pitney bowes

Spectrum™ Technology Platform

Version 2018.2.0

Guía de Machine Learning

Contents

9 - Administración de modelo de aprendizaje automático

10 - Flujos de demostración de ciencia de datos

1 - Introducción

In this section

Modulo Machine Learning

En el módulo Machine Learning de Spectrum™ Technology Platform se ofrece la capacidad de aplicar binning a datos numéricos y ajustar datos de modelos de Machine Learning supervisados y no supervisados en tales modelos.

Nota: El módulo Machine Learning solo se admite en sistemas operativos Windows y Linux.

Binning

Binning divide los registros en grupos (contenedores) para una variable continua sin considerar la información de objetivos. Puede ejecutar binning no supervisado de una de estas dos maneras: usando contenedores del mismo ancho o contenedores de la misma frecuencia.

K-Means Clustering

K-Means Clustering permite crear modelos según agrupaciones en clústeres analíticas, lo cual segmenta un conjunto de registros en clústeres de registros similares a partir de valores de datos.

Regresión lineal

La regresión lineal le permite realizar aprendizaje automático mediante la creación de modelos a partir de conjuntos de datos que usan objetivos continuos con variables de entrada.

Logistic Regression

Logistic Regression crea modelos a partir de conjuntos de datos que usan objetivos binarios con variables de entrada.

Análisis de componentes principales

El análisis de componentes principales es un proceso estadístico mediante el cual se convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables no correlacionadas linealmente, conocido como componentes principales.

Random Forest Classification

Random Forest Classification le permite realizar aprendizaje automático mediante la creación de modelos a partir de conjuntos de datos que usan objetivos continuos con variables de entrada.

Regresión de bosques aleatorios

La regresión de bosques aleatorios le permite realizar aprendizaje automático mediante la creación de modelos a partir de conjuntos de datos que usan objetivos binarios con variables de entrada.

Administración de modelo de aprendizaje automático

En la Administración de modelo de aprendizaje automático se incluye la evaluación de modelo, la cual le permite administrar todos los modelos de aprendizaje automático en su servidor Spectrum™ Technology Platform, y la Administración de binning, la cual le permite administrar cualquier binning de su servidor Spectrum™ Technology Platform.

> **Nota:** En el módulo Machine Learning se utiliza una biblioteca H2O.ai subyacente para el modelado de algoritmos en K-Means Clustering, Regresión lineal, Logistic Regression, Análisis de componentes principales, Random Forest Classification y Regresión de bosques aleatorios.

Un flujo de trabajo de Machine Learning

Un flujo de trabajo de Machine Learning típico implica los siguientes pasos que se deben realizar en uno o más flujos de trabajo:

1. Acceda a los datos usando otros módulos de Spectrum, como Data Integration.
2. Prepare los datos usando etapas de otros módulos de Spectrum, como las de Data Integration, Data Quality y Core.
3. Ajuste un modelo de Machine Learning, ejecute el flujo de datos y, luego, revise el contenido de la pestaña Salida del modelo en la etapa del modelo. Después, puede retocar el modelo, si es necesario, y volver a ejecutar el flujo de datos. A continuación, debe revisar el conjunto completo de datos de salida de evaluación del modelo en la herramienta Gestión de modelos de Machine Learning. Puede revisar un modelo a la vez o comparar dos modelos.
4. (Opcional) Si usará el modelo para evaluar datos, exponga el modelo en la herramienta Gestión de modelos de Machine Learning, que pone el modelo a disposición de la etapa Java Model Scoring.
    a. Cree un flujo de datos de Spectrum™ Technology Platform siguiendo los pasos 1 y 2 anteriores y luego, reemplace el paso 3 por la etapa Java Model Scoring. Configure este flujo de datos para ejecutarlo en modo de lote a fin de completar un archivo con evaluaciones de modelo aplicadas a datos actualizados (los campos utilizados como X o datos de entrada se actualizan en el paso 1-2 como una parte natural de hacer negocios).
    b. De manera alternativa, use un servicio web en Spectrum™ Technology Platform para evaluar datos según demanda. Por ejemplo, acceda al sitio web, obtenga la ID de cliente y los datos de entrada del modelo, evalúelos y devuelva la evaluación a un proceso que personaliza el contenido web para su cliente.
5. (Opcional) También puede implementar evaluaciones de modelo en una base de datos de gráficos Data Hub como una propiedad de entidad, en mapas o en aplicaciones CES.

2 - Binning

In this section

Introducción a Binning

La etapa Binning realiza lo que se conoce como binning supervisado, que divide una variable continua en grupos (contenedores) sin considerar la información de objetivos. Los datos capturados incluyen rangos, cantidades y porcentaje de valores dentro de cada rango.

Las ventajas de ejecutar binning incluyen:

- Permite incluir registros con datos faltantes en el modelo.
- Controla o mitiga el impacto de valores atípicos en el modelo.
- Soluciona el problema de tener escalas diferentes entre las características, permitiendo que las ponderaciones de los coeficientes en el modelo final se puedan comparar.

En binning no supervisado de Spectrum™ Technology Platform, puede usar contenedores del mismo ancho, donde los datos se dividen en contenedores de igual tamaño, o contenedores de la misma frecuencia, donde los datos se dividen en grupos que contienen aproximadamente el mismo número de registros. En la etapa Binning, los contenedores que tienen el mismo ancho se denominan contenedores de Rango igual y los contenedores que tienen la misma frecuencia se denominan contenedores de Completación igual.

Puede ejecutar más funciones binning mediante la herramienta **Administración de binning** de la Administración de modelo de aprendizaje automático.

También puede ver una lista de binning y eliminar binning utilizando instrucciones de línea de comandos. Consulte el “módulo Machine Learning” en la sección **Utilidad de administración** de la Guía de administración.

> **Nota:** Si actualiza Spectrum™ Technology Platform de la versión 12.0 SP1 a 12.0 SP2, deberá dejar de exponer manualmente cualquier binning actualizado en la herramienta Administración de binning de la Administración de modelo de aprendizaje automático antes de utilizarlos para volver a ejecutar binning en 12.0 SP2. No es obligatorio realizar este paso si utiliza binning actualizado en Binning Lookup en vez de binning tradicional.

Definición de las propiedades de binning

1. En **Etapas principales**/**Etapas implementadas**/**Machine Learning**, haga clic en la etapa **Binning** y arrástrela hasta el lienzo, colóquela donde desee en el flujo de datos y conéctela con otras etapas. Tenga en cuenta que la etapa de entrada debe ser el origen de datos que contiene los campos de variables objetivo y de entrada de su modelo. No se requiere una etapa de salida, a menos que seleccione la opción Calificar datos de entrada en la pestaña Opciones

básicas. También puede conectar una etapa de salida si desea capturar su salida, independiente de la herramienta de gestión de modelo Machine Learning.

2. Haga doble clic en la etapa Binning para que aparezca el cuadro de diálogo **Opciones de binning**.
3. Ingrese un **Nombre de binning** si no desea utilizar el nombre predeterminado.
4. Marque la casilla **Sobrescribir** para sobrescribir el modelo existente con datos nuevos.
5. Ingrese una **Descripción** del modelo.
6. Haga clic en **Incluir** para cada campo cuyos datos desea agregar al binning. Tenga en cuenta que solo aparecerán campos numéricos en esta lista.
7. Haga clic en **Aceptar** para guardar la configuración.

Configuración de opciones básicas

1. Seleccione si desea realizar un **estilo de binning** de rango o de población equivalente.
2. En **Intervalo de valor nulo**, seleccione cómo desea manejar los campos bin vacíos, que representan valores desconocidos debido a datos faltantes. Seleccione **El más alto** para asignar valores nulos al bin más alto y seleccione **El más bajo** para asignar valores nulos al bin más bajo. El bin más bajo siempre será el bin 1.
3. Haga clic en **Bines internos objetivo** e ingrese la cantidad de bines que desea completar entre los bines finales. Si realiza binning de rango equivalente, puede seleccionar este tipo de procesamiento o el **Ancho de bin**, pero no ambos. Si realiza binning de población equivalente, solo pueden realizar el procesamiento de bin interno.
4. Si realizar binning de rango equivalente y desea seleccionar este tipo de procesamiento en lugar del procesamiento de bin interno, haga clic en **Ancho de bin** e ingrese la cantidad de unidades que desea en cada bin.
5. Haga clic en **Incluir** para cada campo cuyos datos desea agregar al binning. Tenga en cuenta que solo aparecerán campos numéricos en esta lista.
6. Haga clic en **Aceptar** para guardar la configuración.

Salida Binning

La etapa Binning tiene dos puertos de salida. El primer puerto enviará todos los campos de entrada más un campo de bin por cada campo de entrada seleccionado. Por ejemplo, si la entrada contiene los campos Nombre, Edad e Ingresos y se realiza binning en los campos Edad e Ingresos, la salida del primer puerto contendrá los siguientes campos:

- Nombre

- Edad
- Binned_Age
- Ingresos
- Binned_Income

El segundo puerto envía cuatro tipos de información por cada campo de entrada seleccionado. Por ejemplo, si realiza binning en el campo Edad, la salida del segundo puerto contendrá los siguientes campos:

- Age_Bins
- Age_BinValue
- Age_Count
- Age_Percentage

3 - K-Means Clustering

In this section

Introducción

K-Means Clustering permite crear modelos según agrupaciones en clústeres analíticas, lo cual segmenta un conjunto de registros en clústeres de registros similares a partir de valores de datos.

Para crear su modelo, primero debe completar la ficha Propiedades del modelo. Las fichas Opciones básicas y Opciones avanzadas ofrecen configuraciones predeterminadas suficientes para completar un trabajo, pero puede cambiarlas de acuerdo con sus necesidades. Luego se puede ejecutar el trabajo y una versión limitada de los detalles de salida de modelo resultantes aparece en la pestaña Salida de modelo; el modelo es almacenado en el servidor de Spectrum™ Technology Platform y la salida completa se encuentra disponible en la herramienta de gestión de modelos de Machine Learning.

Definición de las propiedades del modelo

1. En **Etapas principales**/**Etapas implementadas**/**Machine Learning**, haga clic en la etapa **K-Means Clustering** y arrástrela hasta el lienzo, colóquela donde desee en el flujo de datos y conéctela con otras etapas. Tenga en cuenta que la etapa de entrada debe ser el origen de datos que contiene los campos de variables de entrada de su modelo. No se requiere una etapa de salida, a menos que seleccione la opción Calificar datos de entrada en la pestaña Opciones básicas. También puede conectar una etapa de salida si desea capturar su salida, independiente de la herramienta de gestión de modelo Machine Learning.
2. Haga doble clic en la etapa K-Means para que aparezca el cuadro de diálogo **Opciones de K-Means Clustering**.
3. Ingrese un **Nombre de modelo** si no desea utilizar el nombre predeterminado.
4. Opcional: marque la casilla **Sobrescribir** para sobrescribir el modelo existente con datos nuevos.
5. Ingrese el **Número de agrupamiento** que desea en su modelo si no quiere el número predeterminado (5).
6. Opcional: Ingrese una **Descripción** del modelo.
7. Haga clic en **Incluir** para cada campo cuyos datos desea agregar al modelo.
8. Utilice la lista desplegable **Tipo de datos de modelo** para especificar si el campo de entrada se debe utilizar como un campo numérico, categórico, o de fecha y hora.
9. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones básicas

1. Deje marcada la opción **Estandarizar campos de entrada** para estandarizar las columnas numéricas a fin de que la variación media y por unidad sea igual a cero.

