pitney bowes

Spectrum™ Technology Platform

Version 12.0

Guide Machine Learning

Table des matières

1 - Introduction

In this section

Module Machine Learning (introduction à la technologie)

L'équipe de Spectrum a le plaisir de vous présenter un nouveau module puissant : Machine Learning. Cette introduction est une implémentation anticipée de Machine Learning contenant les fonctionnalités fondamentales que, selon nous, vous trouverez les plus utiles. Nous prévoyons d'ajouter de nouvelles fonctionnalités significatives dans les futures versions. Ainsi, pour être certains d'ajouter les fonctionnalités que vous attendez, nous mettons maintenant à votre disposition cette introduction à la technologie. Votre avis permettra de guider l'évolution du module Machine Learning. À mesure que vous explorez Machine Learning, gardez à l'esprit les points suivants :

- Cette introduction à la technologie contient un nombre limité de fonctions. Si vous vous mettez à penser « Si seulement je pouvais... », faites-nous en part en renseignant un formulaire de demande d'amélioration et en l'adressant au support technique. Vos suggestions permettront de déterminer les fonctions à ajouter dans les futures versions. Pour savoir comment contacter le support technique, rendez-vous sur www.pitneybowes.com/us/contact-dcs.html.
- Mêmes les meilleurs logiciels contiennent des bogues. Si vous rencontrez un bogue, faites-le nous savoir en soumettant un rapport de bogue et en l'adressant au support technique. Dans la mesure où il s'agit d'une introduction à la technologie, nous ne pouvons pas garantir que nous résoudront immédiatement votre problème spécifique. Pour savoir comment contacter le support technique, rendez-vous sur www.pitneybowes.com/us/contact-dcs.html.
- N'hésitez pas à utiliser cette introduction à la technologie dans votre environnement de production. Gardez à l'esprit que nous ne pouvons pas adhérer à des contrats de niveau de service (SLA) pour les introductions aux technologies.
- Nous nous attendons à recevoir des avis intéressants et imprévus qui pourraient radicalement affecter la prochaine version de Machine Learning. C'est pourquoi nous ne pouvons pas garantir que vous serez en mesure de conserver l'ensemble des tâches que vous aurez effectuées avec Machine Learning lors de la mise à niveau vers les futures versions.
- Faites preuve de bon sens en matière de prise de décisions commerciales en fonction des découvertes générées via cette introduction à la technologie. Nous ne pouvons pas adhérer à des contrats de niveau de service (SLA) normaux pour les fonctions contenues dans l'introduction à la technologie.

Nous espérons que vous aimerez essayer le module Machine Learning et nous attendons impatiemment de connaître votre avis.

Découverte de Machine Learning

Le module Spectrum™ Technology Platform Machine Learning permet d'adapter des modèles Machine Learning supervisés et non supervisés.

Remarque : Le module Machine Learning est pris en charge uniquement sous les systèmes d'exploitation Windows et Linux.

Binning

Binning divise les enregistrements en groupes (bins) pour une variable continue sans prendre en compte les informations d'objectif. Vous pouvez effectuer un binning sans supervision de deux manières : en utilisant des bins de largeur égale ou des bins de fréquence égale.

K-Means Clustering

K-Means Clustering crée des modèles en fonction d'une mise en cluster analytique, qui segmente un ensemble d'enregistrements en clusters d'enregistrements similaires basés sur les valeurs des données.

Logistic Regression

Logistic Regression crée des modèles à partir de jeux de données qui utilisent des objectifs binaires avec des variables d'entrée.

Java Model Scoring

Cette fonction permet d'évaluer les nouvelles données à l'aide de la formule créée lorsque vous appliquez un modèle Machine Learning.

Machine Learning Model Management

Machine Learning Model Management vous permet de gérer tous les modèles Machine Learning sur votre serveur Spectrum™ Technology Platform. Vous pouvez exposer des modèles, annuler leur exposition ou supprimer des modèles. En outre, vous pouvez afficher des informations détaillées sur chaque modèle et comparer deux modèles du même type.

Remarque : Le module Machine Learning utilise une bibliothèque H2O.ai sous-jacente pour la modélisation des algorithmes dans K-Means Clustering, Logistic Regression et Java Model Scoring.

Workflow de Machine Learning

Un workflow Machine Learning type inclut les étapes suivantes, dans un ou plusieurs flux de données :

1. Accédez aux données à l'aide d'autres modules Spectrum, tels que Data Integration.
2. Préparez les données à l'aide de stages d'autres modules Spectrum tels que Data Integration, Data Quality et Core.

