# Linear Regression 기반의 국내 주택 가격지수 예측 연구

# A Study on the Prediction of Domestic Housing Price Index Based on Linear Regression

최승은

Seung-Eun Choi

g787@naver.com

**Abstract**

In Korean society, real estate is considered an important economic issue considering its share of household assets and social implications. Housing still means more than just a means of housing in Korea. Real estate accounts for a higher proportion of household assets than other developed countries. Of Korea's net assets of households and non-profit organizations, non-financial assets account for 75.4 percent, higher than those of major advanced economies such as Japan (43.3 percent) and the United States (34.8 percent).

The direction of the housing market is of interest to both households and the national economy, but predicting future trends is not simple. This is because a number of factors, including basic physical factors such as demand and supply of housing, such as the government's housing policy, market interest rates and changes in the sentiment of economic players, determine prices in the housing market. The reason why it is difficult to predict even the short-term housing market regardless of the mid- to long-term is that there are not enough leading indicators to predict the housing market in the coming months. Although housing supply and demand indicators, mortgage loans, and interest rates are used on a limited basis, they cannot play a role as leading indicators due to the small number of prior time differences and different timing of the announcement. This requires the development and application of various analytical methods in the real estate sector as well.

Based on the Korea Appraisal Board's housing price index data, the study will be conducted to predict the housing price index for next month. The data were analyzed through Linear Regression through Azure Machine Learning and the housing price index for next month was predicted.

**Keywords** : Azure Machine Learning, Housing Price Index, Linear Regression

## I. 서 론

한국 사회에서 부동산은 가계자산 중 차지하는 비중이나 사회적 함의를 고려해 봤을 때 중요한 경제 이슈로 다뤄진다. 우리나라에서 주택은 아직까지도 단순한 주거 수단 그 이상의 의미가 있다. 대한민국에서는 부동산의 가계자산 가운데 차지하는 비중이 다른 선진국에 비해 높은 편이다. 한국의 가계 및 비영리단체 순 자산 중 비금융자산 비중은 75.4%로, 일본(43.3%), 미국(34.8%) 등 주요 선진국보다 높다.

주택경기의 향방은 가계와 국가 경제 측면 모두의 관심사이지만, 미래 흐름을 예측하는 것은 간단하지 않다. 주택의 수요와 공급 등 기본적인 물리적 요인을 포함하여 정부의 주택정책, 시장금리와 경기 주체들의 심리변화 등 여러 요인이 주택시장에서 가격을 결정하기 때문이다. 중장기를 떠나 단기적인 주택경기의 예측마저도 어려운 이유는 수개월 후의 주택경기를 예측할 수 있는 선행지표가 마땅치 않기 때문이다. 주택수급 동향지수, 주택담보대출, 금리 등이 제한적으로 이용되지만, 선행시차가 적고 발표가 되는 시점도 각기 달라서 선행지표로써 역할을 제대로 하지 못하는 상황이다. 이에 따라 부동산 분야에서도 다양한 분석기법의 개발 및 적용이 요구되고 있다[1].

본 연구는 한국감정원의 주택 가격지수 데이터를 토대로 다음달의 주택 가격지수를 예측하는 연구를 진행한다. Azure Machine Learning을 통한 Linear Regression으로 데이터를 분석하고 다음달의 주택 가격지수를 예측하였다.

## II. 관련연구

### 2.1 마이크로소프트 애저 머신러닝(Microsoft Azure Machine Learning)[2]

마이크로소프트(MS)는 그동안 운영체제, 오피스 등을 기반으로 시장을 유지해왔지만, 최근 모바일 중심의 '클라우드 퍼스트'로 전략과 비전을 수정하였다. 자사의 제품을 클라우드 서비스인 애저(Azure)를 통해 지원하고 여기에 인공지능을 더하여 클라우드 플랫폼 중심으로 서비스로의 변화를 꾀하고 있다.

