**Bike\_Sharing\_Demand**

자전거 공유 시스템은 회원 가입, 대여 및 자전거 반납 과정이 도시 전역의 키오스크 위치 네트워크를 통해 자동화되는 자전거 대여 수단입니다. 이러한 시스템을 사용하여 사람들은 한 위치에서 자전거를 대여하고 필요에 따라 다른 장소로 반납할 수 있습니다. 현재 전 세계적으로 500개 이상의 자전거 공유 프로그램이 있습니다.

이러한 시스템에서 생성된 데이터는 이동 기간, 출발 위치, 도착 위치 및 경과 시간이 명시적으로 기록되기 때문에 연구자에게 매력적입니다. 따라서 자전거 공유 시스템은 도시의 이동성을 연구하는 데 사용할 수 있는 센서 네트워크로 기능합니다. 이 대회에서 참가자들은 워싱턴 D.C.의 Capital Bikeshare 프로그램에서 자전거 대여 수요를 예측하기 위해 과거 사용 패턴과 날씨 데이터를 결합하도록 요청받습니다.

TRAIN\_NEW (1).CSV

**100 노드 CART® 회귀: count 대 temp, atemp, humidity, windspeed, hour, holiday, workingday, season, weather, weekday**

\* 경고 \* Minitab은 기본 제한으로 인해 1024개 노드에서 트리 성장을 중지했습니다.

\* 경고 \* 다음 수준이 weather: 4 의 검정 표본에 없습니다.

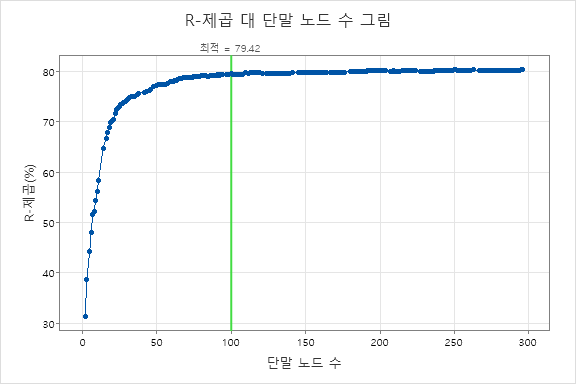
\* 경고 \* 1개 범주형 예측 변수에 교육 및 검정 표본에 모두 없는 수준이 있습니다. 결과를 해석할 때 주의하십시오. 통계량 > 표 > 교차 표 및 카이 제곱을 사용  
하여 어디에 수준이 없는지 확인할 수 있습니다.

**방법**

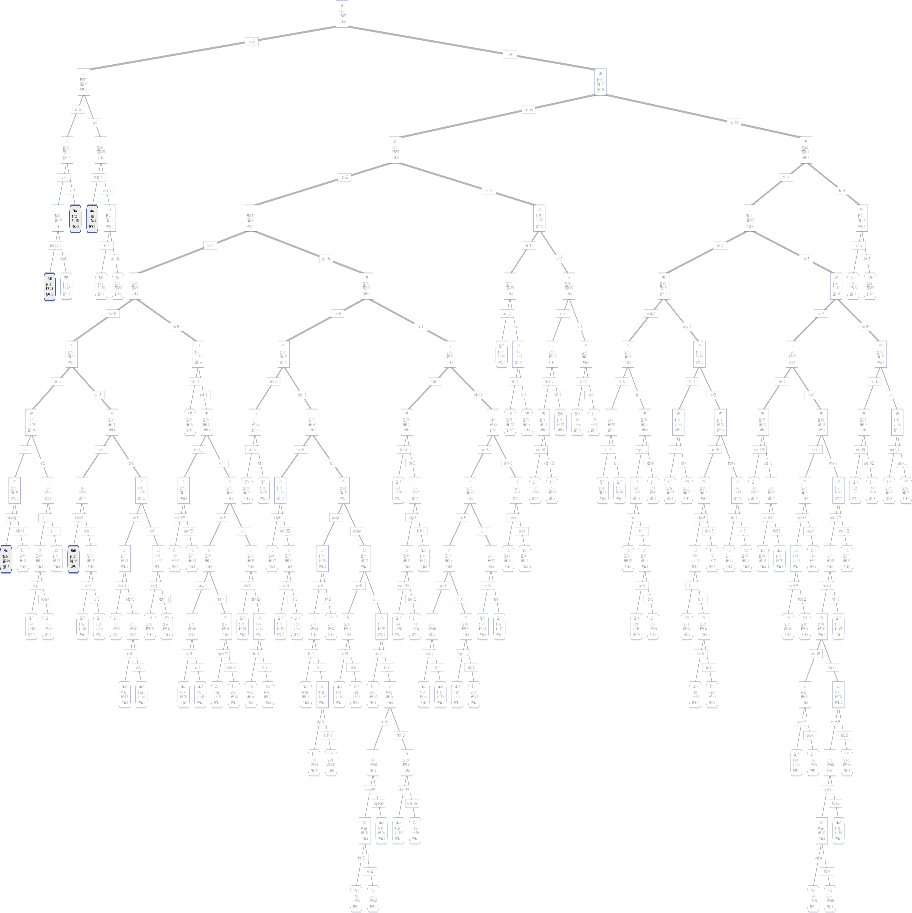
|  |  |
| --- | --- |
| 노드 분할 | 최소 제곱 오차 |
| 최적 트리 | 최대 R-제곱의 1 표준 오차 이내 |
| 모형 검증 | 70/30% 교육/검정 집합 |
| 사용된 행 | 10886 |

