**TreeNet(GB) 모델을 이용한 시멘트 강도 분석**

CONCRETE\_STRENGTH

**하이퍼파라미터 튜닝을 통한 TreeNet® 회귀 분석: Strength 대 Cement, Blast\_Furnace\_Slag, Fly\_Ash, Water, Superplasticizer, Coarse\_Aggregate, Fine\_Aggregate, Age**

**방법**

|  |  |
| --- | --- |
| 손실 함수 | 제곱 오차 |
| 최적 트리 수 선택 기준 | 최대 R-제곱 |
| 모형 검증 | 5-접기 교차 검증 |
| 학습률 | 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3 |
| 하위 표본 부분 | 0.5, 0.7 |
| 트리당 최대 터미널 노드 수 | 6 |
| 최소 단말 노드 크기 | 3 |
| 노드 분할을 위해 선택된 예측 변수 수 | 8 |
| 사용된 행 | 1030 |

**반응 정보**

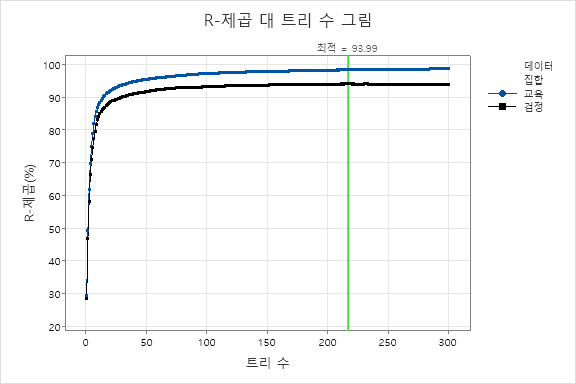
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **평균** | **표준 편차** | **최소값** | **Q1** | **중위수** | **Q3** | **최대값** |
| 35.8178 | 16.7057 | 2.33181 | 23.6950 | 34.4428 | 46.2087 | 82.5992 |

**하이퍼파라미터 최적화**

검정

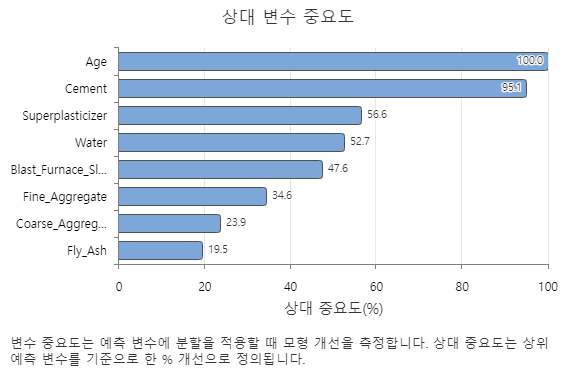
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **모형** | **최적의 트리 수** | **R-제곱(%)** | **평균 절대 편차** | **학습률** | **하위 표본 부분** | **최대 단말 노드** |
| 1 | 300 | 27.39 | 11.4480 | 0.001 | 0.5 | 6 |
| 2 | 300 | 70.27 | 7.2787 | 0.005 | 0.5 | 6 |
| 3 | 300 | 83.55 | 5.3585 | 0.010 | 0.5 | 6 |
| 4 | 300 | 92.44 | 3.2788 | 0.050 | 0.5 | 6 |
| 5 | 300 | 93.30 | 2.9984 | 0.100 | 0.5 | 6 |
| 6 | 298 | 93.69 | 2.8253 | 0.200 | 0.5 | 6 |
| 7 | 269 | 93.45 | 2.9048 | 0.300 | 0.5 | 6 |
| 8 | 300 | 27.27 | 11.4280 | 0.001 | 0.7 | 6 |
| 9 | 300 | 69.82 | 7.3006 | 0.005 | 0.7 | 6 |
| 10 | 300 | 83.22 | 5.4045 | 0.010 | 0.7 | 6 |
| 11 | 300 | 92.36 | 3.2947 | 0.050 | 0.7 | 6 |
| 12 | 300 | 93.60 | 2.8651 | 0.100 | 0.7 | 6 |
| 13 | 300 | 93.98 | 2.7366 | 0.200 | 0.7 | 6 |
| 14\* | 217 | 93.99 | 2.7485 | 0.300 | 0.7 | 6 |

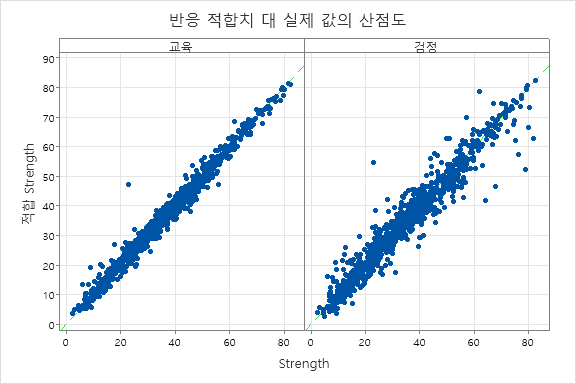
\* 최적의 모형에는 최대 R-제곱이 있습니다. 최적의 모형 출력이 뒤에 나옵니다.



