

Manuel de formation

Février 2017

Mis à jour Mai 2019

Préparé par : Hafed Benteftifa

© Hafed Benteftifa 2016-2021

Ce document ne peut être utilisé dans le cadre d'une formation, publication papier, site internet ou tout support sans mon accord express.

Aucune reproduction, même partielle, ne peut être faite de ce document et de l'ensemble de son contenu : textes, images, etc. sans mon autorisation express. Pour toutes informations, communiquer avec moi à info@degenio.com.

Cahier thématique 7

Machine Learning sur Azure ML Studio

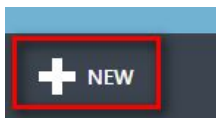
Manipulation 1: Création d'un modèle

Manipulation 1: Création d'un modèle

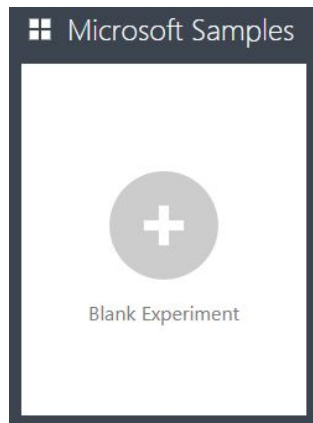
Étape 1 : obtention des données

Tout d'abord, vous devez obtenir les données. Vous pouvez utiliser plusieurs exemples de jeux de données inclus dans Machine Learning Studio ou importer des données à partir de sources diverses. Pour les besoins de cet exemple, nous allons utiliser le jeu de données Données sur le prix des véhicules automobiles (brutes) inclus dans votre espace de travail. Ce jeu de données comprend des entrées relatives à plusieurs véhicules, notamment des informations sur la marque, le modèle, les caractéristiques techniques et le prix.

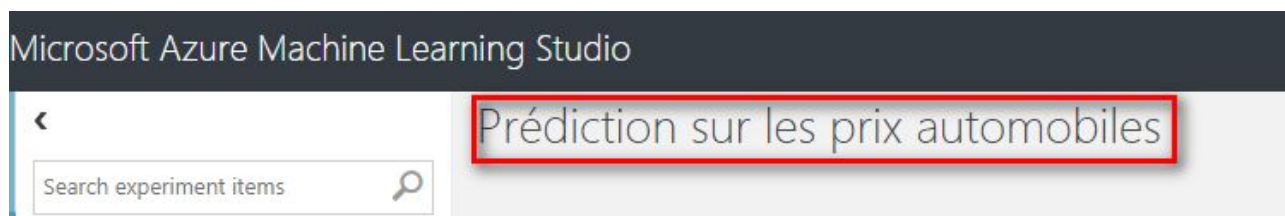
1. Créez une expérience en cliquant sur l'option **+NOUVEAU** située en bas de la fenêtre de Machine Learning Studio, sélectionnez **EXPÉRIENCE**



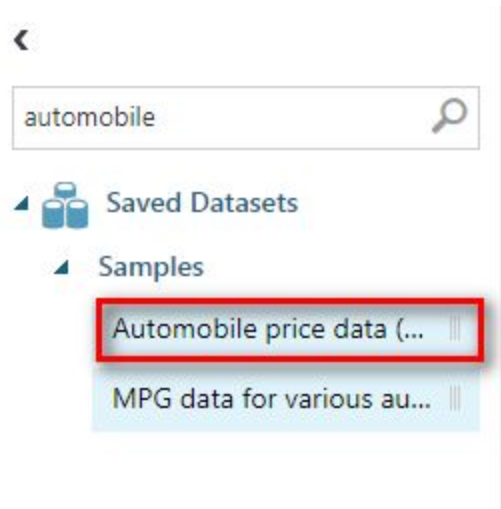
2. Sélectionner -Expérience vide-



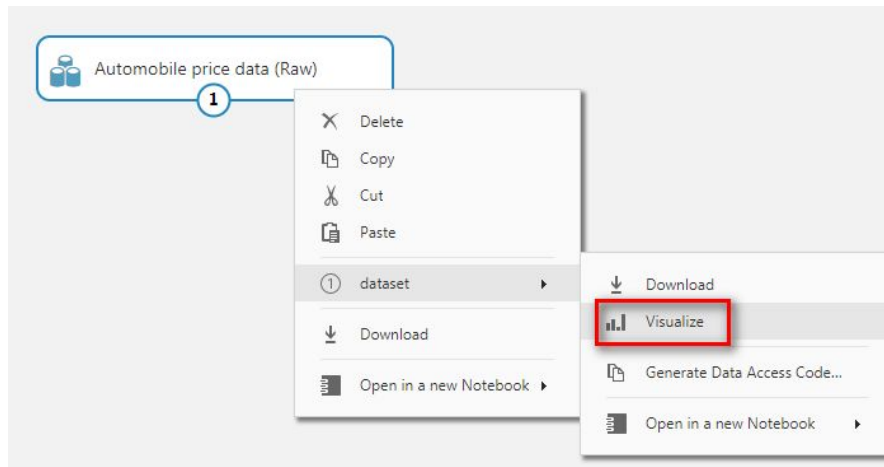
Un nom par défaut est attribué à l'expérience : il apparaît en haut du canevas. Sélectionnez le texte et remplacez-le par un nom plus significatif, par exemple Prédiction sur les prix automobiles. Le nom n'a pas besoin d'être unique.



