이레스 방법을 이용하여 부동산 가격지수를 산출하고 있다 (박현수, 2007). 김원년 등 (2009)은 아파트 임대료 자료를 이용하여 hedonic housing price function (Lancaster, 1966; Rosen, 1974)을 적용하여 임대가격을 산출하고 이를 일반아파트와 공공아파트 간의 비교분석에 적용하였다. 박진우 (2009)는 현재 우리나라에서 구축되어진 주택 가격과 관련된 각종 지수들을 비교하여 그 차이 및 문제점을 지적하고 적절한 활용방안에 대한 연구를 진행하였다.

이상에서 살펴본 바와 같이 부동산 관련 연구들은 대부분 부동산 가격과 직접적 연관된 수치 및 지수들에 대한 연구가 광범위하게 진행되어 왔으며 부동산 경매와 관련된 연구들도 탐색적인 측면에서 경매 낙찰률에 영향을 미치는 요인을 찾고 (이정민과 이성근, 2007; 이정민, 2008) 전반적인 경매 시장의 흐름을 파악할 수 있는 지수의 개발 (최보승 등, 2009)에 관련된 연구에 한정되어 왔다. 본 연구의 목적은 보다 구체적으로 부동산 경매에서 가장 핵심이 되면서 중요한 요인인 경매 물건에 대한 최종 낙찰률을 예측하는데 있다. 현실적으로 경매 정보를 전문적으로 제공하고 있는 업체들은 대부분 과거 부동산 경매 물건 가운데 낙찰이 이루어진 물건을 대상으로 하여 단순한 지역별, 기간별 평균치를 계산하여 이를 경매 낙찰률에 대한 예측치로 제공하고 있는 실정이다. 본 연구에서는 낙찰률 예측을 위한 평균 낙찰률 산출의 방법을 개선하여 보다 정확하고 현실적인 평균 낙찰률 예측 방법을 구축하고자 하였다. 본 논문의 구성은 다음과 같다. 2장에서는 경매 낙찰률 예측모형을 구축하기 위하여 이용된 경매관련 자료에 대한 분석용 데이터 마트의 구축 및 데이터의 특징에 대하여 설명하였고 3장에서는 낙찰률 예측모형의 구현과정에 대하여 논하였다. 보다 구체적으로 단순한 지역별, 기간별 평균치로부터 출발하여 지역별, 기간별 특성에 맞추어 평균치를 계산하는 방법을 고려하였다. 그리고 평균을 기반으로 하여 계산된 경매 낙찰률에 대한 오차를 줄이기 위한 방법으로 의사결정나무분석(decision tree

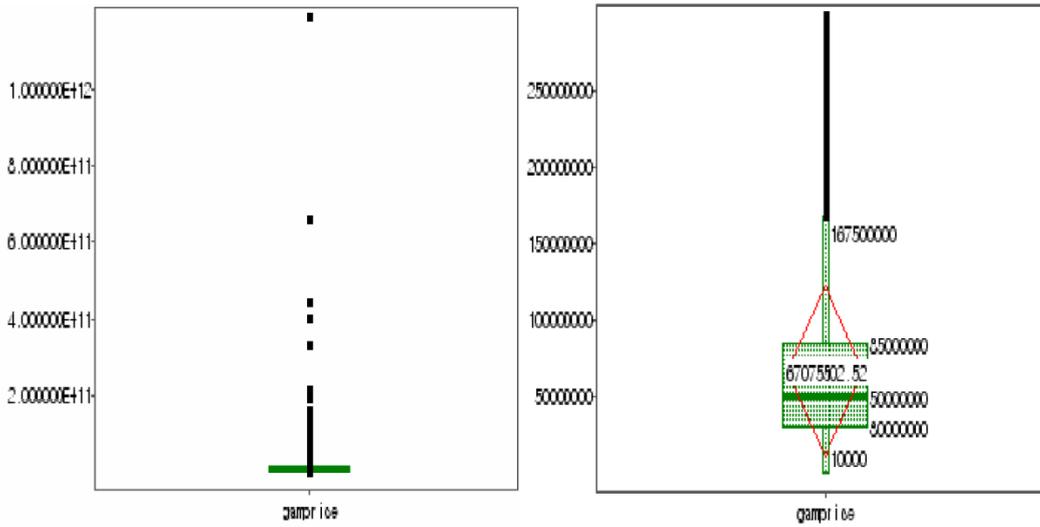


그림 1: 특이치제거 이전과 이후 감정가의 분포

analysis)을 적용하였다. 최종적으로는 선형회귀모형을 이용하여 낙찰률 예측 모형을 구축하였다. 마지막 4장에서는 모형의 활용방안 및 한계점에 대하여 논하였다.

