

LAB 08b – Predicción del riesgo de crédito en Azure Machine Learning Studio

En este laboratorio se explica con detalle el proceso de desarrollo de una solución de análisis predictivo. Va a desarrollar un modelo sencillo en Machine Learning Studio. Después va a implementar el modelo como un servicio web de Azure Machine Learning. Este modelo implementado puede hacer predicciones con datos nuevos.

Paso 1: Pre-Requisitos

1. Area de trabajo de Microsoft Azure Machine Learning Studio.

Paso 2: Carga de datos existentes

Para desarrollar un modelo de predicción de riesgo de crédito, se necesitan datos para entrenar y probar el modelo. Para este laboratorio, se usará el conjunto de datos "**UCI Statlog (German Credit Data)**" del repositorio de Machine Learning de UC Irvine. Puede encontrarlo aquí:

[https://archive.ics.uci.edu/ml/datasets/Statlog+\(German+Credit+Data\)](https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data))

Usaremos el archivo llamado **german.data**. Descargue este archivo en la unidad de disco duro local.

El conjunto de datos german.data contiene filas de 20 variables para 1000 solicitantes de crédito. Estas 20 variables representan el conjunto de características (el vector de características) del conjunto de datos que proporciona características de identificación de cada solicitante de crédito. Una columna adicional en cada fila representa el riesgo de crédito calculado del solicitante, donde 700 solicitantes se identificaron como de bajo riesgo y 300 como de alto riesgo.

Paso 3: Conversión del formato del conjunto de datos

El conjunto de datos original utiliza un formato separado por espacios en blanco. Machine Learning Studio funciona mejor con un archivo de valores delimitados por comas (CSV), así que se va a convertir el conjunto de datos y reemplazar los espacios por comas.




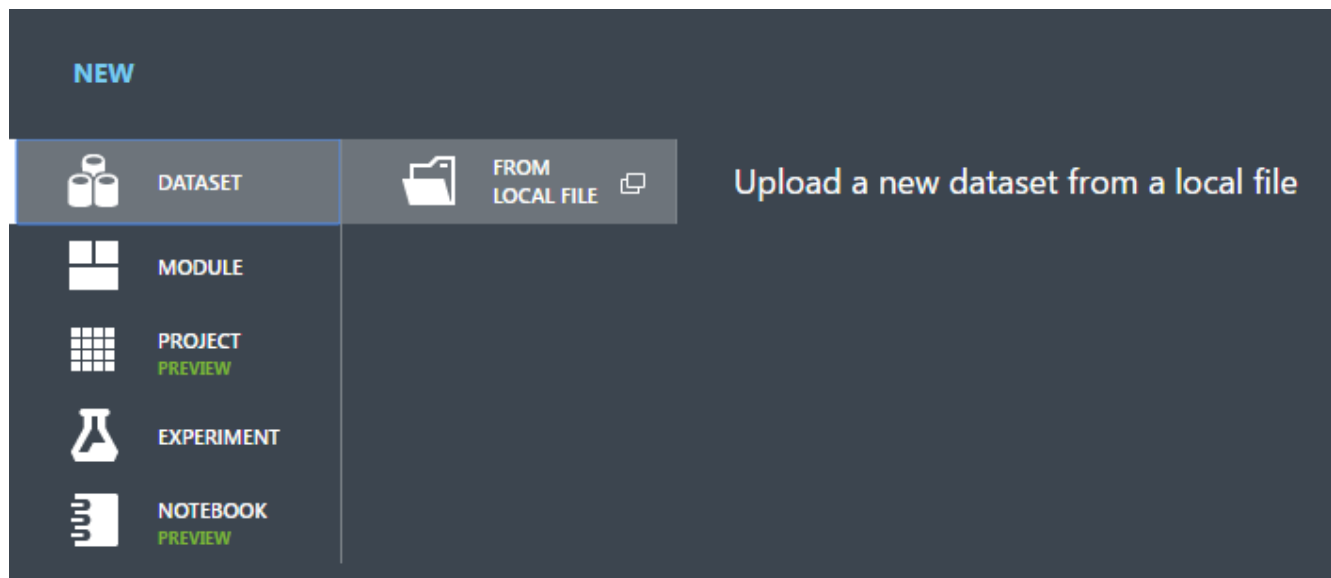
1. En powershell ejecutar el siguiente comando para generar un archivo csv que se puede usar en el experimento:

```
cat german.data | %{$_ -replace '"',','} | sc german.csv
```

Paso 4: Carga del conjunto de datos en Machine Learning Studio

Una vez que los datos se han convertido al formato CSV, hay que cargarlos en Machine Learning Studio.