Sur la gauche de la zone de dessin de l'expérience se trouve une palette de jeux de données et de modules. Renseigner la valeur automobile dans la zone de recherche qui se trouve en haut de cette palette, afin de rechercher le jeu de données sur le prix des véhicules automobiles (brutes). Faites glisser ce jeu de données vers le canevas de l'expérience.





Pour voir à quoi ressemblent ces données, cliquez sur le port de sortie situé en bas du jeu de données d'automobile, puis sélectionnez Visualiser.














3. Dans cet exemple de jeu de données, chaque instance de véhicule automobile apparaît sous la forme d'une ligne, et les variables associées à chaque véhicule apparaissent dans des colonnes. Compte tenu des variables associées à un véhicule automobile spécifique, nous allons essayer de prédire le prix dans la colonne de droite (colonne 26, intitulée « price »).

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

rows: 205
columns: 26

view as:  

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	engine-size	peak-rpm	city-mpg	highway-mpg	price
												
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	convertible	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

Fermez la fenêtre de visualisation en cliquant sur le symbole «x» dans le coin supérieur droit.

Étape 2 : préparation des données

Pour pouvoir être analysé, un jeu de données nécessite généralement un traitement préalable. Vous avez peut-être remarqué l'absence de certaines valeurs dans les colonnes des différentes lignes. Pour que vous puissiez analyser les données correctement, ces valeurs manquantes doivent être nettoyées.

Dans le cas qui nous concerne, nous allons supprimer toutes les lignes dans lesquelles des valeurs sont manquantes. De plus, la colonne « normalized-losses » contient une grande quantité de valeurs manquantes. Nous allons donc l'exclure complètement du modèle.

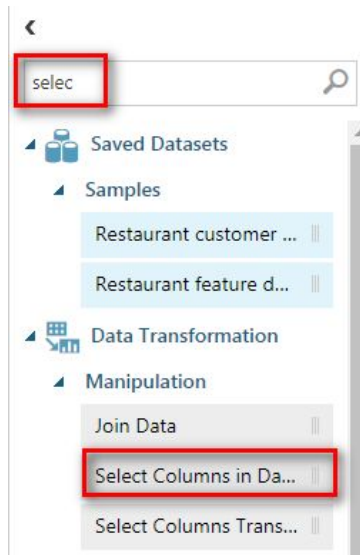
Dans un cas de développement industriel, on devrait évaluer la pertinence de garder cette colonne et faire une opération de remplacement.

Nous commençons par ajouter un module qui supprime la colonne normalized-losses, puis nous ajoutons un module qui supprime toute ligne dans laquelle des données manquent.

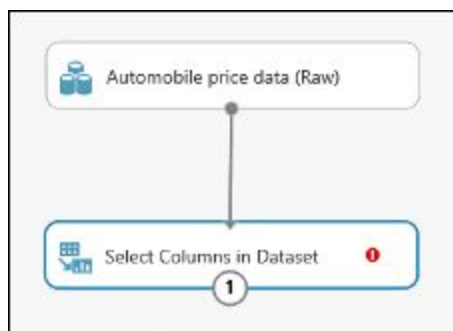
4. Dans la zone de recherche située sur la partie supérieure de la palette de modules, saisissez

la chaîne sélectionner des colonnes afin de rechercher le module Sélectionner des colonnes dans le jeu de données, puis faites glisser ce module vers le canevas de l'expérience.

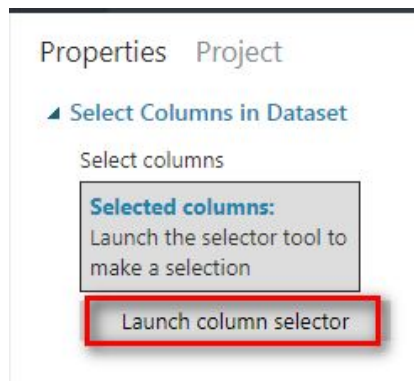
5. Ce module permet de sélectionner les colonnes de données à inclure ou exclure du modèle. Faites un glisser-déposer sur le canevas.