## 2. 분석용 데이터 마트의 구축 및 데이터 탐색

### 2.1. 데이터 추출 및 분석용 데이터 마트의 구축

경매지수 개발 및 낙찰률 예측 모형에 필요한 데이터들은 국내의 한 부동산 경매 정보 제공 업체로부터 제공 받았다. 부동산 경매 자료의 기본 단위는 경매 물건이다. 각 개별 경매 물건은 그 형태에 따라 여러 가지 용도로 구분될 수 있다. 주요 용도로는 주거형태에 따라 아파트, 다세대 주택, 단독 주택 등이 있으며 상업적 용도에 따라 상가, 공장, 복합건물 등으로 구분할 수 있다. 그리고 일반적인 토지도 하나의 용도로써 구분될 수 있다. 본 연구에 이용된 자료는 이 7가지 용도에 동산 부분을 더하여 총 8개의 용도로 구분하였다. 분석용 데이터 마트를 구축하기 위하여 추출된 전체 자료는 1998년 1월 1일부터 2006년 12월 31일 까지 경매가 이루어진 전체 자료로써 총 1,238,480건이다. 다음 표 1은 이 가운데 주요변수에 대한 기초 통계량을 정리한 표이다. 변수들은 크게 3가지 영역으로 구분할 수 있는데 총경매횟수와 총유찰횟수는 빈도로 표현되고 면적과 관련된 변수들은 총건물평을 제외하고 모두 제곱미터 단위이다. 마지막 가격과 관련된 변수들은 모두 천원단위 이다. 각 통계량 가운데 중요하게 살펴 보아야 하는 값들은 먼저 95백분위수와 최대값 간의 차이이다. 대부분의 변수에서 95백분위수에 해당하는 숫자에 비하여 최대값이 지나치게 큰 값을 가지고 있는 것을 확인할 수 있다. 이는 데이터에 매우 큰 특이치가 존재하고 있다고 볼 수 있으며 이에 따라 평균 또한 상대적으로 큰 값을 가지게 된다. 일반적으로 자료의 분포가 대칭인 경우 평균과 중위수가 비슷한 값을 가지게 되나 경매 데이터들은 모두 평균이 중위수보다 매우 큰 것을 확인할 수 있다. 분석결과의 안정성과 대표성을 확보하기 위하여 사전에 특이치에 대한 분석을 수행하였고 분석용 데이터 마트의 구축과정에서 특이치를 제거하도록 하였다. 특이치를 제거하기 위하여 여러가지 기준을 적용하였고 그 가운데 각 변수별로 상위 1% 이상의 값을 특이치로 고려하여 제거하였다. 그림 1은 감정가에 대한 상자그림으로 왼쪽은 전체 데이터를 이용한 상자그림이고 오른쪽은 상위 1%의 값들을 제거한 후 작성한 상자그림이다. 그림에서 보는 바와 같

표 2: 실제 감정가와 표준화 감정가의 비교

기간	최초 감정가	표준화 감정가	차이
1	85,000,000	85,209,117	209,117
2	95,000,000	95,135,262	135,262
3	92,000,000	92,550,688	550,688
4	85,000,000	86,486,490	1,486,490
5	95,000,000	96,629,841	1,629,841
6	90,000,000	91,850,107	1,850,107
7	95,000,000	96,492,273	1,492,273
8	95,000,000	97,009,069	2,009,069
9	95,000,000	96,999,506	1,999,506
10	82,000,000	83,220,813	1,220,813
11	90,000,000	98,596,045	8,596,045
11	70,000,000	75,815,177	5,815,177

이 특이치가 제거됨으로써 여전히 오른쪽 긴 꼬리분포를 보이고 있으나 그 쏠림현상이 상당히 완화된 것으로 보일 수 있다.

다음으로 데이터 정제를 수행하였는데 2000년 이전 데이터는 데이터의 입력에서 다수의 입력오류의 존재 가능성이 있는 것으로 판단되어 분석용 데이터에서 추가적으로 제외하였다. 마지막으로 낙찰률을 기준으로 하여 낙찰률이 20% 미만인 자료와 130% 이상이 되는 자료를 추가적으로 제외하였다. 본 연구의 가장 궁극적인 목적은 각 경매 물건별 낙찰률에 대한 예측이고 일반적인 경매 상황에서 낙찰률이 이 범위를 넘는 경우 특수한 경우로 고려할 수 있다. 이와 같은 단계에 걸친 데이터 추출 및 정제 과정을 거쳐 분석용 데이터 마트로 총 680,989건의 자료가 이용되었다.

## 2.2. 파생변수의 생성

본 연구의 기본적인 목적 가운데 하나는 낙찰률의 예측에 있다. 구축된 데이터 마트로부터 낙찰률 예측을 수행하기 위한 최적의 모형을 구축하기 위하여 낙찰률의 움직임에 영향을 줄 수 있는 파생변수를 추가적으로 생성할 필요가 있다. 본 연구에서는 파생변수로 특수권리와 표준화 감정가를 생성하였다. 특수권리란 개별 물건에 부여된 다양한 법적 권리를 의미한다. 일반적으로 특수권리가 부여된 경매 물건은 낙찰률이 그렇지 않은 물건에 대하여 감소하는 경향이 있다. 먼저 전체 특수권리 총 16건에 대하여 낙찰률간의 영향력 분석을 수행하였다. 16개 특수권리는 각각 유치권, 대지권없음, 지분경매, 재경매, 토지별도등기, 선순위임차인, 선순위가등기, 선순위지상권, 법정지사권, 분묘기지권, 채권자매수신청, 공유자매수신청, 예고등기, 소유가등기, 전소유가압, 선순위가처분으로 구성되었다. 낙찰률간의 관계를 분석한 결과 낙찰률의 변동에 통계적으로 의미 있는 영향을 미치지 않거나 그 영향력이 반대로 나타나는 특수권리를 제외하였고(특수권리가 부여된 경우에 낙찰률이 증가하는 경우) 최종적으로 유치권, 대지권없음, 지분경매, 재경매, 토지별도등기, 선순위임차인, 선순위가등기, 선순위지상권으로 총 8개를 사용하여 특수권리 파생변수를 생성하여 분석용 데이터마트에 추가하였다. 감정가는 낙찰률을 결정하는데 있어서 가장 중요한 변수이다. 그리고 데이터 마트에 구축된 감정가는 경매물건이 최초로 등록된 시점에 부여된 감정가이다. 그러나 최초 경매시점으로부터 실제 낙찰이 이루어지는 시점에는 일정정도 시차가 발생할 수 있다. 그 시차가 길지 않은 경우에는 감정가의 변화가 크지 않을 것이나 시점이 길어지게 되면 그에 따라 감정가의 변화가 발생하게 된다. 따라서 최초 경매 시점에서 고려된 감정가를 시차에 따른 변화를 고려하지 않고 낙찰률 예측이 반영한다면 현 시점에서 부동산 물건에 부여된 시세를 반영할 수 없으며 잘못된 낙찰결과에 대한 예측치를 제공할 수 있게 된다. 따라서 시점간의 차이를 고려한 현실화된 감정가의 사용은 정확한 낙찰률 예측에 매우 중요한 요인이라 할 수

