100 XP

데이터 살펴보기

10분

회귀 모델을 학습하려면, 기존 '특징'(예측의 대상이 되는 엔터티의 특성)과 알려진 '레이블' 값(모델을 학습하여 예측할 숫자 값)이 포함된 데이터 세트가 필요합니다.

파이프라인 만들기

Azure Machine Learning 디자이너를 사용하려면 기계 학습 모델을 학습하는 데 사용할 파이프라인을 만듭니다. 해당 파이프라인은 모델을 학습할 데이터 세트로 시작합니다.

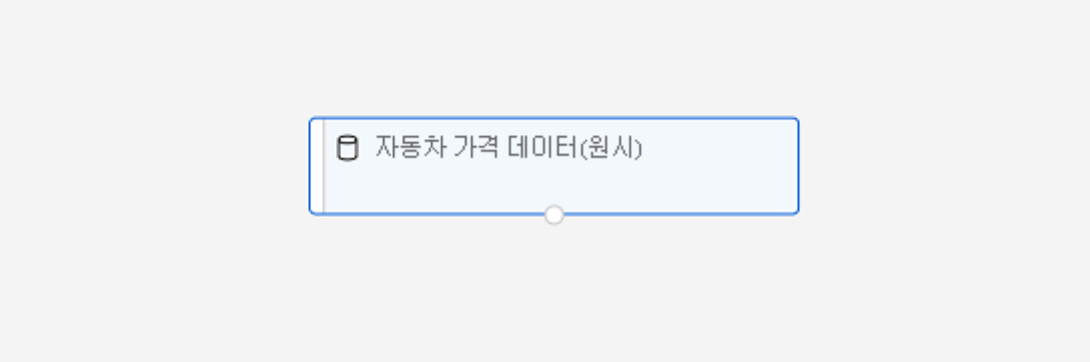
1. Azure Machine Learning Studio 에서 **작성** 아래에 있는 **디자이너** 페이지를 확인하고 + 를 선택하여 새 파이프라인을 만듭니다.
2. **설정** 창에서 기본 파이프라인 이름(*Pipeline-Created-on- date***)을 **자동차 가격 학습** 으로 변경합니다(**설정** 창이 보이지 않으면 상단 파이프라인 이름 옆에 있는 ⚙ 아이콘을 클릭).
3. 파이프라인을 실행할 컴퓨팅 대상을 지정해야 합니다. **설정** 창에서 **컴퓨팅 대상 선택** 을 사용하여 이전에 만든 컴퓨팅 클러스터를 선택합니다.

데이터 세트 추가 및 탐색

이 모듈에서는 특성에 따라 자동차 가격을 예측하는 회귀 모델을 학습합니다. Azure Machine Learning에는 해당 모델에 사용할 수 있는 샘플 데이터 세트가 포함되어 있습니다.

1. 디자이너 왼쪽에서 **샘플 데이터 세트** 섹션을 펼치고 **샘플** 섹션에서 **자동차 가격 데이터(원시)** 데이터 세트를 캔버스에 끌어다 놓습니다.
2. 캔버스에서 **자동차 가격 데이터(원시)** 데이터 세트를 마우스 오른쪽 단추로 클릭(Mac에서는 Ctrl+클릭)하고 **시각화** 메뉴에서 **데이터 세트 출력** 을 선택합니다.
3. 데이터의 스키마를 검토하여 다양한 열의 분포를 히스토그램으로 확인할 수 있습니다.
4. **가격** 열이 표시될 때까지 데이터 세트의 오른쪽으로 스크롤합니다. 해당 열이 모델에서 예측할 레이블입니다.
5. **가격** 열의 열 헤더를 선택하고 오른쪽 창에 표시되는 세부 정보를 확인합니다. 여기에는 열 값에 대한 다양한 통계와 열 값의 분포를 보여 주는 히스토그램이 포함됩니다.
6. 다시 왼쪽으로 스크롤하고 **정규화된 손실** 열 헤더를 선택합니다. 그런 다음 해당 열의 통계를 검토하여 상당 수의 누락된 값이 있는지 확인합니다. 누락된 값이 많으면 **가격** 레이블 예측에 유용하지 않으므로 학습에서 제외하는 것이 좋습니다.
7. **보어**, **스트로크** 및 **마력** 열의 통계에서 누락된 값의 개수를 확인합니다. 해당 열은 **정규화된 손실** 보다 누락된 값이 훨씬 적으므로, 값이 누락된 행을 학습에서 제외하더라도 **가격** 을 예측하는 데 여전히 유용합니다.