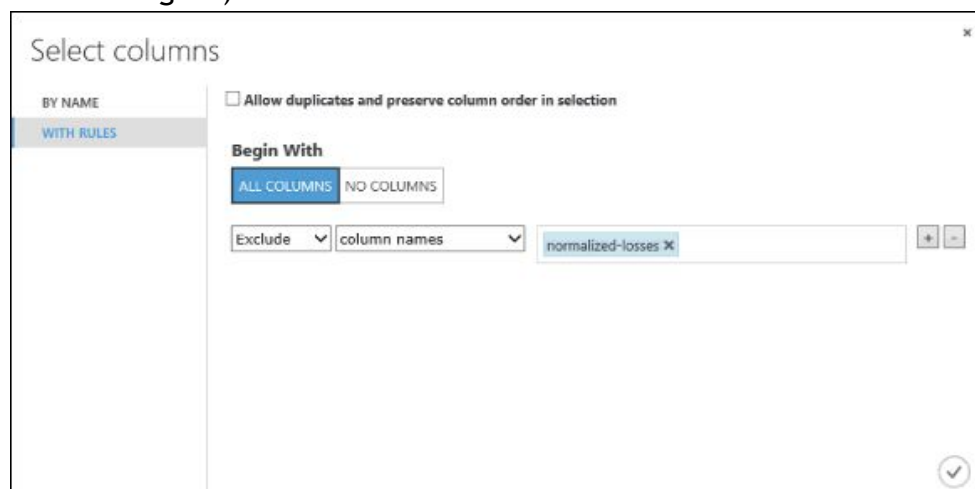
6. Connectez le port de sortie du jeu de données Données sur le prix des véhicules automobiles (brutes) au port d'entrée du module Sélectionner des colonnes dans le jeu de données comme le montre la figure suivante.



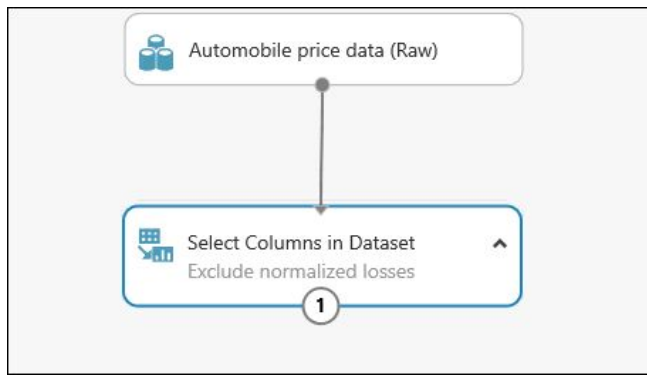
7. Cliquez sur le module Sélectionner des colonnes dans le jeu de données, puis cliquez sur Lancer le sélecteur de colonne dans le volet Propriétés.



- Sur la gauche, cliquez sur **With rules**
- Sous **Commencer par**, cliquez sur **Toutes les colonnes**. Vous indiquez ainsi au module [Sélectionner des colonnes dans le jeu de données](#) de transmettre toutes les colonnes, sauf celles que vous avez à exclure.
- Dans les listes déroulantes, sélectionnez **Exclure** et **Noms des colonnes**, puis cliquez dans la zone de texte. Une liste de colonnes s'affiche. Sélectionnez la colonne **normalized-losses**, qui est alors ajoutée à la zone de texte.
- Cliquez sur le bouton en forme de coche (OK) pour fermer le sélecteur de colonne (en bas à droite de la figure).



8. À présent, le volet de propriétés du module **Sélectionner des colonnes dans le jeu de données** indique qu'il transmettra toutes les colonnes du jeu de données, à l'exception de **normalized-losses**.



9. Faites glisser le module [Nettoyer les données manquantes](#) vers la zone de dessin de l'expérience et connectez-le au module [Sélectionner des colonnes dans le jeu de données](#).
10. Dans le volet **Propriétés**, sélectionnez **Supprimer toute la ligne** sous **Mode de nettoyage**. Cela amène le module [Nettoyage des données manquantes](#) à nettoyer les données en supprimant les lignes pour lesquelles des valeurs manquent. Double- cliquez sur le module et saisissez le commentaire suivant : « Supprimer les lignes de valeur manquantes ».

Properties

▲ Clean Missing Data

Columns to be cleaned

Selected columns:
All columns

Launch column selector

Minimum missing value ratio

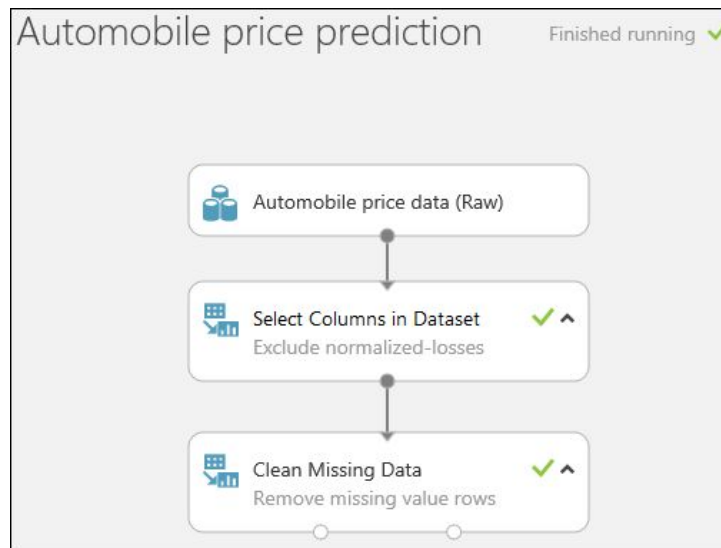
Maximum missing value ratio

Cleaning mode
Remove entire row ▼

11. Exécutez l'expérience en cliquant sur **EXÉCUTER** au bas de la page.

Une fois l'expérience terminée, une coche verte s'affiche en regard de chaque module pour indiquer la réussite de leurs opérations. Notez que le statut **Exécution terminée** s'affiche dans le coin supérieur droit de la fenêtre.

