



# Создание первого эксперимента по обработке и анализу данных в Студии машинного обучения (классической)

Статья • 06.02.2019

Область применения:  Студия машинного обучения (классическая) 

[Машинное обучение Azure](#)

## Важно!

Поддержка Студии машинного обучения (классической) будет прекращена 31 августа 2024 г. До этой даты рекомендуется перейти на [Машинное обучение Azure](#).

Начиная с 1 декабря 2021 года вы не сможете создавать новые ресурсы Студии машинного обучения (классической). Существующие ресурсы Студии машинного обучения (классическая версия) можно будет использовать до 31 августа 2024 г.

- См. [сведения о переносе проектов машинного обучения из Студии машинного обучения \(классическая версия\) в Машинное обучение Azure](#).
- См. дополнительные сведения о [Машинном обучении Azure](#).

Прекращается поддержка документации по Студии машинного обучения (классическая версия). В будущем она может не обновляться.

С помощью этой статьи вы создадите эксперимент по машинному обучению в [Студии машинного обучения \(классической\)](#), который будет прогнозировать цену на автомобили с учетом различных переменных, например марки и технических спецификаций.

Если вы еще не знакомы с машинным обучением, в серии видеороликов [Обработка и анализ данных для начинающих](#) простым языком объясняются базовые принципы машинного обучения.

В этом руководстве для эксперимента используется рабочий процесс по умолчанию:

### 1. Создание модели

- [Получение данных](#)
- [Подготовка данных](#)
- [Определение признаков](#)

### 2. Обучение модели

- [Выбор и применение алгоритма](#)

### 3. Оценка и тестирование модели

- [Прогнозирование цен на новые автомобили](#)

## Получение данных

Для машинного обучения нам прежде всего нужны данные. Студия включает несколько примеров наборов данных (классическая модель). Также данные можно импортировать из других источников. Для этого примера мы воспользуемся примером набора данных **Automobile price data (Raw)** (Данные о ценах на автомобили (необработанные)), который доступен в вашем рабочем пространстве. В этом наборе данных содержится информация о разных автомобилях, включая сведения о производителе, модели, технических характеристиках и цене.

#### Совет

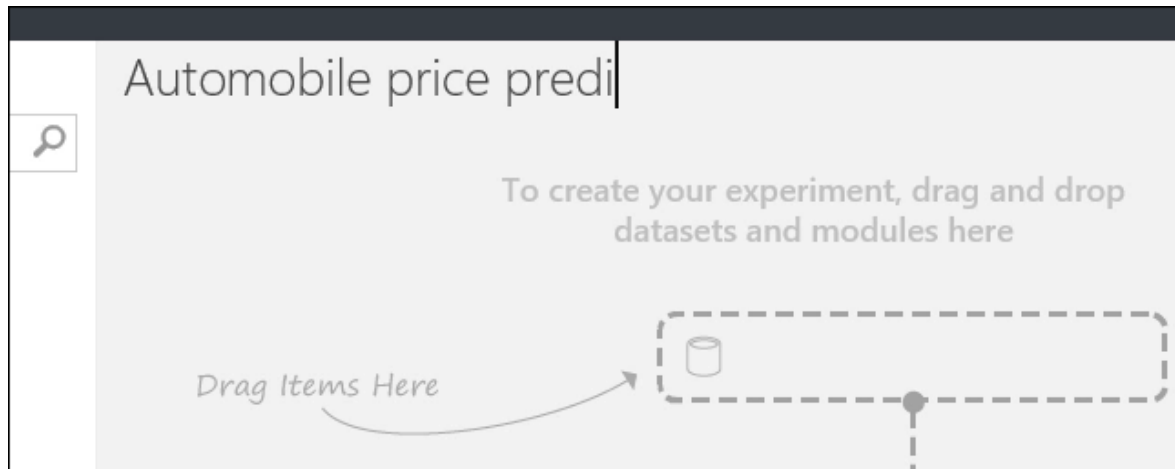
Рабочую копию этого эксперимента вы можете найти в **коллекции решений Azure AI**. Перейдите по ссылке **Ваш первый эксперимент по анализу данных. Прогнозирование цен на автомобили**, затем щелкните **Open in Studio** (Открыть в студии), чтобы загрузить в рабочее пространство классической версии Студии копию этого эксперимента.

Теперь мы узнаем, как поместить этот набор данных в эксперимент.

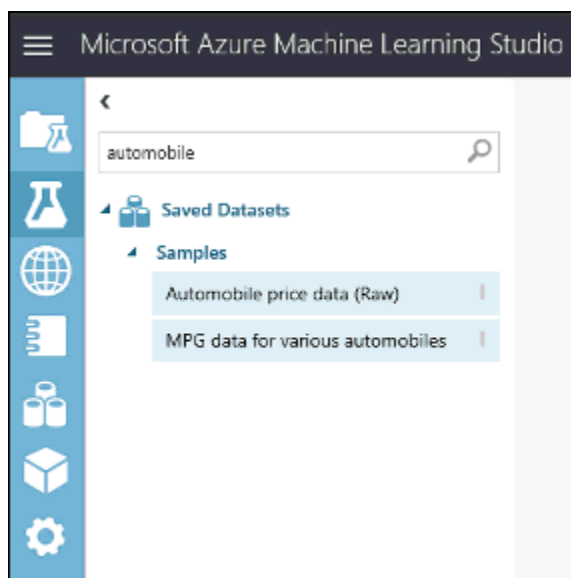
1. Создайте эксперимент, щелкнув **+NEW** (+Создать) в нижней части окна классической версии Студии машинного обучения. Выберите

**EXPERIMENT > Blank Experiment** (Эксперимент > Пустой эксперимент).

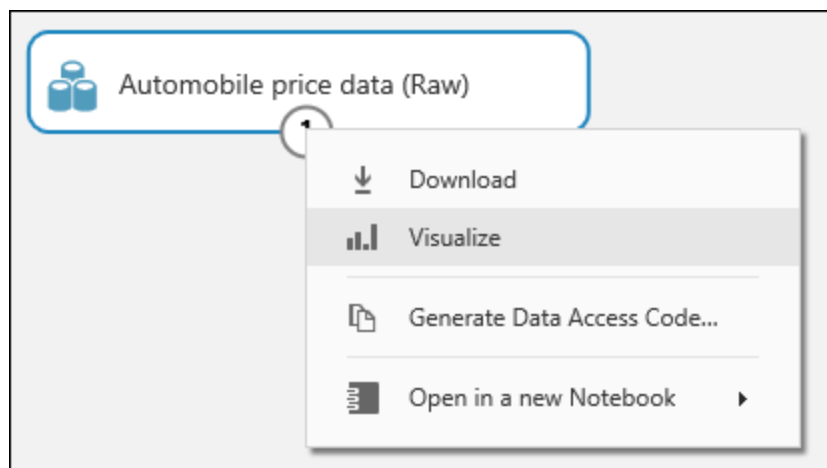
2. Эксперименту будет присвоено имя по умолчанию, которое отображается в верхней части холста. Щелкните этот текст и измените его на любое удобное для вас имя, например **Прогнозирование цен на автомобили**. Имя не обязано быть уникальным.