마이크로소프트 애저 머신러닝은 2014년에 런칭하여 다양한 수준의 과학자들을 위해 전 범위의 단순화된 경험을 제공하는 클라우드 기반의 예측 분석 서비스이다. 애저 머신러닝 스튜디오(Azure Machine Learning Studio)의 가장 큰 장점은 편의성이라 할 수 있다. 웹 기반의 드래그 앤드 드롭 인터페이스를 통해 데이터를 수집/처리하여 머신러닝 모델을 훈련하고 결과를 REST API로 공유한다. 또한, 머신러닝 중간 결과를 쉽게 확인, 재시작, 실행 기능을 제공하여 데이터 과학자가 최소 비용으로 가치 있는 결과를 얻는데 주력할 수 있다.

마이크로소프트의 인공지능 플랫폼은 실용적인 측면에 초점을 두어 실제 비즈니스 수요에 맞춘 솔루션을 내놓고 있다. R과 파이썬(Python)을 지원하고, 표준 머신러닝 알고리즘 및 도구(Text mining, Regression, Classification, Clustering, Anomaly detection) 등을 제공한다.

### 2.2 선형 회귀[3]

통계학에서, 선형 회귀(Linear Regression)는 종속 변수 y와 한 개 이상의 독립 변수 (또는 설명 변수) X와의 선형 상관 관계를 모델링하는 회귀분석 기법이다. 한 개의 설명 변수에 기반한 경우에는 단순 선형 회귀, 둘 이상의 설명 변수에 기반한 경우에는 다중 선형 회귀라고 한다.

선형 회귀는 선형 예측 함수를 사용해 회귀식을 모델링하며, 알려지지 않은 파라미터는 데이터로부터 추정한다. 이렇게 만들어진 회귀식을 선형 모델이라고 한다. 선형 회귀는 깊이 있게 연구되고 널리 사용된 첫 번째 회귀분석 기법이다. 이는 알려지지 않은 파라미터에 대해 선형 관계를 갖는 모델을 세우는 것이, 비선형 관계를 갖는 모델을 세우는 것보다 용이하기 때문이다.

선형 회귀는 여러 사용 사례가 있지만, 대개 아래와 같은 두 가지 분류 중 하나로 요약할 수 있다.

① 값을 예측하는 것이 목적일 경우, 선형 회귀를 사용해 데이터에 적합한 예측 모형을 개발한다. 개발한 선형 회귀식을 사용해 y가 없는 x값에 대해 y를 예측하기 위해 사용할 수 있다.

② 종속 변수 y와 이것과 연관된 독립 변수 X1, ..., Xp가 존재하는 경우에, 선형 회귀 분석을 사용해 Xj와 y의 관계를 정량화할 수 있다. Xj는 y와 전혀 관계가 없을 수도 있고, 추가적인 정보를 제공하는 변수일 수도 있다.