Une fois l'exécution terminée, l'expérience doit ressembler à ceci :



Pour l'instant, nous n'avons exécuté que l'action de nettoyage dans l'expérience. Si vous souhaitez afficher le jeu de données nettoyé, cliquez sur le port de sortie gauche du module Nettoyage des données manquantes et sélectionnez Visualiser. Vous pouvez constater que la colonne normalized-losses n'est plus là et qu'il ne manque plus de données.

Maintenant que les données sont nettoyées, nous pouvons indiquer les fonctionnalités que nous allons utiliser dans le modèle de prévision.

Étape 3 : définition des prédicteurs

Les descripteurs sont des propriétés individuelles mesurables d'un élément qui nous intéresse. Dans notre jeu de données, chaque ligne représente un véhicule et chaque colonne une description de ce véhicule.

La recherche du jeu de descripteurs adéquat pour la création d'un modèle de prévision requiert une certaine expérience et des connaissances sur le problème qui se pose.

Certains prédicteurs sont mieux adaptés à la prévision que d'autres. En outre, certains prédicteurs ont une forte corrélation avec d'autres prédicteurs et peuvent être supprimés.

Par exemple, city-mpg et highway-mpg sont étroitement liées : nous pouvons donc conserver l'une et supprimer l'autre sans trop affecter la prédiction.

Nous allons développer un modèle utilisant un sous-ensemble de ces prédicteurs pour notre jeu de données. Dans ce processus itératif, on peut sélectionner d'autres prédicteurs, relancer l'expérience et voir si l'on obtient de meilleurs résultats. Mais pour commencer, nous allons essayer les prédicteurs suivants avec la cible « price » :

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

12. Faites glisser un autre module [Sélectionner des colonnes dans le jeu de données](#) vers le canevas de l'expérience. Connectez le port de sortie de gauche du module [Nettoyage des données manquantes](#) à l'entrée du module [Sélectionner des colonnes dans le jeu de données](#).



13. Double-cliquez sur le module et saisissez le commentaire suivant : « Sélection des fonctionnalités pour la prévision ».
14. Cliquez sur l'option **Lancer le sélecteur de colonne** figurant dans le volet **Propriétés**.
15. Cliquez sur **With rules**(À l'aide de règles).

16. Sous **Commencer par**, cliquez sur **Aucune colonne**. Dans la ligne de filtre, sélectionnez **Inclure** et **Noms des colonnes**, puis sélectionnez notre liste de noms de colonnes dans la zone de texte. Cela amène le module à ne pas transmettre toutes les colonnes (prédicteurs), à l'exception de celles que nous spécifions.
17. Cliquez sur le bouton en forme de coche (OK) pour continuer.

Cela produit un jeu de données filtré qui contient uniquement les prédicteurs que nous souhaitons transmettre à l'algorithme d'apprentissage utilisé à l'étape suivante. Plus tard, vous pouvez reprendre la procédure en utilisant une autre sélection de prédicteurs.

The screenshot shows a 'Select columns' dialog box. On the left, there are two tabs: 'BY NAME' and 'WITH RULES', with 'WITH RULES' being the active tab. To the right of the tabs is a checkbox labeled 'Allow duplicates and preserve column order in selection'. Below this, the 'Begin With' section has two buttons: 'ALL COLUMNS' and 'NO COLUMNS', with 'NO COLUMNS' being selected. Underneath, there is a filter configuration showing 'Include' selected from a dropdown and 'column names' selected from another dropdown. To the right of this, a list of column names is displayed, each with a close button (X): 'make', 'body-style', 'wh', 'horsepower', 'peak-rpm', and 'price'.

Étape 4 : sélection et application d'un algorithme d'apprentissage

À présent que les données sont prêtes, la construction d'un modèle de prévision passe par la formation et le test. Nous allons utiliser nos données pour former le modèle, puis tester le modèle pour voir dans quelle mesure il peut prédire les prix.

La classification et la régression sont deux types d'algorithmes de machine learning supervisé. La classification permet de prédire une réponse à partir d'un jeu de catégories défini, comme une couleur (rouge, bleu ou vert). La régression est utilisée pour prédire un nombre (continu).

Étant donné que nous voulons prédire un prix, correspondant à un nombre, nous allons utiliser un algorithme de régression. Dans cet exemple, nous allons utiliser un modèle simple de régression linéaire.

Nous formons le modèle en lui fournissant un jeu de données qui inclut le prix. Le modèle analyse les données et recherche les corrélations entre les fonctionnalités d'un véhicule automobile et son prix. Puis nous testons le modèle : nous lui affectons un ensemble de prédicteurs pour véhicules automobiles que nous connaissons et nous étudions la précision du modèle concernant la prédiction des prix.

Nous allons utiliser nos données pour la formation et le test en les divisant en jeux de données distincts d'apprentissage et de test.

