상관관계 기반 파생변수 생성과 AutoML을 활용한 감귤 착과량 예측

서재석[1], 강준혁[1], 장근혁[1], 정아영[1], 김준화[2]
건양대학교 의료인공지능학과[1], 건양대학교 인공지능학과[2]
vhvh1398@naver.com, jh55603364@gmail.com, forren0418@naver.com,
1220jajjay@naver.com, junhwakim@konyang.ac.kr

Correlation-Based Derivative Generation and Mandarin Yield Prediction Using AutoML

Jaeseok Seo[1], Junhyeok Kang[1], Keunhyeok Jang[1], Ayoung Jeong[1], Junhwa Kim[2]
Department of Medical Artificial Intelligence, Konyang University[1]
Department of Artificial Intelligence, Konyang University[2]

요약

감귤 착과량 예측은 농업의 스마트화를 위한 핵심 과제로, 수확량 관리와 재배 전략 수립 등 폭넓은 영역에 영향을 미친다. 특히, 정확한 착과량 예측 모델을 확보하는 일은 효율적인 자원 배분과 생산성 향상을 위해 필수적이다. 본 연구는 새순의 생리적 지표를 활용하고 상관관계 분석을 통해 착과량 예측에 도움이 되는 파생변수를 생성 및 선택하고, 이를 AutoML 기반 기계 학습 모델에 통합함으로써 예측의 정확성과 실용성을 극대화한다.

Abstract

Mandarin yield prediction is crucial for smart agriculture, influencing harvest management and cultivation strategies. Accurate prediction ensures efficient resource allocation and boosts productivity. This study leverages physiological indicators of new shoots and correlation analysis to create and select derived variables, then integrates them into an AutoML-based machine learning model, thereby enhancing both prediction accuracy and practical utility.

1. 서론

감귤 착과량 예측은 농가의 수확·마케팅 전략, 노동력 배분, 자원 관리 등과 직결되어 생산성과 효율성을 높이는 데 핵심적이다. 그러나 복잡한 생리·환경적 요인으로 인해 정확한 예측이 쉽지 않으며, 특히 새순은 나무의 생장 패턴과 직접적으로 연관된 중요한 지표로, 광합성과 꽃 전환 과정에서 착과량에 직접 영향을 미쳐 과도하거나 부족할 경우 낙화 및 품질 저하 등이 발생한다 [1]. 최근 기계 학습 알고리즘을 적용한 감귤 착과량 예측 모델 [2]은 기존 관측 중심 방식보다 높은 정확도를 보였고, 파생변수 활용 [3]이 성능 향상에 효과적인 것으로 확인되었다.

본 연구는 새순의 생리적 특성을 반영한 파생변수를 생성해 AutoML [4] 기반 기계 학습 모델에 통합함으로써, 예측 정확도와 효율성을 극대화하고자 한다. 이를 통해 기존 접근 방식의 한계를 극복하고, 농가 생산성과 효율성 향상에 기여할 것으로 기대한다.

2. 본론

2.1 변수 선택 및 데이터 처리

본 연구에서는 감귤 착과량 예측을 위해 모델 학습에 사용하는 독립 변수로 수고, 수관폭평균, 새순, 엽록소를 대상으로 각 변수와 착과량 간의 상관관계를 분석하였다.

그림 1의 상관관계 분석 결과를 바탕으로, 수고, 수관폭, 엽록소는 다음과 같은 이유로 예측 모델에서 제외하였다. 수고와 수관폭은 나무의 물리적 크기를 나타내는 변수로, 착과량과의 직접적인 상관관계가 각각 0.007, 0.0306으로 매우 낮게 나타났다. 이는 나무의 물리적 크기는 착과량과의 직접적인 연관성이 적으며, 동일한 크기의 나무라 해도 생리적 상태와 환경 요인에 따라 착과량이 크게 달라질 수 있음을 의미한다.

