pitney bowes

Spectrum™ Technology Platform

Version 12.0

Machine Learning-Handbuch

Inhalt

1 - Einführung

In this section

Machine Learning-Modul (Technologievorschau)

Das Spectrum-Team freut sich, die Vorschau auf ein leistungsstarkes neues Modul mit Ihnen zu teilen: Machine Learning. Diese Technologievorschau stellt eine frühe Implementierung des Machine Learning-Moduls dar, das die grundlegenden Funktionen umfasst, die, wie wir finden, für Sie am hilfreichsten sind. In zukünftigen Versionen möchten wir wesentliche neue Funktionen hinzuzufügen. Damit wir sicherstellen können, dass die von Ihnen gewünschten Funktionen hinzugefügt werden, stellen wir Ihnen nun diese Technologievorschau zur Verfügung. Ihr Feedback trägt zur Entwicklung des Machine Learning-Moduls bei. Beachten Sie die folgenden Dinge, wenn Sie Machine Learning erkunden:

- Diese Technologievorschau enthält eine begrenzte Gruppe von Features. Für Verbesserungsvorschläge oder Ideen hinsichtlich neuer Funktionen sind wir stets offen. Senden Sie dem technischen Support einfach eine Verbesserungsanforderung. Ihre Vorschläge helfen uns bei der Bestimmung der Features, die in zukünftigen Releases hinzugefügt werden. Informationen darüber, wie Sie sich an den technischen Support wenden, finden Sie unter www.pitneybowes.com/us/contact-dcs.html.
- Selbst die beste Software enthält Fehler. Wenn Ihnen ein Fehler auffällt, informieren Sie uns darüber, indem Sie einen Fehlerbericht an den technischen Support senden. Da dies eine Technologievorschau ist, können wir leider nicht garantieren, dass wir uns sofort mit Ihrem Problem befassen können. Informationen darüber, wie Sie sich an den technischen Support wenden, finden Sie unter www.pitneybowes.com/us/contact-dcs.html.
- Sie können diese Technologievorschau gerne in Ihrer Produktionsumgebung verwenden. Beachten Sie, dass wir bei Technologievorschauen keine normalen Vereinbarungen zum Servicelevel (Service Level Agreements, SLAs) einhalten können.
- Wir erwarten interessantes und unvorhergesehenes Feedback, das wesentlichen Einfluss auf das neue Release von Machine Learning haben können. Wir können daher nicht garantieren, dass Sie Ihre gesamte Arbeit im Machine Learning-Modul beibehalten können, wenn Sie ein Upgrade auf zukünftige Releases durchführen.
- Nutzen Sie Ihr Urteilsvermögen, wenn es darum geht, Geschäftsentscheidungen zu treffen, die auf mit dieser Technologievorschau generierten Einblicken basieren. Bei Features aus der Technologievorschau können wir die normalen Vereinbarungen zum Servicelevel (Service Level Agreements, SLAs) nicht einhalten.

Wir wünschen Ihnen viel Spaß beim Experimentieren mit dem Machine Learning-Modul und freuen uns auf Ihr Feedback.

Ein erster Einblick in Machine Learning

Das Machine Learning-Modul von Spectrum™ Technology Platform bietet die Möglichkeit, beaufsichtigte und unbeaufsichtigte Machine Learning-Modelle anzupassen.

> **Anmerkung:** Das Machine Learning-Modul wird nur unter Windows- und Linux-Betriebssystemen unterstützt.

Binning

Beim Binning werden Datensätze für eine kontinuierliche Variable in Gruppen (Bins) aufgeteilt, ohne dass dabei Zielinformationen berücksichtigt werden. Sie können das unbeaufsichtigte Binning mit einer der beiden folgenden Methoden durchführen: mit Bins vom Typ „equal-width“ oder mit Bins vom Typ „equal-frequency“.

K-Means Clustering

Beim „K-Means Clustering“ werden Modelle auf der Grundlage des analytischen Clusterings erstellt. Dabei wird eine Reihe von Datensätzen basierend auf Datenwerten in Cluster mit ähnlichen Datensätzen segmentiert.

Logistic Regression

Mit Logistic Regression werden Modelle aus Datasets erstellt, die im Hinblick auf Eingabevariablen binäre Ziele verwenden.

Java Model Scoring

Dieses Feature bewertet mithilfe der Formel, die beim Anpassen eines Machine Learning-Modells erstellt wird, neue Daten.

Machine Learning-Modellverwaltung

Die „Machine Learning“-Modellverwaltung ermöglicht Ihnen die Verwaltung aller Machine Learning-Modelle auf Ihrem Spectrum™ Technology Platform-Server. Sie können Modelle verfügbar machen, die Verfügbarkeit von Modellen aufheben oder Modelle löschen. Zusätzlich können Sie zu jedem Modell detaillierte Informationen anzeigen und zwei beliebige Modelle des gleichen Typs miteinander vergleichen.

> **Anmerkung:** Das Machine Learning-Modul verwendet eine zugrunde liegende H2O.ai-Bibliothek für Modellierungsalgorithmen in „K-Means Clustering“, „Logistic Regression“ und „Java Model Scoring“.