18. Sélectionnez et faites glisser le module [Fractionner les données](#) sur le canevas d'expérience et connectez-le au dernier module [Sélectionner des colonnes dans le jeu de données](#).
19. Cliquez sur le module [Fractionner les données](#) pour le sélectionner.
20. Recherchez **Fraction de lignes dans le premier jeu de données de sortie** (dans le volet **Propriétés** à droite du canevas) et attribuez-lui la valeur 0,75. Ainsi, nous allons utiliser 75 % des données pour entraîner le modèle, et 25 % pour le tester. Par la suite, vous pourrez expérimenter d'autres pourcentages.

Properties Project

▲ Split Data

Splitting mode
Split Rows ▼

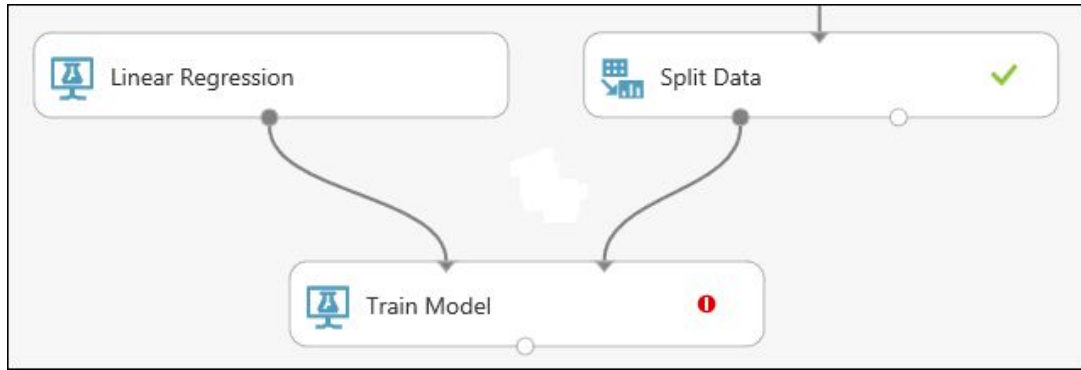
Fraction of rows in the first output dataset
.75

☒ Randomized split

Random seed
0

Stratified split
False ▼

21. Exécutez l'expérience. Lors de l'expérience, les modules [Sélectionner des colonnes dans le jeu de données](#) et [Fractionner les données](#) transmettent des définitions de colonne aux modules que nous allons ajouter par la suite.
22. Pour sélectionner l'algorithme d'apprentissage, développez la catégorie **Machine Learning** dans la palette des modules, à gauche de la zone de dessin, puis développez **Initialiser le modèle**. Différentes catégories de modules s'affichent, permettant d'initialiser des algorithmes d'apprentissage automatique. Pour les besoins de cet exemple, sélectionnez le module [Régression linéaire](#) sous la catégorie **Régression**, puis faites-le glisser vers le canevas de l'expérience. Vous pouvez également rechercher le module en tapant « régression linéaire » dans la zone de recherche de la palette.
23. Recherchez et faites glisser le module [Effectuer le traitement de données pour apprentissage du modèle](#) jusqu'à la zone de dessin de l'expérience. Connectez la sortie du module [Régression linéaire](#) à l'entrée de gauche du module [Former le modèle](#), puis connectez la sortie des données de formation (port gauche) du module [Fractionner les données](#) à l'entrée de droite du module [Former le modèle](#).



24. Cliquez sur le module [Effectuer le traitement de données](#) pour apprentissage du modèle, cliquez sur l'option **Lancer le sélecteur de colonne** du volet **Propriétés** et sélectionnez la colonne **Price**. Il s'agit de la valeur que notre modèle va prévoir.

Vous pouvez sélectionner la colonne **price** dans le sélecteur de colonne en la faisant passer de la liste **Colonnes disponibles** à la liste **Colonnes sélectionnées**.