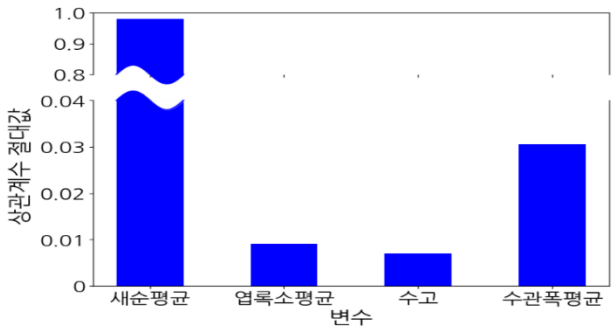

그림 1. 독립 변수와 착과량과의 상관계수 절댓값 비교

엽록소 농도는 광합성 활동과 나무의 건강 상태를 나타내는 지표로 유용하지만, 착과량과의 상관관계는 0.009로 낮았다. 이는 엽록소 농도가 빛, 온도, 수분 등 환경적 요인의 영향을 크게 받으며, 병해나 영양 부족으로 변동할 수 있기 때문이다. 또한, 새순이 잎으로 성장하면 엽록소 농도가 증가하지만, 꽃으로 성장할 경우 엽록소 농도가 감소하여 착과량과의 직접적인 연관성을 약화시킨다.

반면, 새순 평균은 착과량과 상관계수가 0.9755로 감귤 착과량 예측에 있어 가장 높은 연관성을 나타냈다. 이에 따라 새순평균을 주요 독립변수로 선정하여 모델 학습에 활용함으로써 착과량 예측의 정확성을 높이고자 하였다.

2.2 데이터 분석(EDA) 및 파생변수 선정

2.2.1 기본 파생변수 선정

우선, 착과량과의 상관관계가 높은 새순을 선정하여 데이터의 통계량을 바탕으로 표 1 에서 보이는 바와 같이, 새순 값의 평균(Mean), 분산(Var), 차분(Diff), 최댓값(Max), 최솟값(Min), 중앙값(Median)을 기본적인 파생변수 [3]를 만들었다.

표 1. 기본 파생변수

기본 파생변수	의미
mean	새순 값의 평균
var	새순 값의 분산
diff	새순 값의 차분
max	새순 값의 최댓값
min	새순 값의 최솟값
median	새순 값의 중앙값

2.2.2 새순 성장의 흐름 파악을 위한 추가 파생변수

기본 파생변수 [3]만으로는 새순의 세부적인 성장 양상이나 시계열적 패턴을 충분히 반영하기 어렵기에, 성장 불안정성, 불균형 정도, 로그 변환 등을 적용한 추가 파생변수를 생성하였다. 이를 통해 새순 성장 과정에서 나타나는 복잡한 변화와 숨겨진 요인을 더욱 정밀하게 파악하고자 하였다. 추가 파생변수는 표 2 와 같으며, 새순 성장의 불안정성(GII), 새순 초깃값 대비 최종값의 상대적 성장 비율(Growth Ratio), 특정 시점에서의 성장 집중도(Growth Focus), 새순 성장의 비효율성(Inefficient Growth), 새순 성장 중앙값(Median Growth), 새순 성장 불균형성(Growth Imbalance), 그리고 Log-scale 의 성장 안정성(Log Stability)이다. 추가된 파생변수를 얻는 수식은 아래와 같다.

$$\mathrm{GII} = \frac{1}{n}\sum_{i=1}^{n}|\Delta^2 S_i|,\ \Delta^2 S_i = (S_{i+1} - S_i) - (S_i - S_{i-1}) \quad (1)$$

$$\text{Growth Ratio} = \frac{S_{final} - S_{initial}}{S_{final} + 1} \quad (2)$$

$$\text{GrowthFocus} = \frac{\mu_S}{\sigma_S^2 + 10^{-5}} \quad (3)$$

$$\text{Inefficient Growth} = \frac{GII}{\text{Growth Focus} + 10^{-5}} \quad (4)$$

$$\text{Median Growth} = \text{Median}(S_i) - S_{\text{initial}} \quad (5)$$

$$\text{Growth Imbalance} = S_{max} - \text{Median Growth} \quad (6)$$

$$\text{Log Stability} = \ln(1 + \sigma_S^2) \quad (7)$$

S_i 는 새순 데이터값, S_{final} 은 마지막 일의 새순 데이터, $S_{initial}$ 은 시작 일의 새순 데이터, μ_S 은 새순의 평균, σ_S^2 은 새순의 분산, S_{max} 은 새순 데이터의 최댓값을 의미한다.

