**Random Forests 연구**

REGRESSION\_TREE

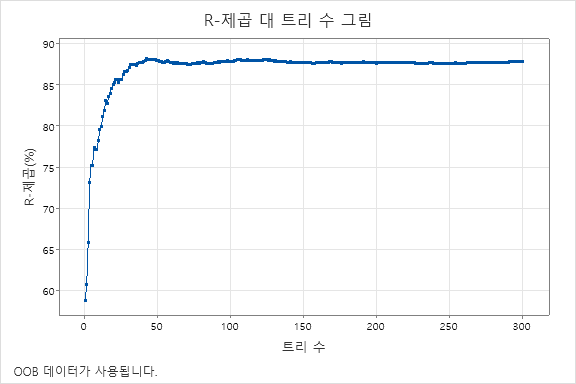
**Random Forests® 회귀: Y 대 X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13**

**방법**

|  |  |
| --- | --- |
| 모형 검증 | OOB 데이터로 검증 |
| 부트스트랩 표본 수 | 300 |
| 표본 크기 | 학습 데이터 크기 506과 동일 |
| 노드 분할을 위해 선택된 예측 변수 수 | 총 예측 변수 수의 제곱근 = 3 |
| 최소 내부 노드 크기 | 2 |
| 사용된 행 | 506 |

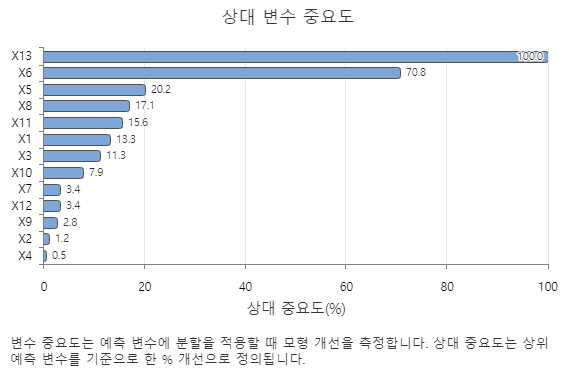
**반응 정보**

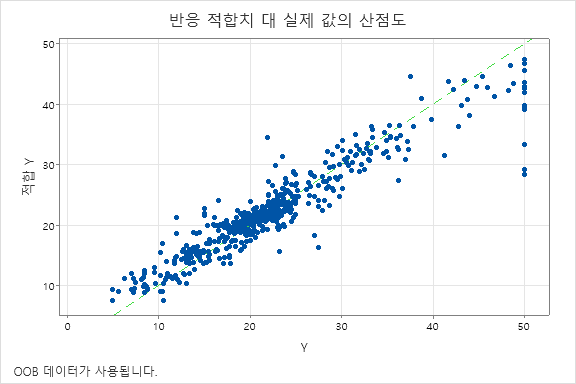
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **평균** | **표준 편차** | **최소값** | **Q1** | **중위수** | **Q3** | **최대값** |
| 22.5328 | 9.19710 | 5 | 16.95 | 21.2 | 25 | 50 |



**모형 요약**

|  |  |
| --- | --- |
| 전체 예측 변수 | 13 |
| 중요 예측 변수 | 13 |
| **통계량** | | | **OOB** |
| R-제곱 | | | 87.81% |
| 루트 평균 제곱 오차(RMSE) | | | 3.2081 |
| 평균 제곱 오차(MSE) | | | 10.2921 |
| 평균 절대 편차(MAD) | | | 2.1478 |
| 평균 절대 백분율 오차(MAPE) | | | 0.1075 |





REGRESSION\_TREE

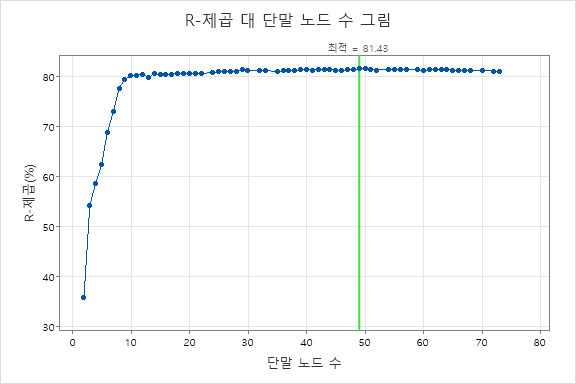
**49 노드 CART® 회귀: Y 대 X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13**

**방법**

|  |  |
| --- | --- |
| 노드 분할 | 최소 제곱 오차 |
| 최적 트리 | 최대 R-제곱 |
| 모형 검증 | 5-접기 교차 검증 |
| 사용된 행 | 506 |