있다. 최보승 등 (2009)은 낙찰 시점의 적절한 감정가를 반영하기 위하여 부동산 매매지수와 지가 변동률을 이용한 감정가를 낙찰이 이루어지는 시점의 감정가로 표준화하는 과정을 진행하였고 생성된 감정가는 표준화 감정가로 명명하였다. 표준화 감정가 생성 절차는 다음과 같다.

1 단계: 감정가가 생성된 시점과 낙찰된 시점의 부동산 매매지수(www.kbstar.com)를 구하여 그 차를 계산한다.

2 단계: 표준화 낙찰가를 계산한다.

$$\text{표준화 감정가} = \text{초기감정가} \times (100 + (\text{낙찰시점 매매지수} - \text{감정시점 매매지수})/100)$$

이와 같은 과정을 통하여 각 물건에 대한 표준화 감정가를 용도별로 생성하였으며 표준화 감정가에 사용된 부동산 매매지수의 생성 시점을 고려하여 2002년 1월을 기준으로 표준화 감정가를 생성하였다. 표 2는 생성된 표준화 감정가의 생성 예로 2006년 12월에 낙찰이 이루어진 경매 자료를 대상으로 하여 감정 시점과 낙찰시점간의 차이를 고려한 표준화 감정가와 최초 감정가를 정리한 표이다. 표의 첫 번째 열은 최초 경매 시점의 감정가를 나타내고 두 번째 열은 표준화 감정가를 나타낸다. 시간이 길어질수록 최초 감정가와 표준화 감정가 간의 차이가 전반적으로 벌어지는 것을 볼 수 있으며 모든 경우에 있어서 표준화 감정가가 최초 감정가보다 큰 것을 볼 수 있다. 이는 낙찰 시점인 2006년 12월을 고려할 때 2006년 한 해 동안 지속적인 부동산 가격의 상승 추세를 고려한다면 일정 정도 당연한 결과 할 수 있다. 마지막 열은 표준화 감정가에서 최초 감정가를 뺀 값이다.

### 3. 낙찰률 예측 모형

낙찰률 예측 모형은 현재 부동산 경매와 관련하여 낙찰률 정보의 정확도를 높이는데 그 목적을 두고 있다. 기존의 부동산 경매 관련 정보 제공 업체들이 이용하는 예측 낙찰률 계산 방식은 지역과 용도를 고려하여 주어진 기간 동안의 평균 낙찰률의 평균치를 계산하여 이를 예측 낙찰률로 산출하는 방식이다. 지역은 광역자치단체를 대 분류로 하고 광역자치단체 내에서 기초자치단체를 소 분류로 하는 2단계로 나눌 수 있다. 그러나 이런 2단계를 지역 분류에서 기초자치단체에 따른 분류는 광역자치단체에 따른 분류에 비하여 세분화 되어 있다. 본 연구에서는 두 지역 구분이 차이를 다소나마 극복하기 위하여 지역구분의 중간단계로써 자립도 단위의 지역구분을 추가하였다. 여기서 자립도란 각 광역시도 별로 각 시군구를 재정자립도에 따라 4개의 등급으로 나누어 산정한 것으로써 예를 들어 서울의 경우 서초구, 중구, 강남구, 송파구, 영등포구, 종로구는 1등급, 양천구, 용산구, 강동구, 마포구, 성북구, 구로구, 동작구는 2등급, 강서구, 서대문구, 광진구, 동대문구, 성동구, 도봉구는 3등급, 마지막으로 관악구, 은평구, 금천구, 중랑구, 노원구, 강북구는 4등급으로 분류된다. 다음으로 용도에 따른 구분이 있는데 용도는 경매 부동산의 유형에 따라 크게 8가지로 구분할 수 있다. 본 연구에서는 8개의 주요 용도 가운데 가장 큰 비중을 차지하고 있는 아파트를 기준으로 하여 낙찰률 예측 모형을 구축하고자 한다. 낙찰률을 계산하는데 있어서 상대적으로 작은 지역에서(도 단위 보다는 시군구 단위) 그리고 비교적 최근 기간(최근 12개월 보다는 최근 3개월)에 낙찰된 물건들을 이용하여 평균을 계산하였다면 이로부터 계산된 낙찰률이 상대적으로 실제 낙찰률에 더 근접해 있을 것이라 생각할 수 있다. 그러나 이와 같이 상대적으로 짧은 기간이나 좁은 지역을 이용하여 낙찰률을 계산하는 데에는 충분한 데이터 확보의 어려움이 따른다. 부동산 시장에서 경매라는 사건은 일반적인 부동산 매매처럼 빈번하게 발생하는 사건이 아니다. 따라서 지역을 시군구 단위로 한정하고 최근 3개월 동안 낙찰된 경매 물건에 대한 낙찰률을 계산하게 되면 대표성을 가지기 위한 충분한 자료의 수를 확보할 수 없게 된다. 충분한 자료를 확보하기 위하여 지역을 시도 단위로 확장하거나 기간을 3개월대신 6개월이나 12개월로 확장하게 되면 충분