표 2. 추가 파생변수 표

추가 파생변수	의미
GII	새순 성장의 불안정성을 나타내는 지표
Growth Ratio	새순 초깃값 대비 최종값의 상대적 성장 비율
Inefficient Growth	새순 성장의 비효율성을 나타내는 지표
Growth Focus	특정 시점에서의 성장 집중도
Median Growth	새순 성장의 중앙값
Growth Imbalance	새순 성장의 불균형성
Log Stability	성장 안정성을 Log 로 변환한 지표

그림 2은 도출된 기본 파생변수와 본 연구에서 추가한 파생변수, 그리고 착과량 간의 상관관계를 시각화한 결과이다. 분석 결과, 새순을 기반으로 생성된 대부분의 파생변수가 상관계수 0.9를 초과하는 강한 상관성을 보였으며, 이는 착과량과 유의미한 관계를 갖는 것으로 해석할 수 있다. 이러한 변수를 예측 모델에 반영함으로써 감귤 착과량 예측 성능을 효과적으로 향상시킬 수 있음을 의미한다.

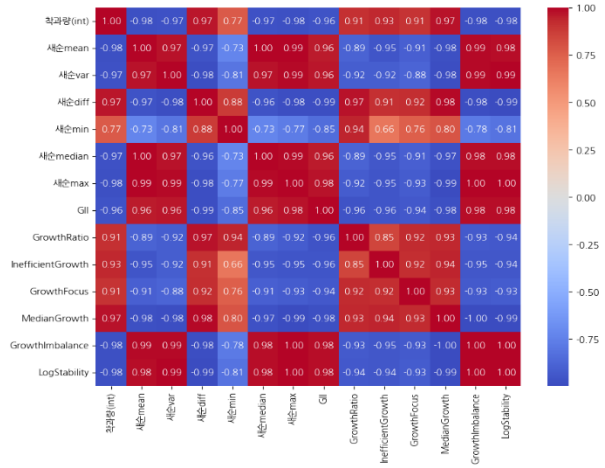

그림 2. 파생변수 간의 상관계수 히트맵

3. 성능 실험

3.1 단일 모델 성능 실험 및 기본 파생변수 중요성 실험

본 연구에서는 데이콘 감귤 착과량 예측 AI 경진대회에서 제공된 데이터 [5]를 사용하였다. 감귤 착과량 예측의 정확도를 향상시키기 위해 Random Forest [6], CatBoost [7], Extra Trees [8] 등을 대표 모델로 선정하였다. 위 알고리즘들은 의사결정 트리 기반으로 시계열 데이터의 복잡한 패턴을 잘 포착하며 MLJAR-Supervised AutoML [4]에서 가장 널리 사용되는 방법이다. 특히, CatBoost는 범주형 및 시계열 데이터에서 우수한 성능을 보여 왔으며, Random Forest와 Extra Trees는 과적합 방지와 변수 중요도 해석에 용이하여 새순 데이터를 다루는 데 적합하다.

앞서 설명한 mean, var, diff, max, min, median은 새순의 전반적인 분포와 변동 수준을 가장 직관적으로 나타내는 주는 지표다. 이들 기본 파생변수 [3]가 감귤 착과량과 매우 높은 상관성을 보였으며, 이전 농업 데이터 분석 연구에서도 평균, 분산, 최솟값, 최댓값 등의 통계량이 작물 생산량 예측에 탁월한 설명력을 제공한다는 사실이 보고된 바 있다 [9].

이에 따라 본 연구는 기본 파생변수를 고정하여 사용하고, 추가 파생변수를 점진적으로 결합하여 각 변수의 기여도를 평가하는 방식으로 모델 성능을 비교하였다. 이를 통해 추가 파생변수가 모델의 성능 향상에 미치는 영향을 체계적으로 분석하고자 하였다.