**반응 정보**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **평균** | **표준 편차** | **최소값** | **Q1** | **중위수** | **Q3** | **최대값** |
| 22.5328 | 9.19710 | 5 | 16.95 | 21.2 | 25 | 50 |



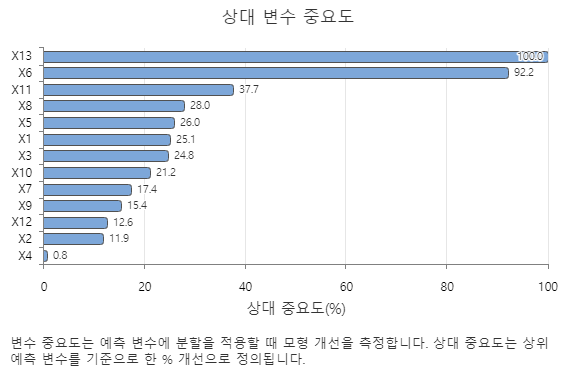
**최적 트리 다이어그램**

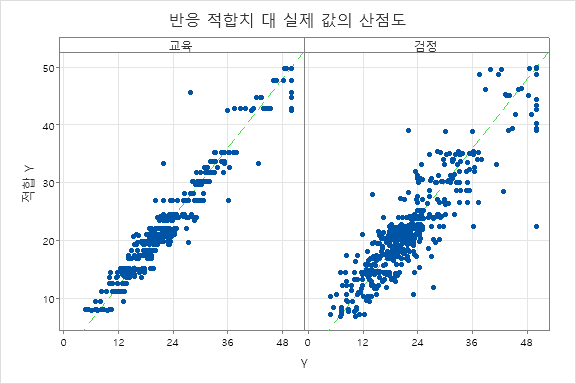
도표, 패턴이(가) 표시된 사진

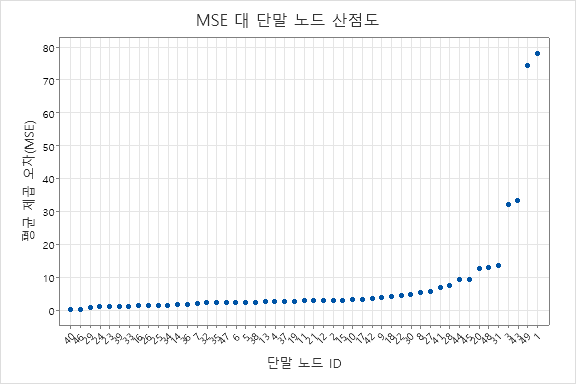
자동 생성된 설명

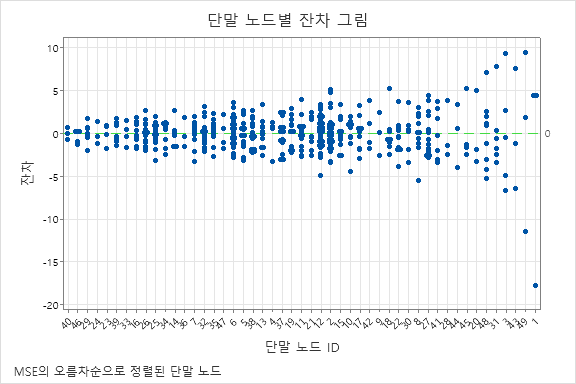
**모형 요약**

|  |  |
| --- | --- |
| 전체 예측 변수 | 13 |
| 중요 예측 변수 | 13 |
| 단말 노드 수 | 49 |
| 최소 단말 노드 크기 | 3 |
| **통계량** | | | **교육** | **검정** |
| R-제곱 | | | 94.18% | 81.43% |
| 루트 평균 제곱 오차(RMSE) | | | 2.2157 | 3.9596 |
| 평균 제곱 오차(MSE) | | | 4.9094 | 15.6781 |
| 평균 절대 편차(MAD) | | | 1.5433 | 2.7870 |
| 평균 절대 백분율 오차(MAPE) | | | 0.0779 | 0.1376 |









