pitney bowes

Spectrum™ Technology Platform

Version 2018.2.0

Machine Learning-Handbuch

Inhalt

9 - Machine Learning-Modellverwaltung

10 - Data Science-Demonstrationsfluss

1 - Einführung

In this section

Machine Learning-Modul

Das Spectrum™ Technology Platform Machine Learning-Modul bietet die Möglichkeit, ein Binning für numerische Daten durchzuführen sowie überwachte und unüberwachte Machine Learning-Modelldaten in diese Modelle einzupassen.

> **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.

Linear Regression

Mit „Linear Regression“ können Sie Machine Learning durchführen, indem Sie Modelle aus Datasets erstellen, die kontinuierliche Ziele mit Eingabevariablen verwenden.

Logistic Regression

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

Principal Component Analysis

„Principal Component Analysis“ ist ein statistisches Verfahren, das einen Beobachtungssatz von möglicherweise korrelierten Variablen in einen Wertesatz von linear nicht korrelierten Variablen (prinzipielle Komponenten) umwandelt.

Random Forest Classification

Mit „Random Forest Classification“ können Sie Machine Learning durchführen, indem Sie Modelle aus Datasets erstellen, die kontinuierliche Ziele mit Eingabevariablen verwenden.

Random Forest Regression

Mit „Random Forest Regression“ können Sie Machine Learning durchführen, indem Sie Modelle aus Datasets erstellen, die kontinuierliche Ziele mit Eingabevariablen verwenden.

Machine Learning-Modellverwaltung

Die Machine Learning-Modellverwaltung umfasst die Modellbewertung, mit der Sie alle Machine Learning Modelle auf Ihrem Spectrum™ Technology Platform-Server verwalten können, und Binning Management, mit dem Sie alle Binnings auf Ihrem Spectrum™ Technology Platform-Server verwalten können.

> **Anmerkung:** Das Machine Learning-Modul verwendet eine zugrunde liegende H2O.ai-Bibliothek für Modellierungsalgorithmen in „K-Means Clustering“, „Linear Regression“, „Logistic Regression“, „Principal Component Analysis“, „Random Forest Classification“ und „Random Forest Regression“.

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.

Sie können weitere Binning-Funktionen mit dem Tool **Binning Management** der Machine Learning-Modellverwaltung durchführen.

Mithilfe von Befehlszeilenanweisungen können Sie eine Binning-Liste anzeigen und Binnings löschen. Weitere Informationen erhalten Sie unter „Machine Learning-Modul“ im Abschnitt **Administrationsumgebung** des Administratorhandbuchs.

> **Anmerkung:** Wenn Sie Spectrum™ Technology Platform von Version 12.0 SP1 auf 12.0 SP2 aktualisieren, müssen Sie für alle aktualisierten Binning-Instanzen im Tool „Binning Management“ der Machine Learning-Modellverwaltung die Verfügbarkeit aufheben, bevor Sie sie in 12.0 SP2 für das Rebinning verwenden. Dieser Schritt ist nicht erforderlich, wenn Sie in Binning Lookup aktualisierte Binnings anstelle von traditionellen Binnings verwenden.

Definieren von Binning-Eigenschaften

1. Klicken Sie unter **Primäre Schritte**/**Bereitgestellte Schritte**/**Machine Learning** auf den **Binning**-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 Binning-Schritt, um das Dialogfeld **Binning-Optionen** anzuzeigen.
3. Geben Sie einen **Binning-Namen** ein, wenn Sie nicht den Standardnamen verwenden möchten.
4. Aktivieren Sie das Kontrollkästchen **Überschreiben**, um das vorhandene Modell mit neuen Daten zu überschreiben.
5. Geben Sie eine **Beschreibung** des Modells ein.
6. 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.
7. Klicken Sie auf **OK**, um Ihre Einstellungen zu speichern.

Konfigurieren von Standardoptionen

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:

- Name
- 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

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. Deaktivieren Sie dieses 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. Lassen Sie **Seed für Algorithmus** 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. Deaktivieren Sie dieses Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.
3. Wählen Sie im Dropdown-Menü **Init** den richtigen Initialisierungsmodus aus.

Furthest	Initialisiert den ersten Mittelpunkt zufällig; den zweiten Mittelpunkt initialisiert der Modus jedoch anschließend so, dass es der davon am weitesten entfernte Datenpunkt ist. Initialisiert die Mittelpunkte 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.

4. 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. Deaktivieren Sie dieses Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.
5. Aktivieren Sie **N-fach** und geben Sie die Anzahl der Folds ein, wenn Sie eine Kreuzvalidierung durchführen.
6. Aktivieren Sie **Faktorzuweisung**, 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.

7. Aktivieren Sie **Maximale Iterationen** und geben Sie die Anzahl der Trainingsiterationen ein, die erfolgen sollen.
8. 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 **Modelldetails**, um in dem Tool für die „Machine Learning“-Modellverwaltung die gesamte Ausgabe anzuzeigen.

Ausgabeport

Der Schritt „K-Means Clustering“ enthält einen optionalen Ausgabeport: den Modellmetrikport. Die Funktion dieses Ports wird durch Ihre Auswahl und Eingabe bestimmt, wenn Sie die grundlegenden und erweiterten Optionen des Schrittes abschließen. Wenn Sie zum Beispiel die N-fache Validierung durchführen, indem Sie das Feld **N-fach** auf der Registerkarte „Erweiterte Optionen“ markieren, wird die Spalte „N-Fach“ in den Ausgabemetriken mit Daten gefüllt. Wenn Sie alternativ keine N-fache Validierung durchführen, ist die Spalte „N-Fach“ leer.

Modellmetrikport

Mit dem **Modellmetrikport** können Sie die Modellbewertungsmetriken in eine Datendatei ausgeben. Dies wird Ihnen helfen, viele Modelle, die innerhalb und außerhalb von Spectrum™ Technology Platform generiert wurden, zu vergleichen und andere Datenverarbeitungsaufgaben an den Metriken auszuführen.

Gehen Sie folgendermaßen vor, um den Modellmetrikport zu verwenden:

1. Öffnen Sie einen Datenfluss, der den Schritt „K-Means Clustering“ verwendet.
2. Hängen Sie einen „Write to File“-Schritt oder einen anderen Datenausgabeschritt an den zweiten Ausgabeport an.
3. Führen Sie den Auftrag aus.
4. Alternative zu Schritt 3: Fügen Sie einen Prüfpunkt zu dem Kanal hinzu, der den Schritt „K-Means Clustering“ mit dem in Schritt 2 hinzugefügten Zielschritt verbindet, indem Sie mit der rechten Maustaste auf den Kanal klicken und „Überprüfungspunkt hinzufügen“ auswählen. Klicken Sie dann in der Enterprise Designer-Symbolleiste auf die Schaltfläche „Aktuellen Fluss überprüfen“ (). Die Überprüfung wird ausgeführt und Sie sollten ähnliche Ergebnisse wie die unten dargestellten sehen.

Creation Time	Flow Name	Metrics	Model Name	Model Type	N Fold	Test	Training
10/11/2018 2:37:00 AM	LinearRegressionTest1...	MSE	LinearRegressionTest1...	Linear Regression	0.0050019841090508	0.00321814876650744	1.31531988837612E-10
10/11/2018 2:37:00 AM	LinearRegressionTest1...	RMSE	LinearRegressionTest1...	Linear Regression	0.0707247064967455	0.0567287296394643	1.14687396359675E-05
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Number of observations	LinearRegressionTest1...	Linear Regression	33	17	33
10/11/2018 2:37:00 AM	LinearRegressionTest1...	R2	LinearRegressionTest1...	Linear Regression	-0.366538763835926	0.0268719304981138	0.999999964065547
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Mean residual deviance	LinearRegressionTest1...	Linear Regression	0.0050019841090508	0.00321814876650744	1.31531988837612E-10

4 - Linear Regression

In this section

Einführung

Mit Linear Regression können Sie Machine Learning durchführen, indem Sie Modelle aus Datasets erstellen, die im Hinblick auf Eingabevariablen kontinuierliche 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 **Linear 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 „Linear Regression“-Schritt, um das Dialogfeld **„Linear Regression“-Optionen** anzuzeigen.
3. Geben Sie einen **Modellnamen** ein, wenn Sie nicht den Standardnamen verwenden möchten.
4. Aktivieren Sie das Kontrollkästchen **Überschreiben**, um das vorhandene Modell mit neuen Daten zu überschreiben.
5. Klicken Sie auf das Dropdown-Menü **Zielfeld**, und wählen Sie ein numerisches Feld aus.
6. Geben Sie eine **Beschreibung** des Modells ein.
7. Klicken Sie bei jedem Feld, dessen Daten zum Modell hinzugefügt werden sollen, auf **Einschließen**. Achten Sie darauf, das Feld einzuschließen, das Sie als Zielfeld ausgewählt haben.
8. Geben Sie über das Dropdown-Menü **Modelldatentyp** an, ob die einzelnen Eingabefelder als numerisches Feld, als kategorisches Feld oder als DateTime-Feld verwendet werden sollen.
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. Wählen Sie eine **Verknüpfungsfunktion** aus der Dropdown-Liste aus. Dies spezifiziert die Verknüpfung zwischen zufälligen und systematischen Komponenten. Es besagt, in welcher Beziehung der erwartete Wert der Antwort mit dem linearen Prädiktor erklärender Variablen steht.