Machine Learning-Workflow

Ein typischer Machine Learning-Workflow umfasst folgende Schritte, die in mindestens einem Datenfluss ausgeführt werden:

1. Greifen Sie über andere Spectrum-Module, z. B. Data Integration, auf die Daten zu.
2. Bereiten Sie die Daten mit Schritten aus anderen Spectrum-Modulen vor, z. B. denen im Data Integration-, im Data Quality- und im Core-Modul.
3. Passen Sie ein Machine Learning-Modell an, führen Sie den Datenfluss aus und überprüfen Sie anschließend die Inhalte auf der Registerkarte „Modellausgabe“ im Modellschritt. Anschließend können Sie das Modell bei Bedarf anpassen und den Datenfluss erneut ausführen. Im Anschluss müssen Sie die vollständige Ausgabe der Modellbewertung im Tool für die „Machine Learning“-Modellverwaltung überprüfen. Sie können jeweils ein Modell überprüfen oder zwei Modelle miteinander vergleichen.
4. (Optional) Wenn das Modell für die Bewertung von Daten verwendet wird, müssen Sie das Modell im Tool für die „Machine Learning“-Modellverwaltung verfügbar machen. Mit diesem Tool wird das Modell für den „Java Model Scoring“-Schritt zur Verfügung gestellt.
   a. Erstellen Sie mithilfe der oben beschriebenen Schritte 1 – 2 einen Spectrum™ Technology Platform-Datenfluss, und ersetzen Sie Schritt 3 dann durch den „Java Model Scoring“-Schritt. Richten Sie diesen Datenfluss so ein, dass er im Batchmodus ausgeführt wird, um eine Datei mit Modellbewertungen aufzufüllen, die auf aktualisierte Daten angewendet werden (die Felder, die als X-Felder oder Eingaben verwendet werden, werden in den Schritten 1– 2 als natürlicher Bestandteil des Handeltreibens aktualisiert).
   b. Verwenden Sie alternativ zum Bewerten von bedarfsgesteuerten Daten einen Webservice in Spectrum™ Technology Platform. Beispiel: Greifen Sie auf die Website zu, rufen Sie die Kunden-ID und die Modelleingaben ab, bewerten Sie diese Eingaben, und geben Sie die Bewertung an einen Prozess zurück, der Webinhalte für Ihren Kunden anpasst.
5. (Optional) Sie können Modellbewertungen auch in einer „Data Hub“-Diagrammdatenbank als Entitätseigenschaft, auf Karten oder in CES-Anwendungen bereitstellen.

2 - Binning

In this section

Einführung in das Binning

Der „Binning“-Schritt führt das so genannte unbeaufsichtigte Binning durch, bei dem eine kontinuierliche Variable in Gruppen (Bins) unterteilt wird. Dabei werden keine objektiven Informationen berücksichtigt. Die erfassten Daten beinhalten Bereiche, Mengen und Prozentsätze von Werten der einzelnen Bereiche.

Zu den Vorteilen bei der Durchführung des Binnings zählen folgende:

- Es lässt zu, dass Datensätze mit fehlenden Daten in das Modell eingeschlossen werden.
- Es steuert bzw. verringert die Auswirkungen von Ausreißern innerhalb des Modells.
- Es löst das Problem, dass die Merkmale verschiedene Skalierungen aufweisen. Dadurch können die Gewichtungen der Koeffizienten im Endmodell miteinander verglichen werden.

Sie können beim unbeaufsichtigten Binning der Spectrum™ Technology Platform Bins vom Typ „equal-width“ verwenden, bei denen die Daten in gleich große Bins unterteilt werden, oder in Bins vom Typ „equal-frequency“, bei denen die Daten in Gruppen aufgeteilt werden, die in etwa die gleiche Anzahl von Datensätzen enthalten. Im „Binning“-Schritt werden Bins vom Typ „equal-width“ als „Equal Range“-Bins und Bins vom Typ „equal-frequency“ als „Equal Population“-Bins bevorzugt.

Konfigurieren von Binning-Optionen

1. Wählen Sie aus, ob Sie den **Binning-Stil** „equal-range“ oder „equal-population“ durchführen möchten.
2. Wählen Sie unter **Nullwert-Bin** aus, wie Sie mit leeren Bin-Feldern umgehen möchten, die unbekannte Werte aufgrund von fehlenden Daten darstellen. Wählen Sie **Höchste** aus, um dem höchsten Bin Nullwerte zuzuweisen, und wählen Sie **Niedrigste** aus, um dem niedrigsten Bin Nullwerte zuzuweisen. Der niedrigste Bin enthält immer 1.
3. Klicken Sie auf **Interne Ziel-Bins** und geben Sie die Anzahl der Bins ein, die Sie zwischen den End-Bins einfügen möchten. Wenn Sie das Binning vom Typ „equal-range“ durchführen, können Sie diesen Verarbeitungstyp oder **Bin-Breite** auswählen, jedoch nicht beides. Wenn Sie das Binning vom Typ „equal-population“ durchführen, können Sie nur interne Bins verarbeiten.
4. Wenn Sie das Binning vom Typ „equal-range“ durchführen und statt der Verarbeitung eines internen Bins diesen Verarbeitungstyp auswählen möchten, klicken Sie auf **Bin-Breite** und geben Sie die Anzahl der Einheiten an, die in den einzelnen Bins enthalten sein sollen.
5. Klicken Sie bei jedem Feld, dessen Daten beim Binning eingeschlossen werden sollen, auf **Einschließen**. Beachten Sie, dass nur numerische Felder in dieser Liste angezeigt werden.
6. Klicken Sie auf **OK**, um Ihre Einstellungen zu speichern.

Binning-Ausgabe

Der „Binning“-Schritt verfügt über zwei Ausgabeports. Am ersten Port werden für jedes ausgewählte Eingabefeld alle Eingabefelder und ein gebinntes Feld ausgegeben. Beispiel: Wenn die Eingabe die Felder „Name“, „Alter“ und „Einkommen“ enthält und Sie das Binning für „Alter“ und „Einkommen“ durchführen, enthält die Ausgabe des ersten Ports folgende Felder:

- Bezeichnung
- Alter
- Binned_Age
- Einkommen
- Binned_Income

Am zweiten Port werden für jedes ausgewählte Eingabefeld vier Informationstypen ausgegeben. Beispiel: Wenn Sie das Binning für „Alter“ durchführen, enthält die Ausgabe des zweiten Ports folgende Felder:

- Age_Bins
- Age_BinValue
- Age_Count
- Age_Percentage

3 - K-Means Clustering

In this section

Einführung in das „K-Means Clustering“

Beim „K-Means Clustering“ werden Modelle auf der Grundlage des analytischen Clusterings erstellt. Dabei wird eine Reihe von Datensätzen basierend auf Datenwerten in Cluster mit ähnlichen Datensätzen segmentiert.