한 자료는 확보되지만 너무나 광범위한 범위를 다루기 때문에 이 역시 예측력과 대표성이 떨어지는 어려움이 따른다.

본 연구에서는 이와 같은 어려움을 극복하고 충분한 대표성을 확보할 수 있는 보다 정교한 평균 낙찰률 계산을 위한 방법을 제안하고자 한다. 그 방안은 지역과 기간 확장을 통한 용도별 평균 낙찰률의 계산이다. 이와 더불어 의사결정나무분석(decision tree analysis) 기법을 이용하여 추가적인 예측오차를 감소시키기 위한 방법을 함께 개발하였다.

### 3.1. 기간별, 지역별 확장을 통한 새로운 낙찰률 예측모형

용도에 따른 예측 낙찰률을 계산하는 방법은 기본적으로 과거 자료의 평균에 의하여 계산한다. 현 시점에서 예측되는 물건에 대한 낙찰률은 동일한 지역의 과거 낙찰된 결과를 기반으로 하는 것이 타당할 것이다. 즉 해당 지역의 과거 평균 낙찰률을 이용하여 현 시점에서 예상되는 낙찰률을 예측해 보고자 하는 것이다. 즉 현재의 시점을 기준으로 하여 최근의 낙찰이 이루어지고 동일한 행정구역에 낙찰이 이루어진 물건을 중심으로 현시점의 경매 물건에 대한 예측 낙찰률을 생성하는 것이다. 그러나 동일한 지역의 과거 자료의 수가 충분히 확보되지 않은 상황 하에서 평균을 계산하는 경우, 특히 소수의 물건에서 지나치게 큰 낙찰률을 보이거나 반대로 지나치게 작은 낙찰률을 보인 경우 이 소수의 물건에 의하여 평균 낙찰률이 영향을 받게 될 것이다. 이러한 경우에 계산된 낙찰률은 결과적으로 현시점의 상황을 정확히 반영하지 못하는 문제가 발생할 수 있으며 전체적으로 예측오차를 증가시키는 문제를 야기할 수 있다. 즉 예측의 관점에서는 가장 최근에 동일한 지역에서 낙찰이 이루어진 물건을 가지고 낙찰률을 예측하는 것이 가장 좋은 방법이겠으나 충분한 자료가 확보되지 않은 상황 하에서 예측 낙찰률을 구하게 되면 대표성이 떨어지는 문제가 발생하게 된다.

본 연구에서는 이와 같은 문제를 동시에 해결하기 위하여 지역 확장과 기간확장을 고려한 낙찰률 계산과정을 구축하였다. 평균 낙찰률을 구축하기 위한 최소 기본단위로써 지역별로는 시군구를 기본단위로 하고 기간으로는 최근 3개월을 기준단위로 이용하였다. 예를 들어 2007년 1월 강남지역의 아파트 용도에 대한 예측 낙찰률을 계산하는데 있어서 2006년 10월부터 2006년 12월까지 강남구에서 낙찰이 이루어진 아파트 경매 물건의 평균 낙찰률을 계산하여 이를 예측 낙찰률로 이용하는 것이다. 단 강남구의 3개월 간 낙찰된 물건의 수가 20건 이상이 되는 경우에만 평균을 계산하였다. 만약 20건이 되지 않는다면 기간을 확장하여 2006년 7월부터 2006년 12월까지 과거 6개월 동안의 낙찰된 물건을 사용하여 평균을 계산한다. 만약 과거 6개월에서도 20건의 물건이 확보되지 않으면 이를 다시 12개월까지 확장을 시도하였다. 즉 기간확장은 3개월을 기준으로 하여 6개월, 12개월로 확장하였다. 또 다른 확장으로는 지역 확장을 고려하였다. 지역 확장은 3가지 단위로 구분하여 시군구별, 자립도별, 광역시도별로 하였다. 기간확장과 지역확장 과정을 통한 평균 낙찰률의 계산과정은 다음과 같다.

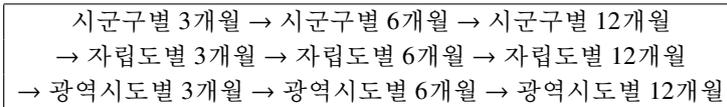
1. 최소단위는 시군구별, 3개월을 기준으로 한다.
2. 기준단위에서 자료의 수가 20이 넘으면 기간확장이나 지역확장을 수행하지 않는다.
3. 기준단위에서 자료의 수가 20보다 작으면 기간확장을 먼저 수행한다. 기간확장을 통하여 자료의 수가 20이 넘으면 기간확장을 중단한다. 만약 6개월로 기간확장을 수행하였을 때 자료의 수가 20이 넘어가게 되면 기간확장을 중단한다.
4. 기간확장을 12개월까지 수행했을 때 자료의 수가 20을 넘어가지 않으면 시군구 단위에서 자립도별 단위로 지역확장을 수행한다.