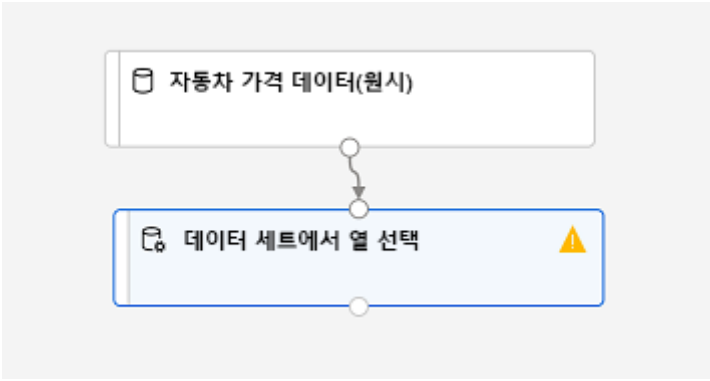
8. **스트로크**, **최고 RPM** 및 **시내 주행 연비** 열의 값을 비교합니다. 관련 값이 모두 다른 기준으로 측정되었으므로 **최고 RPM** 의 큰 값으로 인해 학습 알고리즘이 편중될 수 있으며, **스트로크** 와 같이 값이 더 작은 열보다 해당 열에 과도한 의존도가 발생할 수 있습니다. 일반적으로 데이터 과학자는 숫자 열을 비슷한 기준에 기초하도록 '정규화'하여 해당 편중을 최소화합니다.
9. 다음과 같이 캔버스에서 데이터 세트를 볼 수 있도록 **자동차 가격 데이터(원시) 결과 시각화** 창을 닫습니다.



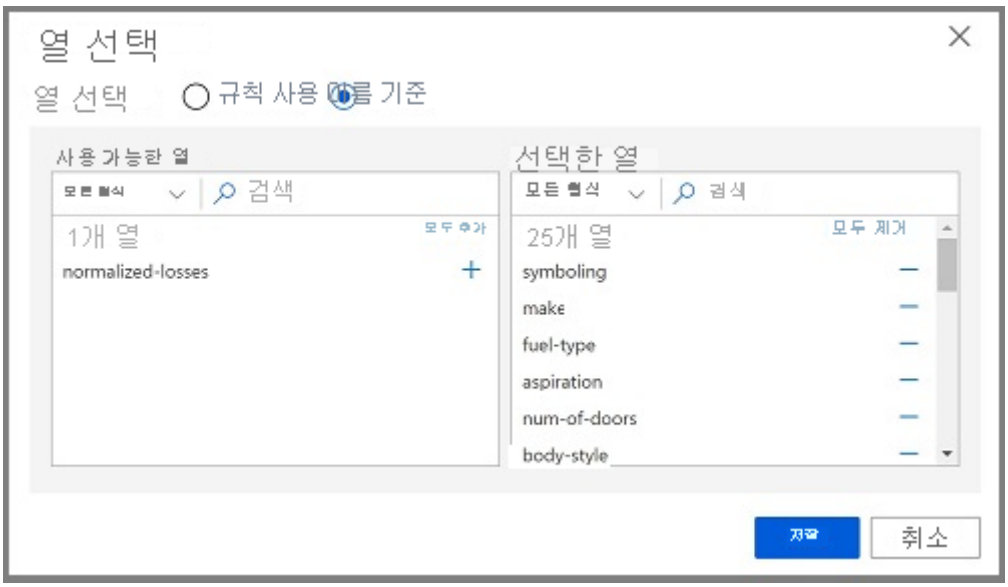
데이터 변환 추가

일반적으로 데이터 변환을 적용하여 모델링용 데이터를 준비합니다. 자동차 가격 데이터의 경우 데이터를 탐색할 때 발견된 문제를 해결하기 위해 변환을 추가합니다.

