100 XP

기계 학습 모델 학습

15분

Azure Machine Learning에는 클라우드 컴퓨팅의 확장성을 활용하여 데이터에 가장 적합한 감독형 기계 학습 모델을 찾기 위해 동시에 여러 전처리 기술과 모델 학습 알고리즘을 자동으로 시도하는 자동화된 Machine Learning 기능이 포함되어 있습니다.

> **참고**
>
> Azure Machine Learning의 자동화된 Machine Learning 기능은 감독형 기계 학습 모델, 즉 학습 데이터가 알려진 레이블 값을 포함하는 모델을 지원합니다. 자동화된 Machine Learning을 사용하여 다음에 대한 모델을 학습시킬 수 있습니다.
>
> - **분류**(범주 또는 클래스 예측)
> - **회귀**(숫자 값 예측)
> - **시계열 예측**(시계열 요소를 사용한 회귀를 통해 미래 시점의 숫자 값 예측 가능)

자동화된 Machine Learning 실험 실행

Azure Machine Learning에서 실행하는 작업은 실험 이라고 합니다. 자동화된 기계 학습을 사용하여 자전거 대여를 예측하는 회귀 모델을 학습시키는 실험을 실행하려면 다음 단계를 수행합니다.

1. Azure Machine Learning Studio 에서 **자동화된 ML** 페이지(**작성자**)를 확인합니다.
2. 다음 설정을 사용하여 새 자동화된 ML 실행을 만듭니다.
    - **데이터 세트 선택**:
        - **데이터 세트**: bike-rentals
    - **실행 구성**:
        - **새 실험 이름**: mslearn-bike-rental
        - **대상 열**: rentals(예측하기 위해 모델을 학습시키는 레이블)
        - **학습 컴퓨팅 대상**: 이전에 만든 컴퓨팅 클러스터
    - **태스크 유형 및 설정**:
        - **태스크 유형**: 회귀(모델에서 숫자 값 예측)
        - **추가 구성 설정:**
            - **기본 메트릭**: **정규화된 제곱 평균 오차***(뒷부분에서 이 메트릭에 대해 자세히 소개함)*

- **최적 모델 설명**: 선택함 - 이 옵션을 선택하면 자동화된 기계 학습 기능이 최적 모델의 기능 중요도를 계산합니다. 이를 통해 예측된 레이블에 각 기능이 미치는 영향을 확인할 수 있습니다.
- **차단된 알고리즘**: ***RandomForest*** 및 ***LightGBM*** 이외의 **모든** 알고리즘를 차단합니다. 일반적으로 최대한 많이 시도하려고 하지만 이렇게 하면 시간이 오래 걸릴 수 있습니다.
- **종료 기준**:
  - **학습 작업 시간(시간)**: 0.25 - 최대 *15*분 후에 실험이 종료됩니다.
  - **메트릭 점수 임계값**: 0.08 - 모델이 정규화된 제곱 평균 오차 메트릭 점수 *0.08* 이하에 도달하는 경우 실험이 종료됩니다.
- **기능화 설정**:
  - **기능화 사용**: 선택함 - *Azure Machine Learning* 기능은 학습 전에 자동으로 기능을 전처리합니다.

3. 자동화된 ML 실행 세부 정보의 제출을 완료하면 자동으로 시작됩니다. 실행 상태가 '준비 중'에서 '실행 중'으로 변경될 때까지 기다립니다.

4. 실행 상태가 '실행 중'으로 변경되면 **모델** 탭에서 학습 알고리즘과 전처리 단계의 가능한 각 조합이 시도되고 결과 모델의 성능이 평가되는지 관측합니다. 이 페이지는 주기적으로 자동으로 새로 고쳐지지만 ↻ **새로 고침** 을 선택할 수도 있습니다. 클러스터 노드를 초기화해야 학습을 시작할 수 있으므로 모델이 표시되는 데 10분 정도 걸릴 수 있습니다.

5. 실험이 완료될 때까지 기다립니다. 시간이 걸릴 수 있으므로 지금 잠깐 쉬어가는 것이 좋을 수 있습니다.

최적 모델 검토

실험을 완료한 후 생성된 최적 모델을 검토할 수 있습니다. 이 경우에는 종료 조건을 사용하여 실험을 중지했습니다. 따라서 실험에서 발견한 "최적" 모델이 가장 적합한 모델이 아니고 이 연습에 허용되는 시간 안에 확인된 최적 모델에 불과할 수 있습니다.

1. 자동화된 기계 학습 실행의 **세부 정보** 탭에서 최적 모델 요약을 확인합니다.