표 3: 기존 방법과 제안 방법간 예측오차 비교

모형	시군구 3개월	시군구 6개월	시군구 12개월	자립도 3개월	자립도 6개월	자립도 12개월	도별 3개월	도별 6개월	도별 12개월
Method 1	15.42	15.48	15.76	15.61	15.70	15.93	15.81	15.95	16.17
Method 2	14.77	14.96	15.11	15.52	15.24	15.26	15.26	15.26	15.25

5. 이와 같은 과정을 반복하여 마지막 광역시도별 12개월까지 확장을 수행한다. 각 단계에서 자료의 수가 20을 넘게 되면 더 이상 확장을 수행하지 않는다.

결과적으로 확장의 과정을 정리하면 다음과 같다.



이와 같은 방식에 의하여 계산된 기간확장과 지역확장을 고려한 예측 낙찰률 구축을 위한 방법은 월 단위를 기준으로 수행되었다. 이 방법에 의한 낙찰률의 예측력을 기존의 단순 평균 계산에 의한 예측 방법과 비교를 수행하기 위하여 다음과 같은 평균절대비율오차(mean absolute percent error; MAPE)을 이용하였다.

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|,$$

여기서  $y_i$ 는 개별 물건의 실제 낙찰률이고  $\hat{y}_i$ 는 예측 낙찰률을 의미한다. 이 값이 작을수록 예측력이 뛰어나다고 할 수 있다. 이제 실제 분석을 통하여 예측의 유용성을 보기로 하자.

비교를 위하여 기존의 단순 평균 계산 방법은 본 연구에서 제안하는 방법과 유사한 과정을 거쳐 지역확장과 기간확장을 고려하였다. 그러나 기존 방법은 각 단계에서 자료의 수를 고려하지 않고 절대적인 지역기준과 기간기준을 고려하여 낙찰률에 대한 평균을 계산하고 이를 낙찰률에 대한 예측값으로 이용하였다. 분석의 편의를 위하여 기존 방법을 Method 1로, 새로이 제안하는 방법은 Method 2로 하였다. 표 3은 기존의 방법과 제안된 방법을 이용하여 2006년 전체 자료에 적용한 후 평균을 계산하여 예측오차를 정리한 표이다. 결과를 비교해 보면 모든 시점의 결과에서 기존 방식의 결과 보다 제안된 방법이 더작은 MAPE를 보임을 알 수 있다. 이는 실제 낙찰률과 예측 낙찰률의 차이가 작다는 것을 의미하며, 제안모델이 기존 방법보다 더욱 정교한 낙찰률을 제공해 준다는 것을 확인할 수 있다.

### 3.2. 의사결정나무분석을 이용한 예측오차 보정

3.1절에서 살펴본 바와 같이 본 연구에서 새롭게 제안하고 있는 월별 낙찰률 예측 방법은 기존의 방법보다 예측오차를 줄일 수 있는 향상된 방법이라 할 수 있다. 그러나 새로이 제안되는 방법에 의해서도 일정 정도 이상의 오차가 발생하고 있다. 이런 예측오차를 추가적으로 줄이기 위하여 본 연구에서는 의사결정나무 분석(Decision tree) 기법을 이용한 예측오차에 대한 보정을 수행하였다. 의사결정나무 분석의 자세한 설명은 강현철 등 (2006)을 참조하라.

의사결정나무분석을 이용하여 예측오차를 줄이기 위한 보정과정은 먼저 과거 데이터를 이용하여 모형을 구축하는 과정과 구축된 모형을 가지고 모형의 평가를 수행하기 위해 검증용 자료를 이용하여 모형에서 생성된 규칙을 적용한 후 예측오차가 얼마나 감소하였는지를 평가하는 평가 과정으로 구분할 수 있다. 먼저 모형 구축을 위하여 2005년 10월부터 2006년 9월까지의 1년간 낙찰된 전체 데이터를