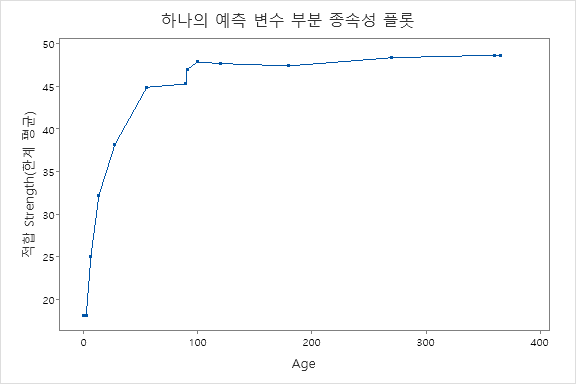
**모형 요약**

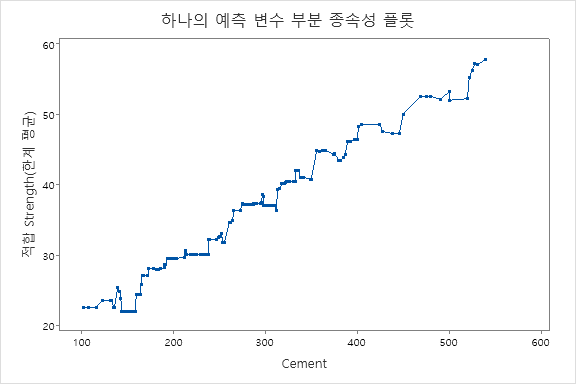
|  |  |
| --- | --- |
| 전체 예측 변수 | 8 |
| 중요 예측 변수 | 8 |
| 성장한 트리 수 | 300 |
| 최적의 트리 수 | 217 |
| **통계량** | | | **교육** | **검정** |
| R-제곱 | | | 98.28% | 93.99% |
| 루트 평균 제곱 오차(RMSE) | | | 2.1919 | 4.0920 |
| 평균 제곱 오차(MSE) | | | 4.8045 | 16.7449 |
| 평균 절대 편차(MAD) | | | 1.5458 | 2.7485 |
| 평균 절대 백분율 오차(MAPE) | | | 0.0584 | 0.0983 |