Sie müssen zunächst die Registerkarte „Modelleigenschaften“ ausfüllen, um Ihr Modell erstellen zu können. Auf den Registerkarten „Standardoptionen“ und „Erweiterte Optionen“ werden genügend Standardeinstellungen zum Abschließen eines Auftrags bereitgestellt. Sie können diese Einstellungen jedoch auch ändern und an Ihre Bedürfnisse anpassen. Anschließend führen Sie Ihren Auftrag aus, und auf der Registerkarte „Modellausgabe“ wird eine eingeschränkte Version der Details zur resultierenden Modellausgabe angezeigt. Das Modell wird auf dem Spectrum™ Technology Platform-Server gespeichert und die vollständige Ausgabe ist im Tool für die „Machine Learning“-Modellverwaltung verfügbar.

Definieren von Modelleigenschaften

1. Klicken Sie unter **Primäre Schritte**/**Bereitgestellte Schritte**/**Machine Learning** auf den **K-Means Clustering**-Schritt und ziehen Sie ihn auf die Arbeitsfläche. Platzieren Sie ihn an einer beliebigen Stelle im Datenfluss und verbinden Sie ihn mit anderen Schritten. Beachten Sie, dass der Eingabeschritt die Datenquelle sein muss, die Eingabevariablenfelder für Ihr Modell enthält. Ein Ausgabeschritt ist nicht erforderlich, es sei denn, Sie wählen auf der Registerkarte „Standardoptionen“ die Eingabedatenoption „Bewerten“ aus. Sie können auch eine Verbindung zu einem Ausgabeschritt herstellen, wenn Sie Ihre Ausgabe unabhängig von dem Tool für die Machine Learning-Modellverwaltung erfassen möchten.
2. Doppelklicken Sie auf den „K-Means“-Schritt, um das Dialogfeld **„K-Means Clustering“-Optionen** anzuzeigen.
3. Geben Sie einen **Modellnamen** ein, wenn Sie nicht den Standardnamen verwenden möchten.
4. Optional: Aktivieren Sie das Kästchen **Überschreiben**, um das vorhandene Modell mit neuen Daten zu überschreiben.
5. Geben Sie die **Anzahl der Cluster** ein, die in Ihrem Modell enthalten sein sollen, wenn diese von der Standardanzahl (5) abweicht.
6. Optional: Geben Sie einen **Beschreibung** des Modells ein.
7. Klicken Sie bei jedem Feld, dessen Daten zum Modell hinzugefügt werden sollen, auf **Einschließen**.
8. Geben Sie über das Dropdown-Menü **Modelldatentyp** an, ob das Eingabefeld als numerisches Feld, als kategorisches Feld oder als DateTime-Feld verwendet werden soll.

9. Klicken Sie auf **OK**, um das Modell und die Konfiguration zu speichern, oder fahren Sie mit der nächsten Registerkarte fort.

Konfigurieren von Standardoptionen

1. Lassen Sie **Eingabefelder standardisieren** aktiviert, um die numerischen Spalten zu standardisieren, damit diese keine Mittelwert- und Einheitenvarianz aufweisen.

   Wenn Sie keine Standardisierung verwenden, können die Ergebnisse Komponenten enthalten, die von Variablen dominiert werden, die statt richtiger Beiträge relativ zu anderen Attributen in der Skalierung eine größere Varianz zu haben scheinen.
2. Aktivieren Sie **Anzahl der Cluster schätzen**, damit der Algorithmus „K-Means“ versucht, die Anzahl der in Ihrem Modell enthaltenen Cluster zu bestimmen. Auch wenn Sie die Anzahl der gewünschten Cluster auf der Registerkarte „Modelleigenschaften“ angeben, kann bei der Routine während der Verarbeitung festgestellt werden, das für die Daten eine andere Anzahl von Clustern geeigneter wäre.
3. Geben Sie für den **Prozentsatz für Trainingsdaten** einen Wert zwischen 1 und 100 an, wenn die Eingabedaten zufällig in Stichproben für Trainings- und Testdaten aufgeteilt werden.
4. Geben Sie den Wert 100 abzüglich der in Schritt 5 eingegebenen Menge als **Prozentsatz für Testdaten** ein.
5. Geben Sie eine Ziffer als **Ausgangswert für Stichprobe** ein, um sicherzustellen, dass die Darstellung der Daten bei jeder Datenflussausführung gleich ist, wenn diese in Test- und Trainingsdaten aufgeteilt werden. Lassen Sie den Wert „0“ in diesem Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.
6. Klicken Sie auf **OK**, um das Modell und die Konfiguration zu speichern, oder fahren Sie mit der nächsten Registerkarte fort.

Konfigurieren erweiterter Optionen

1. Lassen Sie **Konstante Felder ignorieren** aktiviert, damit Felder übersprungen werden, die für die einzelnen Datensätze die gleichen Werte enthalten.
2. Wählen Sie im Dropdown-Menü **Init** den richtigen Initialisierungsmodus aus.

   Furthest — Initialisiert den ersten Zentroid zufällig, den zweiten Zentroid initialisiert der Modus jedoch anschließend so, dass es der davon am weitesten entfernte Datenpunkt ist. Initialisiert die Zentroide so, dass sie gut verteilt sind.