3. Appliquez le modèle Machine Learning, exécutez le flux de données, puis vérifiez le contenu de l'onglet Sortie de modèle dans le stage du modèle. Vous pouvez ensuite affiner le modèle, si nécessaire, et ré-exécuter le flux de données. Ensuite, vous devez vérifier la totalité de la sortie d'évaluation du modèle dans l'outil Machine Learning Model Management. Vous pouvez vérifier un modèle à la fois ou comparer deux modèles.
4. (Facultatif) Si le modèle doit être utilisé pour évaluer des données, exposez le modèle dans l'outil Machine Learning Model Management, ce qui rend le modèle accessible au stage Java Model Scoring.
   a. Créez un flux de données Spectrum™ Technology Platform en suivant les étapes 1 et 2 ci-dessus, puis remplacez l'étape 3 par le stage Java Model Scoring. Configurez ce flux de données de sorte qu'il soit exécuté en mode de traitement par lots pour renseigner un fichier à l'aide des scores du modèle appliqués aux données actualisées (les champs utilisés comme Xs ou entrées sont actualisés à l'étape 1-2 dans le cadre naturel du processus).
   b. Vous pouvez également utiliser un service Web dans Spectrum™ Technology Platform pour évaluer les données à la demande. Par exemple, accédez au site Web, obtenez l'ID client et les entrées du modèle, évaluez-les et renvoyez le score à un processus qui personnalise le contenu Web pour votre client.
5. (Facultatif) Vous pouvez également déployer des scores de modèle dans une base de données graphique Data Hub sous forme de propriété d'entité, sur des cartes ou dans des applications CES.

2 - Binning

In this section

Introduction à Binning

Le stage Binning réalise ce qui est dit un binning sans supervision, qui divise une variable continue en groupes (bins) sans prendre en compte les informations d'objectif. Les données capturées incluent des plages, des quantités et un pourcentage des valeurs de chaque plage.

Lorsque vous exécutez un binning, les avantages sont les suivants :

- Cela permet d'inclure des enregistrements avec des données manquantes dans le modèle.
- Cela contrôle ou atténue l'impact des observations aberrantes sur le modèle.
- Cela résout le problème d'avoir différentes échelles parmi les caractéristiques, permettant ainsi de comparer les poids des coefficients du modèle final.

Dans le binning sans supervision Spectrum™ Technology Platform, vous pouvez utiliser des bins de largeur égale, dans lesquels les données sont divisées en bins de taille identique, ou des bins de fréquence égale, dans lesquels les données sont divisées en groupes contenant à peu près le même nombre d'enregistrements. Dans le stage Binning, les bins de largeur égale sont appelés bins de plage égale et les bins de fréquence égale sont appelés bins de population égale.

Configuration des options Binning

1. Sélectionnez si vous souhaitez effectuer un **Style de binning** de plage égale ou de population égale.
2. Sélectionnez dans **Bin de valeur Null** la manière dont vous souhaitez gérer les champs bin vides, qui représentent des valeurs inconnues à cause de données manquantes. Sélectionnez **Le plus haut** pour affecter les valeurs Null au bin le plus haut et **Le plus bas** pour affecter les valeurs Null au bin le plus bas. Le bin le plus bas est toujours le bin 1.
3. Cliquez sur **Bins internes cibles** et saisissez le nombre de bins que vous souhaitez remplir entre les bins de fin. Si vous effectuez un binning de plage égale, vous pouvez sélectionner ce type de traitement ou **Largeur de bin**, mais pas les deux. Si vous effectuez un binning de population égale, vous pouvez effectuer uniquement un traitement de bin interne.
4. Si vous effectuez un binning de plage égale et que vous souhaitez sélectionner ce type de traitement plutôt que le traitement de bin interne, cliquez sur **Largeur de bin** et saisissez le nombre d'unités de votre choix dans chaque bin.
5. Cliquez sur **Inclure** pour chaque champ dont vous souhaitez inclure les données dans le binning. Notez que seuls les champs numériques apparaîtront dans cette liste.
6. Cliquez sur **OK** pour enregistrer vos paramètres.

Sortie de Binning

Le stage Binning dispose de deux ports de sortie. Le premier port sort tous les champs d'entrée plus un champ mis en bin de chaque champ d'entrée sélectionné. Par exemple, si l'entrée contient les champs Name, Age et Income et si vous effectuez un binning sur Age et Income, la sortie du premier port contiendra les champs suivants :

- Name
- Age
- Binned_Age
- Income
- Binned_Income

Le deuxième port sort quatre types d'informations pour chaque champ d'entrée sélectionné. Par exemple, si vous effectuez un binning sur Age, la sortie du deuxième port contiendra les champs suivants :

- Age_Bins
- Age_BinValue
- Age_Count
- Age_Percentage

3 - K-Means Clustering

In this section

Introduction à K-Means Clustering

K-Means Clustering crée des modèles en fonction d'une mise en cluster analytique, qui segmente un ensemble d'enregistrements en clusters d'enregistrements similaires basés sur les valeurs des données.