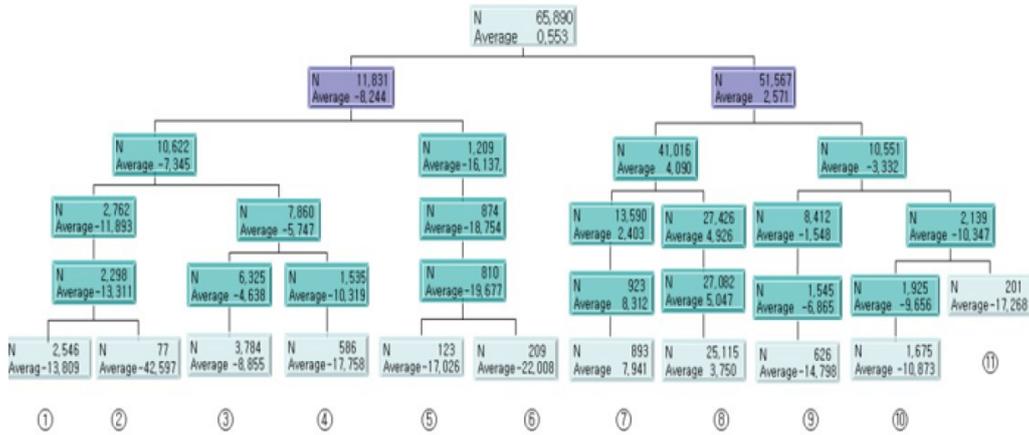


그림 2: 의사결정나무 기법을 이용한 예측오차 분석

이용하였다. 이 자료에 의하여 구축된 모형을 2006년 10월부터 2006년 12월까지의 자료에 적용하여 예측오차의 향상 정도를 평가하였다. 이제 아파트 용도에 대하여 구축된 의사결정나무에 의한 오차 보정 과정 및 결과를 살펴보자. 그림 2는 의사결정나무분석에 의한 트리구조를 나타내고 있다. 여러가지 설명요인 가운데 모형에 선택된 변수는 총 10개로 총 건물 면적, 총 건물 감정가, 총 건물 경매면적, 총 감정가, 총 토지 경매면적, 감정가, 임차횟수, 보증금 합, 월세 합, 특수권리 수가 선택되었다. 아파트 용도 물건의 2005년 9월부터 2006년 9월까지 1년 동안의 건수는 총 65,890건으로 나타났으며 의사결정나무 분석에 의한 분리 규칙은 다음과 같다. 예를 들어 ①의 경우 감정가는 23,400,000보다 작으며, 특수권리 수가 0이거나 1이며, 총 토지 경매면적이 42.9보다 작은 집단으로써 이러한 조건을 만족하는 ①의 경우 경매가 이루어진 건수가 2,546건이며, ①의 예측오차는 -13.809 이다.

1. 감정가 < 23,400,000; 특수권리수 = 0, 1; 총토지경매면적 < 42.9.
2. 감정가 < 23,400,000; 특수권리수 = 0, 1; 42.9 ≥ 총토지경매면적 < 43.203.
3. 23,400,000 ≤ 감정가 < 34,700,000; 특수권리수 = 0, 1; 월세합 < 51,750; 총토지경매면적 < 24.358.
4. 23,400,000 ≤ 감정가 < 34,700,000; 특수권리수 = 0, 1; 51,750 ≤ 월세합; 월세합 < 71,000.
5. 감정가 < 34,700,000; 특수권리수 = 2, 3, 4; 총토지경매면적 < 22.797; 26.8175 ≤ 총건물면적; 보증금합 < 9,100,000.
6. 감정가 < 34,700,000; 특수권리수 = 2, 3, 4; 총토지경매면적 < 22.797; 26.8175 ≤ 총건물면적; 9,100,000 ≤ 보증금합.
7. 34,700,000 ≤ 감정가 < 59,500,000; 특수권리수 = 0; 31,831,000 ≤ 보증금합; 임차횟수 < 2.5.
8. 34,700,000 ≤ 감정가 < 59,500,000; 특수권리수 = 0; 총건물면적 < 169.1745; 월세합 < 95,000.
9. 34,700,000 ≤ 감정가 < 42,986,669; 특수권리수 = 1; 총토지경매면적 < 17.455.
10. 34,700,000 ≤ 감정가; 특수권리수 = 2; 월세합 < 148,000.
11. 34,700,000 ≤ 감정가; 특수권리수 = 3, 4.

표 4: 의사결정나무분석을 통한 예측오차 보정 결과

방법	개체수	MAPE
Method 1(전체)	9,012	14.837
Method 2(전체)	9,012	12.536
의사결정나무 보정(전체)	9,012	12.505

표 5: 모형구축에 이용된 설명변수

변수명	설명
yy_do_12	Method 2에 의해 생성된 12개월 도 단위 평균
imchaprice_sum	임대료로써 건물에 부여된 전세보증금의 합
tot_land	총토지실면적
tot_bd	총건물면적
bd_ext	총건물면적 지시자(면적이 360이상이면 1, 아니면 0)
addr_do_1	새로운 시도 구분(23개 범주, 도별 도시지역과 농촌지역을 구분)
jaripdo	도별 자립도 등급(10등급으로 부여됨, 등급이 높을수록 자립도가 낮음)
sp_num_1	특수권리의 수
log_st_gamprice_1	로그 표준화 감정가
nak_index	아파트 매매지수
imcha_yn	임차인의 존재 유무

이제 의사결정나무분석을 통하여 생성된 규칙을 적용하여 3.1절에 제안된 결과에 대한 보정을 수행하였다. 보정된 결과를 평가하기 위해 과거 1년 동안(2005년 9월 ~ 2006년 9월)의 자료를 이용하여 의사결정나무 모형을 구축하고 이를 2006년 10월부터 12월까지 3개월 동안의 자료에 적용하여 그 결과를 비교하였다. 표 4는 아파트 용도에 대한 예측오차를 보정한 결과이다. 표 4를 살펴보면 3개월 동안 총 경매가 이루어진 물건의 수는 9,012건이다. 기존의 단순 평균 방법(Method 1)에 따라 예측 낙찰률을 산출하고 이를 실제 낙찰률과 비교한 후 MAPE는 14.837이다. 같은 방법으로 제안방법(Method 2)을 적용했을 때 MAPE는 12.536이고 의사결정나무분석에 의한 보정 후 MAPE는 12.505이다. 즉 의사결정나무분석을 이용한 보정을 통하여 추가적으로 예측오차를 줄일 수 있게 되었다.