Select a single column

BY NAME

WITH RULES

AVAILABLE COLUMNS

All Types search columns

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

7 columns available

SELECTED COLUMNS

All Types search columns

price

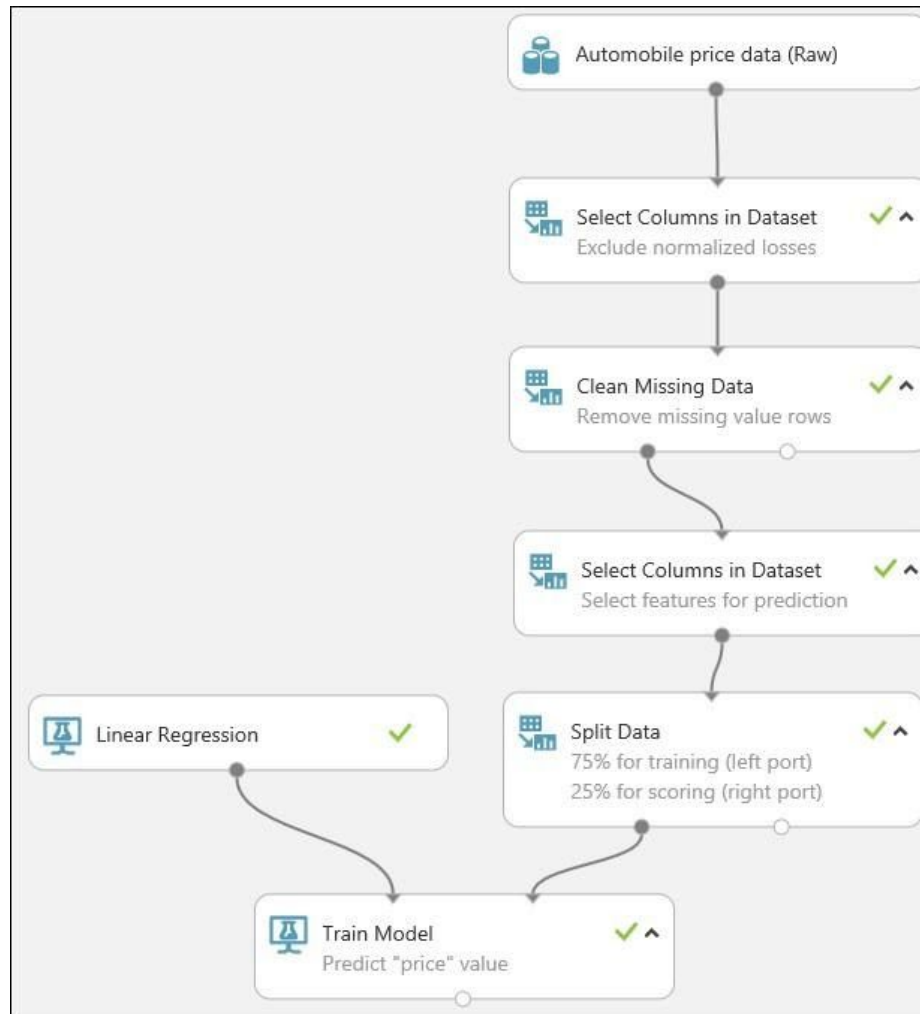
1 columns selected

>

<

25. Exécutez l'expérience.

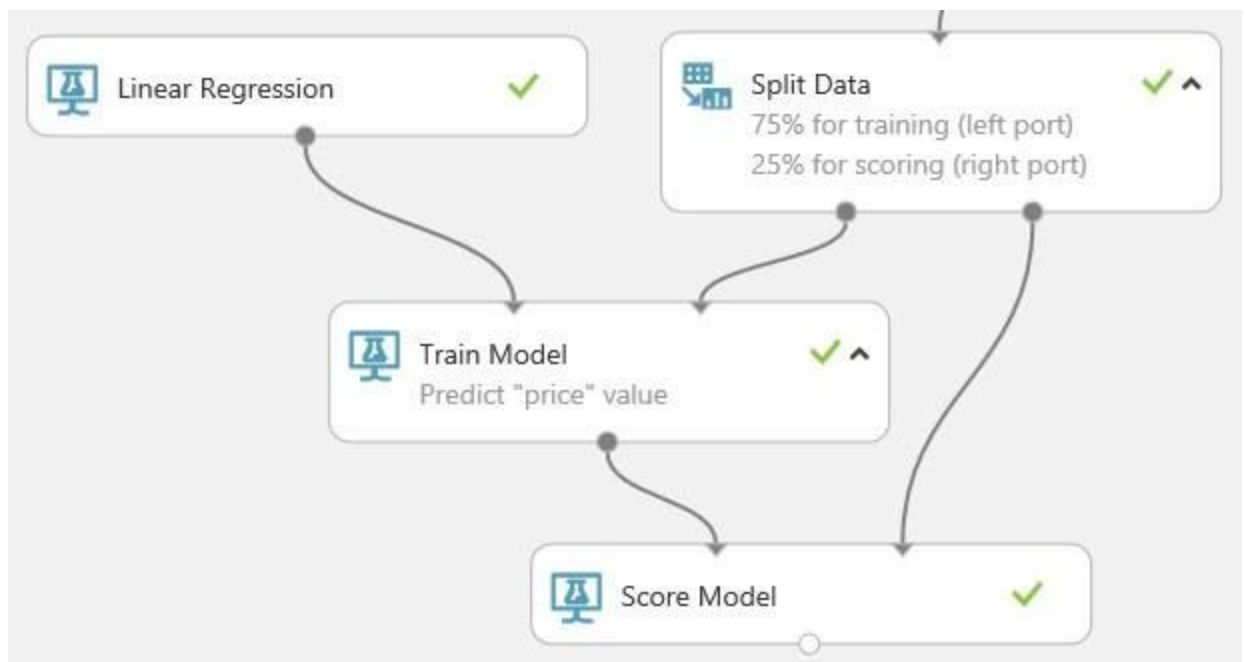
Nous disposons à présent d'un modèle de régression entraîné qui permet d'utiliser de nouvelles données automobiles pour effectuer des prédictions de prix.



Étape 5 : prédiction des nouveaux prix des voitures

À présent que nous avons développé le modèle à l'aide de 75 % de nos données, nous pouvons l'utiliser pour le test du reste de nos données (25 %), afin de voir s'il fonctionne correctement.