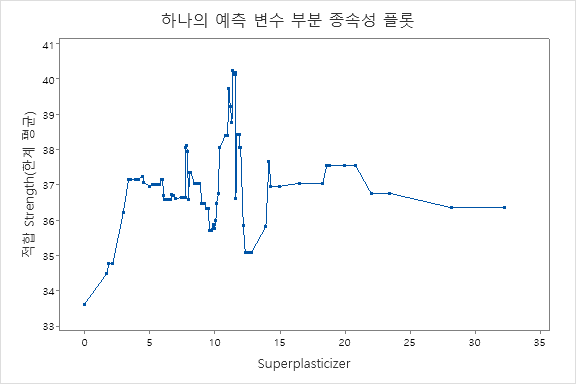


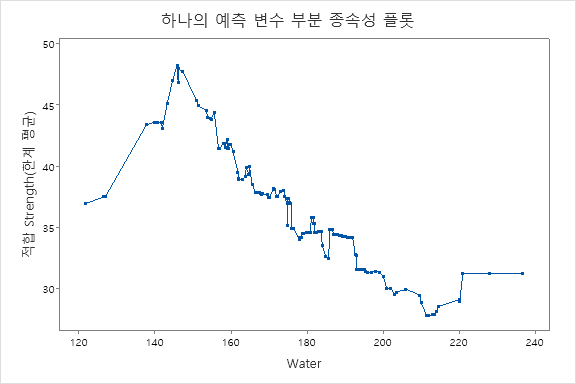


**하나의 예측 변수 부분 종속성 플롯**

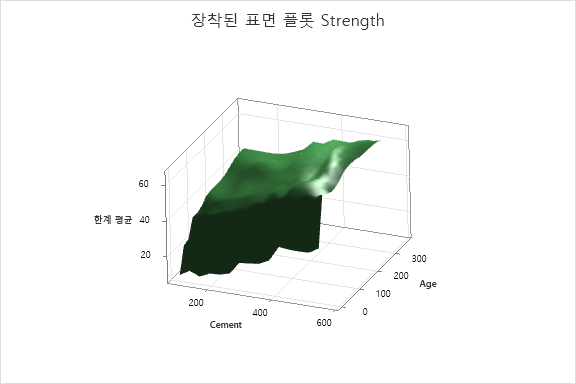


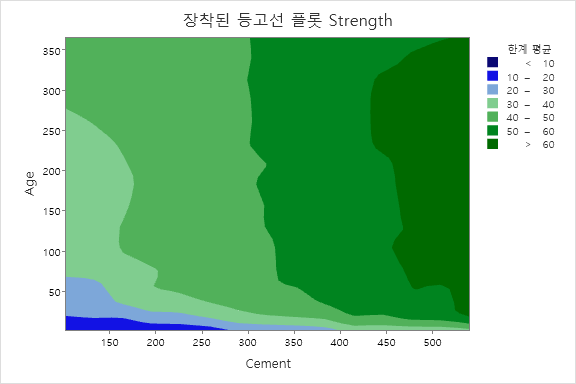


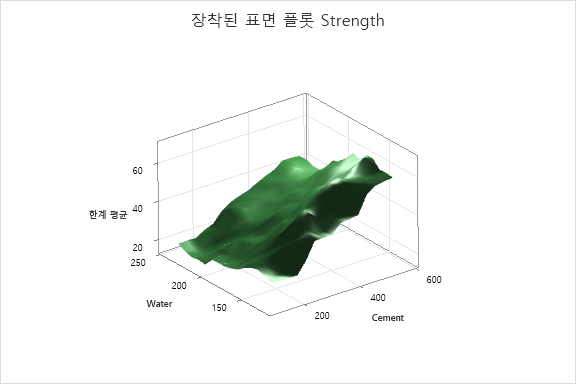


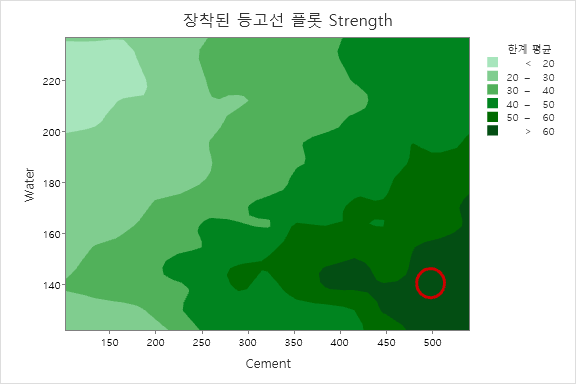


**두 개의 예측 변수 부분 종속성 플롯**









**모델 성능 비교**

CART vs Random Forest vs TreeNet

**CART**

|  |  |
| --- | --- |
| 전체 예측 변수 | 8 |
| 중요 예측 변수 | 8 |
| 단말 노드 수 | 175 |
| 최소 단말 노드 크기 | 3 |
| **통계량** | | | **교육** | **검정** |
| R-제곱 | | | 95.04% | 85.05% |
| 루트 평균 제곱 오차(RMSE) | | | 3.7181 | 6.4558 |
| 평균 제곱 오차(MSE) | | | 13.8245 | 41.6773 |
| 평균 절대 편차(MAD) | | | 2.5649 | 4.5214 |
| 평균 절대 백분율 오차(MAPE) | | | 0.0889 | 0.1549 |

**RandomForest**

|  |  |
| --- | --- |
| 전체 예측 변수 | 8 |
| 중요 예측 변수 | 8 |
| **통계량** | | | **OOB** |
| R-제곱 | | | 92.27% |
| 루트 평균 제곱 오차(RMSE) | | | 4.6425 |
| 평균 제곱 오차(MSE) | | | 21.5530 |
| 평균 절대 편차(MAD) | | | 3.2256 |
| 평균 절대 백분율 오차(MAPE) | | | 0.1208 |

**TreeNet**

|  |  |
| --- | --- |
| 전체 예측 변수 | 8 |
| 중요 예측 변수 | 8 |
| 성장한 트리 수 | 300 |
| 최적의 트리 수 | 217 |
| **통계량** | | | **교육** | **검정** |
| R-제곱 | | | 98.28% | 93.99% |
| 루트 평균 제곱 오차(RMSE) | | | 2.1919 | 4.0920 |
| 평균 제곱 오차(MSE) | | | 4.8045 | 16.7449 |
| 평균 절대 편차(MAD) | | | 1.5458 | 2.7485 |
| 평균 절대 백분율 오차(MAPE) | | | 0.0584 | 0.0983 |