표 3. 단일 모델 성능 실험 표. RF: Random Forest, ET: Extra Trees, CB: CatBoost.

파생변수	RF	ET	CB
Baseline	0.07498	0.07889	0.07565
GII	0.07502	0.07860	0.07519
Imbalance	0.07495	0.07923	0.07571
Log Stability	0.07475	0.07909	0.07518
Growth Ratio	0.07475	0.07915	0.07536
Growth Focus	0.07479	0.07909	0.07528
Inefficient Growth	0.07499	0.07878	0.07576
Median Growth	0.07499	0.078700	0.07586

표 3은 기본 파생변수를 학습데이터에 추가한 상태에서, 추가 파생변수를 단일 모델(Random Forest, Extra Trees, CatBoost)에 개별적으로 추가하여 성능을 비교한 결과를 보여준다. 분석 결과, 세 가지 모델 중 두 가지 모델에서 성능이 개선된 GII, Log Stability, Growth Ratio, Growth Focus 네 가지 변수가 최종적으로 선정되었다.

성능 개선 폭을 살펴보면, GII는 Extra Trees와 CatBoost 모델에서 각각 -0.000286, -0.000459 감소하며 개선되었다. LogStability는 Random Forest에서 -0.000225, CatBoost에서 -0.000469 감소하였다. 또한, Growth Ratio는 Random Forest와 CatBoost에서 각각 -0.000226, -0.000287 감소하였다. Growth Focus는 Random Forest와 Extra Trees에서 각각 -0.000019, -0.00037 감소하였다.

이러한 결과는 일부 변수들이 특정 모델에서 성능 향상에 특히 기여했음을 나타내며, 이러한 변수들은 MLJAR-Supervised AutoML에 포함할 핵심 후보 파생변수로 선택되었으며, 모델의 성능 향상에 기여할 가능성이 높다고 판단된다.

3.2 추가 파생변수 Ablation Study

본 연구에서는 추가 파생변수 [3]의 효과를 검증하기 위하여 Ablation Study를 실시하였다. 본 연구에서는 MLJAR-Supervised AutoML [4]프레임워크를 활용하여, 변수 조합별 NMAE(Normalized Mean Absolute Error) [10]를 기준으로 모델 성능을 비교하였다.

$$\mathrm{NMAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n \cdot (\max(y) - \min(y))} \quad (8)$$

y_i는 실제값, $\hat{y}_i$는 예측값, n은 데이터 개수, $\max(y)$는 실제값의 최댓값, $\min(y)$은 실제값의 최솟값을 의미한다.

표 4는 기본 파생변수와 추가 파생변수를 서로 다른 조합으로 포함했을 때의 성능 지표를 제시한다. 기본 파생변수(mean, var, diff, max, min, median)를 적용하였을 때 NMAE가 0.1016에서 0.0723으로 크게 감소하였으며, 이는 새순 데이터의 통계적 특성이 착과량 예측에 유의미한 정보를 제공함을 보여준다.

표 4. 파생변수 조합 성능 비교

파생변수 조합	NMAE
Baseline	0.101649911
Baseline + 기본	0.072352608
Baseline + 기본 + GII + Log Stability	0.073390663
Baseline + 기본 + GII + Growth Ratio	0.072278048
Baseline + 기본 + GII + Growth Focus	0.076128285
Baseline + 기본 + Growth Ratio + Growth Focus	0.072554203
Baseline + 기본 + Growth Ratio + Log Stability	0.072168847
Baseline + 기본 + Growth Focus + Log Stability	0.072158964
Baseline + 기본 + GII + Growth Ratio + Growth Focus	0.072525546
Baseline + 기본 + GII + Growth Ratio + Log Stability	**0.071490143**
Baseline + 기본 + GII + Growth Focus + Log Stability	0.074134642