1. Abra la página principal de Machine Learning Studio (<https://studio.azureml.net>)
2. Haga clic en el  menú de la esquina superior izquierda de la ventana, haga clic en **Azure Machine Learning**, seleccione **Studio** e inicie sesión.
3. Haga clic en **+NUEVO** en la parte inferior de la página.
4. Seleccione **CONJUNTO DE DATOS**.
5. Seleccione **DE ARCHIVO LOCAL**.



6. En el diálogo **Cargar un nuevo conjunto de datos**, haga clic en Examinar y busque el archivo **german.csv** que ha creado.
7. Escriba un nombre para el conjunto de datos. En este laboratorio, se denominará "UCI German Credit Card Data"



8. Para el tipo de datos, seleccione **Archivo CSV genérico sin encabezado (.nh.csv)**.
9. Agregue una descripción si así lo desea.
10. Haga clic en la marca de verificación **Aceptar**.

×

Upload a new dataset

SELECT THE DATA TO UPLOAD:

german.csv

☐ This is the new version of an existing dataset

ENTER A NAME FOR THE NEW DATASET:

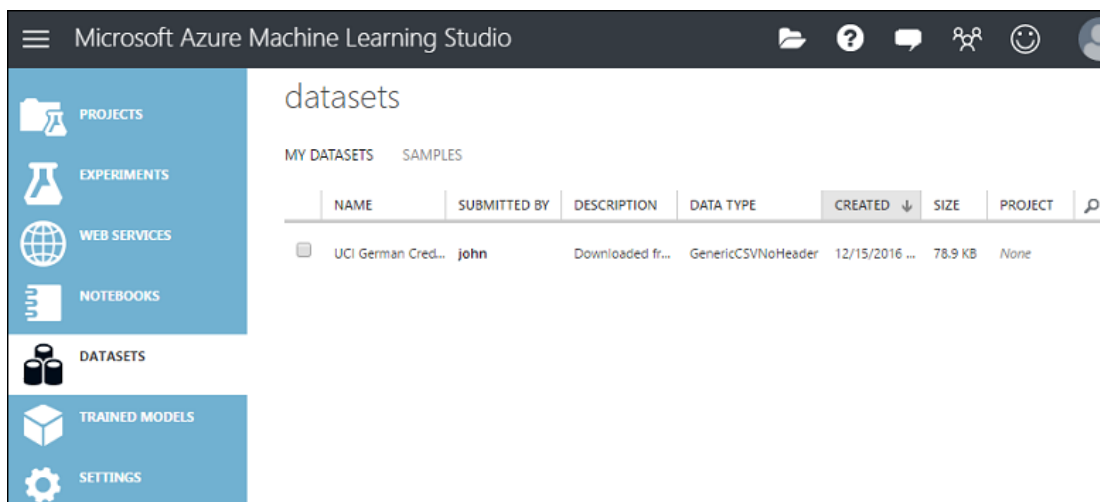
SELECT A TYPE FOR THE NEW DATASET:

▼

PROVIDE AN OPTIONAL DESCRIPTION:

✓

Para administrar los conjuntos de datos que cargó en Studio, haga clic en la pestaña **CONJUNTOS DE DATOS** a la izquierda de la ventana de Studio.



The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left is a navigation pane with icons for PROJECTS, EXPERIMENTS, WEB SERVICES, NOTEBOOKS, DATASETS (highlighted), TRAINED MODELS, and SETTINGS. The main area is titled 'datasets' and contains a tabbed interface with 'MY DATASETS' and 'SAMPLES'. Below the tabs is a table listing datasets.

