선형 회귀를 이용한 쌀 가격 예측 모델의 유의미한 변수 추출

서진경[O], 최다정[*], 고광호[**], 백주련(교신저자)[*]
[O]평택대학교 데이터정보학과,
[*]평택대학교 데이터정보학과,
[**]평택대학교 스마트자동차학과
e-mail: sjk2700@naver.com[O], (dajung2020[*], kwangho[**], jrpaik[*])@ptu.ac.kr

Analyzing Significant Variables from a Linear Regression-Based Prediction Model for Rice Prices

Jin-kyeong Seo[O], Da-jeong Choi[*], Kwang-Ho Ko[**], Juryon Paik(Corresponding Author)[*]
[O]Dept. of Digital Information & Statistics, Pyeongtaek University,
[*]Dept. of Digital Information & Statistics, Pyeongtaek University,
[**]Dept. of Smart Mobility, Pyeongtaek University

● 요 약 ●

쌀을 주식으로 하는 우리나라에서, 쌀의 가격에 영향을 미치는 변수를 찾는 것은 유의미한 연구이다. 본 논문에서는 쌀 가격을 예측하는 모델에 포함되는 여러 변수 가운데 상대적인 중요도가 낮은 변수를 제거하고 유의미한 변수만을 남기고자 한다. 이를 위해 기상, 수확량, 소비자물가의 10년 치 정보를 수집하고 정제한 결과 총 2460일, 7개 지역에서 추출된 17,219개의 데이터를 이용하였다. 모델 평가 결과, 모든 변수를 포함한 모델의 RMSE는 166.0759, 단계적으로 계수가 작은 9개의 변수를 제거한 최종적인 모델의 RMSE는 168.5576으로 유의미한 차이를 보이지 않았다. 최종적으로 남은 변수는 총 10개로 평균 기온, 평균 풍속, 합계 일사, 평균 지면 온도, 0.5M 평균 습도, 4.0M 평균 습도, 10CM 일 토양 수분, 30CM 일 토양 수분, 50CM 일 토양 수분, 전년도 생산량이 포함된다.

키워드: 쌀 가격(Rice price), 유의미한 변수(Significant variables), 선형회귀(Linear regression)

I. Introduction

최근 질병 및 전쟁 등 여러 요인으로 인한 원자재 가격 인상으로 인해 소비자물가가 지속적인 상승세를 나타내고 있다. 이에 한국은행은 올해 2022년의 소비자물가 상승률이 4.7%를 넘어설 수 있다고 발표했는데, 이는 2008년 글로벌 금융위기를 넘어서는 수치이다. 특히 물가 상승률은 생활과 밀접하게 관련된 음식, 즉 농산물과 식품의 가격에서 크게 체감될 수 있다. 따라서 본 논문에서 쌀을 주식으로 하는 우리나라의 문화에 맞게 쌀의 가격을 예측하는 모델을 다뤄볼 것이다. 쌀의 가격에 영향을 주는 여러 변수 가운데 미미한 영향을 미치는 변수는 무엇이며, 해당 변수를 제거했을 때 결과에 큰 영향이 없는지를 살펴본다. 최종적으로 모델에 유의미한 적은 수의 변수를 남기고, 해당 변수들이 무엇인지 살펴본다.

II. The Proposed Scheme

1. 데이터

1.1 데이터 수집 및 정제

본 논문에서는 농산물 유통 정보의 쌀 가격 데이터, 기상청의 기상 데이터 및 수확량 데이터, Kosis의 월별 소비자물가 등락률 데이터를 이용하였다. 이때, 가격 데이터는 도매가격, 중품, 1kg을 기준으로 사용하였다. 기상 데이터는 농업 기상 관측 데이터를 이용했으며, 이에 부족한 자료를 보완하기 위해 종관 기상 관측 데이터를 더하여 사용하였다. 한국의 농업 기상 관측 데이터에는 6개의 도와 7개의 지역이 존재한다. 따라서 종관 기상 관측 데이터는 해당 지역을 기준으로 추출하여 통합하였다. 모든 데이터는 2012년부터 2021년까지의 10년 치의 일별 데이터가 담겨있다. 다만, 수확량에 한해서는 전년도의 자료가 필요하기에 2011년부터 2020년까지의 자료를 수집

하였다. 모든 데이터는 날짜를 기준으로 병합하였으며 가격 데이터와 그 외에 데이터들을 모은 변수 데이터, 총 두 개의 데이터가 완성되었다. 아래의 Table 1은 완성된 변수 데이터에 대한 설명을 담고 있다.