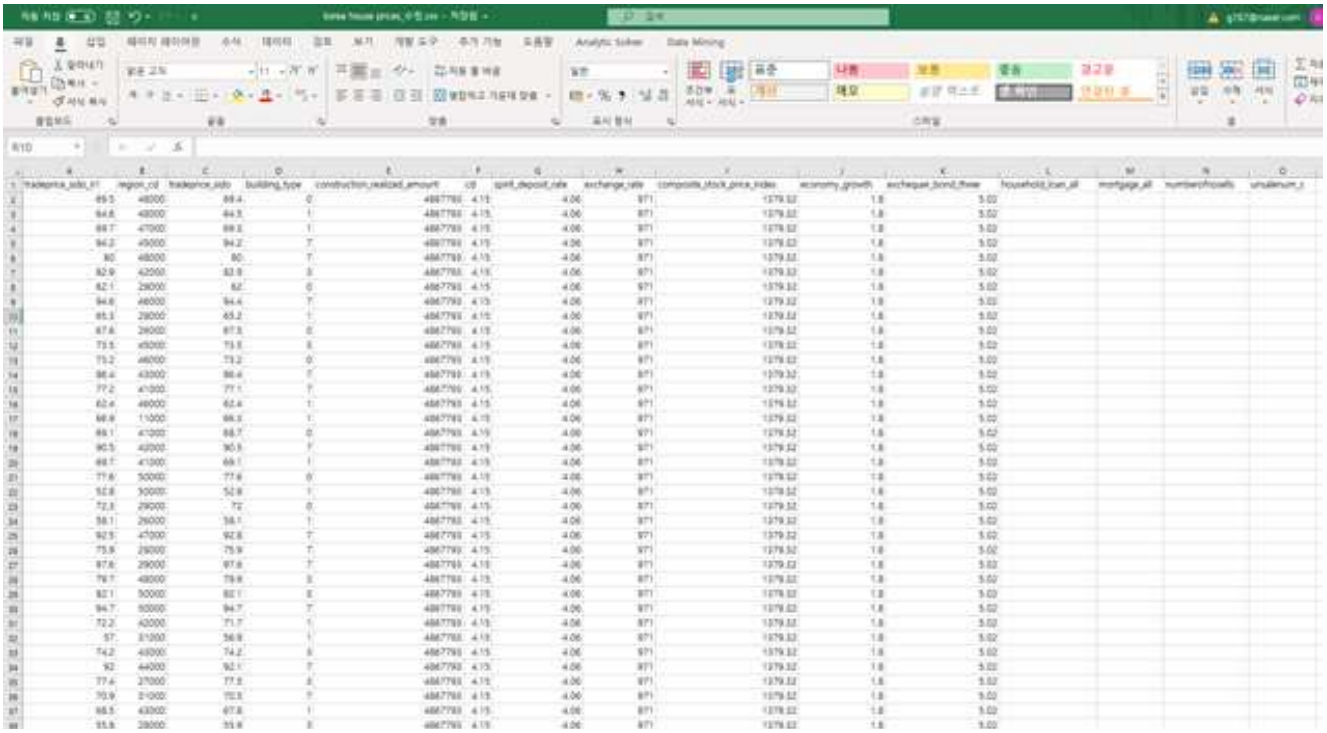
일반적으로 최소제곱법(least square method)을 사용해 선형 회귀 모델을 세운다. 최소제곱법 외에 다른 기법으로도 선형 회귀 모델을 세울 수 있다. 손실 함수(loss fuction)를 최소화 하는 방식으로 선형 회귀 모델을 세울 수도 있다. 최소제곱법은 선형 회귀 모델 뿐 아니라, 비선형 회귀 모델에도 적용할 수 있다. 최소제곱법과 선형 회귀는 가깝게 연관되어 있지만, 그렇다고 해서 동의어는 아니다.

### 2.3 주택 가격지수[4]

주택가격지수란 전국의 주택 매매 및 전세가격을 조사하여 일정시점을 기준 시점으로 한 라스파이레스산식을 적용하여 지역별, 주택유형별, 주택재고 구성비를 가중치 값으로 부여하여 산출하는 지표를 말한다. 전국의 아파트, 단독, 연립주택 중 층화 2단 집락 확률비례추출법으로 표본을 설계하고 매주 또는 매월 조사기준일의 표본주택이 거래가 된 경우에는 실거래가격을 거래가 되지 않은 경우에는 거래사례비교법으로 조사한 가격을 해당지역 부동산중개업소에서 직접 온라인상 조사표에 입력하는 방식으로 조사한다. 이 주택통계는 개별 아파트의 영향력을 동일하게 반영하기 위하여 변동률의 평균값을 채용하기 때문에 전체 주택시장의 동향을 파악하는데 유효하다. 한편 타조사기관의 주택통계는 가격평균의 변동률 또는 시가총액의 변동률을 계산하는 경우가 있고, 실제 신고한 거래가격의 변동률을 측정하는 통계가 있는데, 이들 통계와 성격과 목적이 다른 통계다.