Pour créer votre modèle, vous devez d'abord renseigner l'onglet Propriétés du modèle. Les onglets Options de base et Options avancées fournissent des paramètres par défaut suffisants pour effectuer un job, mais vous pouvez modifier ces paramètres en fonction de vos besoins. Vous exécutez ensuite votre job et une version limitée des détails de sortie du modèle obtenu apparaît dans l'onglet Sortie de modèle ; le modèle est stocké sur le serveur Spectrum™ Technology Platform et la sortie complète est disponible dans l'outil Machine Learning Model Management.

Définition des propriétés du modèle

1. Sous **Stages primaires**/**Stages déployés**/**Machine Learning**, cliquez sur le stage **K-Means Clustering** et faites-le glisser jusqu'au canevas, en le plaçant à l'endroit de votre choix dans le flux de données et en le reliant à d'autres stages. Notez que le stage d'entrée doit être la source de données qui contient les champs de variables d'entrée de votre modèle ; un stage de sortie n'est pas nécessaire, sauf si vous sélectionnez l'option de données d'entrée Score dans l'onglet Options de base. Vous pouvez également connecter un stage de sortie si vous souhaitez capturer votre sortie indépendamment de l'outil Machine Learning Model Management.
2. Double-cliquez sur le stage K-Means pour afficher la boîte de dialogue **Options de K-Means Clustering**.
3. Saisissez un **Nom de modèle** si vous ne souhaitez pas utiliser le nom par défaut.
4. Facultatif : cochez la case **Écraser** pour remplacer le modèle existant par les nouvelles données.
5. Saisissez le **Nombre de clusters** que vous souhaitez dans votre modèle, si vous ne souhaitez pas utiliser le nombre par défaut (5).
6. Facultatif : Saisissez une **Description** du modèle.
7. Cliquez sur **Inclure** pour chaque champ dont vous souhaitez ajouter les données au modèle.
8. Utilisez la liste déroulante **Type de données du modèle** pour spécifier si le champ d'entrée est à utiliser sous forme numérique, catégorique ou de type date/heure.
9. Cliquez sur **OK** pour enregistrer le modèle et la configuration ou pour passer à l'onglet suivant.

Configuration des options de base

1. Laissez la case **Normaliser les champs d'entrée** cochée pour normaliser les colonnes numériques de sorte qu'elles n'aient pas de variance d'unité ni de moyenne.

   Si vous n'utilisez pas la normalisation, les résultats peuvent inclure des composants dominés par des variables apparaissant comme ayant des variances supérieures aux autres attributs comme échelle plutôt que comme véritable contribution.
2. Cochez **Nombre estimé de clusters** pour que l'algorithme K-Means puisse tenter de déterminer le nombre de clusters contenus dans votre modèle. Même si vous indiquez le nombre de clusters souhaité dans l'onglet Propriétés du modèle, la routine peut découvrir lors de son traitement qu'un nombre de clusters différent est plus approprié, vu les données.
3. Spécifiez une valeur comprise entre 1 et 100 comme **Pourcentage de données de formation** lorsque les données d'entrée sont divisées de manière aléatoire en échantillons de données de formation et de test.
4. Saisissez la valeur 100 moins le nombre que vous avez saisi à l'étape 5 comme **Pourcentage de données de test**.
5. Saisissez un nombre pour **Seed for sampling** pour vous assurer que lorsque les données sont divisées en données de test et de formation, cela se produit de la même manière chaque fois que vous exécutez le flux de données. Laissez « 0 » dans ce champ pour obtenir une division aléatoire chaque fois que vous exécutez le flux.
6. Cliquez sur **OK** pour enregistrer le modèle et la configuration ou pour passer à l'onglet suivant.

Configuration des options avancées

1. Laissez la case **Ignorer champs de constante** cochée pour ignorer les champs qui ont la même valeur pour chaque enregistrement.
2. Sélectionnez le mode d'initialisation correct dans la liste déroulante **Init**.

   Furthest — Initialise le premier centroïde de manière aléatoire, mais initialise le deuxième centroïde de sorte qu'il soit le point de données le plus éloigné de lui. Initialise les centroïdes de sorte qu'ils soient bien répartis l'un par rapport à l'autre.

   Plus-Plus — Initialise les centres de cluster avant de procéder aux itérations d'optimisation *k*-means standard. Avec l'initialisation *k*-means++, l'algorithme est sûr de trouver une solution O(log *k*) compétitive par rapport à la solution *k*-means optimale.

Random Par défaut. Sélectionne les clusters K à partir de l'ensemble de N observations de manière aléatoire, de sorte que chaque observation ait autant de chance d'être sélectionnée.