3. В левой части области эксперимента расположена выборка данных и модулей. Введите значение **автомобили** в поле поиска в верхней части палитры и найдите набор данных с названием **Данные о ценах на автомобили (необработанные)**. Перетащите набор данных на холст эксперимента.



Чтобы просмотреть, как выглядят эти данные, щелкните порт вывода в нижней части набора данных об автомобилях и выберите **Visualize** (Визуализировать).



### 💡 Совет

У наборов данных и модулей есть входные и выходные порты, представленные маленькими кружками. Входные порты всегда расположены вверху, а выходные — внизу. Чтобы создать поток данных через эксперимент, подключите выходной порт одного модуля ко входному порту другого. Вы можете щелкнуть выходной порт набора данных или модуля, чтобы увидеть, как выглядят данные в этой точке потока данных.

В этом наборе данных каждая строка представляет автомобиль, а переменные, обозначающие их характеристики, представлены в виде столбцов. Мы спрогнозируем цену автомобиля по представленным характеристикам и отобразим ее в крайнем правом столбце (столбец 26 с названием price (Цена)).

Automobile price prediction > Automobile price data (Raw) > dataset

rows 205 columns 26

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	engine	peak-rpm	city-mpg	highway-mpg	price
view as												
3			alfa-romero	gas	std	two	convertible		5000	21	27	13495
3			alfa-romero	gas	std	two	convertible		5000	21	27	16500
1			alfa-romero	gas	std	two	hatchback	54	5000	19	26	16500
2		164	audi	gas	std	four	sedan	102	5500	24	30	13950
2		164	audi	gas	std	four	sedan	115	5500	18	22	17450
2			audi	gas	std	two	sedan	110	5500	19	25	15250
1		158	audi	gas	std	four	sedan	110	5500	19	25	17710
1			audi	gas	std	four	sedan	110	5500	19	25	18920
1		158	audi	gas	turbo	four	sedan	140	5500	17	20	23875
0			audi	gas	turbo	two	sedan	160	5500	16	22	
2		192	bmw	gas	std	two	sedan	101	5800	23	29	16430
0		192	bmw	gas	std	four	sedan	101	5800	23	29	16925
0		188	bmw	gas	std	two	sedan	121	4250	21	28	20970
0		188	bmw	gas	std	four	sedan	121	4250	21	28	21105
1			bmw	gas	std	four	sedan	121	4250	20	25	24565

Закройте окно визуализации, нажав «x» в правом верхнем углу.

## Подготовка данных

Перед анализом, как правило, требуется определенная предварительная обработка набора данных. Вы могли заметить, что в столбцах некоторых строк отсутствуют значения. Чтобы модель смогла правильно проанализировать данные, необходимо очистить эти недостающие значения. Мы удалим все строки, в которых есть недостающие значения. В столбце **normalized-losses** (нормированные потери) также есть значительная доля недостающих значений, поэтому мы исключим весь этот столбец из модели.

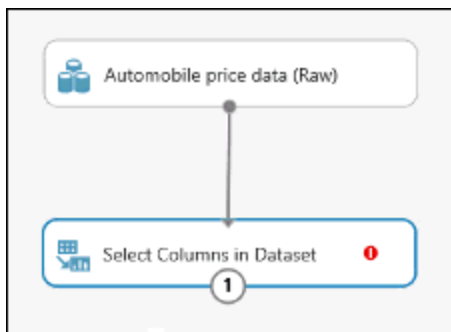
### Совет

Удаление недостающих значений из входных данных является необходимым условием для использования большинства модулей.

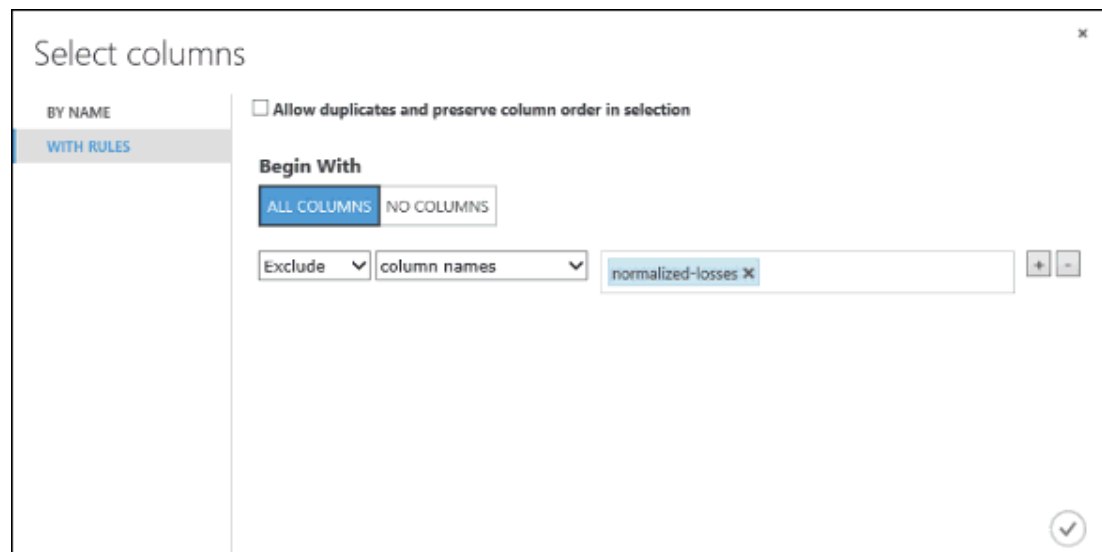
Сначала мы добавим модуль, который полностью удаляет столбец **normalized-**

**losses** (Нормированные потери). Затем мы добавим еще один модуль, который удаляет строки, в которых есть недостающие значения.

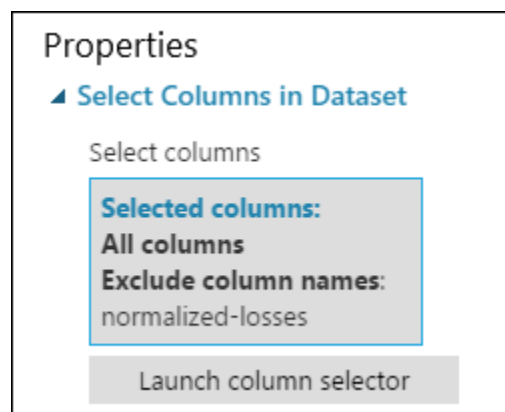
1. Введите **select column** в поле поиска в верхней части палитры модулей, чтобы найти модуль **Select Columns in Dataset** (Выбор столбцов в наборе данных). Перетащите этот модуль на холст эксперимента. Этот модуль позволяет выбрать, какие столбцы данных нужно включить в модель или исключить из нее.
2. Соедините выходной порт набора данных **Automobile price data (Raw)** (Данные о ценах на автомобили (необработанные) с входным портом модуля **Select Columns in Dataset** (Выбор столбцов в наборе данных).