   Si no utiliza la estandarización, los resultados podrían incluir componentes dominados por variables que aparentarán tener variaciones mayores en relación con otros atributos como una cuestión de escala y no como una contribución verdadera.
2. Revise el **Número estimado de agrupamiento** para hacer que el algoritmo de K-Means intente determinar el número de agrupamiento que contendrá el modelo. Aunque designe el número de agrupamiento deseado en la pestaña Propiedades del modelo, la rutina podría descubrir durante su procesamiento que un número de agrupamiento diferente resulta más apropiado en vista de los datos.
3. Especifique un valor entre 1 y 100 como **Porcentaje para datos de capacitación** cuando los datos de entrada se dividen aleatoriamente en muestras de datos de capacitación y de prueba.
4. Ingrese el valor de 100 menos la cantidad que ingresó en el Paso 5 como **Porcentaje para datos de prueba**.
5. Ingrese un número en **Propagar para muestras** para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, esto ocurra siempre de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
6. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones avanzadas

1. Deje marcada la opción **Ignorar campos constantes** para omitir campos que tienen el mismo valor para cada registro.
2. Marque la opción **Inicialización para algoritmo** e ingrese una cantidad de propagaciones para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, siempre se produzca de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
3. Seleccione el modo de inicialización correcto en el menú desplegable **Inicialización**.

   Más lejano Inicializa el primer centroide al azar, pero luego inicializa el segundo centroide para que sea el punto de datos más lejano de él. Inicializa los centroides para que queden bien separados entre sí.

Plus-Plus	Inicializa los centros de clúster antes de proceder con las iteraciones de optimización *k*-means estándar. Con la inicialización *k*-means++, se garantiza que el algoritmo encuentre una solución que es O (registro *k*) competitiva con la solución *k*-means óptima.
Aleatorio	Opción predeterminada. Elige clústeres K del conjunto de observaciones N en forma aleatoria, de manera que cada observación tenga la misma posibilidad de ser elegida.

4. Deje marcada la opción **Propagar para N iteraciones** e ingrese el número de propagación para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, esto ocurra siempre de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
5. Marque la opción **N iteraciones** e ingrese la cantidad de iteraciones si va a realizar una validación cruzada.
6. Marque la opción **Asignación de iteración** y seleccione la lista despegable si está ejecutando una validación cruzada. Este campo solo se aplica si ingresó un valor en **N iteraciones**.

Automático	Opción predeterminada. Permite que el algoritmo seleccione automáticamente una opción; actualmente utiliza Aleatorio.
Módulo	Divide de manera uniforme el conjunto de datos en las iteraciones y no depende de la raíz.
Aleatorio	Divide de manera aleatoria los datos en piezas de n iteraciones; es ideal para grandes conjuntos de datos.

7. Marque la opción **Iteración máxima** e ingrese el número de iteraciones de capacitación que deben ocurrir.
8. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la pestaña siguiente.

Salida de modelo

Esta pestaña muestra las métricas que está usando para evaluar el modelo ajustado. No puede editar estos campos. La columna Capacitación siempre va a contener datos. Si seleccionó una división capacitación/prueba en la pestaña Opciones básicas, también se completará la columna Prueba, a menos que haya seleccionado una validación de N iteraciones en la pestaña Opciones avanzadas, en cuyo caso se completará la columna N iteraciones. Haga clic en el botón **Salida** para regenerar los datos de salida y luego en **Detalles del modelo** para ver los datos de salida completos en la herramienta Administración de modelo Machine Learning.

Puerto de salida

En la etapa K-Means Clustering se proporciona un puerto de salida opcional: Puerto de métricas de modelo. La funcionalidad de este puerto se determina según sus selecciones y entradas cuando se completan las opciones básicas y avanzadas de la etapa. Por ejemplo, si elige realizar una validación de N iteraciones marcando el campo **N iteraciones** en la pestaña Opciones avanzadas, la columna de N iteraciones en las métricas de salida se completará con datos. De forma alternativa, si elige no realizar una validación de N iteraciones, la columna de N iteraciones permanecerá en blanco.

Puerto de métricas de modelo

El **Puerto de métricas de modelo** le permite generar las métricas de la evaluación de modelo en un archivo de datos. Este puerto lo ayudará a comparar muchos modelos generados desde dentro y fuera de Spectrum™ Technology Platform, y a realizar tareas de procesamiento de otros datos en las métricas.

Realice los siguientes pasos para utilizar el Puerto de métricas de modelo:

1. Abra un flujo de datos en el que se utilice la etapa K-Means Clustering.
2. Adjunte una etapa Write to File u otra etapa de salida de datos al puerto de salida secundario.
3. Ejecute el trabajo.
4. Opción alternativa al Paso 3: haga clic derecho en el canal y seleccione “Agregar punto de inspección” para agregar un punto de inspección al canal que permite conectar la etapa K-Means Clustering a la etapa receptora agregada en el Paso 2. Luego, haga clic en el botón Inspeccionar flujo actual () en la barra de herramientas de Enterprise Designer. La inspección se ejecutará y verá resultados similares a aquellos que se muestran a continuación.

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 - Regresión lineal

In this section

Introducción

La regresión lineal le permite realizar aprendizaje automático mediante la creación de modelos a partir de conjuntos de datos que usan objetivos continuos con variables de entrada.

Para crear su modelo, primero debe completar la ficha Propiedades del modelo. Las fichas Opciones básicas y Opciones avanzadas ofrecen configuraciones predeterminadas suficientes para completar un trabajo, pero puede cambiarlas de acuerdo con sus necesidades. Luego se puede ejecutar el trabajo y una versión limitada del modelo resultante aparece en la pestaña Salida de modelo; la salida completa se encuentra disponible en la herramienta de gestión de modelos de Machine Learning.

Definición de las propiedades del modelo

1. En **Etapas principales**/**Etapas implementadas**/**Machine Learning**, haga clic en la etapa **Regresión lineal** y arrástrela hasta el lienzo, colóquela donde desee en el flujo de datos y conéctela con otras etapas. Tenga en cuenta que la etapa de entrada debe ser el origen de datos que contiene los campos de variables objetivo y de entrada de su modelo. No se requiere una etapa de salida, a menos que seleccione la opción Calificar datos de entrada en la pestaña Opciones básicas. También puede conectar una etapa de salida si desea capturar su salida, independiente de la herramienta de gestión de modelo Machine Learning.
2. Haga doble clic en la etapa Regresión lineal para que aparezca el cuadro de diálogo **Opciones de regresión lineal**.
3. Ingrese un **Nombre de modelo** si no desea utilizar el nombre predeterminado.
4. Marque la casilla **Sobrescribir** para sobrescribir el modelo existente con datos nuevos.
5. Haga clic en la lista desplegable **Campo objetivo** y seleccione un campo numérico.
6. Ingrese una **Descripción** del modelo.
7. Haga clic en **Incluir** para cada campo cuyos datos desea agregar al modelo y asegúrese de incluir el campo que seleccionó como el campo Objetivo.
8. Utilice la lista desplegable **Tipo de datos de modelo** para especificar si cada campo de entrada se debe utilizar como un campo numérico, categórico, o de fecha y hora.
9. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones básicas

1. Deje marcada la opción **Estandarizar campos de entrada** para estandarizar las columnas numéricas a fin de que la variación media y por unidad sea igual a cero.

   Si no utiliza la estandarización, los resultados podrían incluir componentes dominados por variables que aparentarán tener variaciones mayores en relación con otros atributos como una cuestión de escala y no como una contribución verdadera.
2. Marque la opción **Calificar datos de entrada** para agregar una columna para la predicción del modelo (calificación) a los datos de entrada.
3. Seleccione una **Función de enlace** de la lista desplegable. Esto especifica el enlace entre los componentes sistemáticos y aleatorios. Indica cómo se relaciona el valor esperado de la respuesta con el predictor lineal de las variables explicativas.

Identidad	Predice las “probabilidades” sin sentido menores que cero o mayores que uno y a veces se utiliza para obtener datos binomiales para producir un modelo de probabilidad lineal. *g(p) = p*
Inverso	Calcula el inverso de las funciones de enlace para obtener estimados reales. *g(μi) = 1μi*
Log	Las apariciones de recuentos en una cantidad fija de tiempo y espacio. *g(μi) = log(μi)*

4. Para especificar cómo manejar los datos faltantes, marque **Omitir** o **Imputar medios**, que agregará el valor medio para cualquier dato faltante.
5. Especifique un valor entre 1 y 100 como **Porcentaje para datos de capacitación** cuando los datos de entrada se dividen aleatoriamente en muestras de datos de capacitación y de prueba.
6. Ingrese el valor de 100 menos la cantidad que ingresó en el Paso 5 como **Porcentaje para datos de prueba**.
7. Ingrese un número en **Propagar para muestras** para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, esto ocurra siempre de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
8. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones avanzadas

1. Deje marcada la opción **Ignorar campos constantes** para omitir campos que tienen el mismo valor para cada registro.
2. Marque la opción **Calcular valores de p** para calcular valores de p a fin de obtener las estimaciones de parámetros.
3. Marque la opción **Quitar columna alineada** para quitar automáticamente las columnas alineadas durante la construcción del modelo. Esto dará como resultado un coeficiente de 0 en el modelo devuelto.

   Esta opción debe estar marcada si la opción **Calcular valores de p** también está marcada.
4. Deje marcada la opción **Incluir término constante (interceptar)** para incluir un término constante (interceptar) en el modelo.

   Debe marcar este campo si también marca la opción **Quitar columna alineada**.
5. Seleccione un **Solucionador** desde la lista desplegable. Tenga en cuenta que CoordinateDescent y CoordinateDescentNaive se encuentran en etapa experimental.

Automático	El solucionador se determinará en función de los datos y parámetros de entrada.
CoordinateDescent	IRLSM con la versión de actualizaciones de covarianza del descenso cíclico por coordenadas en el bucle interior.
CoordinateDescentNaive	IRLSM con la versión de actualizaciones naive del descenso cíclico por coordenadas en el bucle interior.
IRLSM	Ideal para problemas con una pequeña cantidad de predictores o para búsquedas Lambda con penalidad L1.
LBFGS	Ideal para conjuntos de datos con muchas columnas.
6. Deje marcada la opción **Propagar para N iteraciones** e ingrese el número de propagación para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, esto ocurra siempre de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
7. Marque la opción **N iteraciones** e ingrese la cantidad de iteraciones si va a realizar una validación cruzada.
8. Haga clic en **Asignación de iteración** y seleccione de la lista despegable si está ejecutando una validación cruzada. Este campo solo se aplica si ingresó un valor en **N iteraciones** y no se especificó el **Campo de iteración**.

Automático	Permite que el algoritmo seleccione automáticamente una opción; actualmente utiliza Aleatorio.

Módulo	Divide de manera uniforme el conjunto de datos en las iteraciones y no depende de la raíz.
Aleatorio	Divide de manera aleatoria los datos en piezas de n iteraciones; es ideal para grandes conjuntos de datos.

9. Si está ejecutando una validación cruzada, marque la opción **Campo de iteración** y seleccione el campo que contiene la asignación del índice de iteración de validación cruzada en la lista desplegable.

   Este campo solo se aplica si no ingresó un valor en **N iteraciones** y **Asignación de iteración**.
10. Marque la opción **Iteración máxima** e ingrese el número de iteraciones de capacitación que deben ocurrir.
11. Marque la opción **Objetivo épsilon** e ingrese el umbral de convergencia; este debe ser un valor entre 0 y 1. Si el valor objetivo es menor que este umbral, el modelo se convergerá.
12. Marque la opción **Beta épsilon** e ingrese el umbral de convergencia; este debe ser un valor entre 0 y 1. Si el valor objetivo es menor que este umbral, el modelo se convergerá. Si la normalización L1 del cambio beta actual está por debajo de este umbral, considere el uso de la convergencia.
13. Una de las preocupaciones del modelado predictivo es el sobreajuste que ocurre cuando un modelo de análisis se asemeja demasiado (o exactamente) a un conjunto de datos específico y, por ende, podría ocurrir un error cuando se aplica a datos adicionales o a futuras observaciones. Uno de los métodos utilizados para reducir el sobreajuste es la regularización. Seleccione el **Tipo de regularización** que desea utilizar.