# III. 데이터 셋

## 3.1 데이터 셋 설명



<그림 1> Korea House Prices.csv

<표 1> Korea House Prices.csv 변수 내용설명

| 변수 이름 | 변수 내용 |
|---|---|
| region_cd | 지역코드(시도) |
| tradeprice_sido | 주택 가격지수(시도) |
| building_type | 부동산타입 |
| construction_realized_amount | 건설기성액(백만원) |
| cd | cd(91일물) |
| spirit_deposit_rate | 정기예금금리 |
| exchange_rate | 환율 |
| composite_stock_price_index | 종합주가지수 |
| economy_growth | 경제성장률 |
| exchequer_bond_three | 국고채 3년 |
| household_loan_all | 가계대출액(전국) |
| mortgage_all | 주택대출액(전국) |
| numberofnosells | 미분양 가구수(시도) |
| unsalenum_c | 공사완료후 미분양(민간,시도) |
| tradeprice_sido_n1 | 한달 후의 주택 가격지수(시도) |

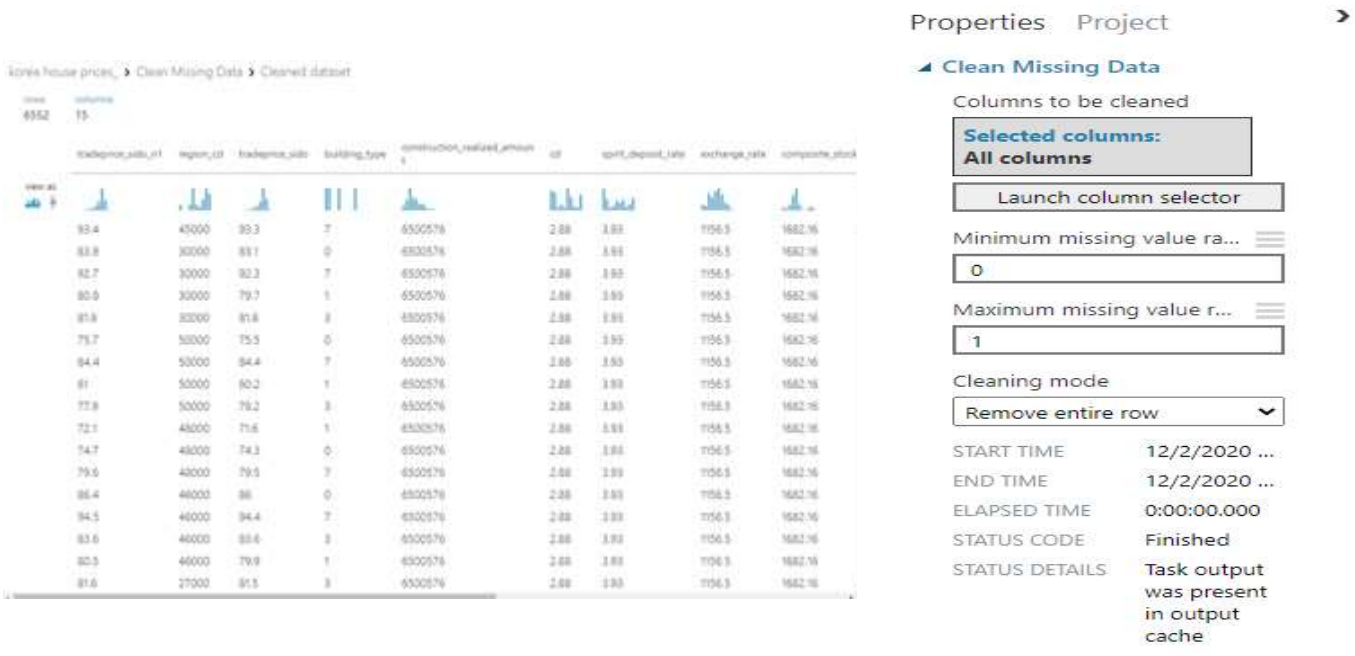
# Ⅳ. 실험 모델

## 4.1 실험 모델


<그림 2> Domestic Housing Price Index Prediction Experiments

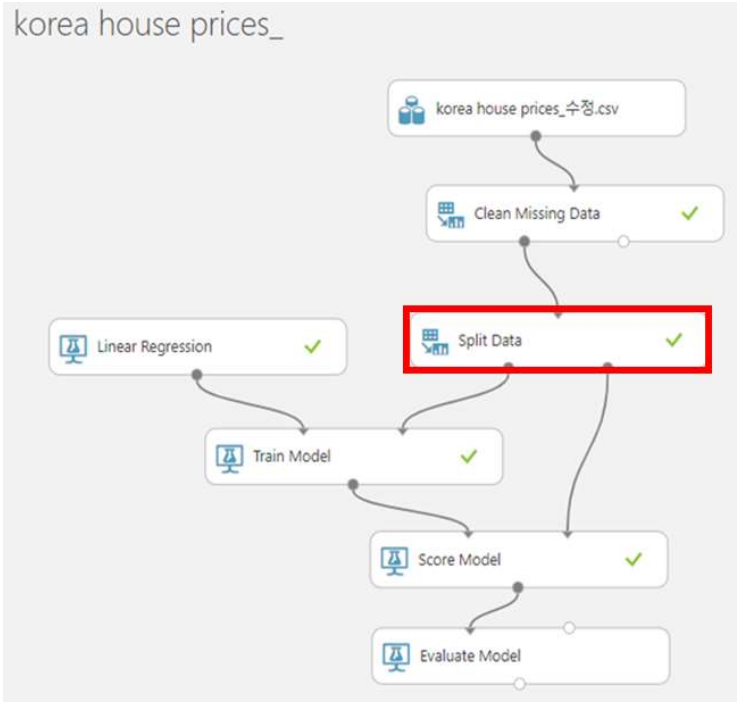
그림 2 는 korea house prices.csv 데이터를 사용해 Microsoft Azure Machine Learning 프로그램을 통해 다음달 주택 가격 지수 예측 모델을 구성한 모습이다.

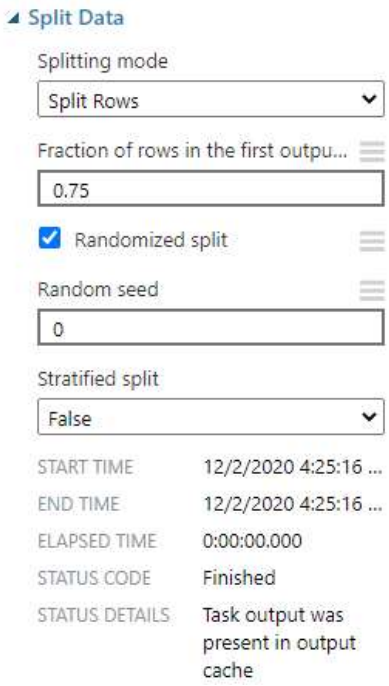
## 4.2 실험 모델 과정




<그림 3> Clean Missing Data

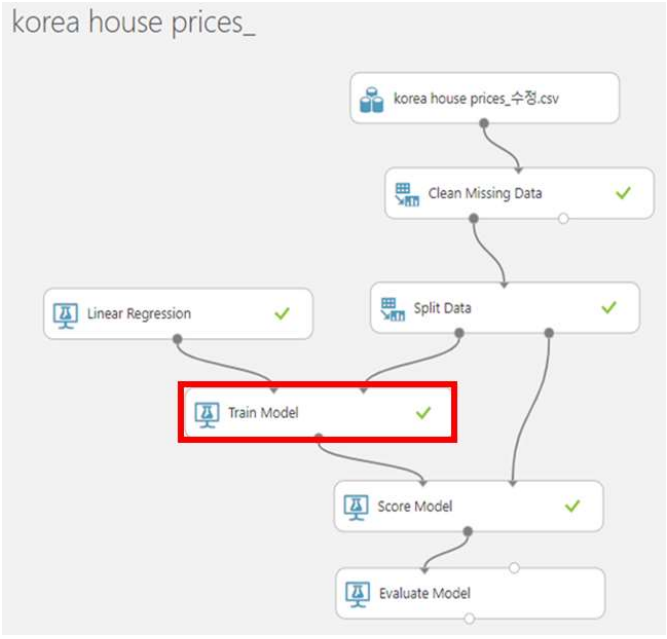
위 그림 3 은 korea house prices.csv 데이터에 대하여 Clean Missing Data를 사용해 전처리를 해주었다. Minimum missing value ratio 를 0 으로 설정, Maximum missing value ratio 를 1 로 설정해주어 결측값이 하나만 있어도 결측값을 정리하게 설정하였고, 선택된 모든 열에 대해 결측된 열이 반환되고 지정된 작업을 수행하게 하였다.