Identität	Sagt sinnlose „Wahrscheinlichkeiten“ von unter null oder größer als eins voraus. Wird manchmal für binominale Daten verwendet, um ein lineares Wahrscheinlichkeitsmodell zu erzielen. $g(p) = p$
Invers	Berechnet die Umkehrung von Verknüpfungsfunktionen für reelle Schätzungen. $g(\mu i)=1\mu i$
Protokoll	Zählt Vorkommen in einer festen Zeitspanne und einem festen Bereich. $g(\mu i)=log(\mu i)$

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. Deaktivieren Sie dieses 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. Aktivieren Sie **p-Werte berechnen**, um p-Werte für die Parameterschätzungen zu berechnen.
3. Aktivieren Sie **Kollineare Spalte entfernen**, 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 „CoordinateDescent“ und „CoordinateDescentNaive“ derzeit experimentell sind.

Auto	Solver wird basierend auf Eingabedaten und Parametern bestimmt.
CoordinateDescent	IRLSM mit der Version der Kovarianzaktualisierungen der zyklischen Koordinate, die aus der innersten Schleife stammt.
CoordinateDescentNaive	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.
LBFGS	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. Deaktivieren Sie dieses 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. Klicken Sie auf **Faktorzuweisung**, 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 **Faktorfeld** 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.

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

   Dieses Feld ist nur anwendbar, wenn Sie unter **N-fach** und **Faktorzuweisung** keinen Wert eingegeben haben.
10. Aktivieren Sie **Maximale 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. Ein häufiges Problem beim prädiktiven Modeling ist die Überanpassung, wenn ein Analytical Model einem bestimmten Dataset zu sehr (oder genau) entspricht und daher bei der Anwendung auf zusätzliche Daten oder künftige Beobachtungen nicht erfolgreich ist. Eine Methode, um Überanpassungen zu vermeiden, ist die Regularisierung. Wählen Sie den zu verwendenden **Regularisierungstyp** aus.

LASSO (Geringster absoluter Schrumpf- und Selektionsoperator)	Wählt eine kleine Teilmenge von Variablen mit einem Wert von Lambda aus, der hoch genug ist, um als entscheidend angesehen zu werden. Dies könnte bei korrelierten Prädiktorvariablen nicht gut funktionieren, da eine Variable der korrelierten Gruppe ausgewählt und alle anderen Variablen entfernt werden. Dies wird auch durch hohe Dimensionalität begrenzt; wenn ein Modell mehr Variablen als Datensätze enthält, ist LASSO darauf beschränkt, wie viele Variablen es auswählen kann. „Ridge Regression“ hat diese Einschränkung nicht. Wenn die Anzahl der im Modell enthaltenen Variablen groß ist oder wenn bekannt ist, dass die Lösung spärlich ist, wird LASSO empfohlen.
Ridge Regression	Behält alle Prädiktorvariablen bei und verkleinert ihre Koeffizienten proportional. Wenn korrelierte Prädiktorvariablen vorhanden sind, reduziert „Ridge Regression“ die Koeffizienten der gesamten Gruppe korrelierter Variablen auf Gleichheit. Wenn Sie nicht möchten, dass korrelierte Prädiktorvariablen aus Ihrem Modell entfernt werden, verwenden Sie „Ridge Regression“.
Elastic Net	Kombiniert LASSO und „Ridge-Regression“, indem es als Variablenselektor fungiert und gleichzeitig den Gruppierungseffekt für korrelierte Variablen beibehält (Koeffizienten der korrelierten Variablen werden gleichzeitig verkleinert). „Elastic Net“ ist nicht durch hohe

Dimensionalität eingeschränkt und kann alle Variablen auswerten, wenn ein Modell mehr Variablen als Datensätze enthält.

14. Überprüfen Sie den **Alpha-Wert**, und ändern Sie den Wert, wenn Sie nicht den Standardwert 0,5 verwenden möchten. Der Alpha-Parameter steuert die Verteilung zwischen den Abzügen ℓ1 und ℓ2. Gültige Werte liegen zwischen 0 und 1; ein Wert von 1,0 stellt LASSO dar, und ein Wert von 0,0 erzeugt „Ridge Regression“. Die folgende Tabelle zeigt, wie Alpha und Lambda die Regularisierung beeinflussen.

`lambda` value	`alpha` value	Result
`lambda` == 0	`alpha` = any value	No regularization. `alpha` is ignored.
`lambda` > 0	`alpha` == 0	Ridge Regression
`lambda` > 0	`alpha` == 1	LASSO
`lambda` > 0	0 < `alpha` < 1	Elastic Net Penalty

Anmerkung: Das einfache Gleichheitszeichen ist ein Zuweisungsoperator, der „ist“ bedeutet. Das doppelte Gleichheitszeichen ist ein Gleichheitsoperator, der „gleich“ bedeutet.

15. Aktivieren Sie den **Wert von Lambda**, und geben Sie einen Wert an, wenn „Linear Regression“ nicht die Standardmethode zur Berechnung des Lambda-Wertes verwenden soll, bei der es sich um eine Heuristik handelt, die auf Trainingsdaten basiert. Der Lambda-Parameter steuert den Umfang der angewendeten Regularisierung. Wenn beispielsweise Lambda 0,0 ist, wird keine Regularisierung angewendet und der Alpha-Parameter wird ignoriert.
16. Aktivieren Sie **Nach optimalem Wert von Lambda suchen**, damit „Linear Regression“ Modelle für den vollständigen Regularisierungspfad berechnet, der beim maximalen Lambda-Wert beginnt (der höchste Lambdawert, der sinnvoll ist; d. h. der niedrigste Wert, der alle Koeffizienten auf Null treibt) und bis zum niedrigsten Lamdba-Wert auf der logarithmischen Skala reicht, wobei die Regularisierungsstärke bei jedem Schritt abnimmt. Das zurückgegebene Modell wird Koeffizienten aufweisen, die dem optimalen Lambda-Wert entsprechen, wie während des Trainings entschieden wurde.
17. Aktivieren Sie **Frühzeitig stoppen**, um die Verarbeitung zu beenden, wenn sich der Trainings- oder Validierungssatz nicht weiter verbessert.
18. Aktivieren Sie **Maximale zu suchende Lambdas**, und geben Sie die maximale Anzahl der Lambdas ein, die während der Lambda-Suche verwendet werden sollen.
19. Aktivieren Sie **Maximale aktive Prädiktoren**, und geben Sie die maximale Anzahl der Prädiktoren ein, die während der Berechnung verwendet werden sollen. Dieser Wert wird als Stoppkriterium verwendet, um einen teuren Modellaufbau mit vielen Prädiktoren zu verhindern.
20. 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 **Modelldetails**, um in dem Tool für die „Machine Learning“-Modellverwaltung die gesamte Ausgabe anzuzeigen.

Ausgabeports

Der Schritt „Linear Regression“ enthält zwei optionale Ausgabeports: den Modellbewertungsport und den Modellmetrikport. Die Funktion dieser Ports hängt von Ihrer Auswahl und Eingabe ab, wenn Sie die grundlegenden und erweiterten Optionen des Schrittes ausführen. Wenn Sie zum Beispiel die n-fache Validierung durchführen, indem Sie das Feld **N-fach** auf der Registerkarte „Erweiterte Optionen“ markieren, wird die Spalte „N-fach“ in den vom Modellmetrikport generierten Ausgabemetriken mit Daten gefüllt. Wenn Sie alternativ keine N-fache Validierung durchführen, ist die Spalte „N-Fach“ leer. Der Modellbewertungsport wird ebenfalls aktiviert, wenn Sie das Feld **Eingabedaten bewerten** auf der Registerkarte „Standardoptionen“ markieren.

Modellbewertungsport

Wenn Sie das Feld **Eingabedaten bewerten** auf der Registerkarte „Standardoptionen“ markieren, teilt dies der „Linear Regression“ die Berechnung vorhergesagter Werte beim Erstellen des Modells mit, wobei wiederum die Spalte **Predicted_Value** für diese Punktzahl in den Ausgabedaten hinzugefügt wird. An diesen Port können Sie ein beliebiges Ziel anhängen: einen Write To File-Schritt, einen Write To Null-Schritt usw.

Modellmetrikport

Mit dem **Modellmetrikport** können Sie die Modellbewertungsmetriken in eine Datendatei ausgeben. Dies wird Ihnen helfen, viele Modelle, die innerhalb und außerhalb von Spectrum™ Technology Platform generiert wurden, zu vergleichen und andere Datenverarbeitungsaufgaben an den Metriken auszuführen.

Gehen Sie folgendermaßen vor, um den Modellmetrikport zu verwenden:

1. Öffnen Sie einen Datenfluss, der den Schritt „Linear Regression“ verwendet.
2. Hängen Sie einen „Write to File“-Schritt oder einen anderen Datenausgabeschritt an den zweiten Ausgabeport an.
3. Führen Sie den Auftrag aus.
4. Alternative zu Schritt 3: Fügen Sie einen Prüfpunkt zu dem Kanal hinzu, der den Schritt „Linear Regression“ mit dem in Schritt 2 hinzugefügten Zielschritt verbindet, indem Sie mit der rechten Maustaste auf den Kanal klicken und „Überprüfungspunkt hinzufügen“ auswählen. Klicken Sie dann in der Enterprise Designer-Symbolleiste auf die Schaltfläche „Aktuellen Fluss überprüfen“ (). Die Überprüfung wird ausgeführt und Sie sollten ähnliche Ergebnisse wie die unten dargestellten sehen.