26. Recherchez et faites glisser le module -Score model- vers le canevas de l'expérience. Connectez la sortie du module -Train Model- au port d'entrée de gauche -Score Model-. Connectez la sortie de données de test (port de droite) du module Split Data au port d'entrée de droite de Score Model.



27. Pour exécuter l'expérience et afficher la sortie du module [Score Model](#) (cliquez sur le port de sortie de [Score Model](#) et sélectionnez **Visualiser**). La sortie affiche les valeurs de prévision associées au prix, ainsi que les valeurs connues des données de test.

Simple test experiment - Copy > Score Model > Scored dataset

rows: 48, columns: 9

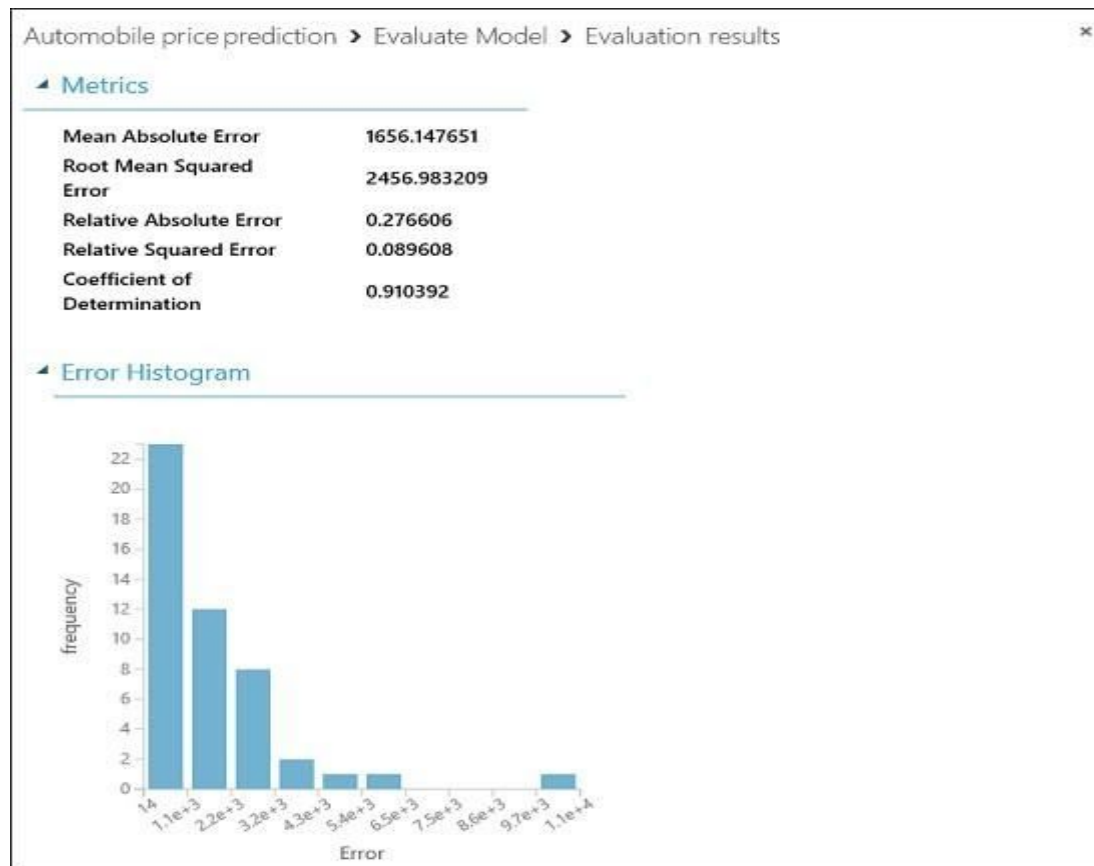
view as:  

Known values (price) and Predicted values (Scored Labels)

make	body-style	wheel-base	engine-size	horsepower	peak-rpm	highway-mpg	price	Scored Labels
subaru	sedan	97	108	111	4800	29	11259	10286.204819
mitsubishi	hatchback	93.7	92	68	5500	38	6669	5446.847864
dodge	hatchback	93.7	90	68	5500	38	6229	6344.800711
honda	hatchback	86.6	92	76	6000	38	6855	5528.302953
alfa-romero	convertible	88.6	130	111	5000	27	16500	13498.476233
volvo	wagon	104.3	141	114	5400	28	16515	16097.608038
isuzu	hatchback	96	119	90	5000	29	11048	8315.257218
dodge	hatchback	93.7	90	68	5500	41	5572	6630.154608
bmw	sedan	101.2	108	101	5800	29	16420	10012.400605

28. Enfin, nous testons la qualité des résultats. Sélectionnez et faites glisser le module [Évaluer le modèle](#) vers le canevas de l'expérience, puis connectez la sortie du module [Noter le modèle](#) à l'entrée de gauche du module [Évaluer le modèle](#).
30. Exécutez l'expérience.

Pour afficher la sortie du module Evaluate Model, cliquez sur le port de sortie, puis sélectionnez Visualiser.