3. Выберите модуль **Select Columns in Dataset** (Выбор столбцов в наборе данных) и нажмите **Launch column selector** (Запустить средство выбора столбцов) в области **Свойства**.
  - Слева щелкните **With rules** (Применяя правила)
  - В разделе **Begin With** (Начинаются с) нажмите кнопку **All columns** (Все столбцы). Такие правила укажут модулю **Select Columns in Dataset** (Выбор столбцов в наборе данных) обрабатывать все столбцы (кроме тех, которые мы собираемся исключить).
  - В раскрывающихся списках выберите **Exclude** (Исключить) и **Column names** (Имена столбцов), а затем щелкните внутри текстового поля. Отобразится список столбцов. Выберите элемент **normalized-losses** (нормированные потери), чтобы добавить его в текстовое поле.
  - Нажмите кнопку ОК (с зеленым флажком) внизу справа, чтобы закрыть средство выбора столбцов.

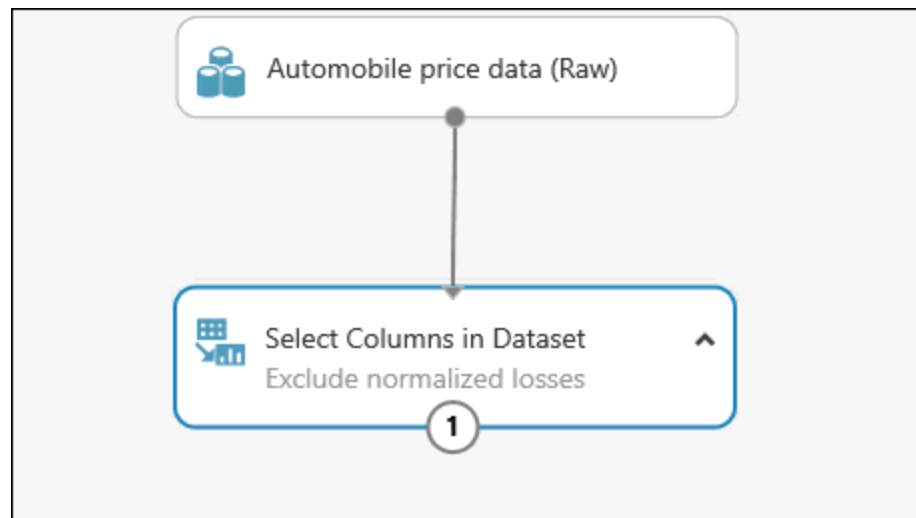


Теперь область свойств модуля **Select Columns in Dataset** (Выбор столбцов в наборе данных) показывает, что модуль пройдет через все столбцы набора данных, за исключением столбца **Normalized-losses**.



#### 💡 Совет

Дважды щелкните модуль и введите текст, чтобы добавить комментарий. Это поможет вам увидеть описание модуля и его действие в рамках эксперимента. В этом случае дважды щелкните модуль **Select Columns in Dataset** (Выбор столбцов в наборе данных) и введите комментарий об исключении столбца нормированных потерь.



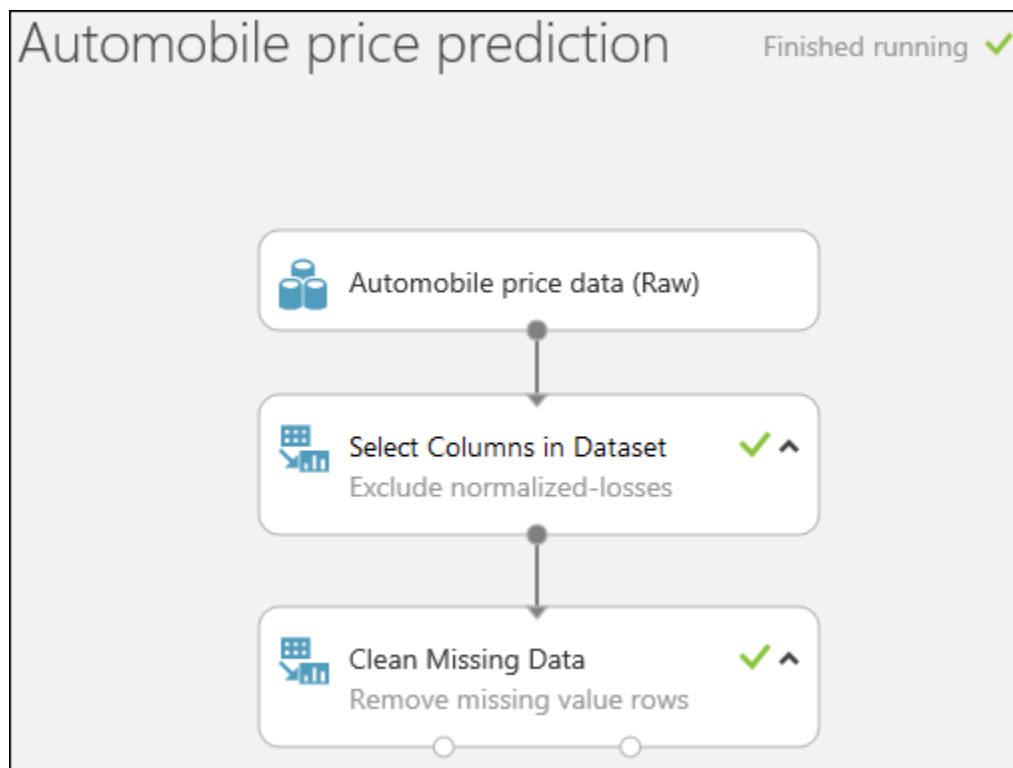
4. Перетащите на холст эксперимента модуль [Clean Missing Data](#) (Очистка недостающих данных) и присоедините его к модулю [Select Columns in Dataset](#) (Выбор столбцов в наборе данных). В панели **Properties** (Свойства) выберите значение **Remove entire row** (Удалить всю строку) для параметра **Cleaning mode** (Режим очистки). Этот параметр указывает модулю [Clean Missing Data](#) (Очистка недостающих данных) полностью удалять те строки, в которых есть недостающие значения. Дважды щелкните модуль и введите комментарий "Удаление строк с недостающими значениями".

The screenshot shows the 'Properties' panel for the 'Clean Missing Data' module. The panel has a title 'Properties' and a sub-header 'Clean Missing Data'. Under 'Columns to be cleaned', there is a 'Selected columns:' section with a dropdown menu showing 'All columns'. Below this is a 'Launch column selector' button. There are two input fields for 'Minimum missing value ratio' (set to 0) and 'Maximum missing value ratio' (set to 1). At the bottom, the 'Cleaning mode' dropdown menu is highlighted with a red rectangle, showing the selected option 'Remove entire row' with a downward arrow.