Creation Time	Flow Name	Metrics	Model Name	Model Type	N Fold	Test	Training
10/11/2018 2:37:00 AM	LinearRegressionTest1...	MSE	LinearRegressionTest1...	Linear Regression	0.005001984109050 8	0.0032181487665074 4	1.31531988837612E-10
10/11/2018 2:37:00 AM	LinearRegressionTest1...	RMSE	LinearRegressionTest1...	Linear Regression	0.0707247064967455	0.0567287296394643	1.14687396359675E-05
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Number of observations	LinearRegressionTest1...	Linear Regression	33	17	33
10/11/2018 2:37:00 AM	LinearRegressionTest1...	R2	LinearRegressionTest1...	Linear Regression	-0.366538763835926	0.0268719304981138	0.999999964065547
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Mean residual deviance	LinearRegressionTest1...	Linear Regression	0.005001984109050 8	0.0032181487665074 4	1.31531988837612E-10

5 - Logistic Regression

In this section

Einführung

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. Aktivieren Sie das Kontrollkästchen **Überschreiben**, um das vorhandene Modell mit neuen Daten zu überschreiben.
5. Klicken Sie auf das Dropdown-Menü **Zielfeld** und wählen Sie „Kategorisch“ aus.
6. Geben Sie eine **Beschreibung** des Modells ein.
7. Klicken Sie bei jedem Feld, dessen Daten zum Modell hinzugefügt werden sollen, auf **Einschließen**. Achten Sie darauf, das Feld einzuschließen, das Sie als Zielfeld ausgewählt haben.
8. Geben Sie über das Dropdown-Menü **Modelldatentyp** an, ob die einzelnen Eingabefelder als numerisches Feld, als kategorisches Feld oder als DateTime-Feld verwendet werden sollen.
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. Deaktivieren Sie dieses 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 „CoordinateDescentNaive“ und „CoordinateDescentNaive“ derzeit experimentell sind.

Auto	Solver wird basierend auf Eingabedaten und Parametern bestimmt.
CoordinateDescentNaive	IRLSM mit der Version der Kovarianzaktualisierungen der zyklischen Koordinate, die aus der innersten Schleife stammt.
CoordinateDescentNaive	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. Deaktivieren Sie dieses 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 **Faktorzuweisung**, 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 **Faktorfeld** 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 **Faktorfeld** und wählen Sie aus der Dropdown-Liste das Feld aus, das die Faktorindexzuweisung für die Kreuzvalidierung enthält.

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

10. Aktivieren Sie **Maximale 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. Ein häufiges Problem beim prädiktiven Modeling ist die Überanpassung, wenn ein Analytical Model einem bestimmten Dataset zu sehr (oder genau) entspricht und daher bei der Anwendung auf zusätzliche Daten oder künftige Beobachtungen nicht erfolgreich ist. Eine Methode, um Überanpassungen zu vermeiden, ist die Regularisierung. Wählen Sie den zu verwendenden **Regularisierungstyp** aus.

LASSO (Geringster absoluter Schrumpf- und Selektionsoperator)	Wählt eine kleine Teilmenge von Variablen mit einem Wert von Lambda aus, der hoch genug ist, um als entscheidend angesehen zu werden. Dies könnte bei korrelierten Prädiktorvariablen nicht gut funktionieren, da eine Variable der korrelierten Gruppe ausgewählt und alle anderen Variablen entfernt werden. Dies wird auch durch hohe Dimensionalität begrenzt; wenn ein Modell mehr Variablen als Datensätze enthält, ist LASSO darauf beschränkt, wie viele Variablen es auswählen kann. „Ridge Regression“ hat diese Einschränkung nicht. Wenn die Anzahl der im Modell enthaltenen Variablen groß ist oder wenn bekannt ist, dass die Lösung spärlich ist, wird LASSO empfohlen.
Ridge Regression	Behält alle Prädiktorvariablen bei und verkleinert ihre Koeffizienten proportional. Wenn korrelierte Prädiktorvariablen vorhanden sind, reduziert „Ridge Regression“ die Koeffizienten der gesamten Gruppe korrelierter Variablen auf Gleichheit. Wenn Sie nicht möchten, dass korrelierte Prädiktorvariablen aus Ihrem Modell entfernt werden, verwenden Sie „Ridge Regression“.
Elastic Net	Kombiniert LASSO und „Ridge-Regression“, indem es als Variablenselektor fungiert und gleichzeitig den Gruppierungseffekt für korrelierte Variablen beibehält (Koeffizienten der korrelierten Variablen werden gleichzeitig verkleinert). „Elastic Net“ ist nicht durch hohe Dimensionalität eingeschränkt und kann alle Variablen auswerten, wenn ein Modell mehr Variablen als Datensätze enthält.

14. Überprüfen Sie den **Alpha-Wert**, und ändern Sie den Wert, wenn Sie nicht den Standardwert 0,5 verwenden möchten. Der Alpha-Parameter steuert die Verteilung zwischen den Abzügen ℓ1 und ℓ2. Gültige Werte liegen zwischen 0 und 1; ein Wert von 1,0 stellt LASSO dar, und ein Wert von 0,0 erzeugt „Ridge Regression“. Die folgende Tabelle zeigt, wie Alpha und Lambda die Regularisierung beeinflussen.

`lambda` value	`alpha` value	Result
`lambda` == 0	`alpha` = any value	No regularization. `alpha` is ignored.
`lambda` > 0	`alpha` == 0	Ridge Regression
`lambda` > 0	`alpha` == 1	LASSO
`lambda` > 0	0 < `alpha` < 1	Elastic Net Penalty

Anmerkung: Das einfache Gleichheitszeichen ist ein Zuweisungsoperator, der „ist“ bedeutet. Das doppelte Gleichheitszeichen ist ein Gleichheitsoperator, der „gleich“ bedeutet.

15. Aktivieren Sie den **Wert von Lambda**, und geben Sie einen Wert an, wenn „Logistic Regression“ nicht die Standardmethode zur Berechnung des Lambda-Wertes verwenden soll, bei der es sich um eine Heuristik handelt, die auf Trainingsdaten basiert. Der Lambda-Parameter steuert den Umfang der angewendeten Regularisierung. Wenn beispielsweise Lambda 0,0 ist, wird keine Regularisierung angewendet und der Alpha-Parameter wird ignoriert.
16. Aktivieren Sie **Nach optimalem Wert von Lambda suchen**, damit „Logistic Regression“ Modelle für den vollständigen Regularisierungspfad berechnet, der beim maximalen Lambda-Wert beginnt (der höchste Lambdawert, der sinnvoll ist; d. h. der niedrigste Wert, der alle Koeffizienten auf Null treibt) und bis zum niedrigsten Lamdba-Wert auf der logarithmischen Skala reicht, wobei die Regularisierungsstärke bei jedem Schritt abnimmt. Das zurückgegebene Modell wird Koeffizienten aufweisen, die dem optimalen Lambda-Wert entsprechen, wie während des Trainings entschieden wurde.
17. Aktivieren Sie **Frühzeitig stoppen**, um die Verarbeitung zu beenden, wenn sich der Trainings- oder Validierungssatz nicht weiter verbessert.
18. Aktivieren Sie **Maximale zu suchende Lambdas**, und geben Sie die maximale Anzahl der Lambdas ein, die während der Lambda-Suche verwendet werden sollen.
19. Aktivieren Sie **Maximale aktive Prädiktoren**, und geben Sie die maximale Anzahl der Prädiktoren ein, die während der Berechnung verwendet werden sollen. Dieser Wert wird als Stoppkriterium verwendet, um einen teuren Modellaufbau mit vielen Prädiktoren zu verhindern.
20. 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 **Modelldetails**, um in dem Tool für die „Machine Learning“-Modellverwaltung die gesamte Ausgabe anzuzeigen.

Ausgabeports

Der Schritt „Logistic Regression“ enthält zwei optionale Ausgabeports: den Modellbewertungsport und den Modellmetrikport. Die Funktion dieser Ports hängt von Ihrer Auswahl und Eingabe ab, wenn Sie die grundlegenden und erweiterten Optionen des Schrittes ausführen. Wenn Sie zum Beispiel die n-fache Validierung durchführen, indem Sie das Feld **N-fach** auf der Registerkarte „Erweiterte Optionen“ markieren, wird die Spalte „N-fach“ in den vom Modellmetrikport generierten Ausgabemetriken mit Daten gefüllt. Wenn Sie alternativ keine N-fache Validierung durchführen, ist die Spalte „N-Fach“ leer. Der Modellbewertungsport wird ebenfalls aktiviert, wenn Sie das Feld **Eingabedaten bewerten** auf der Registerkarte „Standardoptionen“ markieren.

Modellbewertungsport

Wenn Sie auf der Registerkarte „Standardoptionen“ das Eingabefeld **Eingabedaten bewerten** aktivieren, teilt dies der „Logistic Regression“ mit, dass bei der Erstellung des Modells vorhergesagte Werte berechnet werden, die wiederum die Spalten **Predticted_Value**, **Probability_of_class_A**, und **Probability_of_class_B** für dieses Ergebnis in den Ausgabedaten hinzufügt. An diesen Port können Sie ein beliebiges Ziel anhängen: einen Write To File-Schritt, einen Write To Null-Schritt usw.