### 3.3. 경매 물건별 낙찰률 예측 모형

이제 선형 회귀모형을 이용하여 개별 물건의 낙찰률에 영향을 미칠 수 있는 주요한 변수를 찾고 이를 기반으로 하는 물건 낙찰률 예측 모형의 구축 결과를 살펴보도록 하자. 분석에 사용된 데이터는 2002년 1월 1일부터 2006년 12월 31일까지 낙찰이 완료된 데이터 총 476,685건을 대상으로 하였다. 데이터 마트에는 2000년 1월 1일부터 자료가 존재하나 모형 구축을 위한 설명변수로 고려한 파생 변수들 가운데 2002년 1월 1일 이후에만 값을 가지는 변수가 주요한 변수를 사용됨에 따라 2002년 이전 데이터는 모형 구축과정에서 제외하기로 하였다. 모형 구축에 사용된 설명변수들은 데이터 분석 마트에 있는 각종 물건 관련 정보들 가운데에서 개념적으로 낙찰률과 관련성이 있는 변수들의 선정을 일차적으로 수행하였다. 이와 더불어 추가적인 분석 작업을 통하여 생성된 파생변수와 월별 낙찰률 예측 모형에서 사용된 변수들을 추가하여 기본적인 예측모형의 구축작업을 수행하였다. 표 5는 일차적으로 구축된 모형에서 이용되었던 각종 변수들 가운데 최종 모형에 선택된 변수들을 정리한 것이다. 표 6과 7은 회귀분석 결과로써 표 6의 연속형 설명변수에 대한 추정 결과이고 표 7은 범주형 설명변수에 대한 분산분석 결과이다.

모형 구축에 사용된 변수들을 보면 개별 물건 낙찰률에 가장 큰 영향을 주는 요인은 로그 표준화 감정가이다. 즉 경매개시 시점과 낙찰시점간에 감정가의 변동을 고려한 표준화 감정가가 실제 낙찰률에

표 6: 아파트 예측 모형 구축의 회귀분석 결과

Parameter	Estimate	Standard Error	t-Value	Pr >  t
Intercept	-127.7782	4.3541944	-29.35	< .0001
yy_do_12	0.7381562	0.0038189	193.29	< .0001
lmchaprice_sum	$1.0652 \times 10^{-10}$	$4.87433 \times 10^{-11}$	2.19	0.0289
tot_land	$-1.064 \times 10^{-9}$	$2.586964 \times 10^{-9}$	-0.41	0.6809
tot_Bd	0.0007045	0.0001088	6.48	< .0001
Log_st_gamprice_1	7.025171	0.669592	104.92	< .0001
nak_index	-0.113625	0.0032036	-35.47	< .0001

표 7: 아파트 예측 모형 구축의 분산분석 결과

Source	DF	Type III SS	Mean Square	F-Value	Pr >  F
imcha_yn	1	139.097	139.097	1.19	0.2757
bd_ext	1	17214.546	17,214.546	147.05	< .0001
addr_do_1	22	567,698.091	25,804.496	220.43	< .0001
Jaripdo	9	117,610.047	13,067.783	111.63	< .0001
su_num_1	4	657,175.876	164,293.969	1,403.44	< .0001

표 8: 등급에 따른 예측 낙찰률 비교

등급	자료수	최소값	평균값	최대값
1등급	29,967	58.90	91.08	129.62
2등급	59,934	43.86	78.94	111.84
3등급	29,966	20.08	60.23	96.40

가장 큰 영향을 미치는 요인으로 파악되었다. 다음으로 Method 2에 의하여 계산된 도별 12개월 평균 값이 주요한 영향을 미치는 요인으로 파악되었다. 즉 해당 물건이 속한 지역의 예측된 평균 낙찰률이 낙찰률 변화에 주요한 영향을 미치고 있다 할 수 있다. 이와 더불어 아파트의 경우 단위물건과 아파트 전체 건물 여부를 나타내는 bd\_ext, 지역구분(addr\_do), 자립도 구분(Jaripdo) 그리고 특수권리수도 낙찰률의 변화에 주요하게 영향을 미치는 요인으로 파악되었다. 전체 모형에 대한 설명력은 약 56.1%로써 이는 구축된 모형에 의해 반응변수인 물건별 낙찰률의 변화가 약 56.1% 정도 설명된다 할 수 있다.

선형 회귀모형에 의해 예측된 낙찰률의 응용 방법으로 개별 물건 예측 낙찰률을 이용하여 예측 낙찰률에 따라 예측 등급을 생성하는 방법을 고려하였다. 등급은 크게 3가지 구분하였는데 다음 표 8은 각 등급에 따른 예측 낙찰률간의 비교를 수행한 결과이다. 등급별 평균값에 근거하여 1등급을 분류된 경우 예상 유찰횟수를 1회 정도 고려할 수 있으며 2등급의 경우는 2회 정도, 3등급의 경우 3회 이상의 유찰횟수 보인다고 할 수 있다. 이에 근거하여 등급에 따른 실제 유찰횟수의 평균을 계산하면 등급별로 각각 1등급은 0.94번, 2등급은 1.46번 3등급은 2.42번으로 계산되었다. 실제 유찰횟수와 등급에 따른 예상 유찰횟수에는 약간의 차이를 보이고 있으나 그 차이는 크지 않다고 할 수 있다. 이 가운데 예측 낙찰률이 100% 이상은 물건들만을 별도로 분류하여 실제 유찰횟수를 살펴보면 그 평균이 0.57번이었다. 예측 낙찰률이 100% 이상이 되면 최초 경매에서 바로 낙찰이 될 가능성이 매우 높다고 할 수 있으며, 실제 낙찰율의 평균도 0.57번으로써 이에 근사하다고 할 수 있다.