LASSO (Operador de selección y reducción menos absoluto)	Mediante esta regularización se selecciona un subconjunto pequeño de variables con un valor de lambda tan alto que pueda considerarse crucial. Es posible que no pueda ejecutarse correctamente si existen variables predictoras correlacionadas, ya que seleccionará una variable del grupo correlacionado y quitará las demás. También se limita según la dimensionalidad amplia; cuando un modelo contiene más variables que registros, LASSO se limitará según la cantidad de variables que pueda seleccionar. Ridge Regression no tiene esta limitación. Cuando el número de variables incluidas en este modelo es grande, o si la solución es dispersa, se recomienda utilizar LASSO.
Ridge Regression	Mediante esta regularización se retienen todas las variables predictoras y se reducen sus coeficientes proporcionalmente. Cuando existen variables predictoras correlacionadas, Ridge Regression ayuda a reducir los coeficientes del grupo completo de variables correlacionadas para equipararlas. Si no desea quitar las variables predictoras correlacionadas de su modelo, utilice Ridge Regression.
Elastic Net	Combina LASSO y Ridge Regression cuando actúa como un selector de variable mientras ayuda a preservar el efecto grupal en las variables correlacionadas (se reducen simultáneamente coeficientes de variables

correlacionadas). Elastic Net no se limita según la dimensionalidad amplia y ayuda a evaluar todas las variables cuando un modelo contiene más variables que registros.

14. Marque **Valor de alfa** y cambie el valor si no desea utilizar el valor predeterminado 5. Mediante el parámetro alfa se controla la distribución entre las penalizaciones ℓ1 y ℓ2. Rango de valores válidos entre 0 y 1; con un valor de 1.0 se representa a LASSO, mientras que con un valor de 0.0 se produce Ridge Regression. En la tabla anterior se ilustra el efecto de alfa y lambda en la regularización.

`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

Nota: El signo igual único corresponde a una operación de asignación que significa “es”, mientras que el signo igual doble corresponde a un operador de igualdad y significa “igual a”.

15. Marque **Valor de lambda** y especifique un valor si no desea utilizar el método de cálculo del valor de lambda predeterminado mediante Regresión lineal, la cual es una heurística basada en datos de capacitación. Mediante el parámetro lambda se controla la cantidad de regularización aplicada. Por ejemplo, si lambda corresponde a 0.0, no se aplica una regularización y se ignora el parámetro alfa.
16. Marque **Buscar valor óptimo de lambda** para obtener modelos de cálculo de Regresión lineal para una ruta de regularización completa que inicie en lambda max (el valor más alto de lambda que tenga sentido; es decir, el valor más bajo con el que los coeficientes lleguen a cero) y termine en lambda min en la escala de registro, lo que provocaría una reducción de la solidez de la regularización en cada paso. Con el modelo obtenido se tendrán coeficientes que corresponden al valor óptimo de lambda como se decidió durante la capacitación.
17. Marque **Detener antes** para finalizar un procesamiento cuando ya no exista ninguna mejora en la capacitación o en el conjunto de validación.
18. Marque **Máximo de búsqueda de lambdas** e ingrese la cantidad máxima de lambdas que utilizará durante el proceso de búsqueda de lambda.
19. Marque **Máximo de predictores activos** e ingrese la cantidad máxima de predictores que utilizará durante los cálculos. Este valor se utiliza como un criterio de detención para prevenir la creación de un modelo costoso con muchos predictores.
20. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Salida de modelo

Esta pestaña muestra las métricas que está usando para evaluar el modelo ajustado. No puede editar estos campos. La columna Capacitación siempre va a contener datos. Si seleccionó una división capacitación/prueba en la pestaña Opciones básicas, también se completará la columna Prueba, a menos que haya seleccionado una validación de N iteraciones en la pestaña Opciones avanzadas, en cuyo caso se completará la columna N iteraciones.

Después de ejecutar su trabajo, el modelo resultante se guarda en el servidor Spectrum™ Technology Platform. Haga clic en el botón **Salida** para regenerar los datos de salida y luego en **Detalles del modelo** para ver los datos de salida completos en la herramienta Administración de modelo Machine Learning.

Puertos de salida

En la etapa Regresión lineal se proporcionan dos puertos de salida opcionales: Puerto de calificación de modelo y Puerto de métricas de modelo. La funcionalidad de estos puertos se determina según sus selecciones y entradas cuando se completan las opciones básicas y avanzadas de la etapa. Por ejemplo, si elige realizar una validación de N iteraciones marcando el campo **N iteraciones** en la pestaña Opciones avanzadas, la columna de N iteraciones en las métricas de salida generadas en el Puerto de métricas de modelo se completará con datos. De forma alternativa, si elige no realizar una validación de N iteraciones, la columna de N iteraciones permanecerá en blanco. Del mismo modo, el Puerto de calificación de modelo se activa si marca el campo **Calificar datos de entrada** en la pestaña Opciones básicas.

Puerto de calificación de modelo

Cuando marca el campo **Calificar datos de entrada** en la pestaña Opciones básicas, se calculan valores previstos mediante Regresión lineal cuando se crea el modelo con el que se agrega la columna **Predicted_Value** una por una para esa calificación en los datos de salida. Puede adjuntar cualquier tipo de receptor a este puerto: una etapa Write to File, una etapa Write to Null, etc.

Puerto de métricas de modelo

El **Puerto de métricas de modelo** le permite generar las métricas de la evaluación de modelo en un archivo de datos. Este puerto lo ayudará a comparar muchos modelos generados desde dentro y fuera de Spectrum™ Technology Platform, y a realizar tareas de procesamiento de otros datos en las métricas.

Realice los siguientes pasos para utilizar el Puerto de métricas de modelo:

1. Abra un flujo de datos en el que se utilice la etapa Regresión lineal.
2. Adjunte una etapa Write to File u otra etapa de salida de datos al puerto de salida secundario.
3. Ejecute el trabajo.
4. Opción alternativa al Paso 3: haga clic derecho en el canal y seleccione “Agregar punto de inspección” para agregar un punto de inspección al canal que permite conectar la etapa Regresión lineal a la etapa receptora agregada en el Paso 2. Luego, haga clic en el botón Inspeccionar flujo actual () en la barra de herramientas de Enterprise Designer. La inspección se ejecutará y verá resultados similares a aquellos que se muestran a continuación.

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

5 - Logistic Regression

In this section

Introducción

Logistic Regression le permite realizar aprendizaje de máquina mediante la creación de modelos a partir de conjuntos de datos que usan objetivos binarios con variables de entrada.

Para crear su modelo, primero debe completar la ficha Propiedades del modelo. Las fichas Opciones básicas y Opciones avanzadas ofrecen configuraciones predeterminadas suficientes para completar un trabajo, pero puede cambiarlas de acuerdo con sus necesidades. Luego se puede ejecutar el trabajo y una versión limitada del modelo resultante aparece en la pestaña Salida de modelo; la salida completa se encuentra disponible en la herramienta de gestión de modelos de Machine Learning.

Definición de las propiedades del modelo

1. En **Etapas principales**/**Etapas implementadas**/**Machine Learning**, haga clic en la etapa **Logistic Regression** y arrástrela hasta el lienzo, colóquela donde desee en el flujo de datos y conéctela con otras etapas. Tenga en cuenta que la etapa de entrada debe ser el origen de datos que contiene los campos de variables objetivo y de entrada de su modelo. No se requiere una etapa de salida, a menos que seleccione la opción Calificar datos de entrada en la pestaña Opciones básicas. También puede conectar una etapa de salida si desea capturar su salida, independiente de la herramienta de gestión de modelo Machine Learning.
2. Haga doble clic en la etapa Logistic Regression para que aparezca el cuadro de diálogo **Opciones de Logistic Regression**.
3. Ingrese un **Nombre de modelo** si no desea utilizar el nombre predeterminado.
4. Marque la casilla **Sobrescribir** para sobrescribir el modelo existente con datos nuevos.
5. Haga clic en la opción desplegable **Campo objetivo** y seleccione "Categórico".
6. Ingrese una **Descripción** del modelo.
7. Haga clic en **Incluir** para cada campo cuyos datos desea agregar al modelo y asegúrese de incluir el campo que seleccionó como el campo Objetivo.
8. Utilice la lista desplegable **Tipo de datos de modelo** para especificar si cada campo de entrada se debe utilizar como un campo numérico, categórico, o de fecha y hora.
9. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones básicas

1. Deje marcada la opción **Estandarizar campos de entrada** para estandarizar las columnas numéricas a fin de que la variación media y por unidad sea igual a cero.

   Si no utiliza la estandarización, los resultados podrían incluir componentes dominados por variables que aparentarán tener variaciones mayores en relación con otros atributos como una cuestión de escala y no como una contribución verdadera.
2. Marque la opción **Calificar datos de entrada** para agregar una columna para la predicción del modelo (calificación) a los datos de entrada.
3. Marque **Anterior** si se tomaron muestras de los datos y la media de respuesta no refleja la realidad; luego, ingrese la probabilidad anterior para p(y==1) en el campo de texto.
4. Para especificar cómo manejar los datos faltantes, marque **Omitir** o **Imputar medios**, que agregará el valor medio para cualquier dato faltante.
5. Especifique un valor entre 1 y 100 como **Porcentaje para datos de capacitación** cuando los datos de entrada se dividen aleatoriamente en muestras de datos de capacitación y de prueba.
6. Ingrese el valor de 100 menos la cantidad que ingresó en el Paso 5 como **Porcentaje para datos de prueba**.
7. Ingrese un número en **Propagar para muestras** para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, esto ocurra siempre de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
8. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones avanzadas

1. Deje marcada la opción **Ignorar campos constantes** para omitir campos que tienen el mismo valor para cada registro.
2. Deje marcada la opción **Calcular valores de p** para calcular valores de p para las estimaciones de parámetros.
3. Deje marcada la opción **Quitar columna alineada** para quitar automáticamente las columnas alineadas durante la construcción del modelo. Esto dará como resultado un coeficiente de 0 en el modelo devuelto.

   Esta opción debe estar marcada si la opción **Calcular valores de p** también está marcada.
4. Deje marcada la opción **Incluir término constante (interceptar)** para incluir un término constante (interceptar) en el modelo.

Debe marcar este campo si también marca la opción **Quitar columna alineada**.

5. Seleccione un **Solucionador** desde la lista desplegable. Tenga en cuenta que CoordinateDescent y CoordinateDescentNaive se encuentran en etapa experimental.

Automático	El solucionador se determinará en función de los datos y parámetros de entrada.
CoordinateDescentNaive	IRLSM con la versión de actualizaciones de covarianza del descenso cíclico por coordenadas en el bucle interior.
CoordinateDescentNaive	IRLSM con la versión de actualizaciones naive del descenso cíclico por coordenadas en el bucle interior.
IRLSM	Ideal para problemas con una pequeña cantidad de predictores o para búsquedas Lambda con penalidad L1.
L_BFGS	Ideal para conjuntos de datos con muchas columnas.

6. Deje marcada la opción **Propagar para N iteraciones** e ingrese el número de propagación para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, esto ocurra siempre de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
7. Marque la opción **N iteraciones** e ingrese la cantidad de iteraciones si va a realizar una validación cruzada.
8. Marque la opción **Asignación de iteración** y seleccione la lista despegable si está ejecutando una validación cruzada. Este campo solo se aplica si ingresó un valor en **N iteraciones** y no se especificó el **Campo de iteración**.