Table 1. 변수 설명

변수명	설명
avg_temp	평균 기온
avg_wind_spped	평균 풍속
total_sunshine	합계 일조 시간
total_solar_radiation	합계 일사
avg_ground_temp	평균 지면 온도
0.5M_avg_humidity	0.5M 평균 습도
1.5M_avg_humidity	1.5M 평균 습도
4.0M_avg_humidity	4.0M 평균 습도
10CM_daily_soil_moisture	10CM 일 토양 수분
20CM_daily_soil_moisture	20CM 일 토양 수분
30CM_daily_soil_moisture	30CM 일 토양 수분
50CM_daily_soil_moisture	50CM 일 토양 수분
day_rainfall	강수량
pm_total_index	총 물가 지수
pm_living_price_index	생활 물가 지수
pm_fresh_food_index	신선 식품 지수
pm_agriculture_Oil_exclu_index	농산물 및 석유류 제외 지수
pm_food_energy_exclu_index	식료품 및 에너지 제외 지수
yield	전년도 쌀 수확량

1.2 데이터 탐색

농업 기상 관측 데이터와 종관 기상 관측 데이터 사이의 결측일을 제거한 전체 날짜 수는 2460일이다. 두 개의 데이터는 공통적인 날짜인 2019년 10월 9일, 2020년 3월 2일에 대해 결측 일이 존재했다. 다만, 보성 지역에 대해서는 2021년 1월 5일에 대해 추가적인 결측일이 존재한다. 따라서 2460일에 대해 각각 7개의 지역이 존재, 보성 지역에 대한 하루의 결측 일을 뺀 17219일이 전체 데이터의 행 수가 된다. 몇몇 특성에 대한 그래프는 다음과 같다.

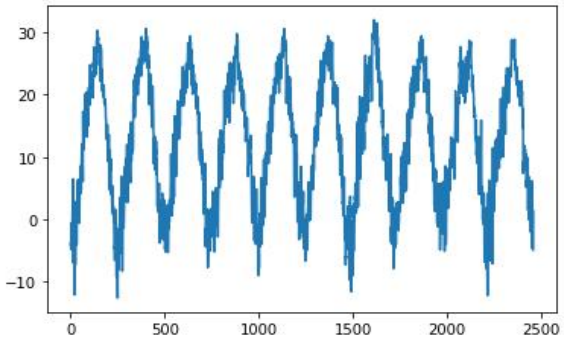

Fig. 1. 평균 기온 그래프

사계절을 가지는 한국의 계절적 특징에 맞게 10년치의 추세가 비슷한 형태로 나타남을 볼 수 있다.

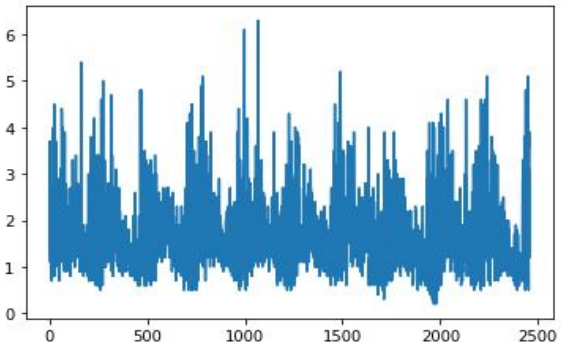

Fig. 2. 평균 풍속 그래프

평균 풍속은 최소 0.1m/s에서 최대 12.6m/s의 값을 가지며 대부분의 값은 1.0m/s에서 5.0m/s 사이에 위치하고 있음을 관찰할 수 있다.

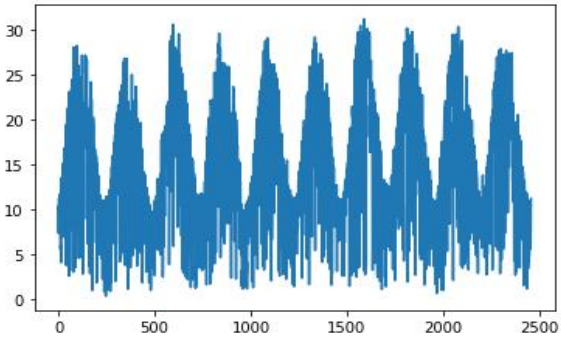

Fig. 3. 합계 일사 그래프

일사는 태양의 복사에너지가 지표면에 닿은 세기를 담은 변수이다. 시기별로 확인했을 때, 여름이 겨울보다 더 높은 일사량을 가짐을 확인할 수 있다.