**반응 정보**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **데이터 집합** | **N** | **N의 비율(%)** | **평균** | **표준 편차** | **최소값** | **Q1** | **중위수** | **Q3** | **최대값** |
| 교육 | 7616 | 70.0 | 191.424 | 181.321 | 1 | 43 | 145 | 281 | 977 |
| 검정 | 3270 | 30.0 | 191.925 | 180.759 | 1 | 42 | 145 | 288 | 943 |

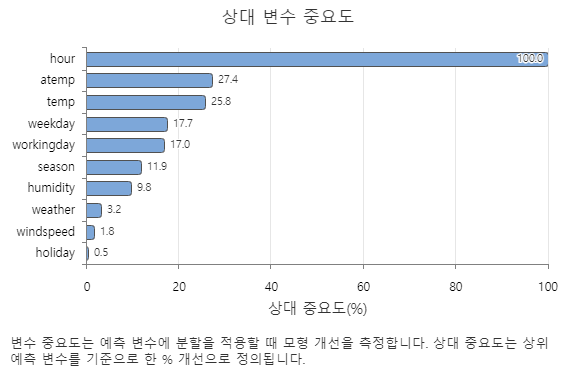


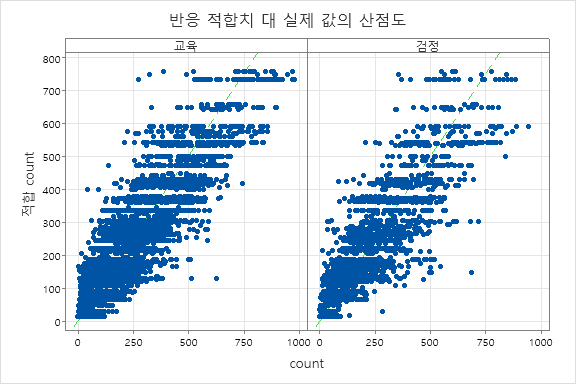
**최적 트리 다이어그램**

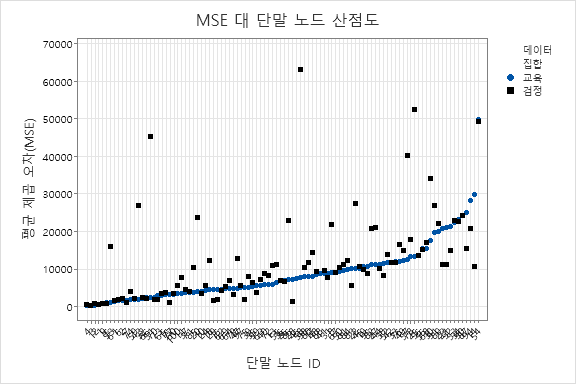


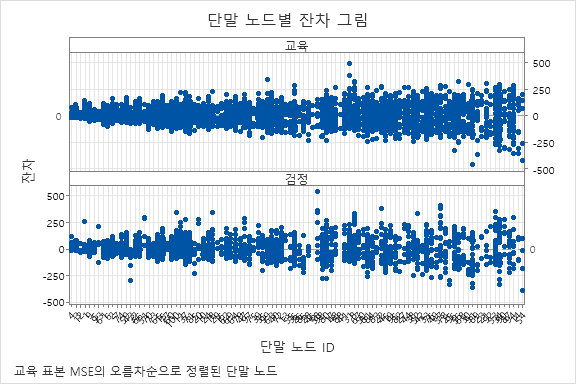
**모형 요약**

|  |  |
| --- | --- |
| 전체 예측 변수 | 10 |
| 중요 예측 변수 | 10 |
| 단말 노드 수 | 100 |
| 최소 단말 노드 크기 | 3 |
| **통계량** | | | **교육** | **검정** |
| R-제곱 | | | 84.14% | 79.42% |
| 루트 평균 제곱 오차(RMSE) | | | 72.2003 | 81.9822 |
| 평균 제곱 오차(MSE) | | | 5212.8876 | 6721.0791 |
| 평균 절대 편차(MAD) | | | 50.8756 | 56.6575 |
| 평균 절대 백분율 오차(MAPE) | | | 0.8232 | 0.8484 |









**count에 대한 예측**

**예측**

예측된 통계량의 열은 워크시트에 저장됩니다.