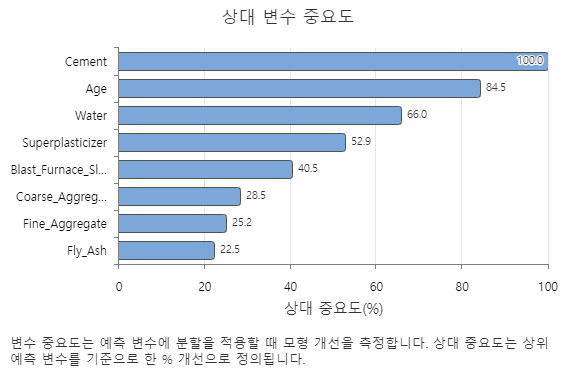
**결과 해석**

- 하이퍼 파라미터 조정

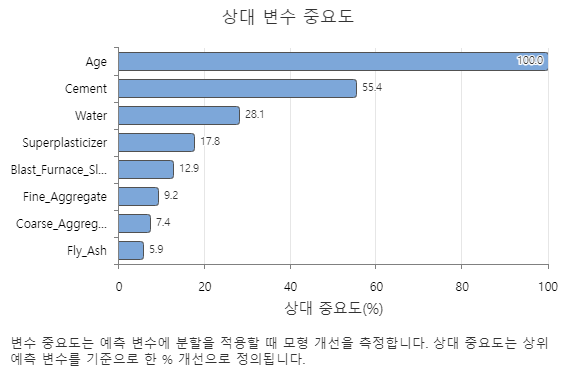
하이퍼 파라미터를 조정하기전 TreeNet의 R2 값은 83.97% 모델 중 가장 낮은 성능을 보였지만 하이퍼 파라미터를 조정한 TreeNet의 R2 은 93.99% 로 다른 모델에 비해 우수한 성능을 보이고 있다. 최초의 학습률을 너무 낮게 설정한 것이 성능 저하의 원인이었으며, 학습률에서 0.1 이상부터 모두 R2 값이 90 이상이 나오는 것으로 보아 최적의 학습률은 0.1 이상으로 보인다. (현재 적용된 학습률 = 0.3)

- 상대적 변수 중요도

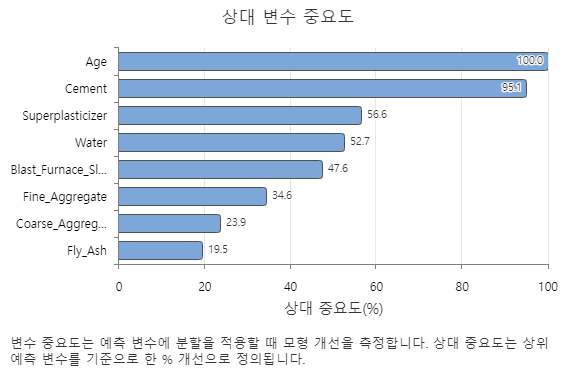
CART



RandomForest



TreeNet

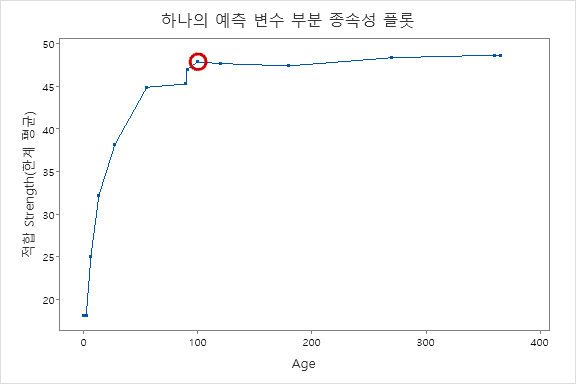


**결과 해석**

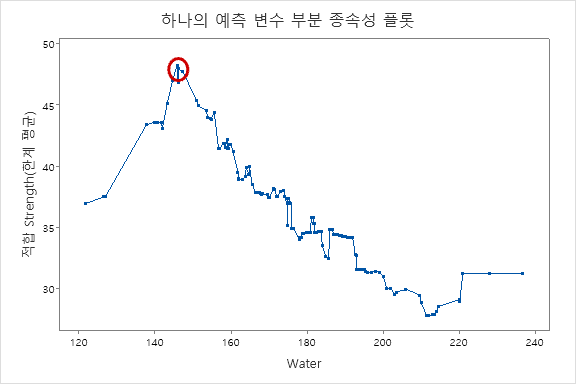
CART 모형에서는 Cement가 가장 중요한 변수로 뽑았지만, RandomForest와 TreeNet에서는 가장 중요한 변수로 Age(양생시간)를 설정하였다. 따라서 시멘트의 강도를 결정하는데 있어서 시멘트의 양보다 중요한 요인으로 양생시간이 중요한 변수로 볼 수 있다.