Automático	Permite que el algoritmo seleccione automáticamente una opción; actualmente utiliza Aleatorio.
Módulo	Divide de manera uniforme el conjunto de datos en las iteraciones y no depende de la raíz.
Aleatorio	Divide de manera aleatoria los datos en piezas de n iteraciones; es ideal para grandes conjuntos de datos.
Estratificado	Estratifica las iteraciones en función de la variable de respuesta para los problemas de clasificación. Distribuye de manera uniforme las observaciones de las diferentes clases en todos los conjuntos mediante la división de un conjunto de datos en datos de capacitación y de prueba. Puede resultar útil si hay muchas clases y el conjunto de datos es relativamente pequeño.

9. Si está ejecutando una validación cruzada, marque la opción **Campo de iteración** y seleccione el campo que contiene la asignación del índice de iteración de validación cruzada en la lista desplegable.

   Este campo solo se aplica si no ingresó un valor en **N iteraciones** y **Asignación de iteración**.

10. Marque la opción **Iteración máxima** e ingrese el número de iteraciones de capacitación que deben ocurrir.
11. Marque la opción **Objetivo épsilon** e ingrese el umbral de convergencia; este debe ser un valor entre 0 y 1. Si el valor objetivo es menor que este umbral, el modelo se convergerá.
12. Marque la opción **Beta épsilon** e ingrese el umbral de convergencia; este debe ser un valor entre 0 y 1. Si el valor objetivo es menor que este umbral, el modelo se convergerá. Si la normalización L1 del cambio beta actual está por debajo de este umbral, considere el uso de la convergencia.
13. Una de las preocupaciones del modelado predictivo es el sobreajuste que ocurre cuando un modelo de análisis se asemeja demasiado (o exactamente) a un conjunto de datos específico y, por ende, podría ocurrir un error cuando se aplica a datos adicionales o a futuras observaciones. Uno de los métodos utilizados para reducir el sobreajuste es la regularización. Seleccione el **Tipo de regularización** que desea utilizar.

LASSO (Operador de selección y reducción menos absoluto)	Mediante esta regularización se selecciona un subconjunto pequeño de variables con un valor de lambda tan alto que pueda considerarse crucial. Es posible que no pueda ejecutarse correctamente si existen variables predictoras correlacionadas, ya que seleccionará una variable del grupo correlacionado y quitará las demás. También se limita según la dimensionalidad amplia; cuando un modelo contiene más variables que registros, LASSO se limitará según la cantidad de variables que pueda seleccionar. Ridge Regression no tiene esta limitación. Cuando el número de variables incluidas en este modelo es grande, o si la solución es dispersa, se recomienda utilizar LASSO.
Ridge Regression	Mediante esta regularización se retienen todas las variables predictoras y se reducen sus coeficientes proporcionalmente. Cuando existen variables predictoras correlacionadas, Ridge Regression ayuda a reducir los coeficientes del grupo completo de variables correlacionadas para equipararlas. Si no desea quitar las variables predictoras correlacionadas de su modelo, utilice Ridge Regression.
Elastic Net	Combina LASSO y Ridge Regression cuando actúa como un selector de variable mientras ayuda a preservar el efecto grupal en las variables correlacionadas (se reducen simultáneamente coeficientes de variables correlacionadas). Elastic Net no se limita según la dimensionalidad amplia y ayuda a evaluar todas las variables cuando un modelo contiene más variables que registros.

14. Marque **Valor de alfa** y cambie el valor si no desea utilizar el valor predeterminado 5. Mediante el parámetro alfa se controla la distribución entre las penalizaciones ℓ1 y ℓ2. Rango de valores válidos entre 0 y 1; con un valor de 1.0 se representa a LASSO, mientras que con un valor de 0.0 se produce Ridge Regression. En la tabla anterior se ilustra el efecto de alfa y lambda en la regularización.

`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

Nota: El signo igual único corresponde a una operación de asignación que significa "es", mientras que el signo igual doble corresponde a un operador de igualdad y significa "igual a".

15. Marque **Valor de lambda** y especifique un valor si no desea utilizar el método de cálculo del valor de lambda predeterminado mediante Logistic Regression, la cual es una heurística basada en datos de capacitación. Mediante el parámetro lambda se controla la cantidad de regularización aplicada. Por ejemplo, si lambda corresponde a 0.0, no se aplica una regularización y se ignora el parámetro alfa.
16. Marque **Buscar valor óptimo de lambda** para obtener modelos de cálculo de Logistic Regression para una ruta de regularización completa que inicie en lambda max (el valor más alto de lambda que tenga sentido; es decir, el valor más bajo con el que los coeficientes lleguen a cero) y termine en lambda min en la escala de registro, lo que provocaría una reducción de la solidez de la regularización en cada paso. Con el modelo obtenido se tendrán coeficientes que corresponden al valor óptimo de lambda como se decidió durante la capacitación.
17. Marque **Detener antes** para finalizar un procesamiento cuando ya no exista ninguna mejora en la capacitación o en el conjunto de validación.
18. Marque **Máximo de búsqueda de lambdas** e ingrese la cantidad máxima de lambdas que utilizará durante el proceso de búsqueda de lambda.
19. Marque **Máximo de predictores activos** e ingrese la cantidad máxima de predictores que utilizará durante los cálculos. Este valor se utiliza como un criterio de detención para prevenir la creación de un modelo costoso con muchos predictores.
20. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Salida de modelo

Esta pestaña muestra las métricas que está usando para evaluar el modelo ajustado. No puede editar estos campos. La columna Capacitación siempre va a contener datos. Si seleccionó una división capacitación/prueba en la pestaña Opciones básicas, también se completará la columna Prueba, a menos que haya seleccionado una validación de N iteraciones en la pestaña Opciones avanzadas, en cuyo caso se completará la columna N iteraciones.

Después de ejecutar su trabajo, el modelo resultante se guarda en el servidor Spectrum™ Technology Platform. Haga clic en el botón **Salida** para regenerar los datos de salida y luego en **Detalles del modelo** para ver los datos de salida completos en la herramienta Administración de modelo Machine Learning.

Puertos de salida

En la etapa Logistic Regression se proporcionan dos puertos de salida opcionales: Puerto de calificación de modelo y Puerto de métricas de modelo. La funcionalidad de estos puertos se determina según sus selecciones y entradas cuando se completan las opciones básicas y avanzadas de la etapa. Por ejemplo, si elige realizar una validación de N iteraciones marcando el campo **N iteraciones** en la pestaña Opciones avanzadas, la columna de N iteraciones en las métricas de salida generadas en el Puerto de métricas de modelo se completará con datos. De forma alternativa, si elige no realizar una validación de N iteraciones, la columna de N iteraciones permanecerá en blanco. Del mismo modo, el Puerto de calificación de modelo se activa si marca el campo **Calificar datos de entrada** en la pestaña Opciones básicas.

Puerto de calificación de modelo

Cuando marca el campo **Calificar datos de entrada** en la pestaña Opciones básicas, se calculan valores previstos mediante Logistic Regression cuando se crea el modelo con el que se agregan las columnas **Predicted_Value**, **Probability_of_class_A** y **Probability_of_class_B** una por una para esa calificación en los datos de salida. Puede adjuntar cualquier tipo de receptor a este puerto: una etapa Write to File, una etapa Write to Null, etc.

Puerto de métricas de modelo

El **Puerto de métricas de modelo** le permite generar las métricas de la evaluación de modelo en un archivo de datos. Este puerto lo ayudará a comparar muchos modelos generados desde dentro y fuera de Spectrum™ Technology Platform, y a realizar tareas de procesamiento de otros datos en las métricas.

Realice los siguientes pasos para utilizar el Puerto de métricas de modelo:

1. Abra un flujo de datos en el que se utilice la etapa Logistic Regression.
2. Adjunte una etapa Write to File u otra etapa de salida de datos al puerto de salida secundario.
3. Ejecute el trabajo.
4. Opción alternativa al Paso 3: haga clic derecho en el canal y seleccione “Agregar punto de inspección” para agregar un punto de inspección al canal que permite conectar la etapa Logistic

Regression a la etapa receptora agregada en el Paso 2. Luego, haga clic en el botón Inspeccionar flujo actual () en la barra de herramientas de Enterprise Designer. La inspección se ejecutará y verá resultados similares a aquellos que se muestran a continuación.

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 - Análisis de componentes principales

In this section

Introducción

Análisis de componentes principales (PCA) es un proceso estadístico que convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables no correlacionadas linealmente, conocido como componentes principales.

Para crear su modelo, primero debe completar la ficha Propiedades del modelo. Las fichas Opciones básicas y Opciones avanzadas ofrecen configuraciones predeterminadas suficientes para completar un trabajo, pero puede cambiarlas de acuerdo con sus necesidades. Luego se puede ejecutar el trabajo y una versión limitada del modelo resultante aparece en la pestaña Salida de modelo; la salida completa se encuentra disponible en la herramienta de gestión de modelos de Machine Learning. Si está satisfecho con los datos de salida de su modelo, puede exponerlo y utilizarlo en un flujo de datos para evaluación.

Definición de las propiedades del modelo

1. En **Etapas principales**/**Etapas implementadas**/**Machine Learning**, haga clic en la etapa **Opciones PCA** y arrástrela hasta el lienzo, colóquela donde desee en el flujo de datos y conéctela con otras etapas. Observe que la etapa de entrada debe ser la fuente de datos que contiene los componentes principales para su modelo. No es necesaria una etapa de salida, pero puede conectar una si desea capturar su salida, independiente de la herramienta Administración de modelo Machine Learning.
2. Haga doble clic en la etapa Opciones PCA para que aparezca el cuadro de diálogo **Opciones PCA**.
3. Ingrese un **Nombre de modelo** si no desea utilizar el nombre predeterminado.
4. Opcional: marque la casilla **Sobrescribir** para sobrescribir el modelo existente con datos nuevos.
5. Ingrese el número de **Componentes principales** que desea que contenga su modelo.
6. Opcional: Ingrese una **Descripción** del modelo.
7. En la tabla **Entradas** haga clic en “Incluir” para cada campo cuyos datos desea agregar al modelo.
8. Utilice la lista desplegable **Tipo de datos de modelo** para especificar si el campo de entrada se debe utilizar como un campo categórico, de fecha y hora, numérico, de cadena o de ID única.
9. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones básicas

1. Deje **Usar todo el nivel de factor** desmarcada para omitir el primer componente principal, el cual tiene la variación más grande en los datos. Marque esta casilla para retener el primer componente principal.
2. Seleccione la opción **Transformar** adecuada para los datos de capacitación.

Degradar	Sustrae la media de cada columna.
Reducir escala	Divide la desviación estándar de cada columna.
Ninguno	
Normalizar	Degrada y divide cada columna por su rango (el máximo menos el mínimo).
Estandarizar	Utiliza varianza media y por unidad igual a cero. Opción predeterminada.

3. Para especificar cómo manejar los **Datos faltantes**, marque **Omitir** o **Imputar medios**, que agregará el valor medio para cualquier dato faltante.
4. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones avanzadas

1. Deje marcada la opción **Ignorar campos constantes** para omitir campos que tienen el mismo valor para cada registro.
2. Seleccione un **Método PCA** de la lista desplegable. Tenga en cuenta que GLRM y Potencia se encuentran en etapa experimental.

GLRM	Se ajusta a un modelo de rango bajo con pérdida de función L2 y sin regularización, soluciona la SVD con álgebra de matriz local. Esta opción está activada solo si marcó **Usar todo el nivel de factor** en la pestaña Opciones básicas.
GramSVD	Utiliza una computación distribuida de la matriz de Gram, seguida de una SVD local con el paquete JAMA.
Potencia	Calcula la SVD con el método de iteraciones de potencia.
Aleatorio	Utiliza el método de iteración de subespacio aleatorio.

3. Deje **Iteración máxima** desmarcada para tener un número ilimitado de iteraciones de capacitación (predeterminado). Marque la casilla e ingrese un número para limitar la cantidad de iteraciones de capacitación.
4. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Salida de modelo

Esta pestaña muestra las métricas que está usando para evaluar el modelo ajustado. No puede editar estos campos.