   Plus-Plus — Initialisiert das Clusterzentrum, bevor mit den standardmäßigen „*k*-means“-Optimierungsiterationen fortgefahren wird. Bei der

„*k*-means++“-Initialisierung wird garantiert, dass der Algorithmus die Lösung „O(log *k*) competitive“ für die optimale „*k*-means“-Lösung findet.

Random Standardeinstellung. Wählt Cluster K zufällig aus der Gruppe der Beobachtungen N aus, damit die einzelnen Beobachtungen gleichermaßen die Möglichkeit haben, ausgewählt zu werden.

3. Lassen Sie **Seed für N-fach** aktiviert, und geben Sie einen numerischen Ausgangswert ein, um sicherzustellen, dass die Darstellung der Daten bei jeder Datenflussausführung gleich ist, wenn diese in Test- und Trainingsdaten aufgeteilt werden. Lassen Sie den Wert „0“ in diesem Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.
4. Aktivieren Sie **N-fach** und geben Sie die Anzahl der Folds ein, wenn Sie eine Kreuzvalidierung durchführen.
5. Aktivieren Sie **Foldzuweisung**, und wählen Sie aus der Dropdown-Liste aus, ob Sie eine Kreuzvalidierung durchführen. Dieses Feld ist nur anwendbar, wenn Sie unter **N-fach** einen Wert eingegeben haben.

   AUTO Standardeinstellung. Lässt zu, dass der Algorithmus automatisch eine Option auswählt; derzeit wird „Random“ verwendet.

   Modulo Teilt das Dataset gleichmäßig auf die Folds auf und ist nicht vom Ausgangswert abhängig.

   Random Teilt die Daten zufällig in „N-fach“-Bestandteile ein; diese Einstellung ist für umfangreiche Datasets am besten geeignet.

   Stratified Schichtet die Folds basierend auf der Antwortvariable für Klassifizierungsprobleme. Verteilt Beobachtungen aus den verschiedenen Klassen gleichmäßig auf alle Datasets, wenn ein Dataset in Trainings- und Testdaten aufgeteilt wird. Dies kann nützlich sein, wenn viele Klassen vorhanden sind und das Dataset relativ klein ist.

6. Aktivieren Sie **Max. Iterationen** und geben Sie die Anzahl der Trainingiterationen ein, die erfolgen sollen.
7. Klicken Sie auf **OK**, um das Modell und die Konfiguration zu speichern, oder fahren Sie mit der nächsten Registerkarte fort.

Modellausgabe

Auf dieser Registerkarte werden die Metriken angezeigt, mit denen Sie das angepasste Modell bewerten. Diese Felder können nicht bearbeitet werden. Die Spalte „Training“ enthält immer Daten. Wenn Sie auf der Registerkarte „Standardoptionen“ eine Aufteilung in Training/Test ausgewählt haben, wird die Spalte „Test“ ebenfalls aufgefüllt. Eine Ausnahme besteht, wenn Sie auf der

Registerkarte „Erweiterte Optionen“ eine Validierung vom Typ „N-fach“ ausgewählt haben. In diesem Fall wird die Spalte „N-fach“ aufgefüllt. Klicken Sie auf die Schaltfläche **Ausgabe**, um die Ausgabe erneut zu generieren, und klicken Sie auf **Für Details hier klicken**, um in dem Toll für die „Machine Learning“-Modellverwaltung die gesamte Ausgabe anzuzeigen.

4 - Logistic Regression

In this section

Einführung in Logistic Regression

Mit Logistic Regression können Sie Machine Learning durchführen, indem Sie Modelle aus Datasets erstellen, die im Hinblick auf Eingabevariablen binäre Ziele verwenden.

Sie müssen zunächst die Registerkarte „Modelleigenschaften“ ausfüllen, um Ihr Modell erstellen zu können. Auf den Registerkarten „Standardoptionen“ und „Erweiterte Optionen“ werden genügend Standardeinstellungen zum Abschließen eines Auftrags bereitgestellt. Sie können diese Einstellungen jedoch auch ändern und an Ihre Bedürfnisse anpassen. Anschließend führen Sie Ihren Auftrag aus, und auf der Registerkarte „Modellausgabe“ wird eine eingeschränkte Version des resultierenden Modells angezeigt. Die vollständige Ausgabe ist im Tool für die „Machine Learning“-Modellverwaltung verfügbar.

Definieren von Modelleigenschaften

1. Klicken Sie unter **Primäre Schritte**/**Bereitgestellte Schritte**/**Machine Learning** auf den **Logistic Regression**-Schritt und ziehen Sie ihn auf die Arbeitsfläche. Platzieren Sie ihn an einer beliebigen Stelle im Datenfluss und verbinden Sie ihn mit anderen Schritten. Beachten Sie, dass der Eingabeschritt die Datenquelle sein muss, die Ziel- und Eingabevariablenfelder für Ihr Modell enthält. Ein Ausgabeschritt ist nicht erforderlich, es sei denn, Sie wählen auf der Registerkarte „Standardoptionen“ die Eingabedatenoption „Bewerten“ aus. Sie können auch eine Verbindung zu einem Ausgabeschritt herstellen, wenn Sie Ihre Ausgabe unabhängig von dem Tool für die Machine Learning-Modellverwaltung erfassen möchten.
2. Doppelklicken Sie auf den „Logistic Regression“-Schritt, um das Dialogfeld **„Logistic Regression“-Optionen** anzuzeigen.
3. Geben Sie einen **Modellnamen** ein, wenn Sie nicht den Standardnamen verwenden möchten.
4. Optional: Aktivieren Sie das Kästchen **Überschreiben**, um das vorhandene Modell mit neuen Daten zu überschreiben.
5. Klicken Sie auf das Dropdown-Menü **Zielfeld** und wählen Sie „Kategorisch“ aus.
6. Optional: Geben Sie einen **Beschreibung** des Modells ein.
7. Klicken Sie bei jedem Feld, dessen Daten zum Modell hinzugefügt werden sollen, auf **Einschließen**.
8. Geben Sie über das Dropdown-Menü **Modelldatentyp** an, ob das Eingabefeld als numerisches Feld, als kategorisches Feld oder als DateTime-Feld verwendet werden soll.
9. Klicken Sie auf **OK**, um das Modell und die Konfiguration zu speichern, oder fahren Sie mit der nächsten Registerkarte fort.