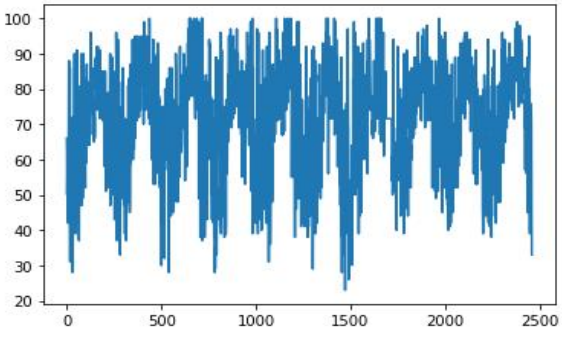

Fig. 4. 1.5M 평균 습도 그래프

1.5M 평균 습도에 대해서도 겨울보다 여름에 더 높은 위치의 값을 가짐을 볼 수 있다. 다만 Fig 3보다는 조금 덜 뚜렷한 형태를 가진다.

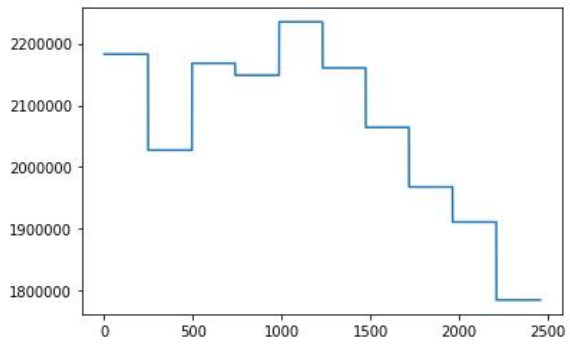

Fig. 5. 전년도 생산량 그래프

쌀에 대한 전년도 생산량 변수 그래프이다. 각각의 해에 대해 같은 전년도 생산량을 가지므로, 다음과 같은 그래프의 추세를 확인할 수 있다.

1.3 데이터 처리

존재하는 변수들은 각기 다른 범위와 단위를 가진다. 따라서 이를 정규화할 필요가 있다. 이미 퍼센트 값을 가지는 물가 관련 변수들을 제외한 총 14개의 변수에 대해 0에서 1의 값을 가지도록 정규화를 진행했다.

최종적으로 모델을 학습하고 평가하기 위해 전체 데이터를 훈련 데이터(80%)와 시험 데이터(20%)로 분리하였으며, 그 결과 훈련 데이터 13,775개와 시험 데이터 3,444개로 나누어졌다.

2. 모델 수립, 예측 및 평가

2.1 모델링

종속 변수인 가격 데이터에 대해 변수 데이터들이 각각 어느 정도의 영향을 주는지 알기 위해 모델의 계수를 관찰한다. 계수의 절댓값의 크기가 클수록, 해당 변수가 가격에 미치는 영향이 크다는 것을 의미한다. 따라서 절댓값의 크기가 작은 순으로 변수를 제거해가며 전체 정확도에 미치는 영향을 관찰한다. 전체 정확도에 큰 영향을 미치지 않는 변수들은 제거하고, 소수의 유의미한 변수만을 남겨 최종 모델을 완성한다.

2.2 모델 평가

학습 데이터를 통해 모델을 학습 후 테스트 데이터에 대해 예측을 수행하였다. 평가 지표는 RMSE(Root Mean Squared Error)를 사용하였으며 이는 평균 제곱근 오차이다. RMSE가 낮을수록 예측력이 좋은 모델이라고 평가한다. 해당 연구에서는 중요도가 낮은 변수들을 제거하고 소수의 유의미한 변수만을 남기고자 한다. 따라서 상대적으로 중요도가 낮은 변수를 제거했을 때 RMSE의 상승 폭을 관찰하여 유의미한 차이가 나지 않는다면 해당 변수를 제거한다. 단계적으로 변수를 제거했을 때의 RMSE는 다음과 같이 나타났다. 변수의 순서는 Table1과 같다.