Después de ejecutar su trabajo, el modelo resultante se guarda en el servidor Spectrum™ Technology Platform. Haga clic en el botón **Salida** para regenerar los datos de salida y luego en **Detalles del modelo** para ver los datos de salida completos en la herramienta Administración de modelo Machine Learning.

Puerto de salida

En la etapa Análisis de componentes principales se proporciona un puerto de salida opcional: Puerto de métricas de modelo. La funcionalidad de este puerto se determina según sus selecciones y entradas cuando se completan las opciones básicas y avanzadas de la etapa. Por ejemplo, …

Puerto de métricas de modelo

El **Puerto de métricas de modelo** le permite generar las métricas de la evaluación de modelo en un archivo de datos. Este puerto lo ayudará a comparar muchos modelos generados desde dentro y fuera de Spectrum™ Technology Platform, y a realizar tareas de procesamiento de otros datos en las métricas.

Realice los siguientes pasos para utilizar el Puerto de métricas de modelo:

1. Abra un flujo de datos en el que se utilice la etapa PCA.
2. Adjunte una etapa Write to File u otra etapa de salida de datos al puerto de salida secundario.
3. Ejecute el trabajo.
4. Opción alternativa al Paso 3: haga clic derecho en el canal y seleccione “Agregar punto de inspección” para agregar un punto de inspección al canal que permite conectar la etapa PCA a la etapa receptora agregada en el Paso 2. Luego, haga clic en el botón Inspeccionar flujo

actual () en la barra de herramientas de Enterprise Designer. La inspección se ejecutará y verá resultados similares a aquellos que se muestran a continuación.

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	0.042617995404117 5	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	0.033104115450634 7	0.90637880520786

7 - Random Forest Classification

In this section

Introducción

Random Forest Classification le permite realizar aprendizaje de máquina mediante la creación de modelos a partir de conjuntos de datos que usan objetivos continuos con variables de entrada.

Para crear su modelo, primero debe completar la ficha Propiedades del modelo. Las fichas Opciones básicas y Opciones avanzadas ofrecen configuraciones predeterminadas suficientes para completar un trabajo, pero puede cambiarlas de acuerdo con sus necesidades. Luego se puede ejecutar el trabajo y una versión limitada del modelo resultante aparece en la pestaña Salida de modelo; la salida completa se encuentra disponible en la herramienta de gestión de modelos de Machine Learning.

> **Nota:** Haga clic **aquí** para obtener información adicional acerca de Random Forest Classification y sus opciones.

Definición de las propiedades del modelo

1. En **Etapas principales**/**Etapas implementadas**/**Machine Learning**, haga clic en la etapa **Random Forest Classification**, arrástrela hasta el lienzo, colóquela en el lugar que desee del flujo de datos y conéctela con otras etapas. Tenga en cuenta que la etapa de entrada debe ser el origen de datos que contiene los campos de variables objetivo y de entrada de su modelo. No se requiere una etapa de salida, a menos que seleccione la opción Calificar datos de entrada en la pestaña Opciones básicas. También puede conectar una etapa de salida si desea capturar su salida, independiente de la herramienta de gestión de modelo Machine Learning.
2. Haga doble clic en la etapa Random Forest Classification y se mostrará el cuadro de diálogo **Opciones de Random Forest Classification**.
3. Ingrese un **Nombre de modelo** si no desea utilizar el nombre predeterminado.
4. Opcional: marque la casilla **Sobrescribir** para sobrescribir el modelo existente con datos nuevos.
5. Haga clic en la lista desplegable **Campo objetivo** y seleccione un campo numérico.
6. Haga clic en **Niveles multinomiales** e ingrese la cantidad máxima de categorías en las que se puede agrupar el campo objetivo. Tenga en cuenta que, si marca esta opción, deshabilitará la opción **Calificar datos de entrada** en la pestaña Opciones básicas.
7. Opcional: Ingrese una **Descripción** del modelo.
8. Haga clic en **Incluir** para cada campo cuyos datos desea agregar al modelo y asegúrese de incluir el campo que seleccionó como el campo Objetivo.
9. Utilice la lista desplegable **Tipo de datos de modelo** para especificar si cada campo de entrada se debe utilizar como un campo numérico, categórico, o de fecha y hora.

10. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones básicas

1. Ingrese la **Cantidad de árboles** máxima en su modelo.
2. Ingrese la **Profundidad máxima** o la cantidad máxima de niveles que desea que tenga su modelo.
3. Ingrese la **Cantidad mínima de filas**; es decir, la cantidad mínima de registros (filas) que desea que tenga su modelo.
4. Ingrese la **Cantidad de contenedores numéricos**; es decir, la cantidad de contenedores que desea que genere el histograma y que luego divida en el mejor punto.
5. Ingrese la **Cantidad de contenedores de nivel superior**; es decir, la cantidad mínima de contenedores que desea tener a nivel de raíz.
6. Ingrese la **Cantidad de contenedores de categoría**; es decir, la cantidad máxima de contenedores que desea que genere el histograma y que luego divida en el mejor punto.
7. Revise la **Tasa de muestra** e ingrese el porcentaje de filas que se usarán como una muestra en cada árbol. Lo anterior puede ser un valor entre 0,0 y 1,0.
8. Revise la **Tasa de muestra de columna por árbol** e ingrese la tasa de muestra de columna para cada árbol. Lo anterior puede ser un valor entre 0,0 y 1,0.
9. Revise las **Columnas en cada nivel** e ingrese el cambio relativo de tasa de muestra de columna para cada nivel. El rango de valores válidos es desde 1,0 hasta el número del predictor de entrada seleccionado. El valor predeterminado es 1.0.
10. Marque la opción **Calificar datos de entrada** para agregar una columna para la predicción del modelo (calificación) a los datos de entrada. Tenga en cuenta que esta opción se deshabilita si marca **Niveles multinominales** en la pestaña Propiedades del modelo.
11. Especifique un valor entre 1 y 100 como **Porcentaje para datos de capacitación** cuando los datos de entrada se dividen aleatoriamente en muestras de datos de capacitación y de prueba.
12. Ingrese el valor de 100 menos la cantidad que ingresó en el Paso 5 como **Porcentaje para datos de prueba**.
13. Utilice el campo **Propagar para datos de prueba** para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, siempre se produzca de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
14. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones avanzadas

1. Deje marcada la opción **Ignorar campos constantes** para omitir campos que tienen el mismo valor para cada registro.
2. Revise **Equilibrar clases** para equilibrar la distribución de clases, ya sea para abarcar una muestra inferior de las clases mayoritarias o una mayor de las minoritarias.
3. Seleccione un **Tipo de histograma**.

   Automático Los depósitos se agrupan de mínimo a máximo en pasos de (máx-mín)/N. Utilice esta opción para especificar el tipo de histograma con el propósito de encontrar los puntos de división óptimos.

   QuantilesGlobal Los depósitos tienen una población equivalente. Lo anterior permite calcular los cuantiles `nbins` de cada columna numérica (no binaria) y luego refinar/rellenar cada depósito (entre dos cuantiles) de manera uniforme (y aleatoria para los elementos restantes) en un total de `nbins_top_level` agrupaciones.

   Aleatorio El algoritmo tomará muestras de los puntos N-1 de mínimo a máximo y utilizará la lista ordenada para encontrar la mejor división.

   RoundRobin El algoritmo pasará en ciclo por todos los tipos de histograma (uno por árbol).

   UniformAdaptive Permite agrupar cada característica en depósitos de tamaño de paso equivalente (no población). Este es el método más rápido, pero puede dar como resultado divisiones menos precisas si la distribución está muy desequilibrada.

4. Seleccione una **Codificación de categorías**.

   Automático Permite realizar una codificación `enum` de forma automática.

   Binario Permite convertir categorías en números enteros, luego en formato binario, y asignar cada dígito a una columna independiente. Codifica los datos en menos dimensiones, pero con algunas distorsiones en cuanto a las distancias.

   > **Nota:** No puede haber más de 32 columnas por característica de categoría.

   Eigen Columnas *k* por característica de categoría, que permiten mantener la proyección de matriz con codificación de asignación de estado activo uno (one-hot) solo en el espacio *k*-dim eigen.

Enum Permite pasar en ciclo por todos los tipos de histograma (uno por árbol).

OneHotExplicit Existe una columna por categoría, con el valor “1” o “0” en cada celda, lo que representa si la fila contiene esa categoría de columna.

5. Marque la opción **Propagar algoritmo y N iteraciones** e ingrese una cantidad de propagaciones para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, siempre se produzca de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
6. Marque la opción **N iteraciones** e ingrese la cantidad de iteraciones si va a realizar una validación cruzada.
7. Marque la opción **Asignación de iteración** y seleccione la lista despegable si ejecuta una validación cruzada. Este campo solo se aplica si ingresó un valor en **N iteraciones** y no se especificó el **Campo de iteración**.

   Automático Permite que el algoritmo seleccione automáticamente una opción; actualmente utiliza Aleatorio.

   Módulo Divide de manera uniforme el conjunto de datos en las iteraciones y no depende de la raíz.

   Aleatorio Divide de manera aleatoria los datos en piezas de n iteraciones; es ideal para grandes conjuntos de datos.

   Estratificado Estratifica las iteraciones en función de la variable de respuesta para los problemas de clasificación. Distribuye de manera uniforme las observaciones de las diferentes clases en todos los conjuntos mediante la división de un conjunto de datos en datos de capacitación y de prueba. Puede resultar útil si hay muchas clases y el conjunto de datos es relativamente pequeño.

8. Si está ejecutando una validación cruzada, marque la opción **Campo de iteración** y seleccione el campo que contiene la asignación del índice de iteración de validación cruzada en la lista desplegable.

   Este campo solo se aplica si no ingresó un valor en **N iteraciones** y **Asignación de iteración**.
9. Revise las **Series de detención** para finalizar la capacitación cuando no se mejore la opción Stopping_metric para la cantidad especificada de series de capacitación e ingrese la cantidad de series de capacitación incorrectas que se producirán antes de la detención. Para desactivar esta característica, ingrese 0. La métrica se calcula según los datos de validación (si se proporcionan); de lo contrario, se utilizarán los datos de capacitación.
10. Seleccione una **Métrica de detención** para determinar cuándo se debe dejar de crear árboles nuevos.

    AUC Área debajo de la curva ROC.

    Nota: Solo se aplica a modelos binomiales.

Automático	El valor predeterminado es la `desviación`.
Lifttopgroup	1 % principal.
Logloss	Pérdida logarítmica.
Meanperclasserror	La tasa promedio de clasificación incorrecta.
Misclassification	El valor de (1 - [predicciones correctas/predicciones totales]) * 100.
MSE	Error cuadrático medio; incluye tanto la varianza como el sesgo de un predictor.
RMSE	Raíz del error cuadrático medio; mide la diferencia entre los valores (de muestra y población) que predijo el modelo o un estimador y aquellos que se observaron en la práctica. También conocida como la raíz cuadrada de MSE.

11. Revise la **Tolerancia de detención** e ingrese un valor para especificar la tolerancia relativa para la detención según la métrica con el propósito de detener la capacitación si la mejora es inferior a este valor. Este campo está habilitado solo si marcó **Series de detención**.
12. Revise la **Mejora mínima de división** e ingrese un valor para especificar la mejora mínima relativa en una reducción de error cuadrático a fin de que se produzca una división. Cuando se ejecuta de forma correcta, esta opción puede ayudar a disminuir el sobreajuste. Los valores óptimos se encuentran en el rango de 1e-10...1e-3. Este campo está habilitado solo si marcó **Series de detención**
13. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Salida de modelo

Esta pestaña muestra las métricas que está usando para evaluar el modelo ajustado. No puede editar estos campos. La columna Capacitación siempre va a contener datos. Si seleccionó una división de capacitación/prueba en la pestaña Opciones básicas, también se completará la columna Prueba, a menos que haya seleccionado una validación de N iteraciones en la pestaña Opciones avanzadas, en cuyo caso se completará la columna N iteraciones.