5. Запустите эксперимент, щелкнув кнопку **RUN** (Запуск) в нижней части страницы.



После завершения эксперимента у всех модулей должен появиться зеленый флажок, означающий успешное выполнение. Обратите также внимание на состояние **Работа завершена** в правом верхнем углу.



#### 💡 Совет

Для чего мы сейчас запускаем эксперимент? После запуска эксперимента все определения столбцов из нашего набора данных передаются в модуль **Select Columns in Dataset** (Выбор столбцов в наборе данных) и проходят через него к модулю **Clean Missing Data** (Очистка недостающих данных). Теперь все модули, которые мы подключим к модулю **Clean Missing Data** (Очистка недостающих данных), смогут получить эту информацию.

Мы получили очищенные данные. Чтобы просмотреть очищенный набор данных, щелкните левый выходной порт модуля **Clean Missing Data** (Очистка отсутствующих данных) и выберите **Visualize** (Визуализировать). Обратите внимание, что столбца **normalized-losses** больше нет. Кроме того, теперь нет недостающих значений.

После очистки данных мы готовы к заданию свойств, которые будут использоваться в прогнозной модели.

# Определение признаков

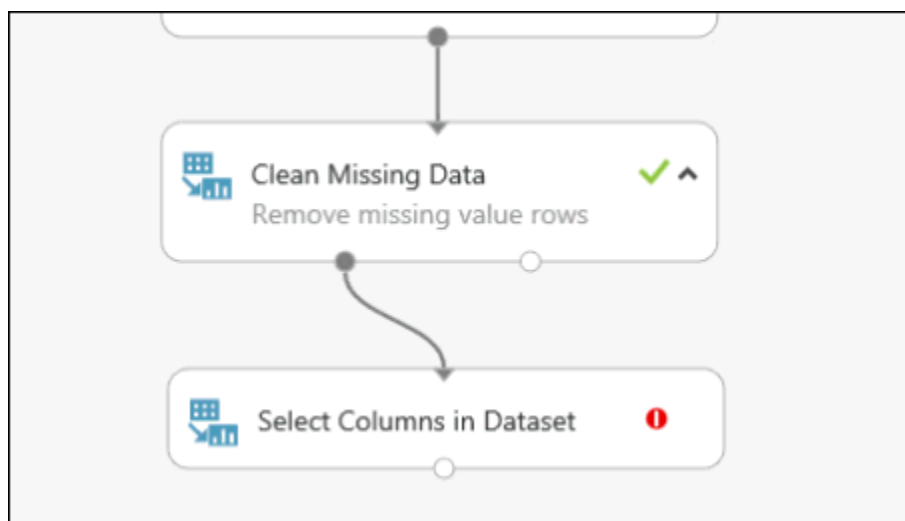
В машинном обучении *признаки* — это отдельные измеримые свойства интересующих вас объектов. В нашем наборе данных каждая строка представляет собой один автомобиль, а каждый столбец — его признак.

Чтобы подобрать подходящий набор признаков для создания прогнозной модели, нужно разбираться в проблеме, которую необходимо решить, и провести ряд экспериментов. Некоторые свойства лучше подходят для прогнозирования цели, чем другие. Некоторые признаки совпадают с другими, и часть из них можно удалить. Например, свойства `city-mpg` и `highway-mpg` действуют почти одинаково, и мы можем удалить любое из них без существенного ухудшения прогноза.

Давайте создадим модель, которая использует подмножество свойств в нашей выборке. Вы можете вернуться сюда позже и выбрать различные свойства, заново запустить эксперимент и оценить результаты. Но для начала давайте попробуем следующие возможности.

make, body-style, wheel-base, engine-size, horsepower, peak-rpm, highway-mpg, price

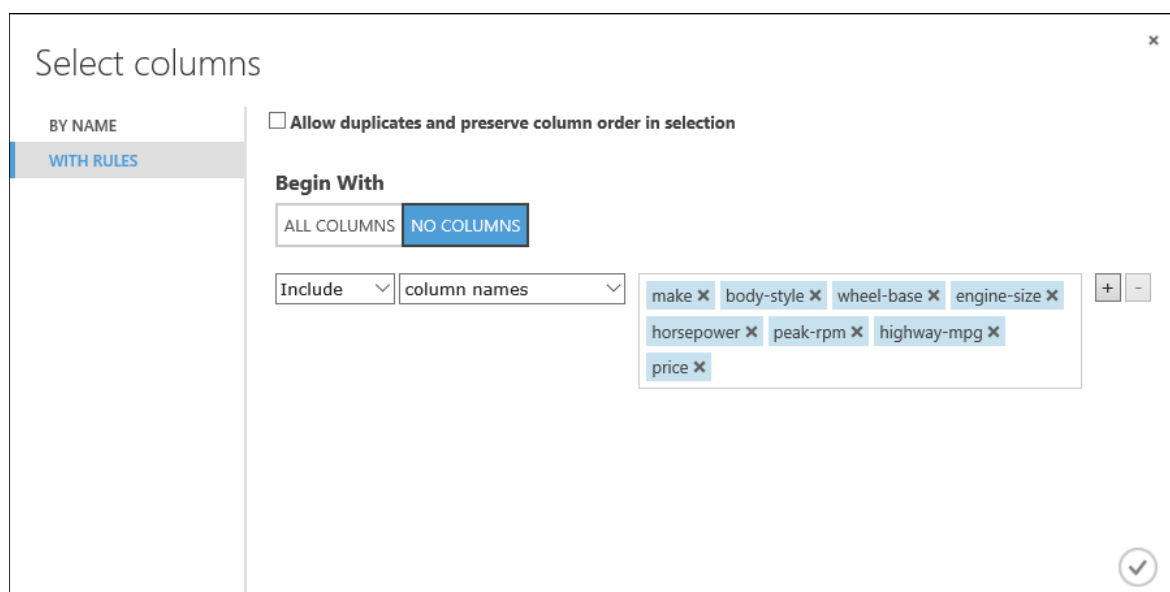
1. Перетащите на холст эксперимента еще один модуль [Select Columns in Dataset](#) (Выбор столбцов в наборе данных). Подключите левый выходной порт модуля [Clean Missing Data](#) (Очистка недостающих данных) к входу модуля [Select Columns in Dataset](#) (Выбор столбцов в наборе данных).



2. Дважды щелкните модуль и введите: "Выбор признаков для

прогнозирования".

3. В области **Свойства** щелкните **Launch column selector** (Запустить средство выбора столбцов).
4. Щелкните **With rules** (С правилами).
5. В разделе **Begin With** (Начальное состояние) нажмите кнопку **No columns** (Нет столбцов). В строке фильтра выберите **Include** (Включить) и **column names** (имена столбцов), а затем выберите в текстовом поле созданный список столбцов. Этот фильтр указывает модулю не проходить никакие столбцы (признаки), кроме выбранных.
6. Щелкните значок с изображением флажка (кнопка "OK").