-모든 변수를 이용한 모델

```
Slope: [ 159.38815107  136.18070629  -69.88004176  115.49515482
-162.44962253
  160.4031078    58.82891803 -222.53440584 -106.224665
-46.86750431
  192.83122717 -215.43818916   61.77044149  -39.77663256
-49.84865636
    3.13537408  -11.558338     58.67976147 -992.03340917]
Intercept: 2746.4288948889525

In [13]: RMSE(y_train,y_train_pred)
Out[13]: 165.1720000681522

In [14]: RMSE(y_test,y_test_pred)
Out[14]: 166.07589782088303
```

-pm_fresh_food_index, pm_agriculture_Oil_exclu_index를 제거한 모델

```
Slope: [ 152.82032296  140.17728101  -68.35358594  111.79354891
-159.5065297
  154.92760205   64.61681983 -218.19439481 -104.32229433
-48.54592488
  198.55249884 -221.83363169   62.68795214   26.74358297
-65.12607564
   21.11867277 -985.4129787 ]
Intercept: 2740.6482899763296

In [35]: RMSE(y_train,y_train_pred)
Out[35]: 165.38689416808018

In [36]: RMSE(y_test,y_test_pred)
Out[36]: 166.45763462164373
```

-pm_total_index, 20CM_daily_soil_moisture 제거한 모델

```
Slope: [ 147.85893872  145.18217829
-68.84918317  113.20110604 -154.64887323
  157.35984912   67.96126654 -221.31447724
-127.93396133  147.29198171
 -218.7413821    62.03384418  -49.92330209
31.42880769 -985.90985739]
Intercept: 2737.560153519233

In [38]: RMSE(y_train,y_train_pred)
Out[38]: 165.4555812634161

In [39]: RMSE(y_test,y_test_pred)
Out[39]: 166.55482391814556
```

-pm_living_price_index, 1.5M_avg_humidity, pm_food_energy_exclu_index를 제거한 모델

```
Slope: [ 178.06043949  131.69434284
-72.94890436  117.51467539 -183.08011436
  181.59283887 -184.88848992 -124.10124605
152.42632003 -231.76844363
   52.60396337 -975.39950143]
Intercept: 2730.8574078759493

In [41]: RMSE(y_train,y_train_pred)
Out[41]: 167.16410634771162

In [42]: RMSE(y_test,y_test_pred)
Out[42]: 167.92925056417977
```

-total_sunshine, day_rainfall을 제거한 모델

```
Slope: [ 263.79817294  148.49978399   46.72716412
-248.28434642  190.31976905
 -154.65871008 -121.26925447  143.53738144 -236.96521776
-975.87615217]
Intercept: 2678.3351850702143

In [9]: RMSE(y_train,y_train_pred)
Out[9]: 167.6934602535293

In [10]: RMSE(y_test,y_test_pred)
Out[10]: 168.55763698476403
```

처음 19개의 변수에서 시작하여 절댓값이 작은 변수들을 단계적으로 2개씩 제거하였다. 다만, 세 번째 모델은 소수점을 제외한 계수의 크기가 같은 변수가 존재하였기에 3개를 제거한 형태로 도출되었다. 모든 변수를 사용한 모델의 RMSE와 총 9개의 변수를 제거하고 10개의 변수만이 남은 모델의 RMSE는 약 2.5의 차이를 보인다. 이는 모델의 예측 정확도에 유의미한 영향을 미치는 정도가 아니라 판단, 9개의 변수를 제거한 모델을 최종 모델로 채택한다.

III. Conclusions

본 연구에서는 회귀 계수의 절댓값을 통해 상대적으로 중요성이 낮은 변수들을 제거하여 유의미한 변수들을 추출하였다. 최종적으로 남은 변수들은 다음과 같다.

[avg_temp,avg_wind_speed, total_solar_radiation, avg_ground_temp,0.5M_avg_humidity, 4.0M_avg_humidity,10CM_daily_soil_moisture, 30CM_daily_soil_moisture, 50CM_daily_soil_moisture, yield]

이에 대한 연장선으로 추후 연구에서는, 추출된 변수들을 여러 다른 기법에 적용해보며 정확도를 높일 수 있는 모델을 수립하는 방향으로 나아가고자 한다.

ACKNOWLEDGEMENT

이 논문은 2021년도 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 이공분야기초연구사업임 (NRF-2021R1F1A1064073).

REFERENCES

[1] https://www.kamis.or.kr/customer/main/main.do
[2] https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36
[3] https://kosis.kr/index/index.do
[4] https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1ET0021&conn_path=I2