Modellmetrikport

Mit dem **Modellmetrikport** können Sie die Modellbewertungsmetriken in eine Datendatei ausgeben. Dies wird Ihnen helfen, viele Modelle, die innerhalb und außerhalb von Spectrum™ Technology Platform generiert wurden, zu vergleichen und andere Datenverarbeitungsaufgaben an den Metriken auszuführen.

Gehen Sie folgendermaßen vor, um den Modellmetrikport zu verwenden:

1. Öffnen Sie einen Datenfluss, der den Schritt „Logistic Regression“ verwendet.

2. Hängen Sie einen „Write to File“-Schritt oder einen anderen Datenausgabeschritt an den zweiten Ausgabeport an.
3. Führen Sie den Auftrag aus.
4. Alternative zu Schritt 3: Fügen Sie einen Prüfpunkt zu dem Kanal hinzu, der den Schritt „Logistic Regression“ mit dem in Schritt 2 hinzugefügten Zielschritt verbindet, indem Sie mit der rechten Maustaste auf den Kanal klicken und „Überprüfungspunkt hinzufügen“ auswählen. Klicken Sie dann in der Enterprise Designer-Symbolleiste auf die Schaltfläche „Aktuellen Fluss überprüfen“ (). Die Überprüfung wird ausgeführt und Sie sollten ähnliche Ergebnisse wie die unten dargestellten sehen.

Creation Time	Flow Name	Metrics	Model Name	Model Type	N Fold	Test	Training
10/11/2018 2:37:00 AM	LinearRegressionTest1...	MSE	LinearRegressionTest1...	Linear Regression	0.0050019841090508	0.00321814876650744	1.31531988837612E-10
10/11/2018 2:37:00 AM	LinearRegressionTest1...	RMSE	LinearRegressionTest1...	Linear Regression	0.0707247064967455	0.0567287296394643	1.14687396359675E-05
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Number of observations	LinearRegressionTest1...	Linear Regression	33	17	33
10/11/2018 2:37:00 AM	LinearRegressionTest1...	R2	LinearRegressionTest1...	Linear Regression	-0.366538763835926	0.0268719304981138	0.999999964065547
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Mean residual deviance	LinearRegressionTest1...	Linear Regression	0.0050019841090508	0.00321814876650744	1.31531988837612E-10

6 - Principal Component Analysis

In this section

Einführung

Principal Component Analysis (PCA) ist ein statistisches Verfahren, das einen Beobachtungssatz von möglicherweise korrelierten Variablen in einen Wertesatz von linear nicht korrelierten Variablen (prinzipielle Komponenten) umwandelt.

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. Wenn Sie mit der Ausgabe Ihres Modells zufrieden sind, können Sie es verfügbar machen und in einem Bewertungsdatenfluss verwenden.

Definieren von Modelleigenschaften

1. Klicken Sie unter **Primäre Schritte**/**Bereitgestellte Schritte**/**Machine Learning** auf den **PCA-Optionen**-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 die prinzipiellen Komponenten Ihres Modells enthält. Ein Ausgabeschritt ist nicht erforderlich, jedoch können Sie 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 „PCA Options“-Schritt, um das Dialogfeld **PCA-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 von **prinzipiellen Komponenten** ein, die Ihr Modell enthalten soll.
6. Optional: Geben Sie einen **Beschreibung** des Modells ein.
7. Klicken Sie in der Tabelle **Eingaben** 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 kategorisches, DateTime-, numerisches, Zeichenfolgen- oder uniqueid-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 **Alle Faktorstufen verwenden** deaktiviert, um die erste prinzipielle Komponente zu überspringen, die über die größte Streuung in den Daten verfügt. Aktivieren Sie dieses Kontrollkästchen, um die erste prinzipielle Komponente zu behalten.
2. Wählen Sie die geeignete **Transformation** für die Trainingsdaten aus.

Demean	Subtrahiert den Mittelwert jeder Spalte.
Descale	Dividiert durch die Standardabweichung jeder Spalte.
Keine	
Normalisieren	Subtrahiert den Mittelwert jeder Spalte und dividiert jede Spalte durch ihren Bereich (Maximum minus Minimum).
Standardisieren	Verwendet keine Mittelwert- und Einheitenvarianz. Standardeinstellung.

3. 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.
4. 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 eine **PCA-Methode** aus der Dropdown-Liste aus. Beachten Sie, dass „GLRM“ und „Potenz“ derzeit experimentell sind.

GLRM	Passt ein generalisiertes niederrangiges Modell mit L2-Verlustfunktion und ohne Regularisierung an. Löst für die SVD mithilfe lokaler Matrixalgebra. Diese Option ist nur aktiviert, wenn Sie auf der Registerkarte „Standardoptionen“ **Alle Faktorstufen verwenden** aktiviert haben.
GramSVD	Verwendet eine verteilte Berechnung der Gram-Matrix, gefolgt von einer lokalen SVD mithilfe des JAMA-Pakets.
Potenz	Berechnet die SVD mithilfe der Potenzmethode.

Randomisiert Verwendet die randomisierte Unterraumiterationsmethode.

3. Lassen Sie **Maximale Iterationen** deaktiviert, um eine unbegrenzte Anzahl von Trainingsiterationen einzustellen (Standard). Aktivieren Sie das Kontrollkästchen, um die Anzahl der Trainingsiterationen zu begrenzen.
4. 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.

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 **Modelldetails**, um in dem Tool für die „Machine Learning“-Modellverwaltung die gesamte Ausgabe anzuzeigen.

Ausgabeport

Der Schritt der Hauptkomponentenanalyse beinhaltet einen optionalen Ausgabeport: den Modellmetrikport. Die Funktion dieses Ports wird durch Ihre Auswahl und Eingabe bestimmt, wenn Sie die grundlegenden und erweiterten Optionen des Schrittes abschließen. Zum Beispiel:

Modellmetrikport

Mit dem **Modellmetrikport** können Sie die Modellbewertungsmetriken in eine Datendatei ausgeben. Dies wird Ihnen helfen, viele Modelle, die innerhalb und außerhalb von Spectrum™ Technology Platform generiert wurden, zu vergleichen und andere Datenverarbeitungsaufgaben an den Metriken auszuführen.

Gehen Sie folgendermaßen vor, um den Modellmetrikport zu verwenden:

1. Öffnen Sie einen Datenfluss, der den PCA-Schritt verwendet.
2. Hängen Sie einen „Write to File“-Schritt oder einen anderen Datenausgabeschritt an den zweiten Ausgabeport an.
3. Führen Sie den Auftrag aus.

4. Alternative zu Schritt 3: Fügen Sie einen Prüfpunkt zu dem Kanal hinzu, der den PCA-Schritt mit dem in Schritt 2 hinzugefügten Zielschritt verbindet, indem Sie mit der rechten Maustaste auf den Kanal klicken und „Überprüfungspunkt hinzufügen“ auswählen. Klicken Sie dann in der Enterprise Designer-Symbolleiste auf die Schaltfläche „Aktuellen Fluss überprüfen“ (). Die Überprüfung wird ausgeführt und Sie sollten ähnliche Ergebnisse wie die unten dargestellten sehen.

Creation Time	Cumulative Proportion	Flow Name	Model Name	Model Type	Principal Component	Proportion of Variance	Standard Deviation
10/11/2018 8:36:00 PM	0.120990707073471	PCA_MODEL_ASSESSMENT	PCA Model Assessment	PCA	PC1	0.120990707073471	1.73278529942543
10/11/2018 8:36:00 PM	0.163608702477588	PCA_MODEL_ASSESSMENT	PCA Model Assessment	PCA	PC2	0.0426179954041175	1.02840755055022
10/11/2018 8:36:00 PM	0.20114715020656	PCA_MODEL_ASSESSMENT	PCA Model Assessment	PCA	PC3	0.0375384477289716	0.965176862701583
10/11/2018 8:36:00 PM	0.236650281720107	PCA_MODEL_ASSESSMENT	PCA Model Assessment	PCA	PC4	0.0355031315135469	0.93864652798561
10/11/2018 8:36:00 PM	0.269754397170741	PCA_MODEL_ASSESSMENT	PCA Model Assessment	PCA	PC5	0.033104154506347	0.90637880520786

7 - Random Forest Classification

In this section

Einführung

Mit Random Forest Classification können Sie Machine Learning durchführen, indem Sie Modelle aus Datasets erstellen, die im Hinblick auf Eingabevariablen kontinuierliche 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.

> **Anmerkung:** Klicken Sie **hier**, um weitere Informationen zu Random Forest Classification und zugehörige Optionen zu erhalten.

Definieren von Modelleigenschaften

1. Klicken Sie unter **Primäre Schritte**/**Bereitgestellte Schritte**/**Machine Learning** auf den **Random Forest Classification**-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 „Random Forest Classification“-Schritt, um das Dialogfeld **„Random Forest Classification“-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 ein numerisches Feld aus.
6. Klicken Sie auf **Multinomiale Ebenen**, und geben Sie die maximale Anzahl von Kategorien ein, in die das Zielfeld gruppiert werden kann. Beachten Sie, dass das Aktivieren dieser Option die Option **Eingabedaten bewerten** auf der Registerkarte „Standardoptionen“ deaktiviert.
7. Optional: Geben Sie einen **Beschreibung** des Modells ein.