Konfigurieren von Standardoptionen

1. Lassen Sie **Eingabefelder standardisieren** aktiviert, um die numerischen Spalten zu standardisieren, damit diese keine Mittelwert- und Einheitenvarianz aufweisen.

   Wenn Sie keine Standardisierung verwenden, können die Ergebnisse Komponenten enthalten, die von Variablen dominiert werden, die statt richtiger Beiträge relativ zu anderen Attributen in der Skalierung eine größere Varianz zu haben scheinen.
2. Aktivieren Sie **Eingabedaten bewerten**, um eine Spalte für die Modellvorhersage (Punktzahl) für Eingabedaten hinzuzufügen.
3. Aktivieren Sie **Vorherig**, wenn die Daten erfasst wurden und die Bedeutung der Antwort nicht die Realität widerspiegelt. Geben Sie anschließend die vorherige Wahrscheinlichkeit für p(y==1) in das Textfeld ein.
4. Geben Sie an, wie mit fehlenden Daten umgegangen werden soll, indem Sie **Überspringen** aktivieren oder **Mittelwerte zuschreiben**, wodurch der Mittelwert für fehlende Daten hinzugefügt wird.
5. Geben Sie für den **Prozentsatz für Trainingsdaten** einen Wert zwischen 1 und 100 an, wenn die Eingabedaten zufällig in Stichproben für Trainings- und Testdaten aufgeteilt werden.
6. Geben Sie den Wert 100 abzüglich der in Schritt 5 eingegebenen Menge als **Prozentsatz für Testdaten** ein.
7. Geben Sie eine Ziffer als **Ausgangswert für Stichprobe** ein, um sicherzustellen, dass die Darstellung der Daten bei jeder Datenflussausführung gleich ist, wenn diese in Test- und Trainingsdaten aufgeteilt werden. Lassen Sie den Wert „0“ in diesem Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.
8. Klicken Sie auf **OK**, um das Modell und die Konfiguration zu speichern, oder fahren Sie mit der nächsten Registerkarte fort.

Konfigurieren erweiterter Optionen

1. Lassen Sie **Konstante Felder ignorieren** aktiviert, damit Felder übersprungen werden, die für die einzelnen Datensätze die gleichen Werte enthalten.
2. Lassen Sie **p-Werte berechnen** aktiviert, um p-Werte für die Parameterschätzungen zu berechnen.
3. Lassen Sie **Kollineare Spalte entfernen** aktiviert, damit kollineare Spalten während der Modellerstellung automatisch entfernt werden. Dies führt zu einem Koeffizienten von 0 im zurückgegebenen Modell.

   Diese Option muss aktiviert werden, wenn **p-Werte berechnen** ebenfalls aktiviert ist.

4. Lassen Sie **Konstanten Begriff einschließen (abfangen)** aktiviert, um einen konstanten Begriff im Modell einzuschließen (abzufangen).

   Dieses Feld muss aktiviert werden, wenn **Kollineare Spalte entfernen** ebenfalls aktiviert ist.

5. Wählen Sie einen **Solver** aus der Dropdown-Liste aus. Beachten Sie, dass COORDINATE_DESCENT und COORDINATE_DESCENT_NAIVE derzeit experimentell sind.

AUTO	Solver wird basierend auf Eingabedaten und Parametern bestimmt.
COORDINATE_DESCENT	IRLSM mit der Version der Kovarianzaktualisierungen der zyklischen Koordinate, die aus der innersten Schleife stammt.
COORDINATE_DESCENT_NAIVE	IRLSM mit der Version der naiven Aktualisierungen der zyklischen Koordinate, die aus der innersten Schleife stammt.
IRLSM	Ideal für Probleme mit einer geringen Anzahl von Prädiktoren oder Lambda-Suchvorgänge mit L1-Penalty.
L_BFGS	Ideal für Datasets mit vielen Spalten.

6. Lassen Sie **Seed für N-fach** aktiviert, und geben Sie einen numerischen Ausgangswert ein, um sicherzustellen, dass die Darstellung der Daten bei jeder Datenflussausführung gleich ist, wenn diese in Test- und Trainingsdaten aufgeteilt werden. Lassen Sie den Wert „0“ in diesem Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.
7. Aktivieren Sie **N-fach** und geben Sie die Anzahl der Folds ein, wenn Sie eine Kreuzvalidierung durchführen.
8. Aktivieren Sie **Foldzuweisung**, und wählen Sie aus der Dropdown-Liste aus, ob Sie eine Kreuzvalidierung durchführen. Dieses Feld ist nur anwendbar, wenn Sie unter **N-fach** einen Wert eingegeben haben und **Foldfeld** nicht angegeben ist.

AUTO	Lässt zu, dass der Algorithmus automatisch eine Option auswählt; derzeit wird „Random“ verwendet.
Modulo	Teilt das Dataset gleichmäßig auf die Folds auf und ist nicht vom Ausgangswert abhängig.
Random	Teilt die Daten zufällig in „N-fach“-Bestandteile ein; diese Einstellung ist für umfangreiche Datasets am besten geeignet.
Stratified	Schichtet die Folds basierend auf der Antwortvariable für Klassifizierungsprobleme. Verteilt Beobachtungen aus den verschiedenen Klassen gleichmäßig auf alle Datasets, wenn ein Dataset in Trainings- und Testdaten aufgeteilt wird. Dies kann nützlich sein, wenn viele Klassen vorhanden sind und das Dataset relativ klein ist.