<그림 4> Split Data

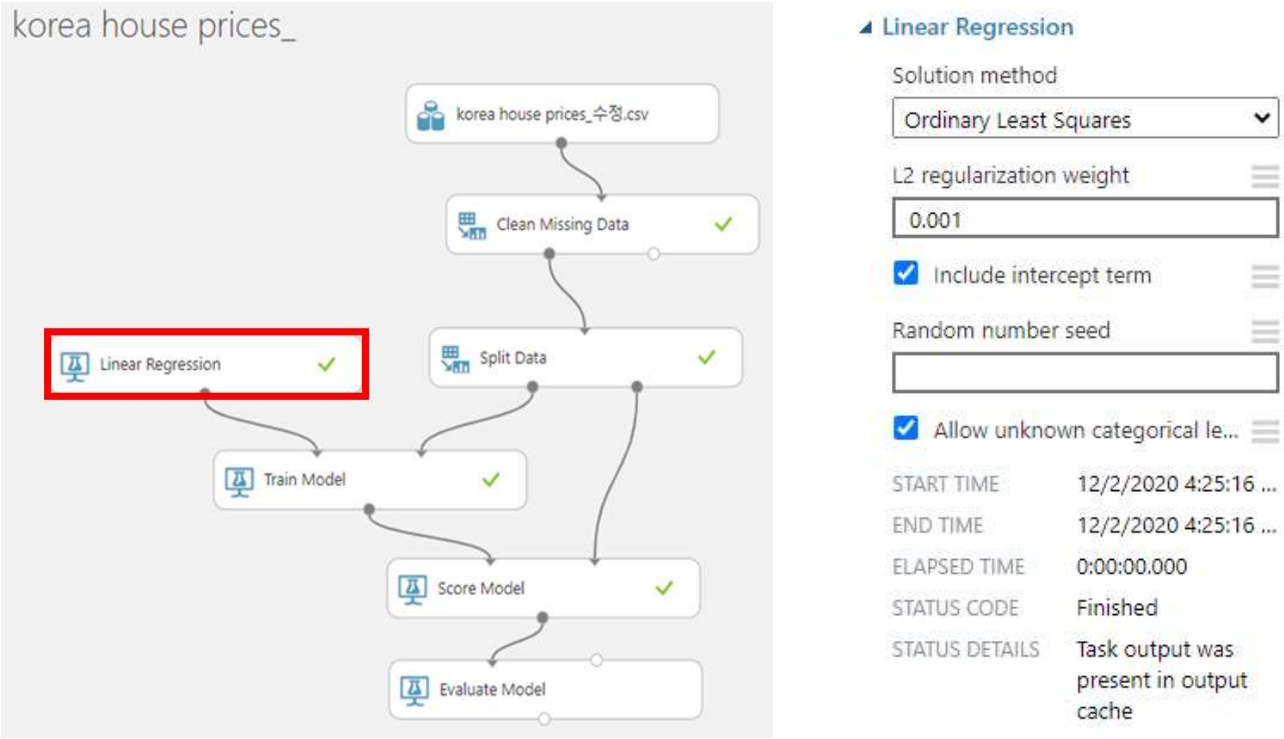
Split Data 를 통해 학습데이터와 훈련데이터를 75%, 25% 로 설정하여 데이터를 분리했다.





<그림 5> Train Model Select Column

선형회귀 알고리즘에 대한 Train Model 에서 tradeprice_sido_n1(다음달 주택가격지수)을 라벨로 하여 선형회귀 알고리즘을 실행하였다.