8. Klicken Sie bei jedem Feld, dessen Daten zum Modell hinzugefügt werden sollen, auf **Einschließen**. Achten Sie darauf, das Feld einzuschließen, das Sie als Zielfeld ausgewählt haben.
9. Geben Sie über das Dropdown-Menü **Modelldatentyp** an, ob die einzelnen Eingabefelder als numerisches Feld, als kategorisches Feld oder als DateTime-Feld verwendet werden sollen.
10. Klicken Sie auf **OK**, um das Modell und die Konfiguration zu speichern, oder fahren Sie mit der nächsten Registerkarte fort.

Konfigurieren von Standardoptionen

1. Geben Sie die maximale **Anzahl von Strukturen** in Ihrem Modell ein.
2. Geben Sie die **Maximale Tiefe** ein: die maximale Anzahl von Ebenen, die Ihr Modell enthalten soll.
3. Geben Sie die **Minimale Anzahl von Zeilen** ein: die minimale Anzahl von Zeilen (oder Datensätzen), die Ihr Modell enthalten soll.
4. Geben Sie die **Anzahl numerischer Bins** ein: die Anzahl von Bins, die das Histogramm erstellen und dann am besten Punkt aufteilen soll.
5. Geben Sie die **Anzahl von Bins auf höchster Ebene** ein: die minimale Anzahl von Bins, die Sie auf der Stammebene haben möchten.
6. Geben Sie die **Anzahl kategorischer Bins** ein: die maximale Anzahl von Bins, die das Histogramm erstellen und dann am besten Punkt aufteilen soll.
7. Aktivieren Sie **Abtastrate**, und geben Sie den Prozentsatz der Zeilen ein, die als Stichprobe in jeder Struktur verwendet werden soll. Dies kann ein Wert zwischen 0,0 und 1,0 sein.
8. Aktivieren Sie **Spaltenabtastrate pro Struktur**, und geben Sie die Spaltenabtastrate für die einzelnen Strukturen ein. Dies kann ein Wert zwischen 0,0 und 1,0 sein.
9. Aktivieren Sie **Spalten in jeder Ebene**, und geben Sie die relative Änderung der Spaltenabtastrate für jede Ebene ein. Gültige Werte liegen zwischen 1,0 und der Zahl des ausgewählten Eingabeprädiktors. Der Standardwert ist 1,0.
10. Aktivieren Sie **Eingabedaten bewerten**, um eine Spalte für die Modellvorhersage (Punktzahl) für Eingabedaten hinzuzufügen. Beachten Sie, dass diese Option deaktiviert ist, wenn Sie auf der Registerkarte „Modelleigenschaften“ die Option **Multinomiale Ebenen** aktiviert haben.
11. 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.
12. Geben Sie den Wert 100 abzüglich der in Schritt 5 eingegebenen Menge als **Prozentsatz für Testdaten** ein.
13. **Ausgangswert für Stichprobe**, um sicherzustellen, dass die Darstellung der Daten bei jeder Datenflussausführung gleich ist, wenn diese in Test- und Trainingsdaten aufgeteilt werden. Deaktivieren Sie dieses Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.

14. 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. Aktivieren Sie **Klassen ausgleichen**, um die Klassenverteilung auszugleichen und entweder für die Mehrheitsklassen ein Undersampling oder für die Minderheitsklassen ein Oversampling durchzuführen.
3. Wählen Sie einen **Histogrammtyp** aus.

Auto	Für Buckets wird ein Binning vom Minimum bis zum Maximum in Schritten von (max-min)/N durchgeführt. Verwenden Sie diese Option, um den Histogrammtyp für das Auffinden optimaler Teilungspunkte anzugeben.
QuantilesGlobal	Buckets haben die gleiche Population. Dies berechnet `nbins` Quantile für jede numerische (nicht binäre) Spalte. Dann wird jeder Bucket (zwischen zwei Quantilen) einheitlich angepasst (zufällig für Reste), sodass sich insgesamt `nbins_top_level` Bins ergeben.
Random	Der Algorithmus nimmt Stichproben von N-1 Punkten von Minimum bis Maximum und verwendet die sortierte Liste, um die beste Teilung zu finden.
RoundRobin	Der Algorithmus wechselt durch alle Histogrammtypen (einer pro Struktur).
UniformAdaptive	Jedes Feature wird per Binning einem Bucket zugeordnet, sodass sich Buckets mit gleicher Schrittgröße (nicht Population) ergeben. Dies ist die schnellste Methode, kann aber zu ungenaueren Aufteilungen führen, wenn die Verteilung sehr verzerrt ist.

4. Wählen Sie eine **Kategorische Codierung** aus.

Auto	Führt automatisch eine `Enum`-Codierung durch.
Binary	Konvertiert Kategorien in Ganzzahlen, dann in Binärwerte, und weist jeder Ziffer eine separate Spalte zu. Codiert die Daten in weniger Dimensionen, jedoch werden Entfernungen etwas verzerrt. **Anmerkung:** Pro kategorischem Feature können nicht mehr als 32 Spalten vorhanden sein.
Eigen	*k* Spalten pro kategorischem Feature, behält nur Projektionen einer 1-aus-n-codierten Matrix auf *k*-dimensionalen Eigen-Raum bei.

Enum	Wechselt durch alle Histogrammtypen (einer pro Struktur).
OneHotExplicit	Pro Kategorie ist eine Spalte vorhanden, wobei „1“ oder „0“ in jeder Zelle anzeigen, ob die Zeile die Kategorie dieser Spalte enthält.

5. Lassen Sie **Seed für Algorithmus und 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. Deaktivieren Sie dieses Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.
6. Aktivieren Sie **N-fach** und geben Sie die Anzahl der Folds ein, wenn Sie eine Kreuzvalidierung durchführen.
7. Aktivieren Sie **Faktorzuweisung**, und wählen Sie aus der Dropdown-Liste aus, ob Sie eine Kreuzüberprüfung durchführen. Dieses Feld ist nur anwendbar, wenn Sie unter **N-fach** einen Wert eingegeben haben und **Faktorfeld** 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.

8. Wenn Sie eine Kreuzvalidierung durchführen, aktivieren Sie **Faktorfeld** und wählen Sie aus der Dropdown-Liste das Feld aus, das die Faktorindexzuweisung für die Kreuzvalidierung enthält.

   Dieses Feld ist nur anwendbar, wenn Sie unter **N-fach** und **Faktorzuweisung** keinen Wert eingegeben haben.
9. Aktivieren Sie **Runden stoppen**, um das Training zu beenden, wenn die Option „Stopping_metric“ sich nicht in der angegebenen Anzahl von Trainingsrunden verbessert, und geben Sie die Anzahl nicht erfolgreicher Trainingsrunden ein, die absolviert werden, bevor gestoppt werden soll. Um diese Funktion zu deaktivieren, geben Sie „0“ an. Die Metrik wird anhand der Überprüfungsdaten berechnet (falls vorhanden), ansonsten werden Trainingsdaten verwendet.
10. Wählen Sie eine **Abbruchmetrik**, um festzulegen, wann die Erstellung neuer Strukturen eingestellt werden soll.

AUC	Fläche unter ROC-Kurve. **Anmerkung:** Gilt nur für binomiale Modelle.
Auto	Standardwert ist `Abweichung`.
Lifttopgroup	Beste 1 %.
Logloss	Logarithmischer Abfall.
Meanperclasserror	Die Fehlklassifizierungsrate.
Misclassification	Der Wert von (1 - (korrekte Vorhersagen/gesamte Vorhersagen)) * 100.
MSE	Mittlerer quadratischer Fehler, berücksichtigt sowohl Streuung als auch Tendenz des Prädiktors.
RMSE	Wurzel aus dem mittleren quadratischen Fehler; misst die Differenz zwischen Werten (Stichproben- und Populationswerte), die von einem Modell oder einem Schätzwert vorhergesagt wurden, und tatsächlich beobachteten Werten. Auch Quadratwurzel von MSE.

11. Aktivieren Sie **Abbruchtoleranz**, und geben Sie einen Wert ein, um die relative Toleranz für den metrikbasierten Abbruch des Trainings zu spezifizieren, wenn die Verbesserung geringer ist als dieser Wert. Dieses Feld ist nur aktiviert, wenn Sie **Runden stoppen** aktiviert haben.
12. Aktivieren Sie **Minimale Aufteilungsverbesserung**, und geben Sie einen Wert ein, um die minimale relative Verbesserung in der Verringerung des quadratischen Fehlers anzugeben, bei der eine Aufteilung durchgeführt werden soll. Wenn diese Option richtig ausgeführt wird, kann die Überanpassung verringert werden. Optimale Werte bewegen sich im Bereich von 1e-10 bis 1e-3. Dieses Feld ist nur aktiviert, wenn Sie **Runden stoppen** aktiviert haben.
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 **Modelldetails**, um in dem Tool für die „Machine Learning“-Modellverwaltung die gesamte Ausgabe anzuzeigen.

Ausgabeports

Der Schritt „Random Forest Classification“ enthält zwei optionale Ausgabeports: den Modellbewertungsport und den Modellmetrikport. Die Funktion dieser Ports hängt von Ihrer Auswahl und Eingabe ab, wenn Sie die grundlegenden und erweiterten Optionen des Schrittes ausführen. Wenn Sie zum Beispiel die n-fache Validierung durchführen, indem Sie das Feld **N-fach** auf der Registerkarte „Erweiterte Optionen“ markieren, wird die Spalte „N-fach“ in den vom Modellmetrikport generierten Ausgabemetriken mit Daten gefüllt. Wenn Sie alternativ keine N-fache Validierung durchführen, ist die Spalte „N-Fach“ leer. Der Modellbewertungsport wird ebenfalls aktiviert, wenn Sie das Feld **Eingabedaten bewerten** auf der Registerkarte „Standardoptionen“ markieren.