3. Laissez la case **Seed pour N fois** cochée et saisissez un numéro de seed pour vous assurer que lorsque les données sont divisées en données de test et de formation, cela se produit de la même manière chaque fois que vous exécutez le flux de données. Laissez « 0 » dans ce champ pour obtenir une division aléatoire chaque fois que vous exécutez le flux.
4. Cochez la case **N fois** et saisissez le nombre de fois si vous effectuez une validation croisée.
5. Cochez **Attribution de fois** et faites votre choix dans la liste déroulante si vous effectuez une validation croisée. Ce champ s'applique uniquement si vous avez saisi une valeur dans **N fois**.

   AUTO Par défaut. Permet à l'algorithme de sélectionner automatiquement une option ; actuellement, il utilise Random (Aléatoire).

   Modulo Distribue le jeu de données de façon égale dans les occurrences N fois et ne dépend pas du seed.

   Random Distribue les données de manière aléatoire dans les occurrences N fois ; recommandé pour les grands jeux de données.

   Stratified Stratifie les occurrences N fois en fonction de la variable de réponse pour les problèmes de classification. Répartit uniformément les observations des différentes classes dans tous les jeux lors de la division d'un jeu de données en données de formation et de test. Cela peut être utile s'il existe de nombreuses classes et si le jeu de données est relativement petit.

6. Cochez **Itération maxi.** et saisissez le nombre d'itérations de formation qui doivent être effectuées.
7. Cliquez sur **OK** pour enregistrer le modèle et la configuration ou pour passer à l'onglet suivant.

Sortie de modèle

Cet onglet affiche la métrique que vous utilisez pour évaluer le modèle adapté. Ces champs ne peuvent pas être modifiés. La colonne Formation contient toujours des données. Si vous avez sélectionné une division formation/test dans l'onglet Options de base, la colonne Test est également renseignée, sauf si vous avez sélectionné une validation N fois dans l'onglet Options avancées, auquel cas la colonne N fois est renseignée. Cliquez sur le bouton **Sortie** pour régénérer la sortie, puis cliquez sur **Pour obtenir des informations détaillées, cliquez ici** pour afficher l'intégralité de la sortie dans l'outil Machine Learning Model Management.

4 - Logistic Regression

In this section

Introduction à Logistic Regression

Logistic Regression vous permet de procéder à un apprentissage machine en créant des modèles à partir de jeux de données qui utilisent des objectifs binaires avec des variables d'entrée.

Pour créer votre modèle, vous devez d'abord renseigner l'onglet Propriétés du modèle. Les onglets Options de base et Options avancées fournissent des paramètres par défaut suffisants pour effectuer un job, mais vous pouvez modifier ces paramètres en fonction de vos besoins. Vous exécutez ensuite votre job et une version limitée du modèle obtenu apparaît dans l'onglet Sortie de modèle ; la sortie complète est disponible dans l'outil Machine Learning Model Management.

Définition des propriétés du modèle

1. Sous **Stages primaires**/**Stages déployés**/**Machine Learning**, cliquez sur le stage **Logistic Regression** et faites-le glisser jusqu'au canevas, en le plaçant à l'endroit de votre choix dans le flux de données et en le reliant à d'autres stages. Notez que le stage d'entrée doit être la source de données qui contient à la fois les champs de variables d'entrée et d'objectif de votre modèle ; un stage de sortie n'est pas nécessaire, sauf si vous sélectionnez l'option de données d'entrée Score dans l'onglet Options de base. Vous pouvez également connecter un stage de sortie si vous souhaitez capturer votre sortie indépendamment de l'outil Machine Learning Model Management.
2. Double-cliquez sur le stage Logistic Regression pour afficher la boîte de dialogue **Logistic Regression Options**.
3. Saisissez un **Nom de modèle** si vous ne souhaitez pas utiliser le nom par défaut.
4. Facultatif : cochez la case **Écraser** pour remplacer le modèle existant par les nouvelles données.
5. Cliquez sur le menu déroulant **Champ Objectif** et sélectionnez « Catégorique ».
6. Facultatif : Saisissez une **Description** du modèle.
7. Cliquez sur **Inclure** pour chaque champ dont vous souhaitez ajouter les données au modèle.
8. Utilisez la liste déroulante **Type de données du modèle** pour spécifier si le champ d'entrée est à utiliser sous forme numérique, catégorique ou de type date/heure.
9. Cliquez sur **OK** pour enregistrer le modèle et la configuration ou pour passer à l'onglet suivant.

Configuration des options de base

1. Laissez la case **Normaliser les champs d'entrée** cochée pour normaliser les colonnes numériques de sorte qu'elles n'aient pas de variance d'unité ni de moyenne.