**결과 해석**

현재 CART 알고리즘으로 평가한 결과 79.42% 의 정확률을 보였습니다.

각 변수들 중 hour에 대한 중요도가 다른 변수들에 비해 약 4배 이상 중요도를 보였습니다.

최 하위의 단말 노드는 평균이 15.22회로 새벽 1시 30분에서 5시 30분 사이였습니다. 전체카운트는 1250으로 전체 데이터의 약 10%를 차지하고 있습니다..

최 상위 단말 노드는 평균이 755.714회로 조건은 다음과 같습니다.

temp <= 26.65, 29.925 < atemp <= 39.7725, humidity <= 45, windspeed <= 21.0012, 16.5 < hour <= 18.5, workingday = {1}, weekday = {0, 1, 2, 3} 전체 카운트는 14회 입니다.

두번째 단말 노드는 평균이 748.429회로 조건은 다음과 같습니다.

temp > 20.09, atemp <= 29.925, humidity <= 74.5, 16.5 < hour <= 18.5, workingday = {1}, season = {2, 3, 4}, weekday = {2} 전체 카운트는 7회 입니다.

세번째 단말 노드는 평균이 731.483으로 조건은 다음과 같습니다.

temp > 26.65, 29.925 < atemp <= 39.7725, humidity <= 77, windspeed <= 21.0012, 16.5 < hour <= 18.5, workingday = {1}, weekday = {0, 1, 2, 3} 전체 카운트는 87회 입니다.

네번째 단말 노드는 평균이 658.267로 조건은 아래와 같습니다.

29.925 < atemp <= 32.955, humidity <= 60.5, 10.5 < hour <= 15.5, workingday = {0}, season = {1, 3} 전체 카운트는 15회 입니다.

다섯번째 단말 노드는 평균이 647.318로 조건은 아래와 같습니다.

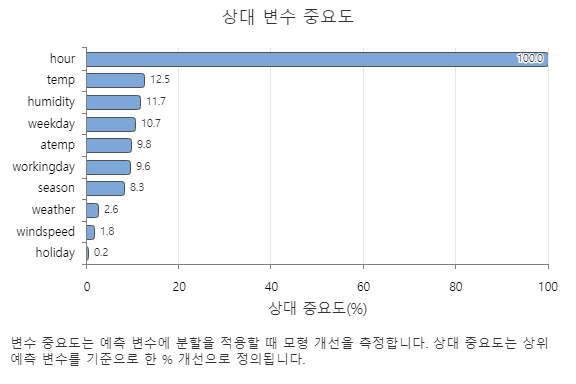
atemp > 29.925, humidity <= 63.5, windspeed <= 21.0012, 16.5 < hour <= 18.5, workingday = {1}, weekday = {4, 5, 6} 전체 카운트는 22회 입니다.

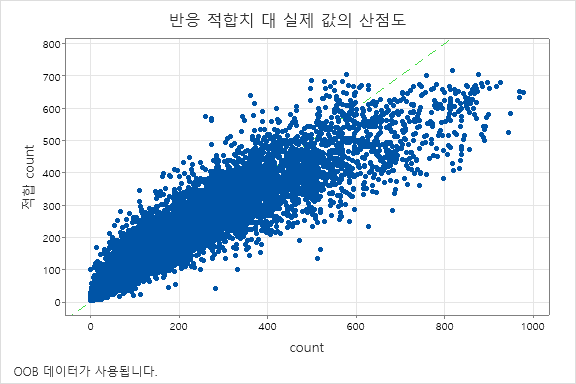
위는 상위 5개 그룹에 대한 상세한 정보이며, 위의 조건이 만족했을 때 자전거 수요가 급증하는 것을 확인하였습니다.

**추가 분석**

**모형 요약**

|  |  |
| --- | --- |
| 전체 예측 변수 | 10 |
| 중요 예측 변수 | 10 |
| **통계량** | | | **OOB** |
| R-제곱 | | | 84.62% |
| 루트 평균 제곱 오차(RMSE) | | | 71.0253 |
| 평균 제곱 오차(MSE) | | | 5044.5884 |
| 평균 절대 편차(MAD) | | | 46.5602 |
| 평균 절대 백분율 오차(MAPE) | | | 0.5802 |





**결과 해석**

추가적으로 랜덤포레스트 모형으로 학습과 검정을 실시한 결과 정확률은 84.62%로 성능이 올라간 것을 확인하였습니다.

상대적 변수 중요도에서도 CART와 비슷하게 hour 의 변수 중요도가 가장 높았으며 , 다른 모든 변수들 보다 약 8 배 이상의 중요도를 보이고 있습니다.

새로운 데이터로 모두 예측을 실시하였지만 데이터의 양이 너무 많아서 생략하겠습니다.