В результате мы получим отфильтрованный набор данных, содержащий только те признаки, которые мы хотим передать в обучающий алгоритм на следующем шаге. Позже вы сможете вернуться назад и заново попробовать выбрать другой набор признаков.

## Выбор и применение алгоритма

Теперь, когда данные готовы, процесс построения прогнозной модели состоит из обучения и тестирования. Сначала мы используем имеющиеся данные для обучения модели, а затем протестируем ее, чтобы увидеть, насколько точно она может прогнозировать цены.

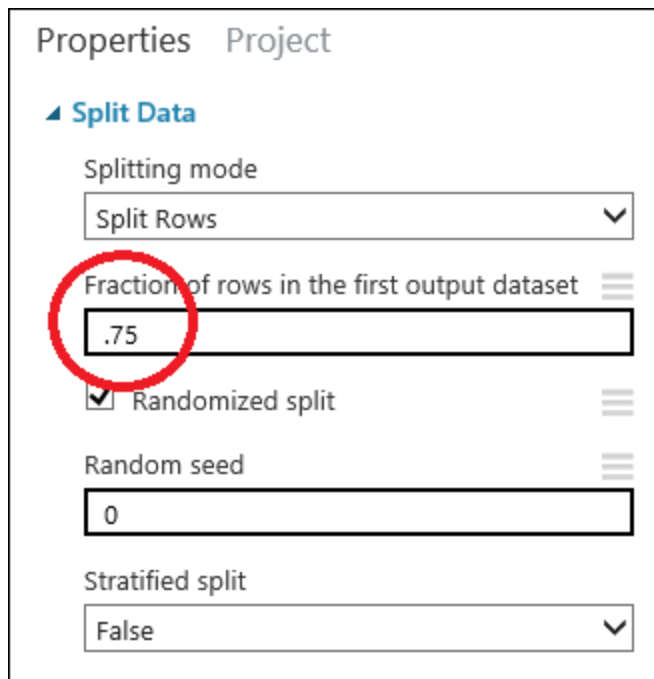
Есть два типа алгоритмов контролируемого машинного обучения — *классификация* и *регрессия*. Классификация используется, чтобы сформировать прогноз по заданному набору категорий, таких как цвет (красный, синий, или зеленый). Регрессия используется для прогнозирования числа.

Мы хотим предсказать цену, которая является числом, поэтому будем использовать модель регрессии. В нашем примере мы воспользуемся моделью *линейной регрессии*.

Для обучения модели мы передаем ей набор данных, содержащий цены. Модель сканирует эти данные и ищет зависимости между свойствами автомобиля и его ценой. После этого мы тестируем модель, то есть передаем ей набор свойств известных нам автомобилей и проверяем, насколько точно модель прогнозирует известные нам цены.

Наш набор данных мы разделим на два набора, один из которых применим для обучения модели, а второй — для тестирования.

1. Выберите и перетащите в область эксперимента модуль [Split Data](#) (Разделение данных), а затем подключите его к последнему модулю [Select Columns in Dataset](#) (Выбор столбцов в наборе данных).
2. Щелчком выберите модуль [Split Data](#) (Разделение данных). В панели **Properties** (Свойства) справа от холста найдите параметр **Fraction of rows in the first output dataset** (Доля строк в первом выходном наборе данных) и установите для него значение 0,75. Таким образом, мы используем 75 процентов данных для обучения модели и оставим 25 процентов для тестирования.



Properties Project

**Split Data**

Splitting mode  
Split Rows

Fraction of rows in the first output dataset  
.75

☒ Randomized split

Random seed  
0

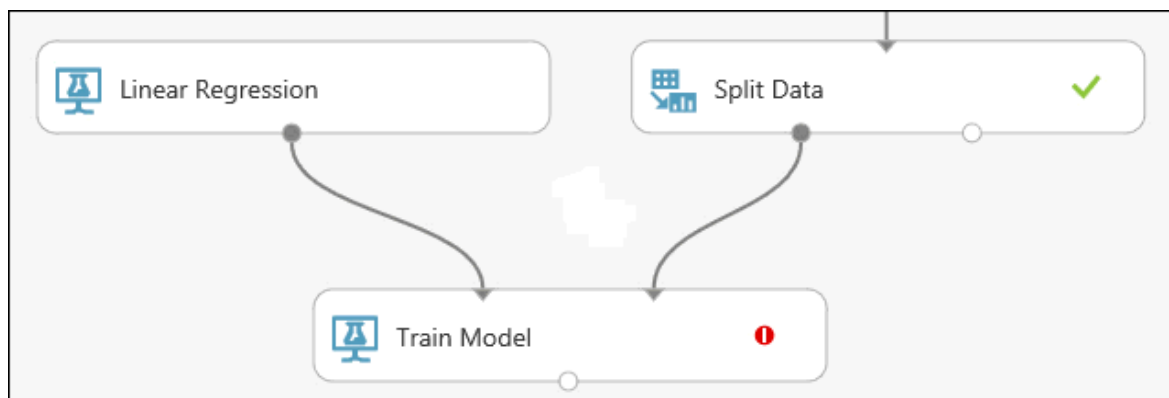
Stratified split  
False

#### 💡 Совет

Изменяя параметр **Псевдослучайные числа**, вы можете создавать различные случайные выборки для обучения и тестирования. Этот параметр задает начальное значение для генератора псевдослучайных чисел.