**결과 해석**

랜덤포레스트

**모형 요약**

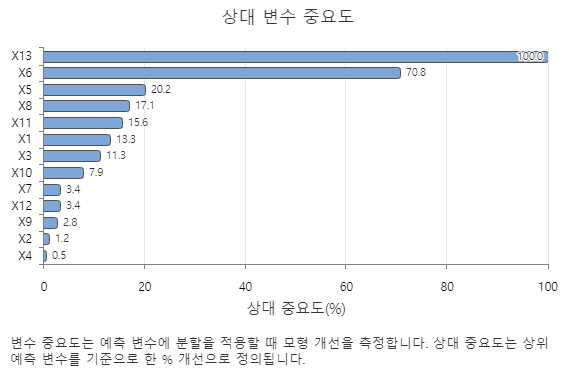
|  |  |
| --- | --- |
| 전체 예측 변수 | 13 |
| 중요 예측 변수 | 13 |
| **통계량** | | | **OOB** |
| R-제곱 | | | 87.81% |
| 루트 평균 제곱 오차(RMSE) | | | 3.2081 |
| 평균 제곱 오차(MSE) | | | 10.2921 |
| 평균 절대 편차(MAD) | | | 2.1478 |
| 평균 절대 백분율 오차(MAPE) | | | 0.1075 |

CART

**모형 요약**

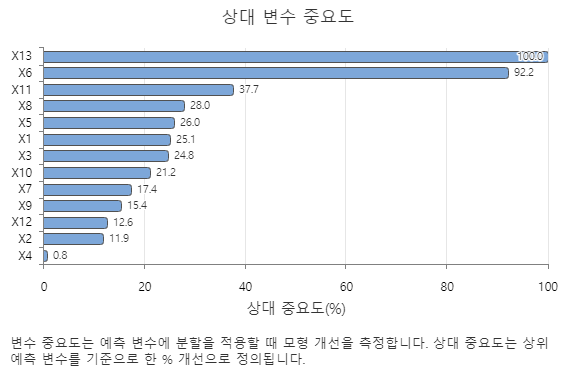
|  |  |
| --- | --- |
| 전체 예측 변수 | 13 |
| 중요 예측 변수 | 13 |
| 단말 노드 수 | 49 |
| 최소 단말 노드 크기 | 3 |
| **통계량** | | | **교육** | **검정** |
| R-제곱 | | | 94.18% | 81.43% |
| 루트 평균 제곱 오차(RMSE) | | | 2.2157 | 3.9596 |
| 평균 제곱 오차(MSE) | | | 4.9094 | 15.6781 |
| 평균 절대 편차(MAD) | | | 1.5433 | 2.7870 |
| 평균 절대 백분율 오차(MAPE) | | | 0.0779 | 0.1376 |

랜덤포레스트



상대 변수 중요도 : X13(100.0) > X6(70.8) > X5(20.2)

CART



상대 변수 중요도 : X13(100.0) > X6(93.2) > X11(37.1)

**하이퍼 파라미터 수정**

1. 트리를 성장시기는 부트스트랩 표본의 수 : 300(고정)

- 총 예측 변수 수의 제곱근 -> (RF Model1)

**RF Model1**

|  |  |
| --- | --- |
| 전체 예측 변수 | 13 |
| 중요 예측 변수 | 13 |
| **통계량** | | | **OOB** |
| R-제곱 | | | 87.81% |
| 루트 평균 제곱 오차(RMSE) | | | 3.2081 |
| 평균 제곱 오차(MSE) | | | 10.2921 |
| 평균 절대 편차(MAD) | | | 2.1478 |
| 평균 절대 백분율 오차(MAPE) | | | 0.1075 |

- 부트스트랩 포레스트를 생성하는 총 예측 변수 수 -> (RF Model2)

**RF Model2**

|  |  |  |
| --- | --- | --- |
| 전체 예측 변수 |  | 13 |
| 중요 예측 변수 |  | 12 |
| **통계량** | | | | **OOB** |
| R-제곱 | | | | 87.93% |
| 루트 평균 제곱 오차(RMSE) | | | | 3.1924 |
| 평균 제곱 오차(MSE) | | | | 10.1911 |
| 평균 절대 편차(MAD) | | | | 2.1693 |
| 평균 절대 백분율 오차(MAPE) | | | | 0.1104 |

- 총 예측 변수 수의 K 백분율 , K=20 -> (RF Model3)

**RF Model3**

|  |  |  |
| --- | --- | --- |
| 전체 예측 변수 |  | 13 |
| 중요 예측 변수 |  | 13 |
| **통계량** | | | | **OOB** |
| R-제곱 | | | | 87.13% |
| 루트 평균 제곱 오차(RMSE) | | | | 3.2957 |
| 평균 제곱 오차(MSE) | | | | 10.8616 |
| 평균 절대 편차(MAD) | | | | 2.2076 |
| 평균 절대 백분율 오차(MAPE) | | | | 0.1099 |

2. 노드 분할에 대한 예측 변수 수 : 총 예측 변수 수의 제곱근(고정)