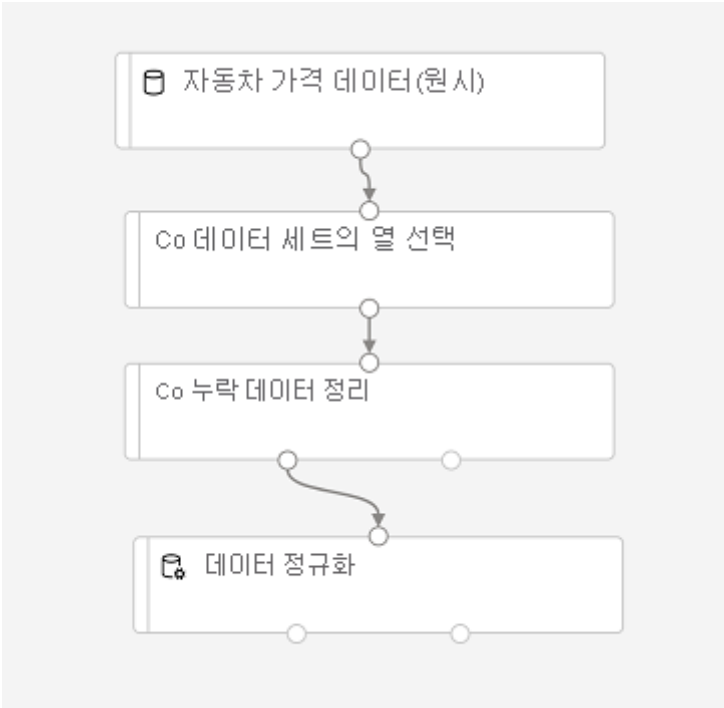
1. 왼쪽 창에서 **데이터 변환** 섹션을 펼칩니다. 여기에는 모델 학습 전에 데이터 변환에 사용할 수 있는 다양한 모듈이 포함되어 있습니다.
2. **데이터 세트에서 열 선택** 모듈을 **자동차 가격 데이터(원시)** 모듈 아래의 캔버스에 끌어다 놓습니다. 그런 다음 다음과 같이 **자동차 가격 데이터(원시)** 모듈 하단의 출력을 **데이터 세트에서 열 선택** 모듈 상단의 입력에 연결합니다.



3. **데이터 세트에서 열 선택** 모듈을 선택하고 오른쪽에 있는 **설정** 창에서 **열 편집** 을 선택합니다. 그리고 나서 다음과 같이 **열 선택** 창에서 **이름순** 을 선택하고 + 링크를 사용하여 **정규화된 손실** 을 제외한 모든 열을 추가합니다.

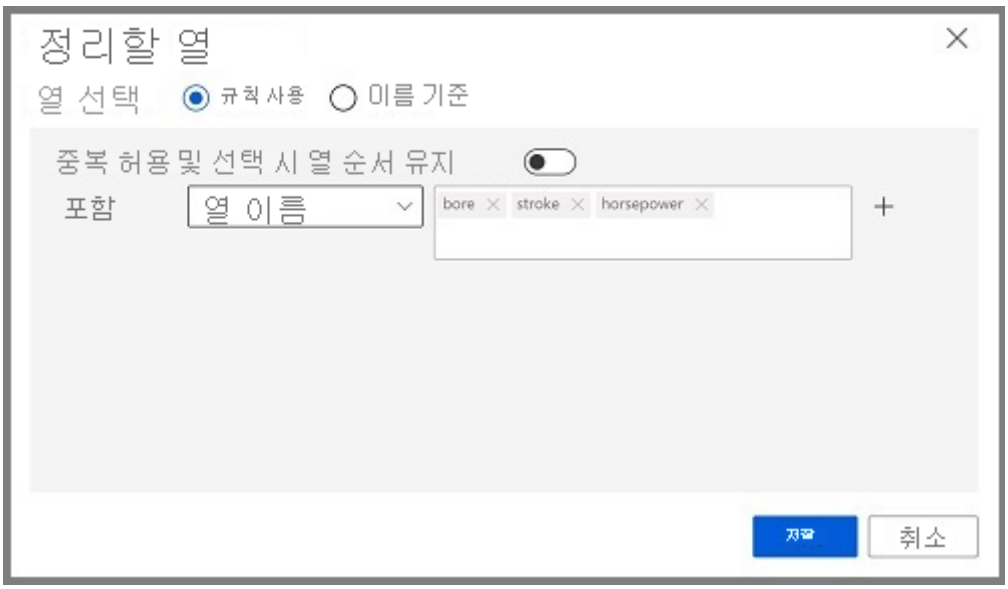


이 연습의 나머지 부분에서는 다음과 같은 파이프라인을 만듭니다.