   Si vous n'utilisez pas la normalisation, les résultats peuvent inclure des composants dominés par des variables apparaissant comme ayant des variances supérieures aux autres attributs comme échelle plutôt que comme véritable contribution.
2. Cochez **Évaluer les données en entrée** pour ajouter une colonne pour la prédiction (l'évaluation) du modèle en fonction des données d'entrée.
3. Cochez **A priori** si les données ont été échantillonnées et que la moyenne de réponse ne reflète pas la réalité ; ensuite, saisissez la probabilité d'a priori pour p(y==1) dans le champ de texte.
4. Spécifiez comment gérer les données manquantes en cochant la case **Ignorer** ou **Imputer les valeurs moyennes**, qui ajoute la valeur moyenne de toutes les données manquantes.
5. Spécifiez une valeur comprise entre 1 et 100 comme **Pourcentage de données de formation** lorsque les données d'entrée sont divisées de manière aléatoire en échantillons de données de formation et de test.
6. Saisissez la valeur 100 moins le nombre que vous avez saisi à l'étape 5 comme **Pourcentage de données de test**.
7. Saisissez un nombre pour **Seed for sampling** pour vous assurer que lorsque les données sont divisées en données de test et de formation, cela se produit de la même manière chaque fois que vous exécutez le flux de données. Laissez « 0 » dans ce champ pour obtenir une division aléatoire chaque fois que vous exécutez le flux.
8. Cliquez sur **OK** pour enregistrer le modèle et la configuration ou pour passer à l'onglet suivant.

Configuration des options avancées

1. Laissez la case **Ignorer champs de constante** cochée pour ignorer les champs qui ont la même valeur pour chaque enregistrement.
2. Laissez la case **Calculer valeurs p** cochée pour calculer les valeurs de p pour les estimations de paramètres.
3. Laissez la case **Supprimer colonne colinéaire** cochée pour supprimer automatiquement les colonnes colinéaires pendant la construction du modèle. Cela permet d'obtenir un coefficient 0 dans le modèle renvoyé.

   Cette option doit être cochée si la case **Calculer valeurs p** est également cochée.

4. Laissez la case **Inclure terme constant (Intercepter)** cochée pour inclure un terme constant (interception) dans le modèle.

   Cette option doit être cochée si la case **Supprimer colonne colinéaire** est également cochée.

5. Sélectionnez un **Solver** dans la liste déroulante. Notez que COORDINATE_DESCENT et COORDINATE_DESCENT_NAIVE sont actuellement des options expérimentales.

AUTO	Solver est déterminé en fonction des paramètres et des données d'entrée.
COORDINATE_DESCENT	IRLSM avec la covariance met à jour la version de descente des coordonnées cycliques dans la boucle la plus interne.
COORDINATE_DESCENT_NAIVE	IRLSM avec la valeur naïve met à jour la version de descente des coordonnées cycliques dans la boucle la plus interne.
IRLSM	Idéal en cas de problèmes avec un petit nombre de prédicteurs ou pour les recherches Lambda avec une pénalité N1.
L_BFGS	Idéal pour les jeux de données avec de nombreuses colonnes.

6. Laissez la case **Seed pour N fois** cochée et saisissez un numéro de seed pour vous assurer que lorsque les données sont divisées en données de test et de formation, cela se produit de la même manière chaque fois que vous exécutez le flux de données. Laissez « 0 » dans ce champ pour obtenir une division aléatoire chaque fois que vous exécutez le flux.
7. Cochez la case **N fois** et saisissez le nombre de fois si vous effectuez une validation croisée.
8. Cochez **Attribution de fois** et faites votre choix dans la liste déroulante si vous effectuez une validation croisée. Ce champ s'applique uniquement si vous avez saisi une valeur dans **N fois** et si **Champ Fois** n'est pas spécifié.

AUTO	Permet à l'algorithme de sélectionner automatiquement une option ; actuellement, il utilise Random (Aléatoire).
Modulo	Distribue le jeu de données de façon égale dans les occurrences N fois et ne dépend pas du seed.
Random	Distribue les données de manière aléatoire dans les occurrences N fois ; recommandé pour les grands jeux de données.
Stratified	Stratifie les occurrences N fois en fonction de la variable de réponse pour les problèmes de classification. Répartit uniformément les observations des différentes classes dans tous les jeux lors de la division d'un jeu de données en données de formation et de test. Cela peut être utile s'il existe de nombreuses classes et si le jeu de données est relativement petit.

9. Si vous effectuez une validation croisée, cochez la case **Champ Fois** et sélectionnez le champ qui contient l'affectation d'index fois la validation croisée dans la liste déroulante.

Ce champ s'applique uniquement si vous n'avez pas saisi de valeur dans **N fois** ni dans **Attribution de fois**.