- 트리 수에 따른 랜덤포레스트 회귀모델 결과 확인

- 트리 수 : 300, 1000, 2000

**트리수 : 300**

|  |  |
| --- | --- |
| 전체 예측 변수 | 13 |
| 중요 예측 변수 | 13 |
| **통계량** | | | **OOB** |
| R-제곱 | | | 87.81% |
| 루트 평균 제곱 오차(RMSE) | | | 3.2081 |
| 평균 제곱 오차(MSE) | | | 10.2921 |
| 평균 절대 편차(MAD) | | | 2.1478 |
| 평균 절대 백분율 오차(MAPE) | | | 0.1075 |

**트리수 : 1000**

|  |  |
| --- | --- |
| 전체 예측 변수 | 13 |
| 중요 예측 변수 | 13 |
| **통계량** | | | **OOB** |
| R-제곱 | | | 88.20% |
| 루트 평균 제곱 오차(RMSE) | | | 3.1568 |
| 평균 제곱 오차(MSE) | | | 9.9652 |
| 평균 절대 편차(MAD) | | | 2.1095 |
| 평균 절대 백분율 오차(MAPE) | | | 0.1057 |

**트리수 : 2000**

|  |  |
| --- | --- |
| 전체 예측 변수 | 13 |
| 중요 예측 변수 | 13 |
| **통계량** | | | **OOB** |
| R-제곱 | | | 88.32% |
| 루트 평균 제곱 오차(RMSE) | | | 3.1397 |
| 평균 제곱 오차(MSE) | | | 9.8579 |
| 평균 절대 편차(MAD) | | | 2.1004 |
| 평균 절대 백분율 오차(MAPE) | | | 0.1055 |

**결과 해석**

부트스트랩 표본의 수를 300으로 고정한 후 노드분할을 결정하는 방법을 바꾼 모형들에서는 총 예측변수의 수 방법인 RF Model2 가 87.93% 로 가장 높았으며, 전체의 트리수를 바꾼 모형들 중에서는 트리수가 가장 많은 2000 모형이 88.32% 로 R2값이 최고로 높았습니다.

**추가 분석**

위에서 가장 높은 모형 정확성을 보인 파라미터의 조건들을 모두 선택한다면 모델의 성능이 가장 좋아질 것 같지만 반드시 그런 것은 아닙니다.

다음은 위에서 R2 값들이 가장 높은 조건들을 모두 사용하여 모델을 적용시킨 결과입니다.

**모형 요약**

|  |  |
| --- | --- |
| 전체 예측 변수 | 13 |
| 중요 예측 변수 | 11 |
| **통계량** | | | **OOB** |
| R-제곱 | | | 87.85% |
| 루트 평균 제곱 오차(RMSE) | | | 3.2023 |
| 평균 제곱 오차(MSE) | | | 10.2547 |
| 평균 절대 편차(MAD) | | | 2.1634 |
| 평균 절대 백분율 오차(MAPE) | | | 0.1107 |

기존의 표본의 수 300일때의 R2 값인 87.93% 와 노드분할 결정방법을 예측변수의 제곱근으로 선택하는 방법의 R2 값인 88.32% 보다 모두 낮아진 것을 알 수 있습니다.

따라서 모형을 적용할 때는 각 파라미터들의 조건을 이해하고 다양한 상황을 검토하여 설정하여야 합니다.

예측 분석

신규 데이터를 활용한 예측치는 다음과 같습니다.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | X12 | X13 |
| 1.46336 | 0 | 19.58 | 0 | 0.605 | 7.489 | 90.8 | 1.9709 | 5 | 403 | 14.7 | 374.43 | 1.73 |
| 1.83377 | 0 | 19.58 | 1 | 0.605 | 7.802 | 98.2 | 2.0407 | 5 | 403 | 14.7 | 389.61 | 1.92 |
| 1.51902 | 0 | 19.58 | 1 | 0.605 | 8.375 | 93.9 | 2.162 | 5 | 403 | 14.7 | 388.45 | 3.32 |
| 2.24236 | 0 | 19.58 | 0 | 0.605 | 5.854 | 91.8 | 2.422 | 5 | 403 | 14.7 | 395.11 | 11.64 |
| 2.924 | 0 | 19.58 | 0 | 0.605 | 6.101 | 93 | 2.2834 | 5 | 403 | 14.7 | 240.16 | 9.81 |

각각의 값들에 대한 실제 값과 예측값은 다음과 같습니다.

|  |  |  |
| --- | --- | --- |
| 실제y | Fit | 잔차 |
| 50 | 48.3317 | 1.6683 |
| 50 | 49.1503 | 0.8497 |
| 50 | 49.098 | 0.902 |
| 22.7 | 22.2173 | 0.4827 |
| 25 | 23.852 | 1.148 |