Después de ejecutar su trabajo, el modelo resultante se guarda en el servidor Spectrum™ Technology Platform. Haga clic en el botón **Salida** para regenerar los datos de salida y luego en **Detalles del modelo** para ver los datos de salida completos en la herramienta Administración de modelo Machine Learning.

Puertos de salida

En la etapa Random Forest Classification se proporcionan dos puertos de salida opcionales: Puerto de calificación de modelo y Puerto de métricas de modelo. La funcionalidad de estos puertos se determina según sus selecciones y entradas cuando se completan las opciones básicas y avanzadas de la etapa. Por ejemplo, si elige realizar una validación de N iteraciones marcando el campo **N iteraciones** en la pestaña Opciones avanzadas, la columna de N iteraciones en las métricas de salida generadas en el Puerto de métricas de modelo se completará con datos. De forma alternativa, si elige no realizar una validación de N iteraciones, la columna de N iteraciones permanecerá en blanco. Del mismo modo, el Puerto de calificación de modelo se activa si marca el campo **Calificar datos de entrada** en la pestaña Opciones básicas.

Puerto de calificación de modelo

Cuando marca el campo **Calificar datos de entrada** en la pestaña Opciones básicas, se calculan valores previstos mediante Random Forest Classification cuando se crea el modelo con el que se agregan las columnas **Predicted_Value**, **Probability_of_class_A** y **Probability_of_class_B** una por una para esa calificación en los datos de salida. Puede adjuntar cualquier tipo de receptor a este puerto: una etapa Write to File, una etapa Write to Null, etc.

> **Nota:** Este puerto no es compatible con modelos multinominales de Random Forest Classification.

Puerto de métricas de modelo

El **Puerto de métricas de modelo** le permite generar las métricas de la evaluación de modelo en un archivo de datos. Este puerto lo ayudará a comparar muchos modelos generados desde dentro y fuera de Spectrum™ Technology Platform, y a realizar tareas de procesamiento de otros datos en las métricas.

Realice los siguientes pasos para utilizar el Puerto de métricas de modelo:

1. Abra un flujo de datos en el que se utilice la etapa Random Forest Classification.
2. Adjunte una etapa Write to File u otra etapa de salida de datos al puerto de salida secundario.
3. Ejecute el trabajo.
4. Opción alternativa al Paso 3: haga clic derecho en el canal y seleccione “Agregar punto de inspección” para agregar un punto de inspección al canal que permite conectar la etapa Random Forest Classification a la etapa receptora agregada en el Paso 2. Luego, haga clic en el botón

Inspeccionar flujo actual () en la barra de herramientas de Enterprise Designer. La inspección se ejecutará y verá resultados similares a aquellos que se muestran a continuación.

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 - Regresión de bosques aleatorios

In this section

Introducción

Random Forest Classification le permite realizar aprendizaje de máquina mediante la creación de modelos a partir de conjuntos de datos que usan objetivos binarios con variables de entrada.

Para crear su modelo, primero debe completar la ficha Propiedades del modelo. Las fichas Opciones básicas y Opciones avanzadas ofrecen configuraciones predeterminadas suficientes para completar un trabajo, pero puede cambiarlas de acuerdo con sus necesidades. Luego se puede ejecutar el trabajo y una versión limitada del modelo resultante aparece en la pestaña Salida de modelo; la salida completa se encuentra disponible en la herramienta de gestión de modelos de Machine Learning.

> **Nota:** Haga clic **aquí** para obtener información adicional acerca de Random Forest Regression y sus opciones.

Definición de las propiedades del modelo

1. En **Primary Stages**/**Deployed Stages**/**Machine Learning**, haga clic en la etapa **Random Forest Regression**, arrástrela hasta el lienzo, colóquela en el lugar que desee del flujo de datos y conéctela con otras etapas. Tenga en cuenta que la etapa de entrada debe ser el origen de datos que contiene los campos de variables objetivo y de entrada de su modelo. No se requiere una etapa de salida, a menos que seleccione la opción Calificar datos de entrada en la pestaña Opciones básicas. También puede conectar una etapa de salida si desea capturar su salida, independiente de la herramienta de gestión de modelo Machine Learning.
2. Haga doble clic en la etapa Regresión de bosques aleatorios y se mostrará el cuadro de diálogo **Opciones de regresión de bosques aleatorios**.
3. Ingrese un **Nombre de modelo** si no desea utilizar el nombre predeterminado.
4. Opcional: marque la casilla **Sobrescribir** para sobrescribir el modelo existente con datos nuevos.
5. Haga clic en la lista desplegable **Campo objetivo** y seleccione un campo numérico.
6. Opcional: Ingrese una **Descripción** del modelo.
7. Haga clic en **Incluir** para cada campo cuyos datos desea agregar al modelo y asegúrese de incluir el campo que seleccionó como el campo Objetivo.
8. Utilice la lista desplegable **Tipo de datos de modelo** para especificar si cada campo de entrada se debe utilizar como un campo numérico, categórico, o de fecha y hora.
9. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Configuración de opciones básicas

1. Ingrese la **Cantidad de árboles** máxima en su modelo. El valor predeterminado es 50.
2. Ingrese la **Profundidad máxima** o la cantidad máxima de niveles que desea que tenga su modelo. El valor predeterminado es 5.
3. Ingrese la **Cantidad mínima de filas**; es decir, la cantidad mínima de registros (filas) que desea que tenga su modelo. El valor predeterminado es 10.
4. Ingrese la **Cantidad de contenedores numéricos**; es decir, la cantidad de contenedores que desea que genere el histograma y que luego divida en el mejor punto. El valor predeterminado es 20.
5. Ingrese la **Cantidad de contenedores de nivel superior**; es decir, la cantidad mínima de contenedores que desea tener a nivel de raíz. El valor predeterminado es 1024.
6. Ingrese la **Cantidad de contenedores de categoría**; es decir, la cantidad máxima de contenedores que desea que genere el histograma y que luego divida en el mejor punto. El valor predeterminado es 1024.
7. Revise la **Tasa de muestra** e ingrese el porcentaje de filas que se usarán como una muestra en cada árbol. Lo anterior puede ser un valor entre 0,0 y 1,0.
8. Revise la **Tasa de muestra de columna por árbol** e ingrese la tasa de muestra de columna para cada árbol. Lo anterior puede ser un valor entre 0,0 y 1,0.
9. Revise las **Columnas en cada nivel** e ingrese el cambio relativo de tasa de muestra de columna para cada nivel. El valor predeterminado de esta opción es 1,0, pero puede tener un valor entre 0,0 y 2,0.
10. Marque la opción **Calificar datos de entrada** para agregar una columna para la predicción del modelo (calificación) a los datos de entrada.
11. Especifique un valor entre 1 y 100 como **Porcentaje para datos de capacitación** cuando los datos de entrada se dividen aleatoriamente en muestras de datos de capacitación y de prueba.
12. Ingrese el valor de 100 menos la cantidad que ingresó en el Paso 5 como **Porcentaje para datos de prueba**.
13. Utilice el campo **Propagar para datos de prueba** para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, siempre se produzca de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
14. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la pestaña siguiente.

Configuración de opciones avanzadas

1. Deje marcada la opción **Ignorar campos constantes** para omitir campos que tienen el mismo valor para cada registro.
2. Seleccione un **Tipo de histograma**.

Automático	Los depósitos se agrupan de mínimo a máximo en pasos de (máx-mín)/N. Utilice esta opción para especificar el tipo de histograma con el propósito de encontrar los puntos de división óptimos.
QuantilesGlobal	Los depósitos tienen una población equivalente. Lo anterior permite calcular los cuantiles `nbins` de cada columna numérica (no binaria) y luego refinar/rellenar cada depósito (entre dos cuantiles) de manera uniforme (y aleatoria para los elementos restantes) en un total de `nbins_top_level` agrupaciones.
Aleatorio	El algoritmo tomará muestras de los puntos N-1 de mínimo a máximo y utilizará la lista ordenada para encontrar la mejor división.
RoundRobin	El algoritmo pasará en ciclo por todos los tipos de histograma (uno por árbol).
UniformAdaptive	Permite agrupar cada característica en depósitos de tamaño de paso equivalente (no población). Este es el método más rápido, pero puede dar como resultado divisiones menos precisas si la distribución está muy desequilibrada.

3. Seleccione una **Codificación de categorías**.

Automático	Permite realizar una codificación `enum` de forma automática.
Binario	Permite convertir categorías en números enteros, luego en formato binario, y asignar cada dígito a una columna independiente. Codifica los datos en menos dimensiones, pero con algunas distorsiones en cuanto a las distancias. **Nota:** No puede haber más de 32 columnas por característica de categoría.
Eigen	Columnas *k* por característica de categoría, que permiten mantener la proyección de matriz con codificación de asignación de estado activo uno (one-hot) solo en el espacio *k*-dim eigen.
Enum	Permite pasar en ciclo por todos los tipos de histograma (uno por árbol).

OneHotExplicit Existe una columna por categoría, con el valor “1” o “0” en cada celda, lo que representa si la fila contiene esa categoría de columna.

4. Marque la opción **Propagar algoritmo y N iteraciones** e ingrese una cantidad de propagaciones para garantizar que cuando los datos se dividan en datos de prueba y de capacitación, siempre se produzca de la misma manera cada vez que ejecute el flujo de datos. Desmarque este campo para obtener una división aleatoria cada vez que ejecuta el flujo.
5. Marque la opción **N iteraciones** e ingrese la cantidad de iteraciones si va a realizar una validación cruzada.
6. Marque la opción **Asignación de iteración** y seleccione la lista despegable si ejecuta una validación cruzada. Este campo solo se aplica si ingresó un valor en **N iteraciones** y no se especificó el **Campo de iteración**.

   Automático Permite que el algoritmo seleccione automáticamente una opción; actualmente utiliza Aleatorio.

   Módulo Divide de manera uniforme el conjunto de datos en las iteraciones y no depende de la raíz.

   Aleatorio Divide de manera aleatoria los datos en piezas de n iteraciones; es ideal para grandes conjuntos de datos.

7. Si está ejecutando una validación cruzada, marque la opción **Campo de iteración** y seleccione el campo que contiene la asignación del índice de iteración de validación cruzada en la lista desplegable.

   Este campo solo se aplica si no ingresó un valor en **N iteraciones** y **Asignación de iteración**.
8. Revise las **Series de detención** para finalizar la capacitación cuando no se mejore la opción Stopping_metric para la cantidad especificada de series de capacitación e ingrese la cantidad de series de capacitación incorrectas que se producirán antes de la detención. Para desactivar esta característica, ingrese 0. La métrica se calcula según los datos de validación (si se proporcionan); de lo contrario, se utilizarán los datos de capacitación.
9. Seleccione una **Métrica de detención** para determinar cuándo se debe dejar de crear árboles nuevos.

   Automático El valor predeterminado es la `desviación`.

   desviación La desviación residual de la media; es igual al MSE.

   EMA Error medio absoluto; la diferencia entre dos variables continuas.

   MSE Error cuadrático medio; incluye tanto la varianza como el sesgo de un predictor.

   RMSE Raíz del error cuadrático medio; mide la diferencia entre los valores (de muestra y población) que predijo el modelo o un estimador y aquellos que se observaron en la práctica. También conocida como la raíz cuadrada de MSE.

RMSLE	Raíz del error logarítmico cuadrático medio; mide la proporción entre los valores predichos y los reales.