10. Cochez **Itération maxi.** et saisissez le nombre d'itérations de formation qui doivent être effectuées.
11. Cochez **Epsilon Objectif** et saisissez le seuil de convergence ; il doit s'agir d'une valeur comprise entre 0 et 1. Si la valeur d'objectif est inférieure à ce seuil, le modèle fait l'objet d'une convergence.
12. Cochez **Epsilon bêta** et saisissez le seuil de convergence ; il doit s'agir d'une valeur comprise entre 0 et 1. Si la valeur d'objectif est inférieure à ce seuil, le modèle fait l'objet d'une convergence. Si la normalisation N1 du changement bêta actuel est inférieure à ce seuil, envisagez de recourir à la convergence.
13. Cliquez sur **OK** pour enregistrer le modèle et la configuration ou pour passer à l'onglet suivant.

Sortie de modèle

Cet onglet affiche la métrique que vous utilisez pour évaluer le modèle adapté. Ces champs ne peuvent pas être modifiés. La colonne Formation contient toujours des données. Si vous avez sélectionné une division formation/test dans l'onglet Options de base, la colonne Test est également renseignée, sauf si vous avez sélectionné une validation N fois dans l'onglet Options avancées, auquel cas la colonne N fois est renseignée.

Une fois que vous avez exécuté votre job, le modèle qui en résulte est stocké sur le serveur Spectrum™ Technology Platform. Cliquez sur le bouton **Sortie** pour régénérer la sortie, puis cliquez sur **Pour obtenir des informations détaillées, cliquez ici** pour afficher l'intégralité de la sortie dans l'outil Machine Learning Model Management.

5 - Java Model Scoring

In this section

Introduction à Java Model Scoring

Java Model Scoring vous permet d'évaluer de nouvelles données à l'aide de la formule créée lorsque vous adaptez un modèle Machine Learning.

> **Remarque :** Les modèles doivent tout d'abord être exposés via Machine Learning Model Management avant de devenir disponibles dans le stage Java Model Scoring. Pour plus d'informations, reportez-vous à la section **Introduction à Machine Learning Model Management** à la page 23.

Pour évaluer vos données, vous devez renseigner deux onglets de la boîte de dialogue **Options de Java Model Scoring**. Commencez par identifier le modèle et son type, puis assurez-vous que les champs du modèle sont correctement mappés vers les champs Spectrum™ Technology Platform. Ensuite, configurez la sortie en sélectionnant les champs que vous souhaitez inclure et en exécutant votre job. L'onglet **Sortie de modèle** contient le mappage des types de données pour Spectrum™ Technology Platform et votre modèle.

Si votre job contient un stage qui capture la sortie dans un fichier ou dans une table, vous pouvez utiliser cette sortie dans un flux de données ultérieur ou un service Web.

Définition des propriétés du modèle

1. Sous **Stages primaires**/**Stages déployés**/**Advanced Analytics**, cliquez sur le stage **Java Model Scoring** et faites-le glisser jusqu'au canevas, en le plaçant à l'endroit de votre choix dans le flux de données et en le reliant aux stages d'entrée et de sortie. Notez que le stage d'entrée doit être la source de données qui contient à la fois les champs d'objectif et de variables d'entrée de votre modèle. Si vous exécutez votre job en mode batch, vous devez également disposer d'un stage de sortie pour capturer les scores du modèle ; sinon, vous pouvez utiliser un service Web Spectrum™ Technology Platform pour évaluer les données en temps réel.
2. Double-cliquez sur le stage Java Model Scoring pour afficher la boîte de dialogue **Options de Model Scoring**.
3. Facultatif : sélectionnez le type d'un modèle que vous évaluez dans la liste déroulante **Filtre de type**.
4. Sélectionnez le **Filtre de type** utilisé pour évaluer le modèle.
5. Sélectionnez le **Nom de modèle** dans la liste déroulante.
6. Sélectionnez le type de modèle que vous évaluez dans le champ **Type de modèle**.
7. Facultatif : Saisissez une **Description** du modèle.

8. La table **Entrées** présente des informations sur les champs d'entrée du modèle. Ces champs et leurs types de données sont automatiquement mappés vers les champs et types de données Spectrum.
9. Cliquez sur **OK** pour enregistrer ces options ou passer à l'onglet suivant.

Sortie de modèle

La table **Sorties** présente des informations sur les champs de sortie du modèle. Ces champs et leurs types de données sont automatiquement mappés vers les champs et types de données Spectrum.

1. Cliquez sur **Inclure** pour chaque champ dont vous souhaitez inclure les données dans la sortie du modèle.
2. Cliquez sur **OK** pour enregistrer le modèle.

6 - Gestion de modèles Machine Learning

In this section

Introduction à Machine Learning Model Management

L'onglet Model Analysis de Machine Learning Model Management affiche une liste de tous les modèles Machine Learning de votre serveur Spectrum™ Technology Platform. Vous pouvez filtrer cette liste en saisissant une chaîne dans la zone de texte ; une recherche de cette chaîne est alors effectuée sur tous les champs de la table.

