데이터 마이닝 7조 모델링 중간 보고서

개체이(가) 표시된 사진

자동 생성된 설명

12132591 정동호

12131820 이건도

12161890 하나영

<목차>

1. 주제
2. 변수 설명
3. Data set 구분
4. 모델링 – Decision Tree

4-1 초기모형 - Train set을 통한 의사결정학습법 결과

4-2 Pruning 적용

4-3 최종 모형 회귀 의사결정 학습법 결과

4-4 결론

4-5 추가) “평균전세가”를 제외한 의사결정 학습법

4-6 한계점

1. 모델링 – Regression
2. 모델링 결론
3. 주제

* **주제**

인천광역시 아파트 단지 데이터를 이용한 아파트 제곱미터당 평균 매매 가격 예측 및 영향을 주는 변수 분석

* **분석 단위**

최근 3년(16년 1월부터 18년 12월까지의 월 기준) 인천광역시 단지별 아파트 평균 매매 가격 기준

1. **변수 설명**
2. **Data Set 구분**

모델링을 학습하고 검증하고 테스트하기 위해서 저희는 Train, validation, test set을 4 : 3 : 3의 비율로 랜덤추출을 통해 구분하였습니다.

|  |  |
| --- | --- |
| Data Set 종류 | Data 수 |
| Train set | 1195 |
| Validation set | 895 |
| Test set | 895 |

1. **의사결정학습법**

* 의사결정 학습법 package 중 rpart package 이용
* 목적변수 연속형에 따른 회귀 의사결정 학습법

**4-1 초기모형 - Train set을 통한 의사결정학습법 결과**

텍스트이(가) 표시된 사진

자동 생성된 설명

<그림 >

초기 모형의 경우, Terminal node의 수는 8개이고, 노드 분리에 있어 가장 영향을 주는 변수가 “평균전세가”임을 알 수 있습니다. 그 다음으로는 “노후도”임을 알 수 있습니다. 그리고 가장 높은 값인 508.4의 경우, 평균 전세가가 206.5보다 크고 교육시설수가 21.5보다 클 때라는 것을 알 수 있습니다.

**4-2 Pruning 적용**

초기모형 그림1의 경우 rpart 함수의 적용 결과만을 판단한 것이기에 정확한 Pruning을 위해 Complexity parameter(오분류값)이 0이라는 조건을 두어 가장 낮은 오분율에서 적절한 terminal node 수가 몇 개인지 판단하기로 했습니다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

<그림 >

그림2를 보시면, terminal node의 수가 증가할수록 오분율 값이 계속 낮아지는 것을 확인할 수 있습니다.

적절한 Pruning을 하기전에 Validation set과 Train set을 통해 terminal node수에 따라 각 MSE의 변동을 알아보았습니다. 먼저 minsplit(노드 분할하기 위해 필요한 최소 데이터의 개수)변동은 결국 terminal node 수와 연관 있기 때문에 minsplit에 따라 validation set과 train set의 MSE을 측정했습니다.

지도, 텍스트이(가) 표시된 사진

자동 생성된 설명

<그림 >

Minsplit이 낮을수록(terminal node 수가 많을수록) validation과 train의 MSE 또한 낮아지는 것을 확인할 수 있습니다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

<그림 >

<그림3>과 <그림4>를 보면 minsplit이 500부터 125까지 validation MSE가 train MSE보다 계속 낮다는 것을 확인할 수 있습니다. 그러므로, <그림2>의 결과와 <그림4>의 분석에 따라 minsplit이 125일 때 지점을 선택하여 Pruning을 했습니다.

**4-3 최종 모형 회귀 의사결정 학습법 결과**

텍스트이(가) 표시된 사진

자동 생성된 설명

<그림 >

<그림 5>에 대해 train MSE, validation MSE, test MSE을 비교한 결과

|  |  |
| --- | --- |
| Train MSE | 1477.903 |
| Validation MSE | 1398.396 |
| Test MSE | 1303.861 |

세 개의 MSE 값 중에서 Test MSE가 가장 낮은 것을 확인할 수 있습니다.