9. Wenn Sie eine Kreuzvalidierung durchführen, aktivieren Sie **Foldfeld** und wählen Sie aus der Dropdown-Liste das Feld aus, das die Foldindexzuweisung für die Kreuzvalidierung enthält.

Dieses Feld ist nur anwendbar, wenn Sie unter **N-fach** und **Foldzuweisung** keinen Wert eingegeben haben.

10. Aktivieren Sie **Max. Iterationen** und geben Sie die Anzahl der Trainingsiterationen ein, die erfolgen sollen.
11. Aktivieren Sie **Ziel-Epsilon** und geben Sie den Schwellenwert für die Konvergenz an. Dieser Wert musst zwischen 0 und 1 liegen. Wenn der Zielwert geringer ist als dieser Schwellenwert, wird das Modell konvergiert.
12. Aktivieren Sie **Beta-Epsilon** und geben Sie den Schwellenwert für die Konvergenz an. Dieser Wert musst zwischen 0 und 1 liegen. Wenn der Zielwert geringer ist als dieser Schwellenwert, wird das Modell konvergiert. Wenn die L1-Normalisierung der aktuellen Beta-Änderung unter diesem Schwellenwert liegt, sollten Sie die Verwendung der Konvergenz in Erwägung ziehen.
13. Klicken Sie auf **OK**, um das Modell und die Konfiguration zu speichern, oder fahren Sie mit der nächsten Registerkarte fort.

Modellausgabe

Auf dieser Registerkarte werden die Metriken angezeigt, mit denen Sie das angepasste Modell bewerten. Diese Felder können nicht bearbeitet werden. Die Spalte „Training“ enthält immer Daten. Wenn Sie auf der Registerkarte „Standardoptionen“ eine Aufteilung in Training/Test ausgewählt haben, wird die Spalte „Test“ ebenfalls aufgefüllt. Eine Ausnahme besteht, wenn Sie auf der Registerkarte „Erweiterte Optionen“ eine Validierung vom Typ „N-fach“ ausgewählt haben. In diesem Fall wird die Spalte „N-fach“ aufgefüllt.

Nachdem Sie Ihren Auftrag ausgeführt haben, wird das resultierende Modell auf dem Spectrum™ Technology Platform-Server gespeichert. Klicken Sie auf die Schaltfläche **Ausgabe**, um die Ausgabe erneut zu generieren, und klicken Sie auf **Für Details hier klicken**, um in dem Toll für die „Machine Learning“-Modellverwaltung die gesamte Ausgabe anzuzeigen.

5 - Java Model Scoring

In this section

Einführung in Java Model Scoring

Java Model Scoring ermöglicht Ihnen die Bewertung neuer Daten mithilfe der Formel, die beim Anpassen eines Machine Learning-Modells erstellt wird.

> **Anmerkung:** Modelle müssen zunächst über die Machine Learning-Modellverwaltung verfügbar gemacht werden, bevor sie im „Java Model Scoring“-Schritt verfügbar werden. Weitere Informationen finden Sie unter **Einführung in die Machine Learning-Modellverwaltung** auf Seite 24.

Sie müssen zwei Registerkarten im Dialogfeld **„Java Model Scoring“-Optionen** ausfüllen, um Ihre Daten bewerten zu können. Identifizieren Sie zunächst das Modell und dessen Typ, und stellen Sie anschließend sicher, dass die Felder des Modells den Spectrum™ Technology Platform-Feldern ordnungsgemäß zugeordnet wurden. Im Anschluss können Sie die Ausgabe konfigurieren, indem Sie auswählen, welche Felder eingeschlossen werden sollen, und Ihren Auftrag ausführen. Die Registerkarte **Modellausgabe** enthält das Mapping für Datentypen der Spectrum™ Technology Platform und Ihr Modell.

Wenn Ihr Auftrag einen Schritt enthält, der die Ausgabe in einer Datei oder Tabelle erfasst, können Sie diese Ausgabe in einem nachfolgenden Datenfluss oder Webservice verwenden.

Definieren von Modelleigenschaften

1. Klicken Sie unter **Primäre Schritte**/**Bereitgestellte Schritte**/**Advanced Analytics** auf den **Java Model Scoring**-Schritt und ziehen Sie ihn auf die Arbeitsfläche. Platzieren Sie ihn an einer beliebigen Stelle im Datenfluss und verbinden Sie ihn mit Eingabe- und Ausgabeschritten. Beachten Sie, dass der Eingabeschritt die Datenquelle sein muss, die Ziel- und Eingabevariablenfelder für Ihr Modell enthält. Wenn Sie Ihren Auftrag im Batchmodus ausführen, benötigen Sie zudem einen Ausgabeschritt für die Erfassung von Modellbewertungen; andernfalls können Sie Daten mithilfe eines Spectrum™ Technology Platform-Webservice in Echtzeit bewerten.
2. Doppelklicken Sie auf den „Java Model Scoring“-Schritt, um das Dialogfenster **„Model Scoring“-Optionen** anzuzeigen.
3. Optional: Wählen Sie im Dropdown-Menü **Typfilter** den Typ eines Modells aus, das Sie bewerten.
4. Wählen Sie den **Typfilter** aus, der für die Bewertung des Modells verwendet wird.
5. Wählen Sie den **Modellnamen** aus dem Dropdown-Menü aus.
6. Geben Sie im Feld **Modelltyp** den Typ des Modells aus, das Sie bewerten.
7. Optional: Geben Sie einen **Beschreibung** des Modells ein.