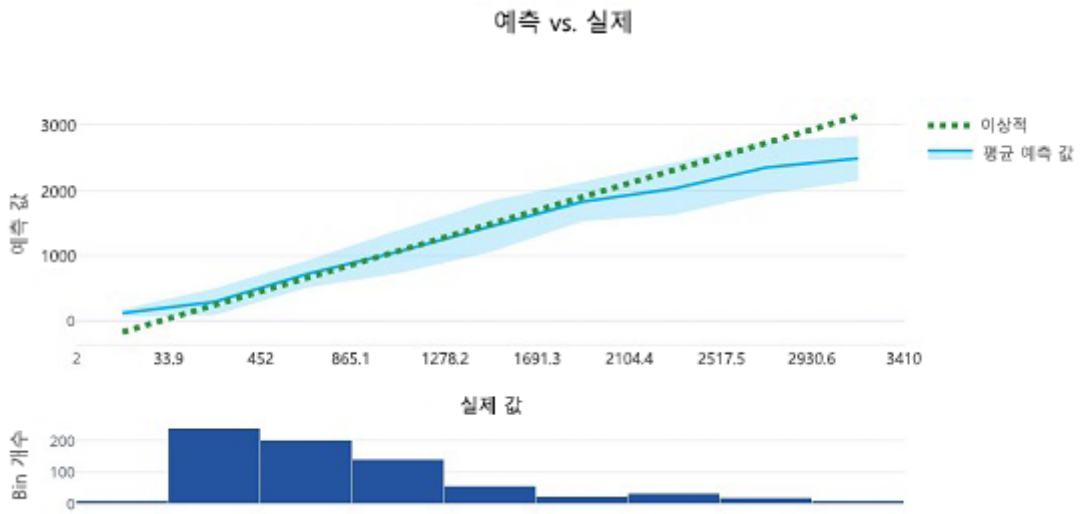
2. 최적 모델의 **알고리즘 이름** 을 선택하여 세부 정보를 확인합니다.

   최적 모델은 지정한 평가 메트릭에 따라 식별됩니다(정규화된 제곱 평균 오차). 이 메트릭을 계산하기 위해 학습 프로세스에서는 일부 데이터를 사용하여 모델을 학습시키고, 교차 유효성 검사 라는 기술을 적용하여 학습시킨 모델을 학습되지 않은 데이터로 반복적으로 테스트하고 예측 값을 실제 알려진 값과 비교했습니다. 예측 값과 실제 값 간 차이(잔차)는 모델의 오차 크기를 나타내며, 이 특정 성능 메트릭은 모든 테스트 사례에서 오차를 제곱하고, 이러한 제곱의 평균을 구한 후 제곱근을 구하는 방식으로 계산됩니다. 이 메트릭은 이 값이 작을수록 모델이 보다 정확하게 예측한다는 것을 의미합니다.

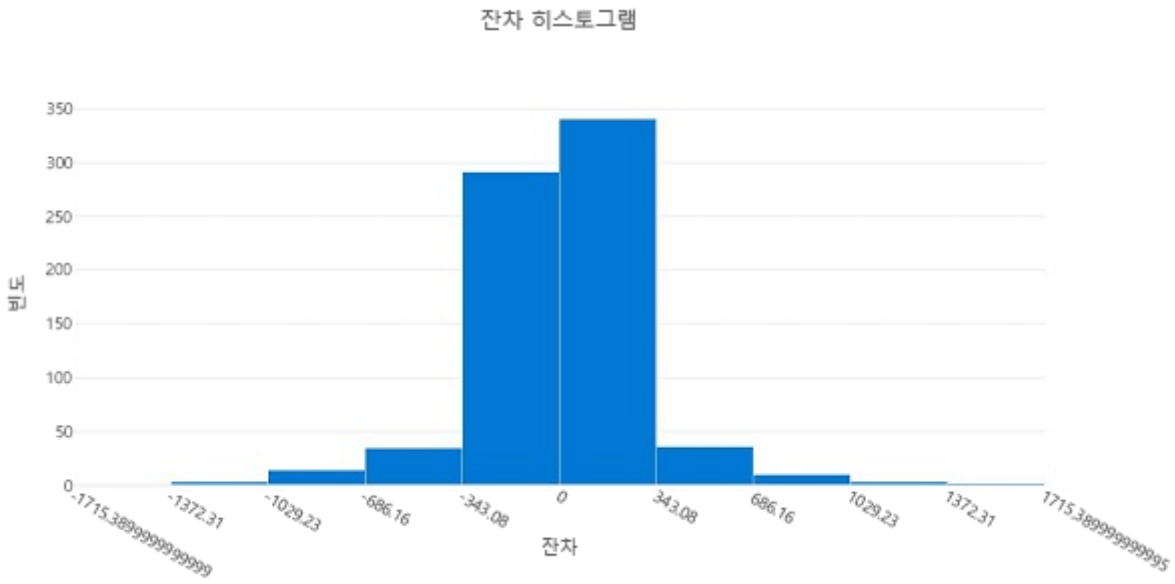
3. 정규화된 제곱 평균 오차 값 옆에 있는 **다른 모든 메트릭 보기** 를 선택하여 회귀 모델에 대한 가능한 다른 평가 메트릭 값을 확인합니다.