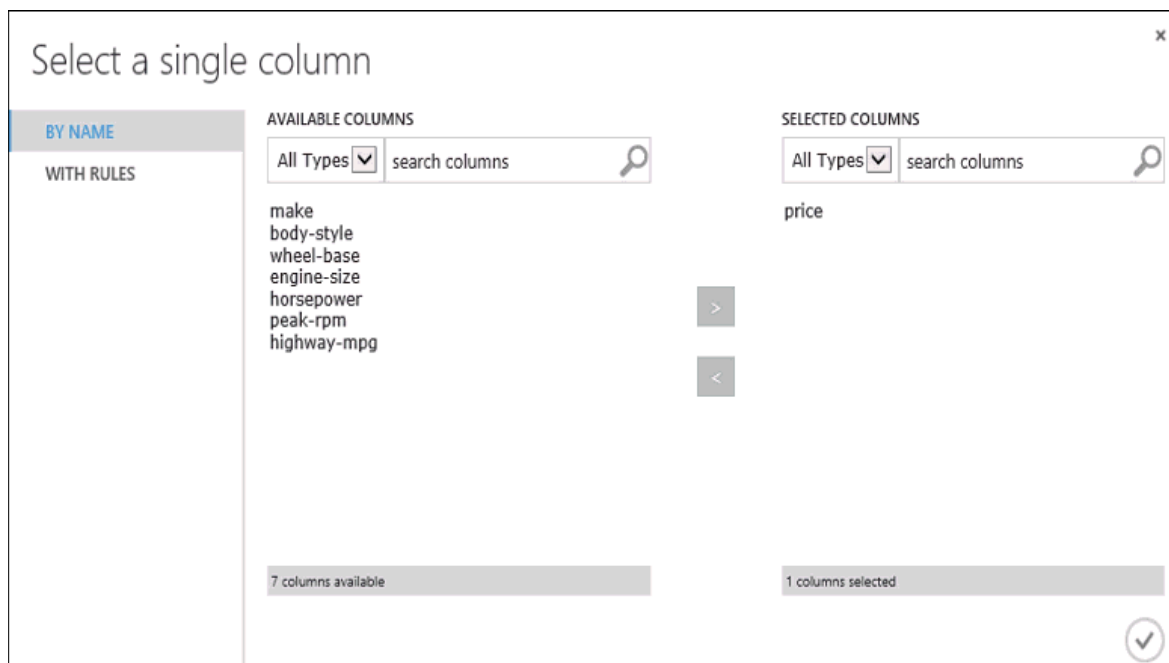
3. Запустите эксперимент. После выполнения эксперимента модули [Select Columns in Dataset](#) (Выбор столбцов в наборе данных) и [Split Data](#) (Разделение данных) смогут передать определения столбцов следующим модулям, которые мы добавим позже.
4. Для выбора алгоритма обучения разверните категорию **Машинное обучение** на палитре модулей слева на холсте эксперимента, а затем разверните **Initialize Model** (Инициализировать модель). Появится несколько категорий модулей, которые можно использовать для инициализации алгоритма обучения. Для этого эксперимента выберите модуль [Linear Regression](#) (Линейная регрессия) из категории **Regression** (Регрессия) и перетащите его на холст. (Можно также найти модуль, введя "linear regression" в поле поиска палитры.)
5. Найдите и переместите модуль [Train Model](#) (Обучение модели) на холст эксперимента. Соедините выход модуля [Linear Regression](#) (Линейная

регрессия) с левым входом модуля **Train Model** (Обучение модели), а затем соедините выход (левый порт) модуля **Split Data** (Разделение данных) с правым входом модуля **Train Model** (Обучение модели).



6. Выберите модуль **Train Model** (Обучение модели), щелкните **Launch column selector** (Запустить средство выбора столбцов) в области **Properties** (Свойства) и выберите столбец **price** (цена). **Price** (Цена) — это значение, которое спрогнозирует наша модель.

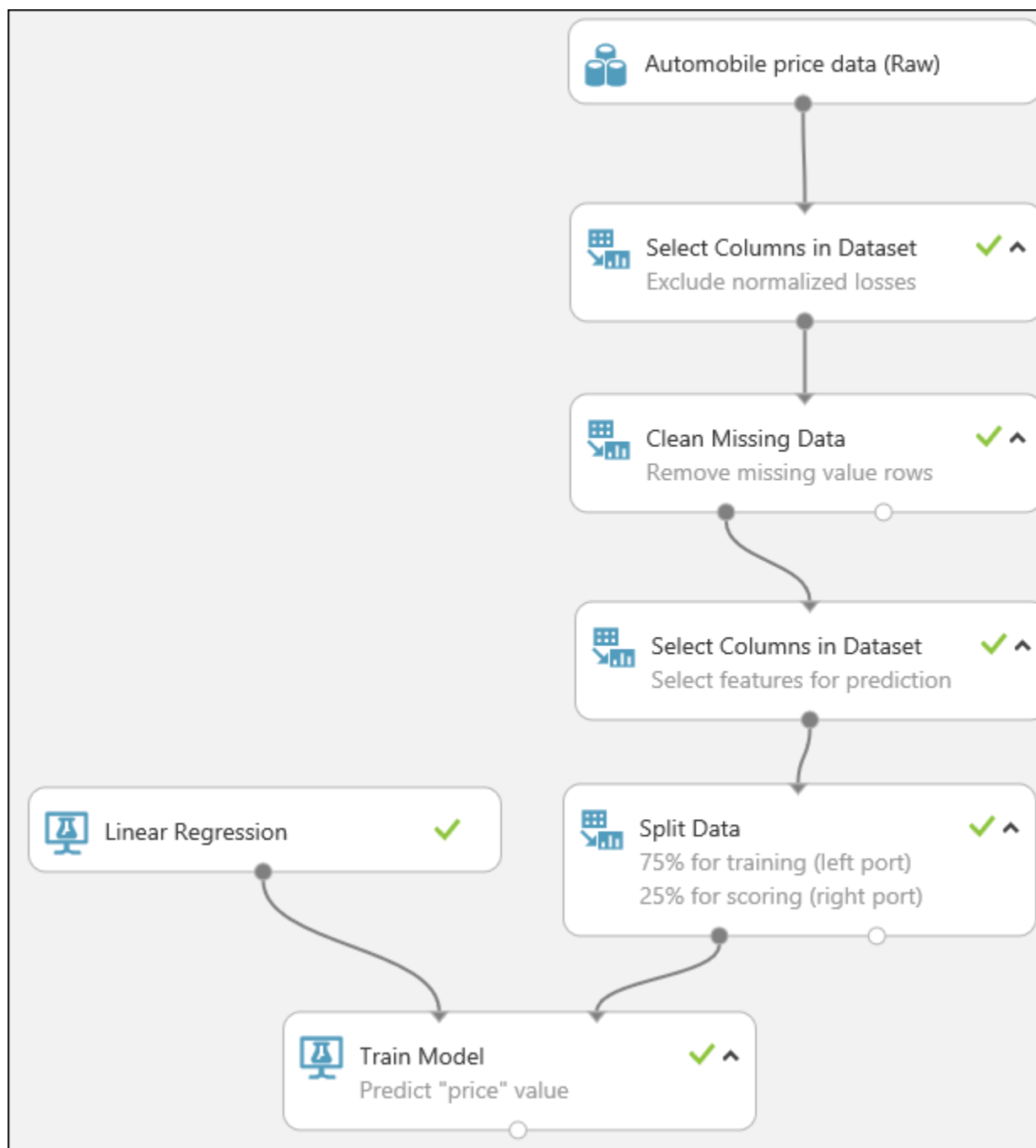
Чтобы выбрать столбец **price** (цена) в средстве выбора столбцов, переместите его из списка **Available columns** (Доступные столбцы) в список **Selected columns** (Выбранные столбцы).



7. Запустите эксперимент.

Теперь у нас есть обученная регрессионная модель, которую можно использовать

для оценки новых данных об автомобилях с целью прогнозирования цен.