종합적인 판단 결과 terminal node수가 15개를 가진 회귀 의사결정 모형으로 선택했습니다. 또한, Root node을 나누는 결정적인 변수가 “평균전세가”인 것을 알 수 있으며, 또한 이후 노드 분할에 있어 가장 많이 기준이 되는 변수는 평균전세가”임을 알 수 있습니다.

|  |  |
| --- | --- |
| 변수명 | 분기 노드 수 |
| 평균전세가 | 5 |
| 노후도 | 3 |
| 단지세대수 | 2 |
| 교육시설수, 스타벅스, 개별세대수, 쇼핑시설수 | 1 |

위의 평균전세가, 노후도, 단지세대수가 인천아파트 평균매매가 형성에 많은 영향을 주는 변수라는 것을 의미한다고 볼 수 있습니다.

**4-4 결론**

회귀 의사결정학습법을 통해서 인천아파트 평균매매가를 분석한 결과, 아파트와 관련된 내부변수, 즉 평균전세가, 노후도, 단지세대수가 다른 외부변수에 비해 인천 아파트 평균매매가에 많은 영향을 준다는 것을 확인할 수 있습니다. 또한 초기모형(그림1)에서 가장 높은 값을 형성하는데 영향을 주는 변수가 “평균전세가”와 “교육시설수” 변수였지만, 최종모형(그림5)에는 모두 “평균전세가”임을 확인할 수 있습니다.

**4-5 추가) “평균전세가”를 제외한 의사결정 학습법**

* + - **초기모형**

스크린샷이(가) 표시된 사진

자동 생성된 설명

<그림 >

초기모형 그림6에 따르면, root node 분기 기준이 “역세권점수”임을 파악할 수 있습니다.

* + - **Pruning 적용**

스크린샷이(가) 표시된 사진

자동 생성된 설명

<그림 >

그림7을 보면 terminal node 수 증가에 따라 오분율이 계속 낮아지는 것을 확인할 수 있습니다.

지도, 텍스트이(가) 표시된 사진

자동 생성된 설명

<그림 >

지도, 스크린샷이(가) 표시된 사진

자동 생성된 설명

<그림 >

그림8과 그림9을 종합하여 판단하면, minsplit이 146에서 선택하는 것이 최적이라는 것을 고려할 수 있습니다.(terminal node 수는 14개)

* + - **최종모형 결과**

스크린샷이(가) 표시된 사진

자동 생성된 설명

<그림 >

최종모형 그림10에 따라 train MSE, validation MSE, test MSE을 비교해보면,

|  |  |
| --- | --- |
| Train MSE | 2058.659 |
| Validation MSE | 2010.161 |
| Test MSE | 2244.963 |

오히려 Train MSE, Validation MSE보다 Test MSE가 높다는 결과를 볼 수 있습니다.

변수에 따른 분기 노드의 수를 파악한다면

|  |  |
| --- | --- |
| 변수명 | 분기 노드 수 |
| 노후도 | 4 |
| 단지세대수, 교육시설수 | 2 |
| 스타벅스, 평균세대수, 개별세대수, 평균혼인건수, 토지면적 | 1 |

가장 영향을 주는 변수는 “노후도”로 내부 변수에 해당합니다.

* + - **결론**

평균전세가 미포함한 초기모형과 최종모형을 비교하면, 초기모형과는 다르게 최종모형에서의 root node가 바뀐 것을 확인할 수 있습니다.

평균전세가를 포함한 결과와 미포함한 결과를 분석하면, 전반적인 data set의 MSE의 변화를 통해 평균전세가가 인천아파트 평균매매가에 많은 영향을 준다는 것을 유추할 수 있습니다. 노후도 변수 또한 인천 아파트 평균매매가 형성에 상당하게 영향을 주는 것을 유추할 수 있습니다.

**4-5 한계점**

Pruning 단계에서 terminal node수가 많을수록 오분율이 계속 낮아지는 추세를 갖기에 이 부분에 있어 인터넷 검색을 통해 정보를 얻으려고 했으나 거의 전무했기에 적정한 수준의 terminal node를 선정하는데 어려움이 있었습니다. 추후 학습이 필요한 부분이라 생각이 됩니다.