8. Die Tabelle **Eingaben** enthält Informationen zu den Eingabefeldern des Modells. Diese Felder und die zugehörigen Datentypen ordnen Spectrum automatisch Felder und Datentypen zu.
9. Klicken Sie auf **OK**, um diese Optionen zu speichern, oder fahren Sie mit der nächsten Registerkarte fort.

Modellausgabe

Die Tabelle **Ausgaben** enthält Informationen zu den Ausgabefeldern des Modells. Diese Felder und die zugehörigen Datentypen ordnen Spectrum automatisch Felder und Datentypen zu.

1. Klicken Sie bei jedem Feld, dessen Daten im Modell eingeschlossen werden sollen, auf **Einschließen**.
2. Klicken Sie auf **OK**, um das Modell zu speichern.

6 - Machine Learning-Modellverwaltung

In this section

Einführung in die Machine Learning-Modellverwaltung

Auf der Registerkarte „Modellanalyse“ bei der Machine Learning-Modellverwaltung wird eine Liste mit allen Machine Learning-Modellen auf Ihrem Spectrum™ Technology Platform-Server angezeigt. Sie können diese Liste filtern, indem Sie eine Zeichenfolge in das Textfeld eingeben. Jedes Feld in der Tabelle wird dann nach dieser Zeichenfolge durchsucht.

In diesen Modellen können mehrere Vorgänge durchgeführt werden. Sie können Modelle verfügbar machen, die Verfügbarkeit von Modellen aufheben oder Modelle löschen. Mit den verfügbar gemachten Modellen werden im „Java Model Scoring“-Schritt mithilfe der beim Anpassen der Machine Learning-Modelle erstellten Formeln neue Daten bewertet. Zusätzlich können Sie zu jedem Modell detaillierte Informationen anzeigen. Welche Details zurückgegeben werden, hängt von dem Modelltyp ab, dessen Daten Sie anzeigen. Schließlich können Sie zwei beliebige Modelle des gleichen Typs miteinander vergleichen. Bei diesem Vergleich werden für alle Modelle, die Sie miteinander vergleichen, die gleichen Informationen nebeneinander angezeigt, die auf der Registerkarte „Modelldetail“ enthalten sind.

Zugreifen auf die Modellanalyse bei der Machine Learning-Modellverwaltung

Es gibt drei Möglichkeiten für den Zugriff auf die Machine Learning-Modellverwaltung:

- Verwenden Sie die Begrüßungsseite der Spectrum™ Technology Platform:
  - Öffnen Sie einen Webbrowser und navigieren Sie zur Spectrum™ Technology Platform-Begrüßungsseite unter:

    http://<servername>:<port>

    Wenn Sie beispielsweise Spectrum™ Technology Platform auf einem Computer mit dem Namen „myspectrumplatform“ installiert haben und dieser den HTTP-Standardport 8080 verwendet, navigieren Sie zu:

    http://myspectrumplatform:8080

  - Klicken Sie auf **Spectrum Machine Learning**.
  - Klicken Sie auf **„Machine Learning“-Datenbank öffnen**.
- Klicken Sie bei einem der Schritte zur Modellerstellung auf **Für Modelldetails hier klicken**.
- Verwenden Sie einen Webbrowser:
  - Öffnen Sie einen Webbrowser und navigieren Sie zur Seite „Machine Learning-Modellverwaltung“ der Spectrum™ Technology Platform unter:

    http://<Servername>:<Port>/machinelearning

Wenn Sie beispielsweise Spectrum™ Technology Platform auf einem Computer mit dem Namen „myspectrumplatform“ installiert haben und dieser den HTTP-Standardport 8080 verwendet, navigieren Sie zu:

http://myspectrumplatform:8080/machinelearning

- Geben Sie einen gültigen Benutzernamen und das zugehörige Kennwort für die Spectrum™ Technology Platform ein.
- Klicken Sie auf die Registerkarte **Modellanalyse**, wenn das Tool geöffnet wird.

Vorgänge zur Modellverwaltungsanalyse

Führen Sie die folgenden Vorgänge durch, indem Sie ein Modell auswählen und auf die entsprechende Schaltfläche klicken:

	Machen Sie das Modell verfügbar, damit es für den „Java Model Scoring“-Schritt verfügbar ist. Wenn ein Modell nicht verfügbar gemacht wird, kann es nicht für Bewertungen verwendet werden.
	Heben Sie die Verfügbarkeit des Modells auf.
	Löschen Sie das Modell. **Anmerkung:** Ein verfügbar gemachtes Modell kann nicht gelöscht werden. Zu diesem Zeitpunkt ist jedoch keine inhärente Sicherheit vorhanden, durch die verhindert wird, dass ein Benutzer die Modelle eines anderen Benutzers löscht.
	Zeigen Sie Details zur Modellausgabe an. Sie können auch über die „K-Means Clustering“- und „Logistic Regression“-Schritte auf diese Informationen zugreifen, indem Sie auf der Registerkarte „Modellausgabe“ auf „Für Modelldetails hier klicken“ klicken.
	Vergleichen Sie die Modelle miteinander.

Die Registerkarte „Modelldetail“

Auf dem Bildschirm „Modelldetail“ werden folgende Informationen für ein Modell angezeigt:

- **Modellname**: Der Name des Modells
- **Modelltyp**: Der Typ des Machine Learning-Modells

- **Benutzer**: Der Benutzername der Person, die das Modell erstellt hat
- **Beschreibung**: Die Beschreibung des Modells, wenn bei der Erstellung des Modells eine Beschreibung angegeben wurde
- **Status**: Gibt an, ob das Modell verfügbar gemacht wurde oder ob die Verfügbarkeit aufgehoben wurde
- **Datenflussname**: Der Name des Datenflusses, der das Modell erzeugt hat
- **Erstellungszeit**: Das Datum und die Uhrzeit der Modellerstellung

Zusätzliche Details werden auf Grundlage des Modelltyps bereitgestellt.