<그림 6> Linear Regression

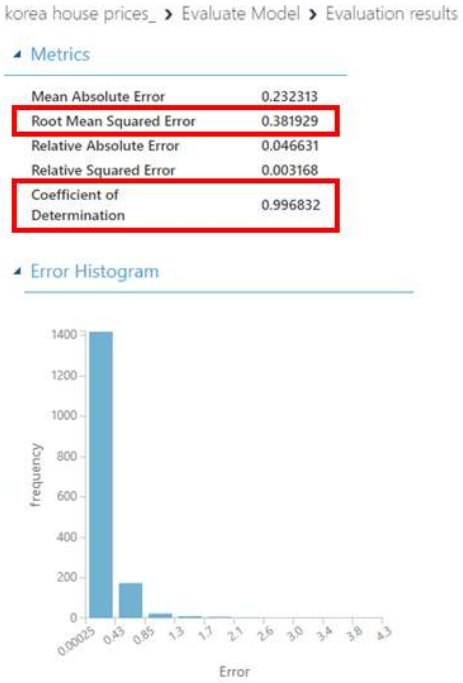
그림 6 은 선형회귀 알고리즘에서 최소제곱법을 사용하였고, L2 정규화 가중치를 0.001로 설정했다.


| Feature | Weight |
|---|---|
| tradeprice_sido | 0.978281 |
| Bias | -0.181076 |
| spirit_deposit_rate | 0.0942917 |
| economy_growth | 0.0741057 |
| exchequer_bond_three | -0.03904 |
| cd | 0.0160269 |
| building_type | -0.00850676 |
| composite_stock_price_index | 0.000618599 |
| exchange_rate | 0.000532418 |
| numberofnosells | -0.0000240036 |
| mortgage_all | 0.0000125797 |
| household_loan_all | -0.00000730477 |
| unsalenum_c | -0.00000214097 |
| region_cd | 0.00000200431 |
| construction_realized_amount | 3.37149e-9 |

<그림 7> Linear Regression Test data result

그림 6 은 선형 회귀 알고리즘에 대한 test data result이다. Tradeprice_sido(해당 월의 주택 가격지수)가 0.98로 매우 강한 양의 상관관계를 보여줬고, spirit_deposit_rate(정기예금금리)와 economy_growth(경제성장률)가 0.09, 0.07로 약한 양의 상관관계를 보여줬다.

### 4.3 실험 모델 평가


<그림 8> Linear Regression

그림 8은 Linear Regression에 대한 평가 모델이다. Linear Regression은 Mean Absolute Error = 0.232313, Root Mean Squared Error = 0.381929, Relative Absolute Error = 0.046631, Relative Squared Error = 0.003168, Coefficient of Determination = 0.996832를 기록했다.

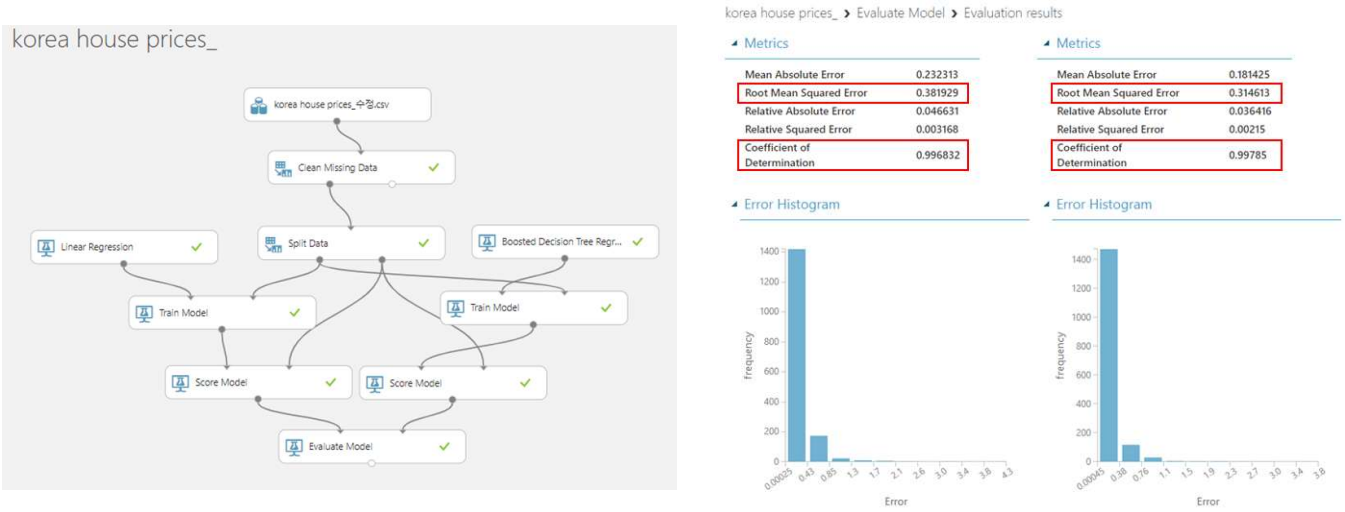
평균 제곱근 편차(RMSE)는 정밀도를 표현하는데 적합한 척도이며 낮을수록 좋다. RMSE가 0.38 정도로 정밀도는 높다고 할 수 있다.