10. Revise la **Tolerancia de detención** e ingrese un valor para especificar la tolerancia relativa para la detención según la métrica con el propósito de detener la capacitación si la mejora es inferior a este valor.
11. Revise la **Mejora mínima de división** e ingrese un valor para especificar la mejora mínima relativa en una reducción de error cuadrático a fin de que se produzca una división. Cuando se ejecuta de forma correcta, esta opción puede ayudar a disminuir el sobreajuste. Los valores óptimos se encuentran en el rango de 1e-10...1e-3. Este campo está habilitado solo si marcó **Series de detención**.
12. Haga clic en **Aceptar** para guardar el modelo y la configuración, o continúe a la ficha siguiente.

Salida de modelo

Esta pestaña muestra las métricas que está usando para evaluar el modelo ajustado. No puede editar estos campos. La columna Capacitación siempre va a contener datos. Si seleccionó una división de capacitación/prueba en la pestaña Opciones básicas, también se completará la columna Prueba, a menos que haya seleccionado una validación de N iteraciones en la pestaña Opciones avanzadas, en cuyo caso se completará la columna N iteraciones.

Después de ejecutar su trabajo, el modelo resultante se guarda en el servidor Spectrum™ Technology Platform. Haga clic en el botón **Salida** para regenerar los datos de salida y luego en **Detalles del modelo** para ver los datos de salida completos en la herramienta Administración de modelo Machine Learning.

Puertos de salida

En la etapa Random Forest Regression se proporcionan dos puertos de salida opcionales: Puerto de calificación de modelo y Puerto de métricas de modelo. La funcionalidad de estos puertos se determina según sus selecciones y entradas cuando se completan las opciones básicas y avanzadas de la etapa. Por ejemplo, si elige realizar una validación de N iteraciones marcando el campo **N iteraciones** en la pestaña Opciones avanzadas, la columna de N iteraciones en las métricas de salida generadas en el Puerto de métricas de modelo se completará con datos. De forma alternativa, si elige no realizar una validación de N iteraciones, la columna de N iteraciones permanecerá en blanco. Del mismo modo, el Puerto de calificación de modelo se activa si marca el campo **Calificar datos de entrada** en la pestaña Opciones básicas.

Puerto de calificación de modelo

Cuando marca el campo **Calificar datos de entrada** en la pestaña Opciones básicas, se calculan valores previstos mediante Random Forest Regression cuando se crea el modelo con el que se agrega la columna **Predicted_Value** una por una para esa calificación en los datos de salida. Puede adjuntar cualquier tipo de receptor a este puerto: una etapa Write to File, una etapa Write to Null, etc.

Puerto de métricas de modelo

El **Puerto de métricas de modelo** le permite generar las métricas de la evaluación de modelo en un archivo de datos. Este puerto lo ayudará a comparar muchos modelos generados desde dentro y fuera de Spectrum™ Technology Platform, y a realizar tareas de procesamiento de otros datos en las métricas.

Realice los siguientes pasos para utilizar el Puerto de métricas de modelo:

1. Abra un flujo de datos en el que se utilice la etapa Random Forest Regression.
2. Adjunte una etapa Write to File u otra etapa de salida de datos al puerto de salida secundario.
3. Ejecute el trabajo.
4. Opción alternativa al Paso 3: haga clic derecho en el canal y seleccione “Agregar punto de inspección” para agregar un punto de inspección al canal que permite conectar la etapa Random Forest Regression a la etapa receptora agregada en el Paso 2. Luego, haga clic en el botón Inspeccionar flujo actual () en la barra de herramientas de Enterprise Designer. La inspección se ejecutará y verá resultados similares a aquellos que se muestran a continuación.

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 - Administración de modelo de aprendizaje automático

In this section

Acceso a Machine Learning Model Management

Hay tres maneras de acceder a la gestión de modelos de Machine Learning:

- Use la página de bienvenida de Spectrum™ Technology Platform:
  - Abra un navegador web y acceda a la página de bienvenida de Spectrum™ Technology Platform:

    <servername>:<port>

    Por ejemplo, si instaló Spectrum™ Technology Platform en una computadora denominada "myspectrumplatform" y se utiliza el puerto predeterminado 8080, accederá a:

    myspectrumplatform:8080
  - Haga clic en **Spectrum Machine Learning**.
  - Haga clic en **Abrir Machine Learning Model Management**.
- Haga clic en **Para obtener más detalles haga clic aquí** desde una de las etapas de la generación de modelos.
- Use un navegador web:
  - Abra un navegador web y vaya a la página de gestión de modelos de Machine Learning de Spectrum™ Technology Platform:

    <servername>:<port>/machinelearning

    Por ejemplo, si instaló Spectrum™ Technology Platform en una computadora denominada "myspectrumplatform" y se utiliza el puerto predeterminado 8080, accederá a:

    myspectrumplatform:8080/machinelearning
  - Ingrese un nombre de usuario y contraseña de Spectrum™ Technology Platform válidos.

Evaluación de modelo

Introducción a Evaluación de modelo

En la pestaña Evaluación de modelo en Machine Learning Model Management aparece una lista de todos los modelos de Machine Learning en su servidor Spectrum™ Technology Platform. Puede

filtrar esta lista ingresando una cadena en el cuadro de texto; se buscará esa cadena en cada campo de la tabla.

Puede realizar varias operaciones en estos modelos. Puede importar, exportar, exponer, anular exposiciones o eliminar modelos. Los modelos expuestos se usan en la etapa Java Model Scoring para evaluar nuevos datos usando fórmulas creadas cuando ajusta los modelos de Machine Learning. Además, puede ver información detallada de cada modelo; los detalles devueltos dependen del tipo de modelo cuyos datos está visualizando. Para terminar, puede comparar cualquier par de modelos del mismo tipo. Esta comparación muestra, lado a lado, la misma información que aparece en la pestaña Detalle de modelo de cada uno de los modelos que está comparando.

Operaciones de la evaluación de modelo

Realice estas operaciones seleccionando un modelo y haciendo clic en el botón correspondiente:

	Vea los detalles de salida del modelo. También puede acceder a esta información desde las etapas K-Means Clustering y Logistic Regression haciendo clic en "Para obtener más detalles del modelo, haga clic aquí" en la pestaña Salida del modelo.
	Compare los modelos.
	Importar un modelo a partir de una ruta específica. Seleccione si desea sobrescribir un modelo existente del mismo nombre, si procede.
	Exportar un modelo a una ruta específica. Seleccione si desea sobrescribir un modelo existente del mismo nombre, si procede.
	Exponga el modelo para ponerlo a disposición de la etapa Java Model Scoring. Si no expone el modelo, no lo puede usar para evaluación.
	Anule la exposición del modelo.
	Elimine el modelo. **Nota:** No puede eliminar un modelo expuesto; sin embargo, en este momento, no existe seguridad inherente que impida a un usuario eliminar los modelos de otro usuario.

Ficha Detalles de modelo

La pantalla Detalle de modelo muestra la siguiente información para todos los modelos:

- **Nombre de modelo**: el nombre del modelo

- **Tipo de modelo**: el tipo de modelo de Machine Learning
- **Usuario**: el nombre de usuario de la persona que creó el modelo
- **Descripción**: la descripción del modelo en caso de que se haya proporcionado una cuando se creó
- **Estado**: si el modelo se expuso o si se anuló la exposición
- **Nombre de flujo de datos**: el nombre del flujo de datos que produce el modelo
- **Tiempo de creación**: la fecha y la hora en que se creó el modelo

Se proporcionan detalles adicionales en función del tipo de modelo.

Detalles de K-Means Clustering

La pantalla Detalle de modelo incluye la siguiente información para modelos K-Means Clustering:

Resumen de modelo

Proporciona datos de capacitación para lo siguiente:

- Número de filas
- Número de clústeres
- Número de columnas categóricas
- Número de iteraciones
- Suma de cuadrados dentro del clúster
- Suma total de cuadrados
- Suma de cuadrados entre el clúster

Métricas

Proporciona datos de capacitación, prueba y N subconjuntos para lo siguiente:

- Suma total de cuadrados dentro del clúster
- Suma total de cuadrados
- Suma de cuadrados entre el clúster

Estadísticas de centroide

Proporciona datos de capacitación, prueba y N subconjuntos para cada centroide:

- Tamaño
- Suma de cuadrados dentro del clúster

Agrupamiento de medias

Proporciona información detallada de cada centroide. El contenido varía según los datos de entrada. Un clúster es un grupo de observaciones de un conjunto de datos identificado como similar según un algoritmo de agrupamiento específico

Agrupamiento estandarizado de medias

Proporciona información estandarizada de cada centroide. El contenido varía según los datos de entrada.

Detalles de Logistic Regression

La pantalla Detalle de modelo incluye la siguiente información para modelos Logistic Regression:

Métricas

Proporciona datos de capacitación, prueba y N subconjuntos para lo siguiente:

- Error cuadrático medio (MSE)
- Error cuadrático medio de raíz (RMSE)
- Número de observaciones
- R-cuadrado (R2)
- Pérdida logarítmica (Logloss)
- Área bajo la curva (AUC)
- Recuperación precisa de área bajo la curva (PR AUC)
- Coeficiente Gini
- Error medio por clase
- Criterio de información Akaike (AIC)
- Lambda
- Desviación residual
- Desviación nula
- Grado de libertad nulo
- Grado de libertad residual

Umbral de métricas máximas

Proporciona el Umbral de métricas máximas de capacitación para datos de capacitación, prueba, N subconjuntos usando las métricas siguientes:

- 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

Matriz de confusión

Ilustra el rendimiento de un modelo en un conjunto de datos de capacitación, prueba y N subconjuntos para los que se conocen los valores verdaderos.

Gráfico de coeficiente estándar

Muestra los predictores más importantes proporcionando el valor relativo de los coeficientes, lo que indica cuánto cambia el objetivo por un cambio en la entrada.

Coeficientes de GLM

Muestra los coeficientes para un modelo lineal generalizado, que estiman los modelos de regresión para resultados que siguen distribuciones exponenciales.

Curvas AUC

Área bajo la curva; determina cuál de los modelos usados predice las clases que mejor usan los datos de capacitación, prueba y N subconjuntos.

Curvas de ganancia/elevación

Evalúan la capacidad de predicción de un modelo de clasificación binaria usando datos de capacitación, prueba y N subconjuntos.

Detalles de Linear Regression

La pantalla Detalle de modelo incluye la siguiente información para modelos Linear Regression:

Métricas

Proporciona datos de capacitación, prueba y N subconjuntos para lo siguiente:

- Error cuadrático medio (MSE)
- Error cuadrático medio de raíz (RMSE)
- Número de observaciones
- R-cuadrado (R2)
- Desviación residual de la media
- Error medio absoluto (MAE)
- Error logarítmico cuadrático medio de raíz (RMSLE)
- Criterio de información Akaike (AIC)
- Lambda
- Desviación residual
- Desviación nula
- Grado de libertad nulo
- Grado de libertad residual

Gráfico de coeficiente estándar

Muestra los predictores más importantes proporcionando el valor relativo de los coeficientes, lo que indica cuánto el cambio de un valor de coeficiente de un predictor específico cambia el valor objetivo en forma positiva o negativa. Además, grafica los 25 coeficientes principales en el modelo.

Coeficientes de GLM

Muestra los coeficientes para un modelo lineal generalizado, que estiman los modelos de regresión para resultados que siguen distribuciones exponenciales.

Detalles de Random Forest Regression

La pantalla Detalle de modelo incluye la siguiente información para modelos Random Forest Regression:

Métricas

Proporciona datos de capacitación, prueba y N subconjuntos para lo siguiente:

- Error cuadrático medio (MSE)
- Error cuadrático medio de raíz (RMSE)
- Número de observaciones
- R-cuadrado (R2)
- Desviación residual de la media
- Error medio absoluto (MAE)
- Error logarítmico cuadrático medio de raíz (RMSLE)

Importancias variables

Proporciona valores de importancia para cada variable usando las siguientes métricas:

- Importancia relativa
- Importancia escalada
- Porcentaje

Además, grafica las 25 variables principales en el modelo.