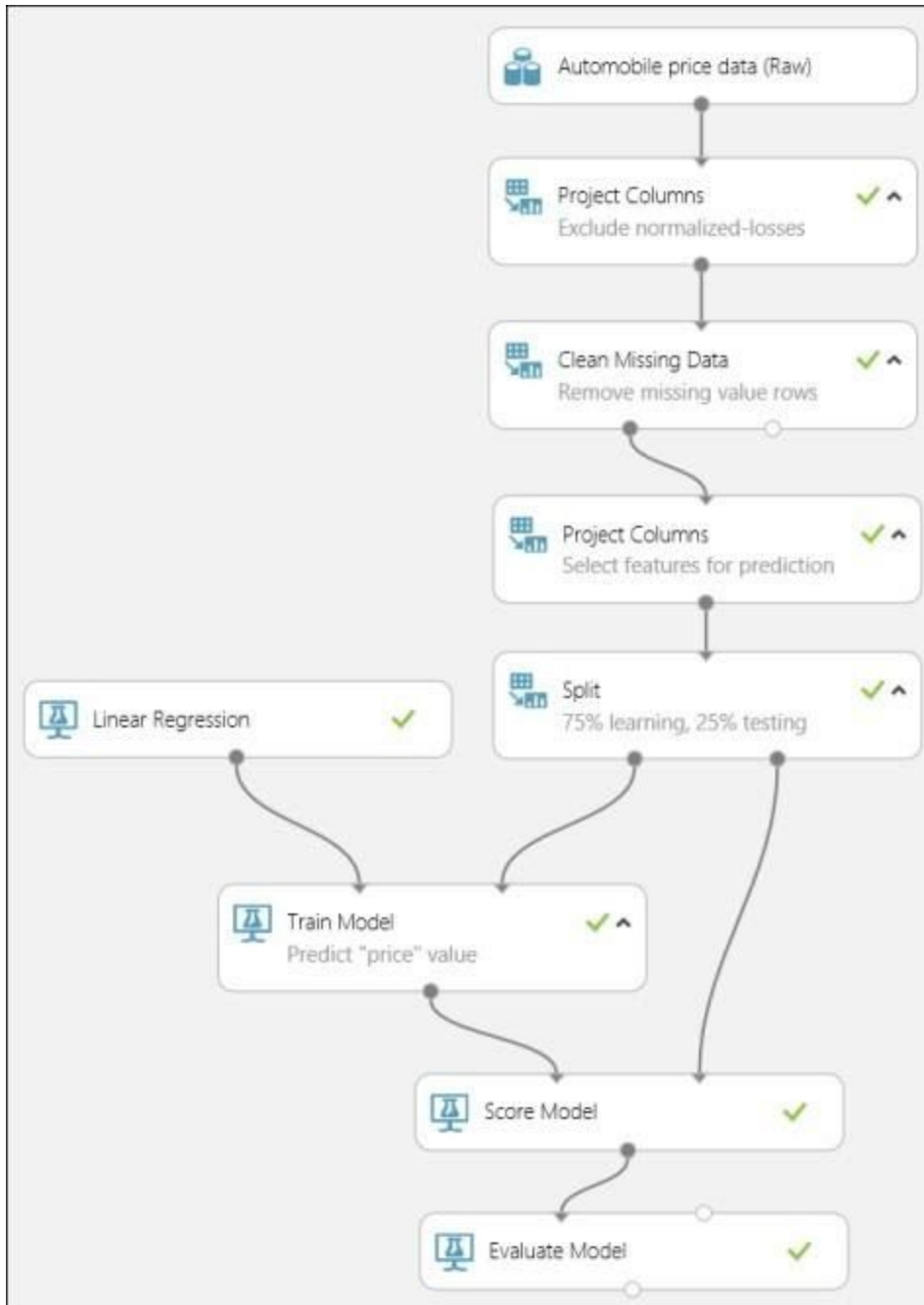
Les statistiques suivantes s'affichent pour notre modèle :

- **Erreur d'absolue moyenne (EAM)** : la moyenne des erreurs absolues (une *erreur* correspond à la différence entre la valeur prévue et la valeur réelle).
- **Racine de l'erreur quadratique moyenne (RMSE)** : la racine carrée de la moyenne des erreurs carrées des prévisions effectuées sur le jeu de données de test.
- **Erreur absolue relative**: la moyenne des erreurs absolues relative à la différence absolue entre les valeurs réelles et la moyenne de toutes les valeurs réelles.
- **Erreur carrée relative**: la moyenne des erreurs carrées relative à la différence carrée entre les valeurs réelles et la moyenne de toutes les valeurs réelles.
- **Coefficient de détermination** : aussi nommé **valeur R au carré**, il s'agit d'une mesure statistique indiquant à quel point un modèle correspond aux données.

Pour chacune des statistiques liées aux erreurs, les valeurs les plus petites sont privilégiées. En effet, une valeur plus petite indique un degré de correspondance plus étroit avec la valeur réelle. Par contre, pour la valeur du Coefficient de détermination, plus elle est proche de un (1.0), plus la prévision est correcte.

Expérience finale

L'expérience finale doit ressembler à ceci :



Note : ce document a été adapté pour une grande partie des tutoriels Azure ML.