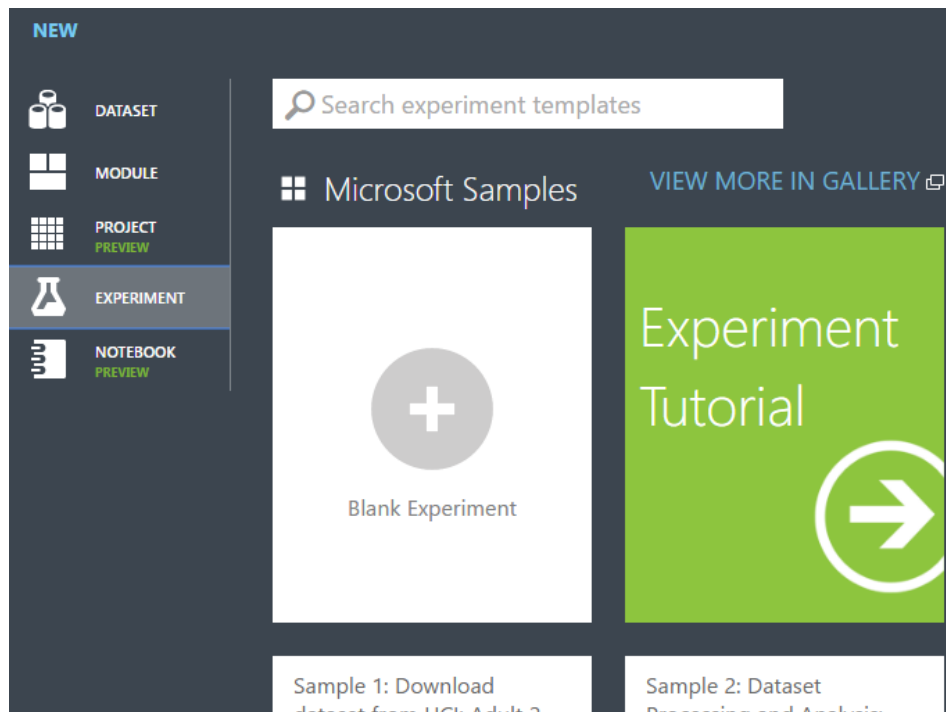
	NAME	SUBMITTED BY	DESCRIPTION	DATA TYPE	CREATED ↓	SIZE	PROJECT	
<input type="checkbox"/>	UCI German Cred...	john	Downloaded fr...	GenericCSVNoHeader	12/15/2016 ...	78.9 KB	None	



Paso 5: Creación de un experimento

El siguiente paso de este laboratorio es crear un experimento en Machine Learning Studio que utilice el conjunto de datos cargado..

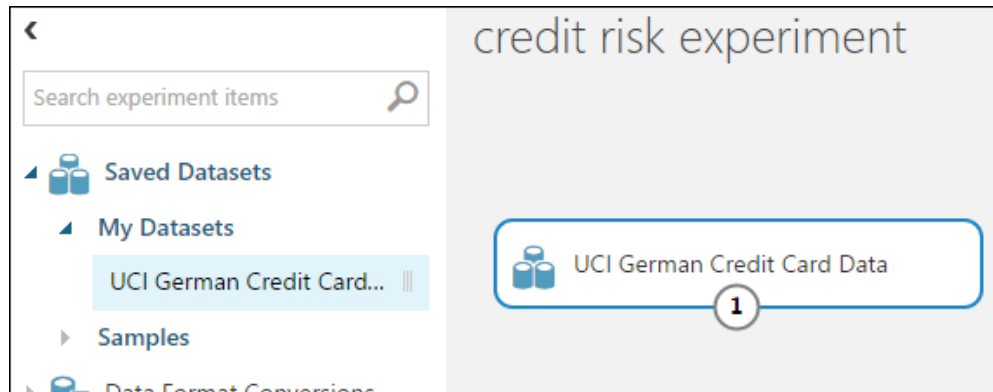
1. En Studio, haga clic en **+NUEVO** en la parte inferior de la ventana.
2. Seleccione **EXPERIMENTO** y luego "Experimento en blanco".



3. Seleccione el nombre del experimento predeterminado en la parte superior del lienzo y cámbielo por uno significativo



4. En la paleta de módulos, a la izquierda del lienzo de experimentos, expanda **Conjuntos de datos guardados**.
5. Busque el conjunto de datos que ha creado en **Mis conjuntos de datos** y arrástrelo al lienzo. También puede buscar el conjunto de datos escribiendo su nombre en el cuadro **Buscar** que está encima de la paleta.