필요한 모듈을 추가하고 구성할 때 위의 이미지를 참조로 사용하여 나머지 단계를 수행합니다.

4. **데이터 변환** 섹션에서 **누락된 데이터 정리** 모듈을 **데이터 세트에서 열 선택** 모듈 아래에 끌어다 놓습니다. 그런 다음 **데이터 세트에서 열 선택** 모듈의 출력을 **누락된 데이터 정리** 모듈의 입력에 연결합니다.
5. **누락된 데이터 정리** 모듈을 선택하고 오른쪽의 설정 창에서 **열 편집** 을 클릭합니다. 그리고 나서 다음과 같이 **열 선택** 창에서 **규칙 사용** 을 선택하고 **포함** 목록에서 **열 이름** 을 선택한 다음 열 이름 상자에 **보어**, **스트로크** 및 **마력** 을 입력합니다(철자와 대문자 표시가 정확히 일치해야 함).

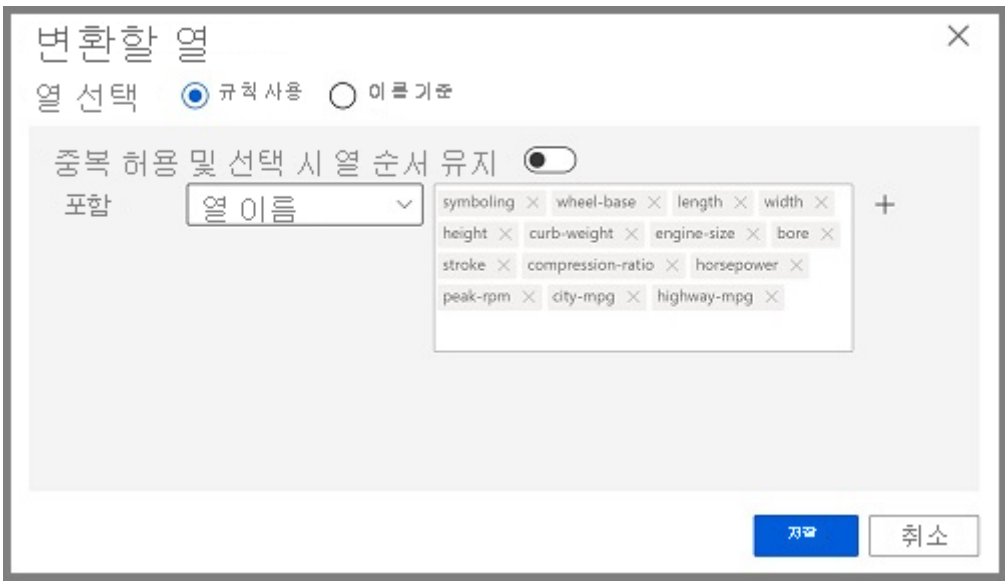


6. **누락된 데이터 정리** 모듈을 선택한 상태로 설정 창에서 다음 구성을 설정합니다.

   - **최소 누락 값 비율**: 0.0
   - **최대 누락 값 비율**: 1.0
   - **정리 모드**: 전체 행 제거

7. **데이터 정규화** 모듈을 **누락된 데이터 정리** 모듈 아래에 있는 캔버스로 끌어다 놓습니다. 그런 다음 **누락된 데이터 정리** 모듈의 가장 왼쪽에 있는 출력과 **데이터 정규화** 모듈의 입력을 연결합니다.
8. **데이터 정규화** 모듈을 선택하고 설정을 확인합니다. 여기서 변환 방법 및 변환할 열을 지정해야 한다는 것을 알 수 있습니다. 그런 다음 변환을 **MinMax** 로 설정하고 다음 **열 이름** 을 포함하도록 규칙을 적용하여 열을 편집합니다(철자, 대문자 표시 및 하이픈 넣기가 정확히 일치하는지 확인).