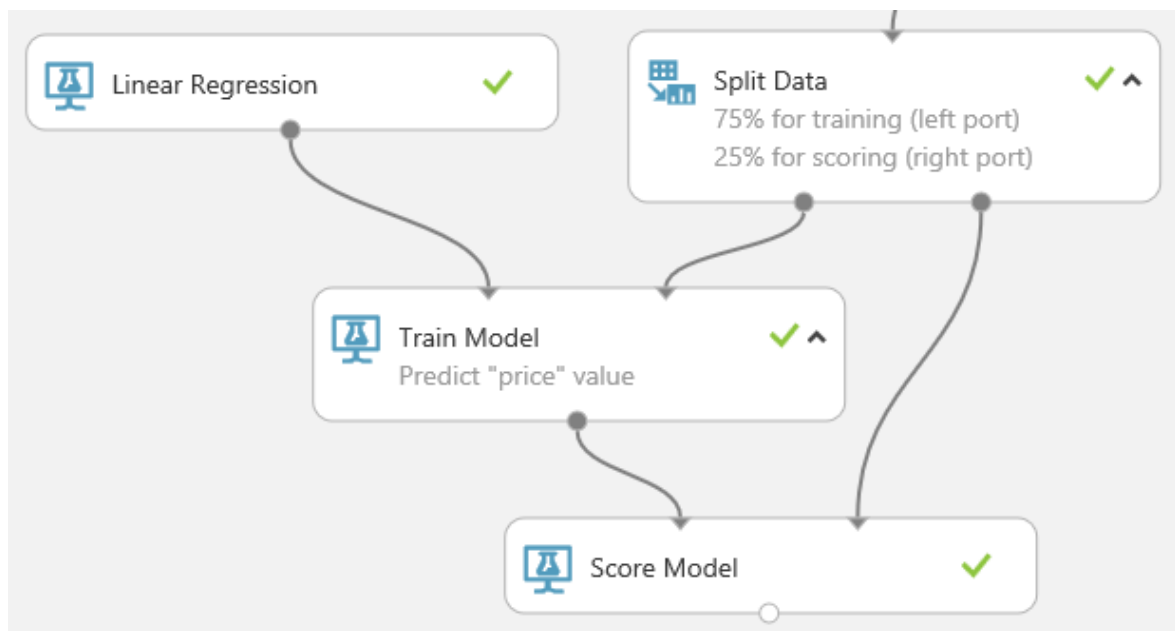


## Прогнозирование цен на новые автомобили

После того как мы обучили модель, используя 75 процентов наших данных, ее можно использовать для оценки оставшихся 25 процентов данных, чтобы проверить, насколько хорошо она работает.

1. Найдите модуль [Score Model](#) (Оценка модели) и перетащите его на холст эксперимента. Соедините выход модуля [Train Model](#) (Обучение модели) с

левым входным портом модуля [Score Model](#) (Оценка модели). Соедините выход тестовых данных (правый порт) модуля [Split Data](#) (Разделение данных) с правым входным портом [Score Model](#) (Оценка модели).



2. Запустите эксперимент и проверьте выходные данные модуля [Score Model](#) (Оценка модели), щелкнув порт вывода модуля [Score Model](#) (Оценка модели) и выбрав **Visualize** (Визуализировать). На порту вывода будут показаны прогнозируемые значения цены вместе с известными значениями проверочных данных.

Simple test experiment - Copy > Score Model > Scored dataset

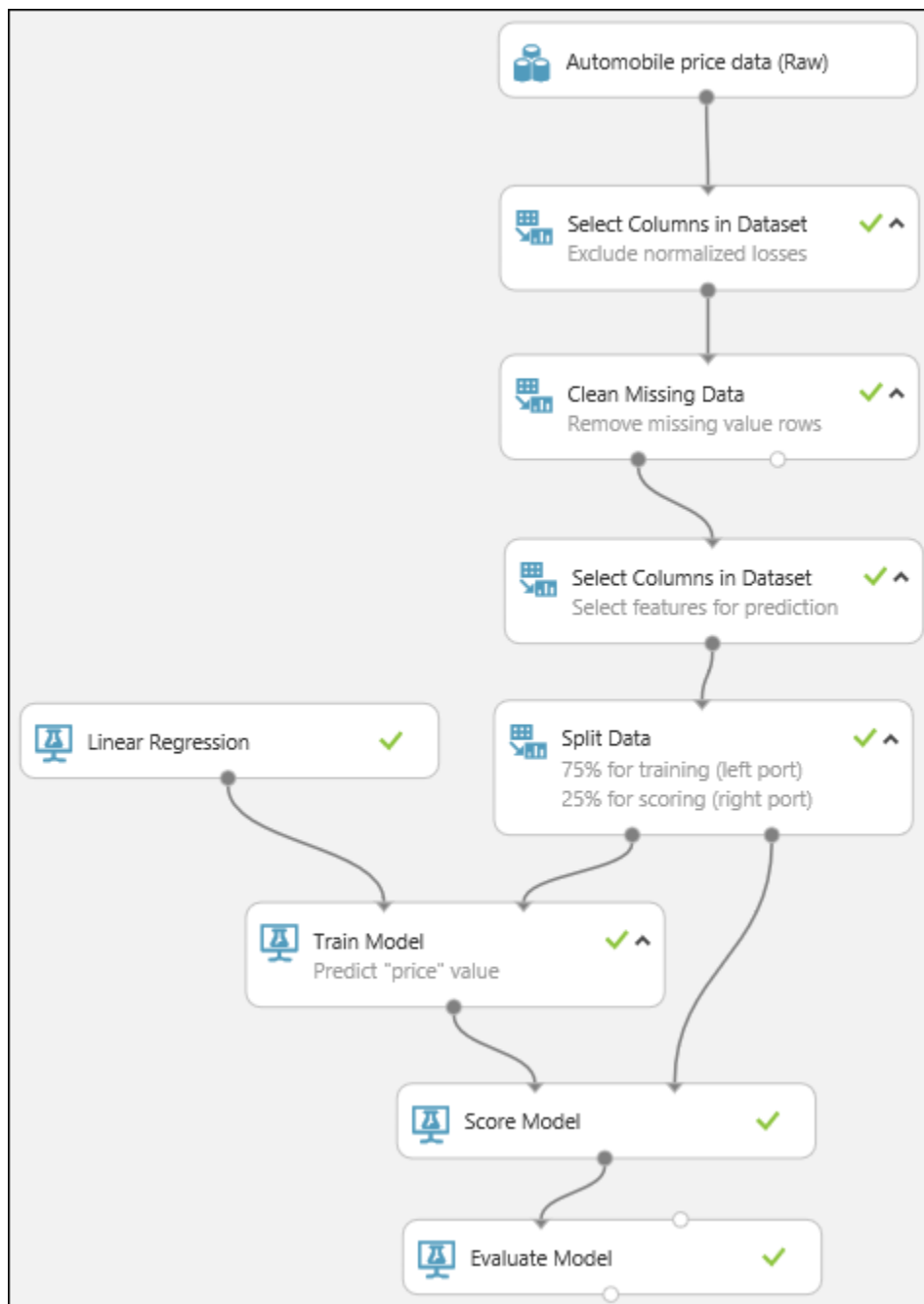
rows	columns								
48	9								
		make	body-style	wheel-base	engine-size	horsepower	peak-rpm	highway-mpg	price
view as									Scored Labels
		subaru	sedan	97	108	111	4800	29	11259
		mitsubishi	hatchback	93.7	92	68	5500	38	6669
		dodge	hatchback	93.7	90	68	5500	38	6229
		honda	hatchback	86.6	92	76	6000	38	6855
		alfa-romero	convertible	88.6	130	111	5000	27	16500
		volvo	wagon	104.3	141	114	5400	28	16515
		isuzu	hatchback	96	119	90	5000	29	11048
		dodge	hatchback	93.7	90	68	5500	41	5572
		honda	sedan	101.2	108	101	5800	29	16430