GII, Growth Ratio, Log Stability를 포함한 변수 조합이 NMAE 0.0714로 최고 성능을 보였으며, 이는 성장 비율 변화와 로그 변환이 데이터 잡음을 줄이고 핵심 패턴을 강조해 모델의 예측 정확도를 높인 결과다. 특히 Growth Ratio와 Log Stability는 복잡한 양상을 단순화하면서도 중요한 변화를 반영해 비선형적 관계 학습에 기여한다. 한편, 모든 파생변수가 성능 향상에 기여하지는 않으며, Growth Focus는 중복 설명 또는 잡음 추가로 예측력을 저해할 수 있다. 이는 변수 생성 시 데이터 본질과 타깃 변수 간 관계를 명확히 이해하는 것이 중요함을 시사한다. Ablation Study 결과, 추가 파생변수를 포함하면 모델 성능이 전반적으로 향상되며, 이는 MLJAR-Supervised AutoML의 자동 변수 최적화가 연구 효율성을 높이고 변수 조합 평가에 유용함을 보여준다.

4. 결론

본 논문은 상관관계 기반 파생변수를 생성하고 AutoML [4]을 활용하여 감귤 착과량을 예측하는 모델을 연구하였다. 상관도가 높은 새순 생리 지표를 중심으로 GII, Growth Ratio, Log Stability 등의 파생변수를 설계한 결과, 모델의 NMAE가 0.0714까지 낮아지며 초기 Baseline 대비 약 29%의 성능 향상을 확인할 수 있었다. 이는 새순 데이터의 복잡한 특성을 정교하게 반영하고, MLJAR-Supervised AutoML을 통한 자동화된 최적화 전략이 예측 정확도 제고에 효과적임을 보여준다.

이 연구는 감귤 스마트 농업 사용자의 주요 목표인 정확한 착과량 예측을 통해 수확·마케팅 전략, 자원 관리 효율성을 높이는 데 기여했다. 설계된 파생변수와 모델은 데이터의 핵심 패턴을 효과적으로 반영해 사용자 기대를 충족시켰으며, NMAE 감소로 의사결정 신뢰도를 향상시켰음을 보여준다. 특히, 상관관계가 높은 지표를 기반으로 파생변수를 도출함으로써 착과량 예측의 핵심 패턴을 효과적으로 포착할 수 있었음이 시사된다. 향후에는 기후, 토양 등 외부 환경 요인을 추가로 결합하고, 딥러닝 기반 모델을 도입해 착과량 변동의 세밀한 양상을 정밀하게 포착할 계획이다. 이를 통해 감귤뿐만 아니라 사과, 배, 포도 등 다양한 작물에도 적용할 수 있는 스마트 농업 솔루션으로 확장할 수 있을 것으로 기대한다.

ACKNOWLEDGEMENT

본 과제(결과물)는 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다. (2021RIS-004)

본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음 (2024-0-00047)

참고문헌

[1] Jeju Special Self-Governing Province Agricultural Technology Institute. 2021 Housgamgyul Cultivation Techniques. Jeju Special Self-Governing Province, https://agri.jeju.go.kr.

[2] 이봉규 "기계학습 기반의 감귤 착과량 예측 시스템" 한국소프트웨어감정평가학회 논문지 20.2 pp.29-34 (2024) : 29.

[3] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods." Computers & electrical engineering 40.1 (2014): 16-28.

[4] MLJAR. MLJAR-supervised: AutoML for Humans. https://supervised.mljar.com/. Accessed 23 Dec. 2024.

[5] DACON. 감귤 착과량 예측 AI 경진대회. DACON, https://dacon.io/competitions/official/236038/overview/description. Accessed 14 Dec. 2022

[6] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

[7] Prokhorenkova, Liudmila, et al. "CatBoost: unbiased boosting with categorical features." Advances in neural information processing systems 31 (2018).

[8] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." Machine learning 63 (2006): 3-42.

[9] Hong, Seong-Eun, et al. "Tomato Production and Growth Prediction Using ConvLSTM Model: The Role of Statistical Features like Mean and Maximum Values in Enhancing Prediction Accuracy." KI-IT Journal, 2020,https://ki-it.com/_PR/view/?aidx=22690&bidx=1843.Accessed 23 Dec. 2024.

[10] Berisha, F. J. O. L. L. A. "Quality of the predictions: mean absolute error, accuracy and coverage." Preprint (Sept. 2017). doi 10 (2017).