결정계수(coefficient of Determination)은 데이터 분석 모델의 신뢰도를 나타내며, 결정계수가 0.9968이 나온 것을 신뢰도 99.68%로 해석된다. 이 데이터 분석의 신뢰도는 매우 높다고 할 수 있다.

## Ⅴ. 결론

본 연구에서는 '이번달 주택 가격지수'와 다른 경제 지표들을 알고 있을 때, 다음달 주택 가격지수 예측을 도출해냈다. Tradeprice_sido(해당 월의 주택 가격지수)가 0.98로 매우 강한 양의 상관관계를 보여줬고, spirit_deposit_rate(정기예금금리)와 economy_growth(경제성장률)가 0.09, 0.07로 약한 양의 상관관계를 보여줬다. 이외 변수들을 미비한 상관관계로 제외했다.

이 연구에 따른 개선방안은 한달 후의 주택 가격지수가 아닌 6개월, 1년 또는 그 이상 이후의 주택 가격지수를 예측하는 것이다. 단순히 한달 후의 주택 가격지수는 활용하기에 부족한 지표이기 때문에 더 긴 기간 이후의 주택 가격지수를 예측할 수 있도록 발전시켜야 할 것이다. 그리고 이러한 예측 내용에 대한 웹 서비스를 많은 사람들이 이용할 수 있게 제공한다면 공익을 이끌어낼 수 있을 것이다.




<그림 9> Linear Regression과 Boosted Decision Tree Regression 모델 평가 비교

또한 Linear Regression은 피처는 많지만 너무 단순하다는 단점이 있다. 따라서 비선형성 모델인 Boosted Decision Tree Regression(강화된 의사결정 트리 회귀)를 통해 두 개의 모델을 비교했을 때 강화된 의사결정 트리 회귀의 RMSE가 더 낮고, 결정 계수가 더 높았기 때문에 더 정확한 분석 모델이라고 할 수 있다. 이를 통해 Linear Regression 이외의 다른 분석 모델들을 탐색하여 최적의 분석 모델을 찾아내는 것 또한 앞으로의 과제가 될 것이다.

## References


[1] 박재수, (2020), "주택시장 예측을 위한 부동산 감성지수 개발 연구 - 뉴스와 방송 빅데이터에 대한 AI 기술 적용", 강원대학교 대학원 박사학위 논문, 1-2

[2] 최정란, 송영미, 김철홍, 김선자. (2017). "인공지능을 위한 클라우드 컴퓨팅 산업 동향", 전자통신동향분석 제32권 제5호, 110.

[3] 위키피디아, "선형 회귀", https://ko.wikipedia.org/wiki/선형_회귀, (2020.12.1)

[4] 네이버 지식백과, "주택가격지수", https://terms.naver.com/entry.nhn?cid=42094&docId=586835&categoryId=42094, (2020.12.1)