Modellbewertungsport

Wenn Sie auf der Registerkarte „Standardoptionen“ das Eingabefeld **Eingabedaten bewerten** aktivieren, teilt dies der „Random Forest Regression“ mit, dass bei der Erstellung des Modells vorhergesagte Werte berechnet werden, die wiederum die Spalten **Predticted_Value**, **Probability_of_class_A**, und **Probability_of_class_B** für dieses Ergebnis in den Ausgabedaten hinzufügt. An diesen Port können Sie ein beliebiges Ziel anhängen: einen Write To File-Schritt, einen Write To Null-Schritt usw.

> **Anmerkung:** Dieser Port ist nicht für multinomiale „Random Forest Classification“-Modelle geeignet.

Modellmetrikport

Mit dem **Modellmetrikport** können Sie die Modellbewertungsmetriken in eine Datendatei ausgeben. Dies wird Ihnen helfen, viele Modelle, die innerhalb und außerhalb von Spectrum™ Technology Platform generiert wurden, zu vergleichen und andere Datenverarbeitungsaufgaben an den Metriken auszuführen.

Gehen Sie folgendermaßen vor, um den Modellmetrikport zu verwenden:

1. Öffnen Sie einen Datenfluss, der den Schritt „Random Forest Classification“ verwendet.

2. Hängen Sie einen „Write to File“-Schritt oder einen anderen Datenausgabeschritt an den zweiten Ausgabeport an.
3. Führen Sie den Auftrag aus.
4. Alternative zu Schritt 3: Fügen Sie einen Prüfpunkt zu dem Kanal hinzu, der den Schritt „Random Forest Classification“ mit dem in Schritt 2 hinzugefügten Zielschritt verbindet, indem Sie mit der rechten Maustaste auf den Kanal klicken und „Überprüfungspunkt hinzufügen“ auswählen. Klicken Sie dann in der Enterprise Designer-Symbolleiste auf die Schaltfläche „Aktuellen Fluss überprüfen“ (). Die Überprüfung wird ausgeführt und Sie sollten ähnliche Ergebnisse wie die unten dargestellten sehen.

Creation Time	Flow Name	Metrics	Model Name	Model Type	N Fold	Test	Training
10/11/2018 2:37:00 AM	LinearRegressionTest1...	MSE	LinearRegressionTest1...	Linear Regression	0.0050019841090508	0.00321814876650744	1.31531988837612E-10
10/11/2018 2:37:00 AM	LinearRegressionTest1...	RMSE	LinearRegressionTest1...	Linear Regression	0.0707247064967455	0.0567287296394643	1.14687396359675E-05
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Number of observations	LinearRegressionTest1...	Linear Regression	33	17	33
10/11/2018 2:37:00 AM	LinearRegressionTest1...	R2	LinearRegressionTest1...	Linear Regression	-0.366538763835926	0.0268719304981138	0.999999964065547
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Mean residual deviance	LinearRegressionTest1...	Linear Regression	0.0050019841090508	0.00321814876650744	1.31531988837612E-10

8 - Random Forest Regression

In this section

Einführung

Mit Random Forest 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.

> **Anmerkung:** Klicken Sie **hier**, um weitere Informationen zu Random Forest Regression und zugehörige Optionen zu erhalten.

Definieren von Modelleigenschaften

1. Klicken Sie unter **Primäre Schritte**/**Bereitgestellte Schritte**/**Machine Learning** auf den **Random Forest 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 „Random Forest Regression“-Schritt, um das Dialogfeld **„Random Forest 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 ein numerisches Feld 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**. Achten Sie darauf, das Feld einzuschließen, das Sie als Zielfeld ausgewählt haben.
8. Geben Sie über das Dropdown-Menü **Modelldatentyp** an, ob die einzelnen Eingabefelder als numerisches Feld, als kategorisches Feld oder als DateTime-Feld verwendet werden sollen.

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. Geben Sie die maximale **Anzahl von Strukturen** in Ihrem Modell ein. Der Standardwert ist 50.
2. Geben Sie die **Maximale Tiefe** ein: die maximale Anzahl von Ebenen, die Ihr Modell enthalten soll. Der Standardwert ist 5.
3. Geben Sie die **Minimale Anzahl von Zeilen** ein: die minimale Anzahl von Zeilen (oder Datensätzen), die Ihr Modell enthalten soll. Der Standardwert ist 10.
4. Geben Sie die **Anzahl numerischer Bins** ein: die Anzahl von Bins, die das Histogramm erstellen und dann am besten Punkt aufteilen soll. Der Standardwert ist 20.
5. Geben Sie die **Anzahl von Bins auf höchster Ebene** ein: die minimale Anzahl von Bins, die Sie auf der Stammebene haben möchten. Der Standardwert ist 1024.
6. Geben Sie die **Anzahl kategorischer Bins** ein: die maximale Anzahl von Bins, die das Histogramm erstellen und dann am besten Punkt aufteilen soll. Der Standardwert ist 1024.
7. Aktivieren Sie **Abtastrate**, und geben Sie den Prozentsatz der Zeilen ein, die als Stichprobe in jeder Struktur verwendet werden soll. Dies kann ein Wert zwischen 0,0 und 1,0 sein.
8. Aktivieren Sie **Spaltenabtastrate pro Struktur**, und geben Sie die Spaltenabtastrate für die einzelnen Strukturen ein. Dies kann ein Wert zwischen 0,0 und 1,0 sein.
9. Aktivieren Sie **Spalten in jeder Ebene**, und geben Sie die relative Änderung der Spaltenabtastrate für jede Ebene ein. Der Standardwert dieser Option ist 1,0 und kann ein Wert zwischen 0,0 und 2,0 sein.
10. Aktivieren Sie **Eingabedaten bewerten**, um eine Spalte für die Modellvorhersage (Punktzahl) für Eingabedaten hinzuzufügen.
11. 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.
12. Geben Sie den Wert 100 abzüglich der in Schritt 5 eingegebenen Menge als **Prozentsatz für Testdaten** ein.
13. **Ausgangswert für Stichprobe**, um sicherzustellen, dass die Darstellung der Daten bei jeder Datenflussausführung gleich ist, wenn diese in Test- und Trainingsdaten aufgeteilt werden. Deaktivieren Sie dieses Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.
14. 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 einen **Histogrammtyp** aus.

Auto	Für Buckets wird ein Binning vom Minimum bis zum Maximum in Schritten von (max-min)/N durchgeführt. Verwenden Sie diese Option, um den Histogrammtyp für das Auffinden optimaler Teilungspunkte anzugeben.
QuantilesGlobal	Buckets haben die gleiche Population. Dies berechnet `nbins` Quantile für jede numerische (nicht binäre) Spalte. Dann wird jeder Bucket (zwischen zwei Quantilen) einheitlich angepasst (zufällig für Reste), sodass sich insgesamt `nbins_top_level` Bins ergeben.
Random	Der Algorithmus nimmt Stichproben von N-1 Punkten von Minimum bis Maximum und verwendet die sortierte Liste, um die beste Teilung zu finden.
RoundRobin	Der Algorithmus wechselt durch alle Histogrammtypen (einer pro Struktur).
UniformAdaptive	Jedes Feature wird per Binning einem Bucket zugeordnet, sodass sich Buckets mit gleicher Schrittgröße (nicht Population) ergeben. Dies ist die schnellste Methode, kann aber zu ungenaueren Aufteilungen führen, wenn die Verteilung sehr verzerrt ist.

3. Wählen Sie eine **Kategorische Codierung** aus.

Auto	Führt automatisch eine `Enum`-Codierung durch.
Binary	Konvertiert Kategorien in Ganzzahlen, dann in Binärwerte, und weist jeder Ziffer eine separate Spalte zu. Codiert die Daten in weniger Dimensionen, jedoch werden Entfernungen etwas verzerrt. **Anmerkung:** Pro kategorischem Feature können nicht mehr als 32 Spalten vorhanden sein.
Eigen	*k* Spalten pro kategorischem Feature, behält nur Projektionen einer 1-aus-n-codierten Matrix auf *k*-dimensionalen Eigen-Raum bei.
Enum	Wechselt durch alle Histogrammtypen (einer pro Struktur).
OneHotExplicit	Pro Kategorie ist eine Spalte vorhanden, wobei „1“ oder „0“ in jeder Zelle anzeigen, ob die Zeile die Kategorie dieser Spalte enthält.

4. Lassen Sie **Seed für Algorithmus und 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. Deaktivieren Sie dieses Feld, damit die Aufteilung bei jeder Datenflussausführung beliebig erfolgt.

5. Aktivieren Sie **N-fach** und geben Sie die Anzahl der Folds ein, wenn Sie eine Kreuzvalidierung durchführen.
6. Aktivieren Sie **Faktorzuweisung**, und wählen Sie aus der Dropdown-Liste aus, ob Sie eine Kreuzüberprüfung durchführen. Dieses Feld ist nur anwendbar, wenn Sie unter **N-fach** einen Wert eingegeben haben und **Faktorfeld** 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.