Plusieurs opérations peuvent être réalisées sur ces modèles. Vous pouvez exposer des modèles, annuler leur exposition ou supprimer des modèles. Les modèles exposés sont utilisés dans le stage Java Model Scoring pour évaluer les nouvelles données à l'aide de formules créées lorsque vous adaptez des modèles Machine Learning. En outre, vous pouvez afficher des informations détaillées sur chaque modèle ; les détails renvoyés dépendent du type de modèle dont vous visualisez les données. Enfin, vous pouvez comparer deux modèles du même type. Cette comparaison affiche côte-à-côte les mêmes informations que celles qui figurent dans l'onglet Détail du modèle pour chacun des modèles que vous comparez.

Accès à Machine Learning Model Management Model Analysis

Il existe trois méthodes pour accéder à Machine Learning Model Management :

- Utilisez la page d'accueil Spectrum™ Technology Platform :
  - Ouvrez un navigateur internet et allez sur la page d'accueil Spectrum™ Technology Platform à :

    http://<nomduserveur>:<port>

    Par exemple, si vous avez installé Spectrum™ Technology Platform sur un ordinateur appelé « monspectrumplateform » et qu'il utilise le port HTTP par défaut, 8080, vous devrez aller sur :

    http://monspectrumplatform:8080

  - Cliquez sur **Spectrum Machine Learning**.
  - Cliquez sur **Ouvrir le référentiel Machine Learning**.
- Cliquez sur **Pour obtenir des détails sur le modèle, cliquez ici** à partir de l'un des stages de création de modèles.
- Utilisez un navigateur Web :
  - Ouvrez un navigateur Web et accédez à la page Spectrum™ Technology Platform Machine Learning Model Management sur :

    http://<nom du serveur>:<port>/machinelearning

Par exemple, si vous avez installé Spectrum™ Technology Platform sur un ordinateur appelé « monspectrumplateform » et qu'il utilise le port HTTP par défaut, 8080, vous devrez aller sur :

http://myspectrumplatform:8080/machinelearning

- Saisissez un nom d'utilisateur et un mot de passe Spectrum™ Technology Platform valides.
- Lorsque l'outil s'ouvre, cliquez sur l'onglet **Model Analysis**.

Opérations Analyse du modèle Model Management

Effectuez ces opérations en sélectionnant un modèle, puis en cliquant sur le bouton approprié :

	Exposez le modèle pour le rendre accessible au stage Java Model Scoring. Si un modèle n'est pas exposé, il ne peut pas être utilisé pour l'évaluation.
	Annulez l'exposition du modèle.
	Supprimez le modèle. **Remarque :** Vous ne pouvez pas supprimer un modèle exposé ; cependant, pour l'instant, il n'existe aucune sécurité inhérente qui empêche un utilisateur de supprimer des modèles d'un autre utilisateur.
	Affichez les détails de sortie du modèle. Vous pouvez également accéder à ces informations depuis les stages K-Means Clustering et Logistic Regression en cliquant sur « Pour obtenir des détails sur le modèle, cliquez ici » dans l'onglet Sortie de modèle.
	Comparez des modèles.

Onglet Détail du modèle

L'écran Détail du modèle affiche les informations suivantes pour tous les modèles :

- **Nom du modèle** : nom du modèle
- **Type de modèle** : type de modèle Machine Learning
- **Utilisateur** : nom d'utilisateur de la personne qui a créé le modèle
- **Description** : description du modèle si une description a été fournie lors de sa création
- **État** : indique si le modèle est exposé ou non exposé
- **Nom de flux de données** : nom du flux de données qui a produit le modèle

- **Heure de création** : date et heure de création du modèle

Des détails supplémentaires sont fournis en fonction du type de modèle.

Détails de K-Means Clustering

L'écran Détail du modèle affiche les informations suivantes pour les modèles K-Means Clustering :

Synthèse du modèle

- Nombre de lignes
- Nombre de clusters
- Nombre de colonnes catégoriques
- Nombre d'itérations
- Au sein de la somme de clusters de carrés
- Somme totale des carrés
- Entre la somme de clusters de carrés

Mesures

Fournit des données de formation, de test et N fois pour les éléments suivants :

- Total au sein de la somme de clusters de carrés
- Somme totale des carrés
- Entre la somme de clusters de carrés

Statistiques des centroïdes

Fournit des données de formation, de test et N fois pour chaque centroïde :

- Taille
- Au sein de la somme de clusters de carrés

Moyennes des clusters

Fournit des informations détaillées sur chaque centroïde. Le contenu varie suivant les données d'entrée. Un cluster est un groupe d'observations formulées à partir d'un jeu de données identifiées comme similaires selon un algorithme de mise en cluster donné.