Detalles de Random Forest Classification: Binomial

La pantalla Detalle de modelo incluye la siguiente información para modelos **binomiales** Random Forest Classification:

Métricas

Proporciona datos de capacitación, prueba y N subconjuntos para lo siguiente:

- Error cuadrático medio (MSE)
- Error cuadrático medio de raíz (RMSE)
- Número de observaciones
- R-cuadrado (R2)
- Logloss
- Área bajo la curva (AUC)
- Recuperación precisa de área bajo la curva (PR AUC)
- Gini
- Error medio por clase

Umbral de métricas máximas

Proporciona el Umbral de métricas máximas de capacitación para datos de capacitación, prueba, N subconjuntos usando las métricas siguientes:

- 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

Matriz de confusión

Ilustra el rendimiento de un modelo en un conjunto de datos de capacitación, prueba y N subconjuntos para los que se conocen los valores verdaderos.

Importancias variables

Proporciona valores de importancia para cada variable usando las siguientes métricas:

- Importancia relativa
- Importancia escalada
- Porcentaje

Además, grafica las 25 variables principales en el modelo.

Curvas AUC

Área bajo la curva; determina cuál de los modelos usados predice las clases que mejor usan los datos de capacitación, prueba y N subconjuntos.

Curvas de ganancia/elevación

Se evalúa la capacidad de predicción de un modelo de clasificación binaria mediante datos de capacitación, de pruebas y de N subconjuntos.

Detalles de Random Forest Classification: Multinomial

La pantalla Detalle de modelo incluye la siguiente información para modelos **multinomiales** Random Forest Classification:

Métricas

Proporciona datos de capacitación, prueba y N subconjuntos para lo siguiente:

- Error cuadrático medio (MSE)
- Error cuadrático medio de raíz (RMSE)
- Número de observaciones
- R-cuadrado (R2)
- Logloss

- Error medio por clase

Matriz de confusión

Ilustra el rendimiento de un modelo en un conjunto de datos de capacitación, prueba y N subconjuntos para los que se conocen los valores verdaderos.

Importancias variables

Proporciona valores de importancia para cada variable usando las siguientes métricas:

- Importancia relativa
- Importancia escalada
- Porcentaje

Además, grafica las 25 variables principales en el modelo.

Detalles de análisis de componentes principales

La pantalla Detalle de modelo incluye la siguiente información para modelos PCA:

Importancia de los componentes

Muestra los componentes principales en orden de importancia en función de las siguientes métricas:

- Desviación estándar
- Proporción de varianza
- Proporción acumulativa

Rotación

Grafica la matriz de cargas variables, el peso por el cual se debe multiplicar cada variable original estandarizada para obtener la calificación del componente.

Administración de binning

Introducción a Binning Management

En la pestaña Binning Management de Machine Learning Model Management aparece una lista de todos los Binning en su servidor Spectrum™ Technology Platform. Puede filtrar esta lista ingresando una cadena en el cuadro de texto; se buscará esa cadena en cada campo de la tabla.

Puede realizar varias operaciones en binning. Puede importar, exportar, exponer, anular exposiciones o eliminar binnings. Los elementos bin expuestos son utilizados por la etapa Binning Lookup para aplicar elementos binning previamente definidos a datos nuevos.

Operaciones de administración de binning

Para realizar estas operaciones, seleccione un elemento binning y haga clic en el botón correspondiente:

	Importar un elemento binning. Seleccione si desea sobrescribir un elemento binning existente del mismo nombre, si procede.
	Exportar un elemento binning. Seleccione si desea sobrescribir un elemento binning existente del mismo nombre, si procede.
	Exponga el elemento binning para ponerlo a disposición de la etapa Binning Lookup. Si un elemento binning no se expone, no se podrá usar para búsquedas.
	Dejar de exponer un elemento binning.
	Eliminar un elemento binning. **Nota:** No puede eliminar un elemento binning expuesto; sin embargo, en este momento, no existe seguridad inherente que impida a un usuario eliminar los elementos binning de otros usuarios.

10 - Flujos de demostración de ciencia de datos

In this section

Introducción

Los módulos Machine Learning y Analytics Scoring, junto con los módulos para la preparación de datos para modelado, son parte de lo que ofrece la ciencia de datos de Spectrum. Mediante estas demostraciones se presentan ejemplos de preparación de datos, modelado y calificación de modelos. Puede crear sus propios flujos de datos si sigue las instrucciones paso a paso o puede utilizar los flujos de datos proporcionados como referencia.

Aprendizaje supervisado: Predicción de probabilidad de incumplimiento

Descargue la demostración de aprendizaje supervisado

La demostración de aprendizaje supervisado de ciencia de datos le permite realizar la predicción predeterminada de la probabilidad de incumplimiento con los datos de Lending Club. Se utilizan numerosos archivos con los que se demuestra, en conjunto, la funcionalidad de la Solución de ciencia de datos de Spectrum™ Technology Platform en Enterprise Designer.

En Spectrum_DataScience_Supervised_Learning.zip se incluyen los siguientes archivos:

- Spectrum_DataScience_Supervised_Learning.pdf: documentación en la que se entrega información sobre cómo crear y utilizar el flujo de datos único para categorización, el flujo de datos para calificación y todos los archivos compatibles.
- Data.zip: los archivos de entrada, de prueba y de capacitación requeridos para cada flujo de datos incluido.
  - 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: el flujos de datos para Spectrum™ Technology Platform 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: el flujos de datos para Spectrum™ Technology Platform 12.2.
  - LendingClub_2007_2016Q12_v122_MultipleCategorizers.df
  - LendingClub_2007_2016Q1Q2_v122_SingleCategorizer.df
  - LendingClub_2016Q3_v122_SingleCategorizer_Scoring.df
- ReadMe.txt: descripciones e instrucciones de alto nivel para los archivos mencionados anteriormente.

Puede crear su propio flujo de datos si sigue las instrucciones de la documentación paso a paso o puede utilizar los flujos de datos que se incluyen como referencias para confirmar cómo deberían lucir las etapas individuales finalizadas y los flujos de datos como un todo.

Aprendizaje no supervisado: Segmentación

Descargue la demostración del aprendizaje no supervisado

La demostración del aprendizaje no supervisado de ciencia de datos le permite realizar la segmentación mediante los datos Gastos de consumo. Se utilizan numerosos archivos con los que se demuestra, en conjunto, la funcionalidad de la Solución de ciencia de datos de Spectrum™ Technology Platform en Enterprise Designer.

En Spectrum_DataScience_Unsupervised_Learning.zip se incluyen los siguientes archivos:

- Spectrum_DataScience_Unsupervised_Learning.pdf: documentación en la que se entrega información sobre cómo crear y utilizar el flujo de datos principal, el subflujo, el flujo de datos para calificación y todos los archivos compatibles
- Data.zip: los archivos de entrada y de salida requeridos para cada uno de los flujos de datos incluidos
  - Carpeta Input: los archivos de entrada requeridos para cada uno de los flujos de datos incluidos
  - Carpeta Output: los archivos de salida requeridos para cada uno de los flujos de datos incluidos

    - Carpeta PythonBased: archivos de entrada y de salida requeridos para utilizar el procesamiento Phyton opcional en vez de utilizar Group Statistics y las etapas Transformer en el flujo de datos principal
- Consumer_Expenditure_Demo_DF_(v12.1).zip: los flujos de datos para Spectrum™ Technology Platform 12.1
    - ConsumerExpenditure_v121_sampleandcluster.df
    - ConsumerExpenditure_v121_sampleandcluster_subflow.df
    - ConsumerExpenditure_v121_score.df
    - ConsumerExpenditure_v121_subflow.df
    - Carpeta PythonBased: flujos de datos, flujos de procesos, secuencia de comandos bat, secuencia de comandos y documentación de Python requeridos para utilizar el procesamiento Phyton opcional en vez de utilizar Group Statistics y las etapas Transformer en el flujo de datos principal
- Consumer_Expenditure_Demo_DF_(v12.2).zip: los flujos de datos para Spectrum™ Technology Platform 12.2
    - ConsumerExpenditure_v122_sampleandcluster.df
    - ConsumerExpenditure_v122_sampleandcluster_subflow.df
    - ConsumerExpenditure_v122_score.df
    - ConsumerExpenditure_v122_subflow.df
    - Carpeta PythonBased: flujos de datos, flujos de procesos, secuencia de comandos bat, secuencia de comandos y documentación de Python requeridos para utilizar el procesamiento Phyton opcional en vez de utilizar Group Statistics y las etapas Transformer en el flujo de datos principal
- ReadMe.txt: descripciones e instrucciones de alto nivel para los archivos mencionados anteriormente.

Puede crear su propio flujo de datos si sigue las instrucciones de la documentación paso a paso o puede utilizar los flujos de datos que se incluyen como referencias para confirmar cómo deberían lucir las etapas individuales finalizadas y los flujos de datos como un todo.

Notices

Avisos de USPS®

Pitney Bowes Inc. posee una licencia no exclusiva para publicar y vender bases de datos ZIP + 4® en medios magnéticos y ópticos. Las siguientes marcas comerciales son propiedad del Servicio Postal de los Estados Unidos: CASS, CASS Certified, DPV, eLOT, FASTforward, First-Class Mail, Intelligent Mail, LACS^{Link}, NCOA^{Link}, PAVE, PLANET Code, Postal Service, POSTNET, Post Office, RDI, Suite^{Link} , United States Postal Service, Standard Mail, United States Post Office, USPS, ZIP Code, y ZIP + 4. Esta lista no es exhaustiva de todas las marcas comerciales que pertenecen al servicio postal.

Pitney Bowes Inc. es titular de una licencia no exclusiva de USPS® para el procesamiento NCOA^{Link}®.

Los precios de los productos, las opciones y los servicios del software de Pitney Bowes no los establece, controla ni aprueba USPS® o el gobierno de Estados Unidos. Al utilizar los datos RDI™ para determinar los costos del envío de paquetes, la decisión comercial sobre qué empresa de entrega de paquetes se va a usar, no la toma USPS® ni el gobierno de Estados Unidos.

Proveedor de datos y avisos relacionados

Los productos de datos que se incluyen en este medio y que se usan en las aplicaciones del software de Pitney Bowes Software, están protegidas mediante distintas marcas comerciales, además de un o más de los siguientes derechos de autor:

© Derechos de autor, Servicio Postal de los Estados Unidos. Todos los derechos reservados.

© 2014 TomTom. Todos los derechos reservados. TomTom y el logotipo de TomTom son marcas comerciales registradas de TomTom N.V.

© 2016 HERE

Fuente: INEGI (Instituto Nacional de Estadística y Geografía)

Basado en los datos electrónicos de © National Land Survey Sweden.

© Derechos de autor Oficina del Censo de los Estados Unidos

© Derechos de autor Nova Marketing Group, Inc.

Algunas partes de este programa tienen © Derechos de autor 1993-2007 de Nova Marketing Group Inc. Todos los derechos reservados

© Copyright Second Decimal, LLC

© Derechos de autor Servicio de correo de Canadá

Este CD-ROM contiene datos de una compilación cuyos derechos de autor son propiedad del servicio de correo de Canadá.

© 2007 Claritas, Inc.

El conjunto de datos Geocode Address World contiene datos con licencia de GeoNames Project (**www.geonames.org**) suministrados en virtud de la licencia de atribución de Creative Commons (la “Licencia de atribución”) que se encuentra en **http://creativecommons.org/licenses/by/3.0/legalcode**. El uso de los datos de GeoNames (según se describe en el manual de usuario de Spectrum™ Technology Platform) se rige por los términos de la Licencia de atribución. Todo conflicto entre el acuerdo establecido con Pitney Bowes Software, Inc. y la Licencia de atribución se resolverá a favor de la Licencia de atribución exclusivamente en cuanto a lo relacionado con el uso de los datos de GeoNames.

3001 Summer Street
Stamford CT 06926-0700
USA

www.pitneybowes.com