7. Wenn Sie eine Kreuzvalidierung durchführen, aktivieren Sie **Faktorfeld** und wählen Sie aus der Dropdown-Liste das Feld aus, das die Faktorindexzuweisung für die Kreuzvalidierung enthält.

   Dieses Feld ist nur anwendbar, wenn Sie unter **N-fach** und **Faktorzuweisung** keinen Wert eingegeben haben.
8. Aktivieren Sie **Runden stoppen**, um das Training zu beenden, wenn die Option „Stopping_metric“ sich nicht in der angegebenen Anzahl von Trainingsrunden verbessert, und geben Sie die Anzahl nicht erfolgreicher Trainingsrunden ein, die absolviert werden, bevor gestoppt werden soll. Um diese Funktion zu deaktivieren, geben Sie „0“ an. Die Metrik wird anhand der Überprüfungsdaten berechnet (falls vorhanden), ansonsten werden Trainingsdaten verwendet.
9. Wählen Sie eine **Abbruchmetrik**, um festzulegen, wann die Erstellung neuer Strukturen eingestellt werden soll.

Auto	Standardwert ist `Abweichung`.
Abweichung	Mittlere Restabweichung; identisch mit MSE.
MAE	Mittlerer absoluter Fehler; die Differenz zwischen zwei kontinuierlichen Variablen.
MSE	Mittlerer quadratischer Fehler, berücksichtigt sowohl Streuung als auch Tendenz des Prädiktors.
RMSE	Wurzel aus dem mittleren quadratischen Fehler; misst die Differenz zwischen Werten (Stichproben- und Populationswerte), die von einem Modell oder einem Schätzwert vorhergesagt wurden, und tatsächlich beobachteten Werten. Auch Quadratwurzel von MSE.

RMSLE	Wurzel des mittleren quadratischen logarithmischen Fehlers; misst das Verhältnis zwischen vorhergesagtem und tatsächlichem Wert.

10. Aktivieren Sie **Abbruchtoleranz**, und geben Sie einen Wert ein, um die relative Toleranz für den metrikbasierten Abbruch des Trainings zu spezifizieren, wenn die Verbesserung geringer ist als dieser Wert.
11. Aktivieren Sie **Minimale Aufteilungsverbesserung**, und geben Sie einen Wert ein, um die minimale relative Verbesserung in der Verringerung des quadratischen Fehlers anzugeben, bei der eine Aufteilung durchgeführt werden soll. Wenn diese Option richtig ausgeführt wird, kann die Überanpassung verringert werden. Optimale Werte bewegen sich im Bereich von 1e-10 bis 1e-3. Dieses Feld ist nur aktiviert, wenn Sie **Runden stoppen** aktiviert haben.
12. 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 **Modelldetails**, um in dem Tool für die „Machine Learning“-Modellverwaltung die gesamte Ausgabe anzuzeigen.

Ausgabeports

Der Schritt „Random Forest Regression“ enthält zwei optionale Ausgabeports: den Modellbewertungsport und den Modellmetrikport. Die Funktion dieser Ports hängt von Ihrer Auswahl und Eingabe ab, wenn Sie die grundlegenden und erweiterten Optionen des Schrittes ausführen. Wenn Sie zum Beispiel die n-fache Validierung durchführen, indem Sie das Feld **N-fach** auf der Registerkarte „Erweiterte Optionen“ markieren, wird die Spalte „N-fach“ in den vom Modellmetrikport generierten Ausgabemetriken mit Daten gefüllt. Wenn Sie alternativ keine N-fache Validierung durchführen, ist die Spalte „N-Fach“ leer. Der Modellbewertungsport wird ebenfalls aktiviert, wenn Sie das Feld **Eingabedaten bewerten** auf der Registerkarte „Standardoptionen“ markieren.

Modellbewertungsport

Wenn Sie das Feld **Eingabedaten bewerten** auf der Registerkarte „Standardoptionen“ markieren, teilt dies der „Random Forest Regression“ die Berechnung vorhergesagter Werte beim Erstellen des Modells mit, wobei wiederum die Spalte **Predicted_Value** für diese Punktzahl in den Ausgabedaten hinzugefügt wird. An diesen Port können Sie ein beliebiges Ziel anhängen: einen Write To File-Schritt, einen Write To Null-Schritt usw.

Modellmetrikport

Mit dem **Modellmetrikport** können Sie die Modellbewertungsmetriken in eine Datendatei ausgeben. Dies wird Ihnen helfen, viele Modelle, die innerhalb und außerhalb von Spectrum™ Technology Platform generiert wurden, zu vergleichen und andere Datenverarbeitungsaufgaben an den Metriken auszuführen.

Gehen Sie folgendermaßen vor, um den Modellmetrikport zu verwenden:

1. Öffnen Sie einen Datenfluss, der den Schritt „Random Forest Regression“ verwendet.
2. Hängen Sie einen „Write to File“-Schritt oder einen anderen Datenausgabeschritt an den zweiten Ausgabeport an.
3. Führen Sie den Auftrag aus.
4. Alternative zu Schritt 3: Fügen Sie einen Prüfpunkt zu dem Kanal hinzu, der den Schritt „Random Forest Regression“ mit dem in Schritt 2 hinzugefügten Zielschritt verbindet, indem Sie mit der rechten Maustaste auf den Kanal klicken und „Überprüfungspunkt hinzufügen“ auswählen. Klicken Sie dann in der Enterprise Designer-Symbolleiste auf die Schaltfläche „Aktuellen Fluss überprüfen“ (). Die Überprüfung wird ausgeführt und Sie sollten ähnliche Ergebnisse wie die unten dargestellten sehen.

Creation Time	Flow Name	Metrics	Model Name	Model Type	N Fold	Test	Training
10/11/2018 2:37:00 AM	LinearRegressionTest1	MSE	LinearRegressionTest1...	Linear Regression	0.0050019841090508	0.00321814876650744	1.31531988837612E-10
10/11/2018 2:37:00 AM	LinearRegressionTest1...	RMSE	LinearRegressionTest1...	Linear Regression	0.0707247064967455	0.0567287296394643	1.14687396359675E-05
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Number of observations	LinearRegressionTest1...	Linear Regression	33	17	33
10/11/2018 2:37:00 AM	LinearRegressionTest1...	R2	LinearRegressionTest1...	Linear Regression	-0.366538763835926	0.0268719304981138	0.999999964065547
10/11/2018 2:37:00 AM	LinearRegressionTest1...	Mean residual deviance	LinearRegressionTest1...	Linear Regression	0.0050019841090508	0.00321814876650744	1.31531988837612E-10

9 - Machine Learning-Modellverwaltung

In this section

Zugreifen auf die 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:

    <Servername>:<Port>

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

    myspectrumplatform:8080

  - Klicken Sie auf **Spectrum Machine Learning**.
  - Klicken Sie auf **Machine Learning-Modellverwaltung 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:

    <Servername>:<Port>/machinelearning

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

    myspectrumplatform:8080/machinelearning

  - Geben Sie einen gültigen Benutzernamen und das zugehörige Kennwort für die Spectrum™ Technology Platform ein.

Modellbewertung

Einführung in die Modellbewertung

Auf der Registerkarte „Modellbewertung“ 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.

Modellbewertungsvorgänge

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

	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.
	Importieren Sie ein Modell aus einem bestimmten Pfad. Wählen Sie aus, ob ein vorhandenes Modell mit demselben Namen überschrieben werden soll.
	Exportieren Sie ein Modell in einen bestimmten Pfad. Wählen Sie aus, ob ein vorhandenes Modell mit demselben Namen überschrieben werden soll.

	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.

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

Bietet Trainingsdaten für Folgendes:

- 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)
- Precision-Recall-Bereich unter der Kurve (PR AUC)
- Gini-Koeffizient
- Mean Per Class Error
- Akaike-Informationskriterium (AIC)
- Lambda
- 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

Zeigt 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.

Details zu Linear Regression

Auf dem Bildschirm „Modelldetail“ werden folgende Informationen für Linear 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)
- Mittlere Restabweichung
- Mittlerer absoluter Fehler (MAE)
- Wurzel aus dem mittleren quadratischen Fehler (RMSE)
- Akaike-Informationskriterium (AIC)
- Lambda
- Restabweichung
- Abweichung von Null
- Freiheitsgrad von Null
- Restfreiheitsgrad

Standardisiertes Koeffizientendiagramm

Zeigt die wichtigsten Prädiktoren an, indem der relative Wert der Koeffizienten angegeben wird. Dieser gibt an, wie stark eine Änderung des jeweiligen Prädiktorkoeffizientenwerts den Zielwert positiv oder negativ verändert. Zudem werden die besten 25 Koeffizienten im Modell dargestellt.

GLM-Koeffizienten

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

Details zu Random Forest Regression

Auf dem Bildschirm „Modelldetail“ werden folgende Informationen für Random Forest 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)
- Mittlere Restabweichung
- Mittlerer absoluter Fehler (MAE)
- Wurzel aus dem mittleren quadratischen Fehler (RMSE)

Variablenwichtigkeiten

Stellt Wichtigkeitswerte für jede Variable unter Verwendung der folgenden Metriken bereit:

- Relative Wichtigkeit
- Skalierte Wichtigkeit
- Prozentsatz

Zudem werden die besten 25 Variablen im Modell dargestellt.