#### 4. 결론

본 연구의 기본 목적은 경매자료를 이용하여 보다 정확하고 활용 가능한 정보의 산출에 있다. 이를 위하여 본 연구에서는 크게 2가지 부분에 대한 분석을 수행하였다. 첫 번째 분석은 월별 낙찰률의

예측이다. 이는 기존의 경매 물건 지역별 낙찰률 예측을 위한 방법에 기반하여 예측오차를 보다 효과적으로 줄일 수 있는 방법을 제안하고자 하는 방향으로부터 출발하여 지역확장과 기간확장을 함께 고려한 예측 방법을 제안하였고 그 결과를 다시 의사결정나무 분석(decision tree analysis) 기법을 이용하여 예측오차를 추가적으로 줄일 수 있는 방법에 제안하였다. 이 두 가지 방법을 통하여 기존의 방법보다 예측오차를 줄일 수 있는 낙찰률의 산출 방법을 구축하였다. 두 번째 분석은 개별 물건별 낙찰률 예측 모형의 구축하는 것이다. 현재 경매 정보 시장에서는 대부분 경매물건에 대한 예측 낙찰률을 제공하는 방법은 지역별 평균을 계산하고 이를 이용하여 예측 낙찰률로 고려하는 방법으로써 개별 물건단위 예측 낙찰률을 제공하는 방안은 전무한 실정이다. 본 연구에서 선형회귀모형을 이용하여 개별 물건에 대한 예측 낙찰률 산정을 위한 모형을 개발하였다. 개별 물건들이 가지고 있는 여러 데이터적 특성을 파악하고 이를 모형화 하여 낙찰률 예측에 이용하고자 하였다. 이와 더불어 부동산 경매에 영향을 줄 수 있는 각종 파생변수를 함께 생성하였다. 부동산 경매에 있어서 가장 중요한 요인 가운데 하나인 감정가를 경매 생성 시점과 낙찰 시점간의 차이를 보정할 수 있는 표준화된 감정가의 구축을 함께 수행하였다. 생성된 표준화 감정가를 이용하여 보다 현실화된 감정가의 반영이 추가적인 데이터의 수집 없이 가능하게 되었다. 향후 구축된 모형을 실제 일어나게 될 경매 결과에 반영하여 발생하는 문제점들을 수정하고 추가적인 정보 수집을 통하여 보다 현실화되고 부동산 관련 외부적인 상황(예, 재개발 사업, 정부시책)을 반영할 수 있는 모형 개발을 수행 하여야 할 것이다.

**참고 문헌**

강현철, 한상태, 최중후, 이성건, 김은석, 엄익현, 김미경 (2006). <고객관계관리(CRM)를 위한 데이터 마이닝 방법론>, 자유아카데미, 서울.

김원년, 조무상, 양현석 (2009). Hedonic 가격모형을 이용한 일반아파트와 공공아파트의 주거특성별 잠재임대료추정에 관한 연구, *Journal of the Korean Data Analysis Society*, **11**, 773-792.

박유성, 김기환 (2002). <시계열 자료분석 I>, 자유아카데미, 서울.

박진우 (2009). 통계기반 정책사례 연구: 주택가격지수 통계의 구축, 개선, 활용을 중심으로, <응용통계연구>, **33**, 635-651.

박현수 (2007). 거래빈도가 낮은 시장에서의 실거래 부동산 가격지수 작성에 관한 연구, <부동산학연구>, **13**, 187-200.

이정민 (2008). 부동산정책변화와 주택경매 낙찰가율 변동추세에 관한 연구, <한국부동산학보>, **33**, 101-124.

이정민, 이성근 (2007). 부동산경매시장의 낙찰가율 변동요인에 관한 연구, <한국부동산학보>, **30**, 76-94.

최보승, 강현철, 최호식, 한상태 (2009). 부동산 경매지수의 개발, *Journal of the Korean Data Analysis Society*, **11**, 3115-3125.

Lancaster, K. J. (1966). A new approach to consumer theory, *Journal of Political Economy*, **74**, 132-157.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition, *Journal of Political Economy*, **82**, 34-55.

# A Study for the Development of a Bid Price Rate Prediction Model

Boseung Choi<sup>a</sup>, Hyuncheol Kang<sup>b</sup>, Sang-Tae Han<sup>1,b</sup>

<sup>a</sup>Department of Computer Science and Statistics, Daegu University

<sup>b</sup>Department of Informational Statistics, Hoseo University

---

## Abstract

Property auctions have become a new method for real estate investment because the property auction market grows in tandem with the growth of the real estate market. This study focused on the statistical model for predicting bid price rates which is the main index for participants in the real estate auction market. For estimating the monthly bid price rate, we proposed a new method to make up for the mean of regions and terms as well as to reduce the prediction error using a decision tree analysis. We also proposed a linear regression model to predict a bid price rate for individual auction property. We applied the proposed model to apartment auction property and tried to predict the bid price rate as well as categorize individual auction property into an auction grade.

**Keywords:** Real estate auction, bid price rate prediction, linear regression model, decision tree analysis.

---

---

<sup>1</sup> Corresponding author: Professor, Department of Informational Statistics, Hoseo University, Asan-Si, Chungnam 336-795, Korea. E-mail: sthan@hoseo.edu