4. **메트릭** 탭을 선택하고 **잔차** 및 **predicted_true** 차트를 아직 선택하지 않은 경우 선택합니다. 그런 다음, 예측 값을 참 값과 비교하고 잔차(예측 값과 실제 값의 차이)를 히스토그램으로 표시하여 모델의 성능을 보여 주는 차트를 검토합니다.

   예측 값 및 참 값 차트는 예측 값이 참 값과 밀접하게 상호 연관되는 대각선 추세를 표시합니다. 점선은 완벽한 모델의 작동 성능을 보여 주며, 모델의 평균 예측 값 줄이 이 점선에 가까울수록 성능이 좋은 것입니다. 꺾은선형 차트 아래의 히스토그램은 참 값의 분포를 보여 줍니다.

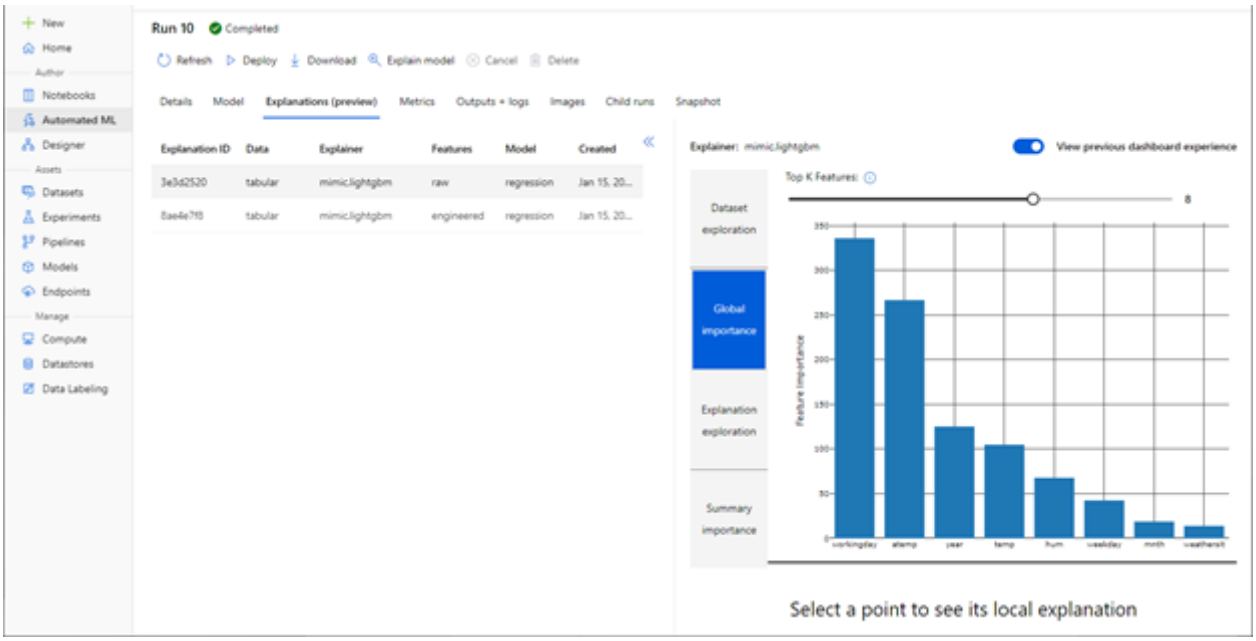


   잔차 히스토그램 은 잔차 값 범위의 빈도를 보여 줍니다. 잔차는 모델로 설명할 수 없는 예측 값과 실제 값 간의 차이를 나타냅니다. 즉, 오차를 나타냅니다. 따라서 가장 자주 발생하는 잔차 값은 0 주변에 클러스터링되고(즉, 대부분의 오차는 작음), 오차가 아주 클 가능성은 훨씬 작습니다.



5. **설명** 탭을 선택합니다. **설명 ID** 옆의 화살표 >>를 클릭하여 설명 목록을 펼칩니다. 설명 ID를 선택하고 오른쪽에서 **이전 대시보드 환경 보기** 를 선택합니다. 그런 다음 **글로벌 중**

요도 를 선택합니다. 이 차트는 다음과 같이 데이터 세트의 각 기능이 레이블 예측에 영향을 미치는 정도를 보여 줍니다.



다음 단원: 모델을 서비스로 배포

계속 >