„K-Means Clustering“-Details

Auf dem Bildschirm „Modelldetail“ werden folgende Informationen für „K-Means Clustering“-Modelle angezeigt:

Modellübersicht

- Anzahl der Zeilen
- Anzahl der Cluster
- Anzahl der kategorischen Spalten
- Anzahl der Iterationen
- Innerhalb der Cluster-Summe von Quadraten
- Gesamtsumme von Quadraten
- Zwischen der Cluster-Summe von Quadraten

Metriken

Stellt für folgende Informationen Trainings-, Test- und „N-fach“-Daten bereit:

- Innerhalb der Cluster-Gesamtsumme von Quadraten
- Gesamtsumme von Quadraten
- Zwischen der Cluster-Summe von Quadraten

Zentroidstatistik

Stellt für die einzelnen Zentroide folgende Trainings-, Test- und „N-fach“-Daten bereit:

- Größe
- Innerhalb der Cluster-Summe von Quadraten

Cluster-Mittelwerte

Stellt für jeden Zentroid detaillierte Informationen bereit. Der Inhalt variiert je nach Eingabedaten. Ein Cluster stellt eine Gruppe von Beobachtungen aus einem Dataset dar, die gemäß eines bestimmten Clustering-Algorithmus als ähnlich identifiziert wurden

Standardisierte Cluster-Mittelwerte

Stellt für jeden Zentroid standardisierte Informationen bereit. Der Inhalt variiert je nach Eingabedaten.

Details zu „Logistic Regression“

Auf dem Bildschirm „Modelldetail“ werden folgende Informationen für „Logistic Regression“-Modelle angezeigt:

Metriken

Stellt für folgende Informationen Trainings-, Test- und „N-fach“-Daten bereit:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Anzahl der Beobachtungen
- R-Quadrat (R2)
- Logarithmischer Abfall (Logloss)
- Area under the curve (AUC)
- Gini-Koeffizient
- Mean Per Class Error
- AIC
- Restabweichung
- Abweichung von Null
- Freiheitsgrad von Null
- Restfreiheitsgrad

Maximum Metrics Threshold

Gibt den Training Maximum Metrics Threshold für Trainings-, Text- und „N-fach“-Daten mithilfe der folgenden Metriken an:

- 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

Konfusionsmatrix

Stellt die Leistung eines Modells in einer Reihe von Trainings-, Test- und „N-fach“-Daten bereit, bei denen die tatsächlichen Werte bekannt sind.

Standardisiertes Koeffizientendiagramm

Zeigt die wichtigsten Prädiktoren an, indem der relative Wert der Koeffizienten angegeben wird. Dieser gibt an, wie stark sich das Ziel durch eine Änderung der Eingabe verändert.

GLM-Koeffizienten

Koeffizienten für ein Generalized Linear-Modell, das Regressionsmodelle für Ergebnisse nach Exponentialverteilungen schätzt.

AUC-Kurven

Area under the curve (Fläche unter der Kurve); bestimmt, welches der Modelle die Klassen mithilfe der Trainings-, Test- und „N-fach“-Daten am besten vorhersagt.

Anhebungs-/Verstärkungskurven

Wertet die Fähigkeit des binären Klassifizierungsmodells aus, mithilfe der Trainings-, Test- und „N-Fold“-Daten Vorhersagen zu treffe.

Notices

USPS® Urheberrechtshinweise

Pitney Bowes Inc. wurde eine nicht-ausschließliche Lizenz erteilt, die die Veröffentlichung und den Verkauf von ZIP + 4® Postleitzahl-Datenbanken auf optischen und magnetischen Medien genehmigt. Folgende Marken sind Markenzeichen des 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, und ZIP + 4. Hierbei handelt es sich jedoch nicht um eine vollständige Liste der Marken, die zum United States Postal Service gehören.

Pitney Bowes Inc. ist nicht-exklusiver Lizenznehmer von USPS® für die Verarbeitungsprozesse von NCOA^{Link}®.

Die Preisgestaltung jeglicher Pitney Bowes Softwareprodukte, -optionen und -dienstleistungen erfolgt nicht durch USPS® oder die Regierung der Vereinigten Staaten. Es wird auch keine Regulierung oder Genehmigung der Preise durch USPS® oder die US-Regierung durchgeführt. Bei der Verwendung von RDI™-Daten zur Berechnung von Paketversandkosten wird die Entscheidung, welcher Paketlieferdienst genutzt wird, nicht von USPS® oder der Regierung der Vereinigten Staaten getroffen.

Datenbereitstellung und Hinweise

Hier verwendete Datenprodukte und Datenprodukte, die in Software-Anwendungen von Pitney Bowes verwendet werden, sind durch verschiedene Markenzeichen und mindestens eines der folgenden Urheberrechte geschützt:

© Copyright United States Postal Service. Alle Rechte vorbehalten.

© 2014 TomTom. Alle Rechte vorbehalten. TomTom und das TomTom Logo sind eingetragene Marken von TomTom N.V.

© 2016 HERE

Fuente: INEGI (Instituto Nacional de Estadística y Geografía)

Basierend auf elektronischen Daten © National Land Survey Sweden.

© Copyright United States Census Bureau

© Copyright Nova Marketing Group, Inc.

Teile dieses Programms sind urheberrechtlich geschützt durch © Copyright 1993-2007 Nova Marketing Group Inc. Alle Rechte vorbehalten.

© Copyright Second Decimal, LLC

© Copyright Canada Post Corporation

3001 Summer Street
Stamford CT 06926-0700
USA

www.pitneybowes.com