Details zu Random Forest Classification – Binomial

Auf dem Bildschirm „Modelldetail“ werden folgende Informationen für **binomiale** Random Forest Classification-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)
- Logloss
- Area under the curve (AUC)
- Precision-Recall-Bereich unter der Kurve (PR AUC)
- Gini
- Mean Per Class Error

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.

Variablenwichtigkeiten

Stellt Wichtigkeitswerte für jede Variable unter Verwendung der folgenden Metriken bereit:

- Relative Wichtigkeit
- Skalierte Wichtigkeit
- Prozentsatz

Zudem werden die besten 25 Variablen im Modell dargestellt.

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 treffen.

Details zu Random Forest Classification – Multinomial

Auf dem Bildschirm „Modelldetail“ werden folgende Informationen für **multinomiale** Random Forest Classification-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)
- Logloss
- Mean Per Class Error

Konfusionsmatrix

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

Variablenwichtigkeiten

Stellt Wichtigkeitswerte für jede Variable unter Verwendung der folgenden Metriken bereit:

- Relative Wichtigkeit
- Skalierte Wichtigkeit
- Prozentsatz

Zudem werden die besten 25 Variablen im Modell dargestellt.

Details zu Principal Component Analysis

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

Bedeutung der Komponenten

Zeigt die prinzipiellen Komponenten in der Reihenfolge der Wichtigkeit an, die auf folgenden Metriken basieren:

- Standardabweichung

- Streuungsverhältnis
- Kumulatives Verhältnis

Rotation

Stellt die Matrix der Variablenbelastungen dar, die Gewichtung mit der jede standardisierte Originalvariable multipliziert werden sollte, um die Komponentenpunktzahl zu erhalten.

Binning Management

Einführung in das Binning Management

Auf der Registerkarte „Binning Management“ bei der Machine Learning-Modellverwaltung wird eine Liste mit allen Binnings 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 Binnings können mehrere Vorgänge durchgeführt werden. Sie können Binnings importieren, exportieren, verfügbar machen, ihre Verfügbarkeit aufheben oder sie löschen. Verfügbar gemachte Bins werden im „Binning Lookup“-Schritt verwendet, um zuvor definierte Binnings auf neue Daten anzuwenden.

Binning Management-Vorgänge

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

	Binning importieren. Wählen Sie aus, ob ein vorhandenes Binning mit demselben Namen überschrieben werden soll.
	Binning exportieren. Wählen Sie aus, ob ein vorhandenes Binning mit demselben Namen überschrieben werden soll.
	Machen Sie das Binning verfügbar, damit es für den „Binning Lookup“-Schritt verfügbar ist. Wenn ein Binning nicht verfügbar gemacht wird, kann es nicht für Lookup verwendet werden.
	Verfügbarkeit des Binnings aufheben.

Binning löschen.

Anmerkung: Ein verfügbar gemachtes Binning kann nicht gelöscht werden. Zu diesem Zeitpunkt ist jedoch keine inhärente Sicherheit vorhanden, durch die verhindert wird, dass ein Benutzer die Binnings eines anderen Benutzers löscht.

10 - Data Science-Demonstrationsfluss

In this section

Einführung

Das Machine Learning-Modul und die Analytics Scoring-Module sind zusammen mit den Modulen zur Vorbereitung von Modeling-Daten Teil des Angebots von Spectrum Data Science. Diese Demonstrationen zeigen Beispiele für Datenaufbereitung, -modellierung und Model Scoring. Sie können Ihre eigenen Datenflüsse mithilfe der Schritt-für-Schritt-Anweisungen erstellen oder die bereitgestellten Datenflüsse als Referenz verwenden.

Überwachtes Lernen: Kreditausfallvorhersage

Laden Sie die Demonstration für das überwachte Lernen herunter.

Die Demonstration für das überwachte Lernen von Data Science führt die Kreditausfallvorhersage mithilfe von Lending Club-Daten durch. Es nutzt mehrere Dateien, die zusammen die Funktion der Data Science Solution der Spectrum™-Technologieplattform in Enterprise Designer demonstrieren.

„Spectrum_DataScience_Supervised_Learning.zip“ beinhaltet folgende Dateien:

- Spectrum_DataScience_Supervised_Learning.pdf: Eine Dokumentation, die Sie durch die Erstellung und Verwendung des Datenflusses des einzelnen Kategorisierungsmoduls, des Bewertungsdatenflusses und aller unterstützenden Dateien führt.
- Data.zip: Die erforderlichen Eingabedateien, Testdateien und Trainingsdateien für jeden der enthaltenen Datenflüsse.
  - loan.csv
  - LoanStats_2016Q1.csv
  - LoanStats_2016Q2.csv
  - LoanStats_2016Q3.csv
  - testData.txt
  - testDataCollege.txt
  - testDataStable.txt
  - testDataThankful.txt
  - trainData.txt
  - trainDataCollege.txt
  - trainDataStable.txt
  - trainDataThankful.txt
  - training.xml
  - trainingCollege.xml

  - trainingStable.xml
  - trainingThanks.xml
- Lending_Club_Demo_DF_(V12.1).zip: Die Datenflüsse für die Spectrum™-Technologieplattform 12.1:
  - LendingClub_2007_2016Q12_v121_MultipleCategorizers.df
  - LendingClub_2007_2016Q1Q2_v121_SingleCategorizer.df
  - LendingClub_2016Q3_v121_SingleCategorizer_Scoring.df
- Lending_Club_Demo_DF_(V12.2).zip: Die Datenflüsse für die Spectrum™-Technologieplattform 12.2:
  - LendingClub_2007_2016Q12_v122_MultipleCategorizers.df
  - LendingClub_2007_2016Q1Q2_v122_SingleCategorizer.df
  - LendingClub_2016Q3_v122_SingleCategorizer_Scoring.df
- ReadMe.txt: Umfangreiche Beschreibungen und Anweisungen für die zuvor genannten Dateien.

Sie können Ihren eigenen Datenfluss erstellen, indem Sie die Schritt-für-Schritt-Anweisungen in der Dokumentation befolgen, oder Sie können die enthaltenen Datenflüsse als Referenzen verwenden, um zu bestätigen, wie die einzelnen abgeschlossenen Schritte und Datenflüsse insgesamt aussehen sollten.

Unüberwachtes Lernen: Segmentierung

Laden Sie die Demonstration für das unüberwachte Lernen herunter.

Die Demonstration für das unüberwachte Lernen von Data Science führt die Segmentierung anhand von Consumer Expenditure-Daten durch. Es nutzt mehrere Dateien, die zusammen die Funktion der Data Science Solution der Spectrum™-Technologieplattform in Enterprise Designer demonstrieren.

„Spectrum_DataScience_Unsupervised_Learning.zip“ beinhaltet folgende Dateien:

- Spectrum_DataScience_Unsupervised_Learning.pdf: Eine Dokumentation, die Sie durch die Erstellung und Verwendung des primären Datenflusses, des Unterflusses, des Bewertungsdatenflusses und aller unterstützenden Dateien führt.
- Data.zip: Die erforderlichen Eingabe- und Ausgabedateien für jeden der enthaltenen Datenflüsse
  - Eingabeordner: Die erforderlichen Eingabedateien für jeden der enthaltenen Datenflüsse
  - Ausgabeordner: Die erforderlichen Ausgabedateien für jeden der enthaltenen Datenflüsse
  - PythonBased-Ordner: Die erforderlichen Eingabe- und Ausgabedateien zur Verwendung der optionalen Python-Verarbeitung anstelle von Group Statistics- und Transformer-Schritten im primären Datenfluss

- Consumer_Expenditure_Demo_DF_(v12.1).zip: Die Datenflüsse für die Spectrum™ Technology Platform 12.1
  - ConsumerExpenditure_v121_sampleandcluster.df
  - ConsumerExpenditure_v121_sampleandcluster_subflow.df
  - ConsumerExpenditure_v121_score.df
  - ConsumerExpenditure_v121_subflow.df
  - PythonBased-Ordner: Die erforderlichen Datenflüsse und Prozessflüsse sowie das Bat-Skript, das Python-Skript und die Dokumentation zur Verwendung der optionalen Python-Verarbeitung anstelle von Group Statistics- und Transformer-Schritten im primären Datenfluss
- Consumer_Expenditure_Demo_DF_(v12.2).zip: Die Datenflüsse für die Spectrum™ Technology Platform 12.2
  - ConsumerExpenditure_v122_sampleandcluster.df
  - ConsumerExpenditure_v122_sampleandcluster_subflow.df
  - ConsumerExpenditure_v122_score.df
  - ConsumerExpenditure_v122_subflow.df
  - PythonBased-Ordner: Die erforderlichen Datenflüsse und Prozessflüsse sowie das Bat-Skript, das Python-Skript und die Dokumentation zur Verwendung der optionalen Python-Verarbeitung anstelle von Group Statistics- und Transformer-Schritten im primären Datenfluss
- ReadMe.txt: Umfangreiche Beschreibungen und Anweisungen für die zuvor genannten Dateien.

Sie können Ihren eigenen Datenfluss erstellen, indem Sie die Schritt-für-Schritt-Anweisungen in der Dokumentation befolgen, oder Sie können die enthaltenen Datenflüsse als Referenzen verwenden, um zu bestätigen, wie die einzelnen abgeschlossenen Schritte und Datenflüsse insgesamt aussehen sollten.

Notices

USPS® Urheberrechtshinweise

Datenbereitstellung und Hinweise

3001 Summer Street
Stamford CT 06926-0700
USA

www.pitneybowes.com