- 하나의 예측변수 부분 종속성 플롯

Age



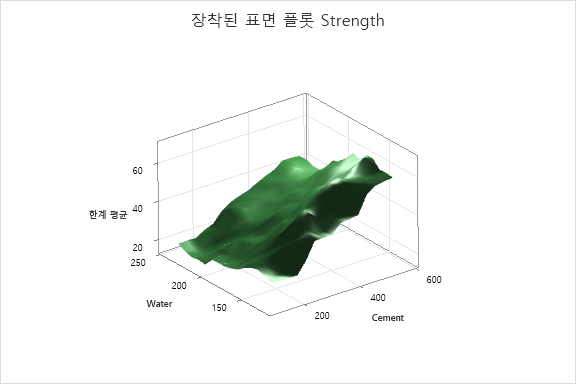
Watet

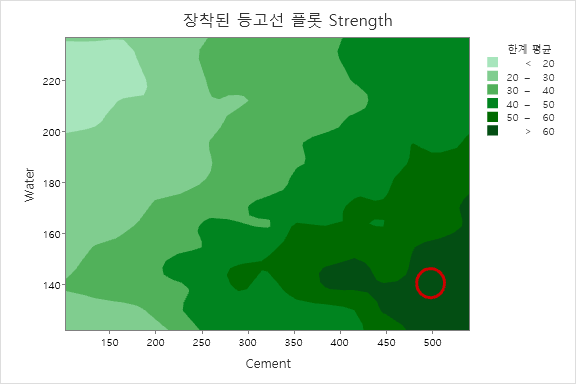


**결과 해석**

적합 강도 대비 최적의 양생시간은 100시간으로 볼 수 있으며, 최적의 물의 양 또한 145.9 로 알 수 있다.

- 두개의 예측변수 부분 종속성 플롯

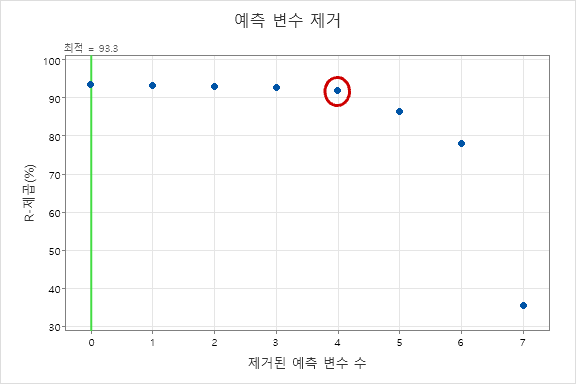




**결과 해석**

한 개의 예측변수에서 최적의 시멘트양에 대한 정확한 수치를 얻을 수 없기 때문에, 최적의 강도 대비 두개의 변수를 비교 한 결과 최적의 시멘트와 물의 양은 각각 500, 140 으로 볼 수 있습니다. (안전 수치)

- 주요 변수 추출



**중요하지 않은 예측 변수를 제거하여 모형 선택**

검정

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **모형** | **최적의 트리 수** | **R-제곱(%)** | **예측 변수 수** | **제거된 예측 변수** |
| 1\* | 300 | 93.30 | 8 | 없음 |
| 2 | 300 | 93.17 | 7 | Fly\_Ash |
| 3 | 300 | 92.98 | 6 | Coarse\_Aggregate |
| 4 | 299 | 92.60 | 5 | Fine\_Aggregate |
| 5 | 300 | 91.91 | 4 | Superplasticizer |
| 6 | 299 | 86.28 | 3 | Blast\_Furnace\_Slag |
| 7 | 299 | 78.10 | 2 | Water |
| 8 | 258 | 35.43 | 1 | Age |

\* 선택한 모형에는 최대 R-제곱이 있습니다. 선택한 모형의 출력이 뒤에 나옵니다.

**결과 해석**

중요하지 않은 변수를 제거하여 변화하는 R2를 측정한 결과 중요하지 않은 변수 4개를 제거하여도 R2 는 90이상을 유지 하였다. 따라서 하위 4개의 변수는 시멘트 강도를 결정하는데 있어 중요하지 않은 변수로 해석할 수 있다. (Fly\_Ash, Coarse\_Aggregate, Fine\_Aggregate, Superplasticizer)