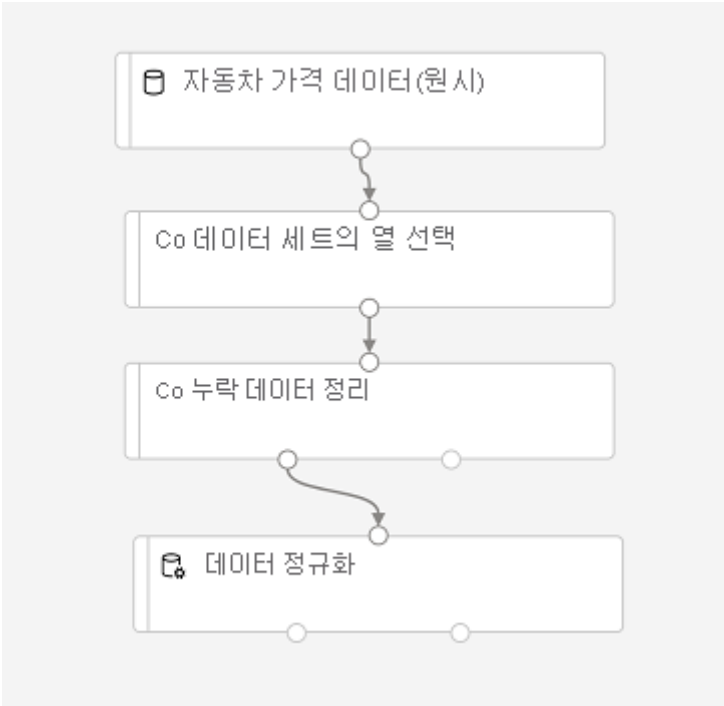
   - **symboling(수치 변환 값)**
   - **wheel-base**
   - **length**
   - **width**
   - **height(높이)**
   - **curb-weight(정비 중량)**
   - **engine-size**
   - **bore(보어)**
   - **stroke(스트로크)**
   - **compression-ratio(압축비)**
   - **horsepower(마력)**
   - **peak-rpm**
   - **city-mpg(시내 주행 연비)**
   - **highway-mpg**



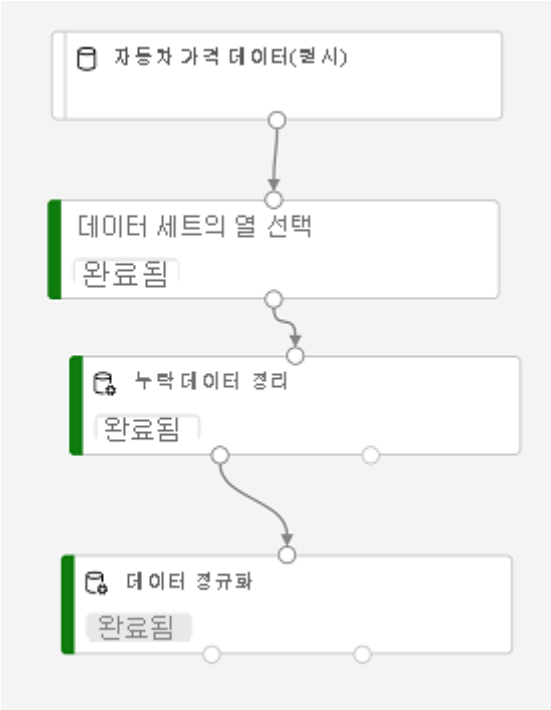
파이프라인 실행

데이터 변환을 적용하려면 파이프라인을 실험으로 실행해야 합니다.

1. 파이프라인이 다음과 유사해야 합니다.



2. **제출** 을 선택하고 컴퓨팅 클러스터에서 **mslearn-auto-training** 이라는 새 실험으로 파이프라인을 실행합니다.
3. 실행이 끝날 때까지 기다립니다. 5분 이상 걸릴 수 있습니다. 실행이 완료되면 모듈은 다음과 같습니다.



변환된 데이터 보기

이제 모델 학습을 위한 데이터 세트가 준비되었습니다.

1. 완료된 **데이터 정규화** 모듈을 선택하고 오른쪽 **설정** 창의 **출력 + 로그** 탭에서 **변환된 데이터 세트** 에 대한 **시각화** 아이콘을 선택합니다.
2. 데이터를 보고 **정규화된 손실** 열이 제거된 것을 확인하고, 모든 행에 **보어**, **스트로크** 및 **마력** 데이터가 포함되어 있으며 선택한 숫자 열이 공통 기준으로 정규화된 것을 확인합니다.
3. 정규화된 데이터의 결과 시각화를 닫습니다.

다음 단원: 학습 파이프라인 만들기 및 실행

계속 >