Known values  
Predicted values

3. Теперь мы готовы проверить качество результатов. Выберите модуль [Evaluate Model](#) (Анализ модели) и перетащите его на холст эксперимента, а затем соедините выход модуля [Score Model](#) (Оценка модели) с левым входом

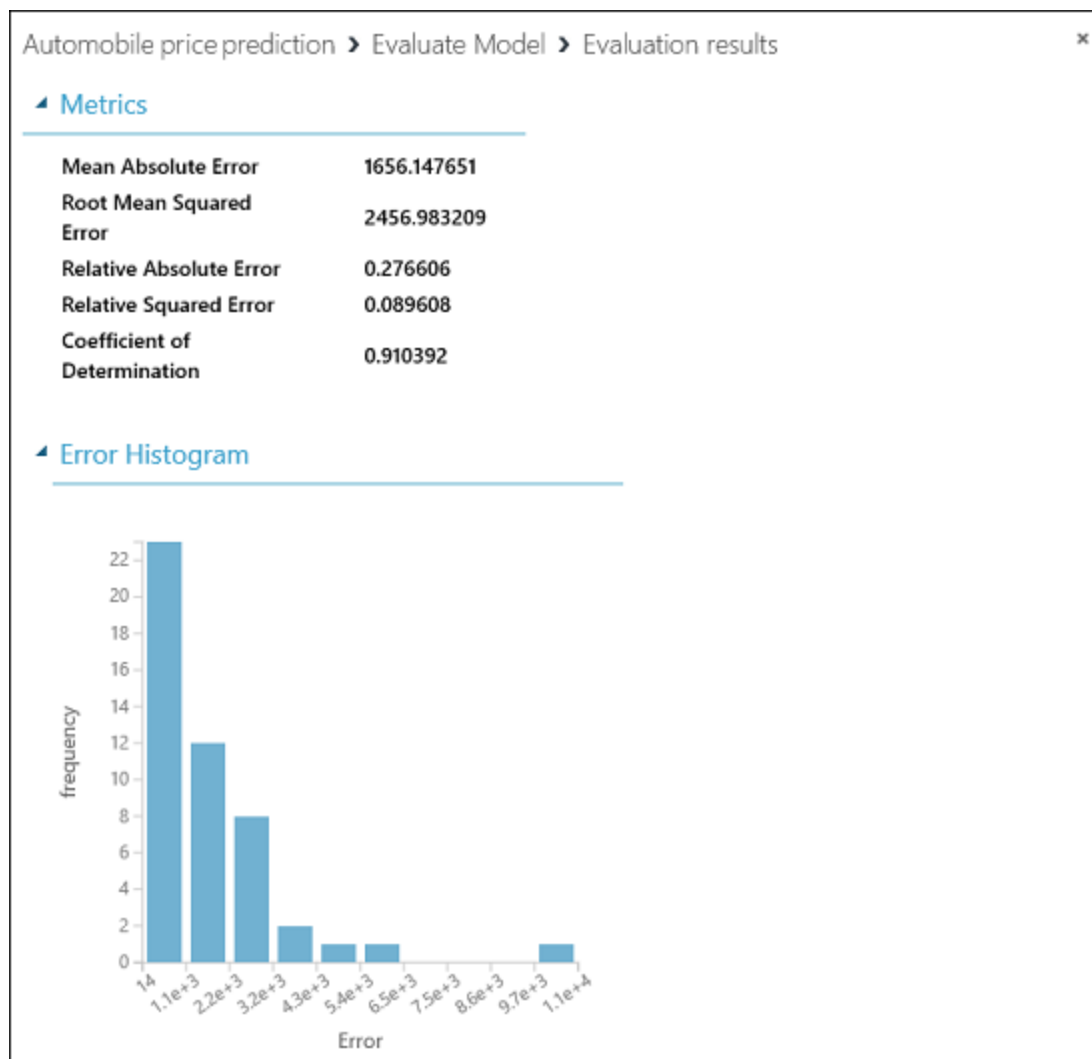


модуля [Evaluate Model](#) (Анализ модели). Окончательная схема нашего эксперимента должна выглядеть следующим образом:



#### 4. Запустите эксперимент.

Чтобы проверить выходные данные модуля [Evaluate Model](#) (Анализ модели), щелкните порт вывода и выберите элемент **Visualize** (Визуализировать).



Для нашей модели будет выведена следующая статистика.

- **Средняя абсолютная погрешность.** Среднее значение абсолютной погрешности (*погрешность* — это разница между спрогнозированным и фактическим значением).
- **Среднеквадратичное отклонение.** Квадратный корень из среднего значения возведенных в квадрат арифметических отклонений спрогнозированных значений тестового набора данных.
- **Относительное арифметическое отклонение.** Среднее арифметическое отклонение по отношению к абсолютной разнице между фактическими значениями и средним арифметическим всех фактических значений.
- **Относительное среднеквадратичное отклонение.** Среднее арифметическое среднеквадратичных отклонений по отношению к абсолютной разнице между фактическими значениями и средним арифметическим всех фактических значений.
- **Коэффициент смешанной корреляции (R в квадрате).** Статистический

показатель, который оценивает соответствие модели данным.

Чем меньше значение каждой погрешности, тем лучше. Меньшее значение указывает на то, что прогноз лучше соответствует фактическим значениям. Для показателя **коэффициент смешанной корреляции** чем ближе его значение к единице (1,0), тем точнее прогноз.

## Очистка ресурсов

Если вы не планируете дальше использовать ресурсы, созданные при работе с этой статьей, удалите их, чтобы плата не взималась. О том, как это сделать, см. в статье [Экспорт и удаление встроенных в продукт данных пользователей из Студии машинного обучения Azure](#).

## Дальнейшие действия

В этом кратком руководстве описано, как создать простой эксперимент с использованием примера набора данных. Чтобы узнать больше о создании и развертывании модели, перейдите к руководству по прогнозному решению.

**Руководство. Разработка прогнозного решения в Студии (классической)**