Moyennes des clusters standardisés

Fournit des informations standardisées sur chaque centroïde. Le contenu varie suivant les données d'entrée.

Détails de Logistic Regression

L'écran Détail du modèle affiche les informations suivantes pour les modèles Logistic Regression :

Mesures

Fournit des données de formation, de test et N fois pour les éléments suivants :

- Mean squared error (MSE)
- Root mean squared error (RMSE)
- Nombre d'observations
- R-squared (R2)
- Logarithmic loss (Logloss)
- Area under the curve (AUC)
- Coefficient Gini
- Moyenne par erreur de classe
- AIC
- Déviance résiduelle
- Déviance nulle
- Degré de liberté nul
- Degré de liberté résiduel

Seuil de métrique maximum

Fournit le seuil de métrique maximum de formation pour les données de formation, de test et N fois à l'aide des métriques suivantes :

- max f1
- max f2
- max f0point5
- max accuracy
- max precision
- max recall
- max specificity
- max absolute_mcc
- max min_per_class_accuracy
- max mean_per_class_accuracy

Matrice de confusion

Illustre les performances d'un modèle sur un ensemble de données de formation, de test et N fois dont les valeurs true sont connues.

Graphique de coefficients standardisé

Affiche les prédicteurs les plus importants en fournissant la valeur relative des coefficients, qui indique dans quelle mesure une modification de l'entrée modifie l'objectif.

Coefficients GLM

Coefficients d'un Generalized Linear Model, qui évalue les résultats des modèles de régression suivant des répartitions exponentielles.

Courbes AUC

Area Under the Curve ; détermine lequel des modèles utilisés prédit le mieux les classes à l'aide des données de formation, de test et N fois.

Courbes de montée/gain

Évalue la capacité de prédiction d'un modèle de classification binaire à l'aide des données de formation, de test et N fois.

Notices

Avis USPS®

Pitney Bowes Inc. détient une licence non exclusive pour la publication et la vente de bases de données ZIP + 4® sur des supports optiques et magnétiques. Les marques de commerce suivantes appartiennent à United States Postal Service : CASS, CASS Certified, DPV, eLOT, FASTforward, First-Class Mail, Intelligent Mail, LACS^{Link}, NCOA^{Link}, PAVE, PLANET Code, Postal Service, POSTNET, Post Office, RDI, Suite^{Link} , United States Postal Service, Standard Mail, United States Post Office, USPS, ZIP Code et ZIP + 4. Cette liste de marques de commerce appartenant à U.S. Postal Service n'est pas exhaustive.

Pitney Bowes Inc. détient une licence non exclusive de USPS® pour le traitement NCOA^{Link}®.

Les prix des produits, des options et des services de Pitney Bowes Software ne sont pas établis, contrôlés ni approuvés par USPS® ni par le gouvernement des États-Unis. Lors de l'utilisation de données RDI™ pour déterminer les frais d'expédition de colis, le choix commercial de l'entreprise de distribution de colis à utiliser n'est pas fait par USPS® ni par le gouvernement des État-Unis.

Fournisseur de données et avis associés

Les produits de données contenus sur ce support et utilisés au sein des applications Pitney Bowes Software sont protégés par différentes marques de commerce et par un ou plusieurs des copyrights suivants :

© Copyright United States Postal Service. Tous droits réservés.

© 2014 TomTom. Tous droits réservés. TomTom et le logo TomTom logo sont des marques déposées de TomTom N.V.

© 2016 HERE

Source : INEGI (Instituto Nacional de Estadística y Geografía)

Basées sur les données électroniques © National Land Survey Sweden.

© Copyright United States Census Bureau

© Copyright Nova Marketing Group, Inc.

Des portions de ce programme sont sous © Copyright 1993-2007 de Nova Marketing Group Inc. Tous droits réservés.

© Copyright Second Decimal, LLC

© Copyright Canada Post Corporation

Ce CD-ROM contient des données provenant d'une compilation dont Canada Post Corporation possède le copyright.

© 2007 Claritas, Inc.

Le jeu de données Geocode Address World contient des données distribuées sous licence de GeoNames Project (**www.geonames.org**) fournies sous la licence Creative Commons Attribution License (« Attribution License ») à l'adresse : **http://creativecommons.org/licenses/by/3.0/legalcode**. Votre utilisation des données GeoNames (décrites dans le Manuel de l'utilisateur Spectrum™ Technology Platform) est régie par les conditions de la licence Attribution License et tout conflit entre votre accord avec Pitney Bowes Software, Inc. et la licence Attribution License sera résolu en faveur de la licence Attribution License uniquement s'il concerne votre utilisation des données GeoNames.

3001 Summer Street
Stamford CT 06926-0700
USA

www.pitneybowes.com