Paso 6: Preparación de los datos

Puede ver los 100 primeros registros de los datos e información estadística de todo el conjunto: Haga clic en el puerto de salida del conjunto de datos (el círculo pequeño de la parte inferior) y seleccione **Visualizar**.

Dado que el archivo de datos no incluye encabezados de columna, Estudio de aprendizaje automático ha proporcionado encabezados genéricos (Col1, Col2, etc.). No es esencial que los encabezados sean perfectos para crear un modelo, pero facilitan el trabajo con los datos del experimento. Además, cuando finalmente se publique este modelo en un servicio web, los encabezados ayudarán al usuario del servicio a identificar las columnas.

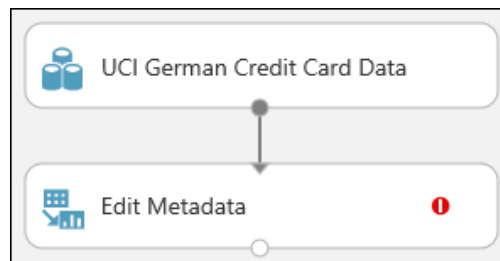
Se pueden agregar encabezados de columna mediante el módulo **Edit Metadata** (Editar metadatos).

1. En la paleta de módulos, escriba "metadatos" en el cuadro **Buscar**. El módulo **Edit Metadata** (Editar metadatos) aparece en la lista de módulos.
2. Haga clic en el módulo **Edit Metadata** (Editar metadatos), arrástrelo al lienzo y colóquelo bajo el conjunto de datos agregado anteriormente.
3. Conecte el conjunto de datos al módulo **Edit Metadata** (Editar metadatos): haga clic en el puerto de salida del conjunto de datos (el círculo pequeño de la parte inferior del conjunto de datos), arrástrelo al puerto de entrada



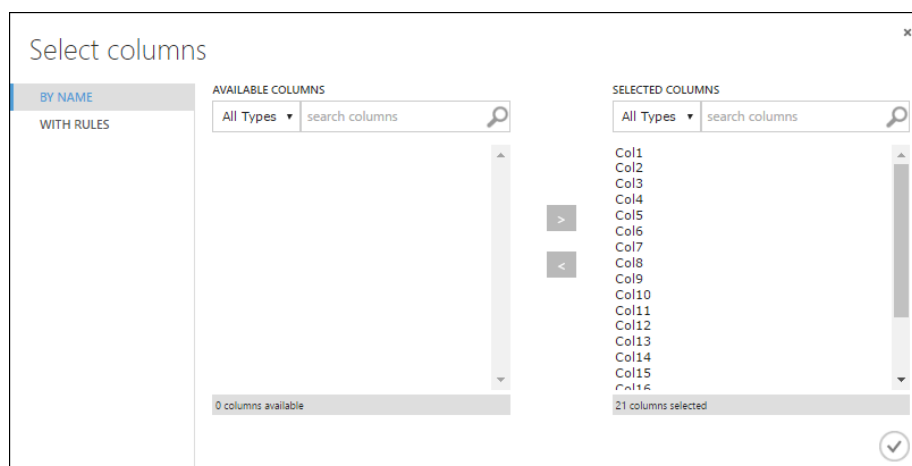
del módulo **Edit Metadata** (Editar metadatos) (el círculo pequeño de la parte superior del módulo) y luego suelte el botón del ratón. El conjunto de datos y el módulo permanecen conectados aunque se desplace por el lienzo.

El experimento debería tener ahora un aspecto similar al siguiente:



El signo de exclamación rojo indica que no se han configurado aún las propiedades de este módulo. Se hará a continuación.

4. Seleccione **Editar metadatos** y, en el panel **Propiedades** a la derecha del lienzo, haga clic en **Launch column selector** (Iniciar el selector de columnas).
5. En el cuadro de diálogo **Seleccionar columnas**, elija todas las filas de **Columnas disponibles** y haga clic en > para moverlas a **Columnas seleccionadas**. El cuadro de diálogo debe ser similar al siguiente:



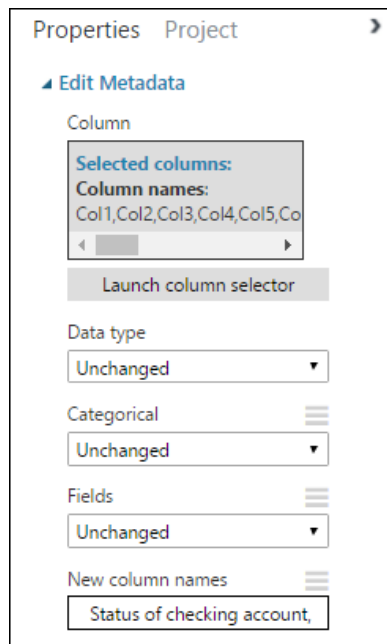
6. Haga clic en la marca de verificación **Aceptar**
7. En el panel **Propiedades**, busque el parámetro **Nuevo nombre de columna**. En este campo, escriba la lista de nombres de las 21 columnas del conjunto



de datos, separadas por comas y en el orden de las columnas. Copie y pegue la siguiente lista:

Status of checking account, Duration in months, Credit history, Purpose, Credit amount, Savings account/bond, Present employment since, Installment rate in percentage of disposable income, Personal status and sex, Other debtors, Present residence since, Property, Age in years, Other installment plans, Housing, Number of existing credits, Job, Number of people providing maintenance for, Telephone, Foreign worker, Credit risk

El panel Propiedades tiene un aspecto similar al siguiente:



Paso 7: Creación de conjuntos de datos de entrenamiento y prueba

Se necesitan algunos datos para entrenar el modelo y otros tantos para probarlo. De este modo, en el siguiente paso del experimento, se divide el conjunto de datos en dos conjuntos de datos independientes: uno para el entrenamiento de nuestro modelo y el otro para probarlo.

Para ello, se usa el módulo **Split Data** (Dividir datos).

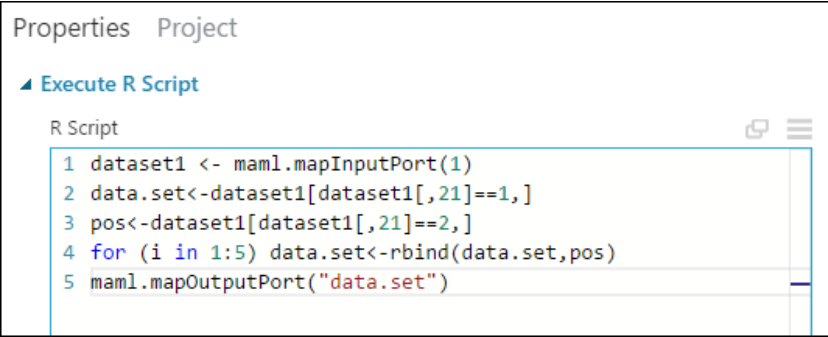


1. Busque el módulo **Split Data** (Dividir datos), arrástrelo al lienzo y conéctelo al módulo **Edit Metadata** (Editar metadatos).
2. De manera predeterminada, la proporción de división es 0,5 y se establece el parámetro **División aleatoria** . Esto significa que una mitad aleatoria de los datos sale a través de un puerto del módulo **Split Data** (Dividir datos) y la otra mitad, por el otro. Puede cambiar estos parámetros, así como el parámetro **Valor de inicialización aleatorio**, para cambiar la división entre datos de entrenamiento y de prueba. En este ejemplo, se dejan tal cual.
3. Haga doble clic en el módulo **Split Data** (Dividir datos) y escriba el comentario "Dividir 50% de los datos de entrenamiento y pruebas".

A tener en cuenta, el costo de clasificar erróneamente un riesgo de crédito alto como bajo es cinco veces más alto que el costo de clasificar erróneamente un riesgo de crédito bajo como alto. Para tener esto en cuenta, se debe generar un nuevo conjunto de datos que refleje esta función de costo. En el nuevo conjunto de datos, cada ejemplo de alto riesgo se replica cinco veces, mientras que los ejemplos de bajo riesgo no se replican.

Podemos conseguir esta replicación mediante el código R:

4. Busque el módulo **Execute R Script** (Ejecutar script R) y arrástrelo al lienzo del experimento.
5. Conecte el puerto de salida de la izquierda del módulo **Split Data** (Dividir datos) al primer puerto de entrada ("Dataset1") del módulo **Execute R Script** (Ejecutar script R).
6. Haga doble clic en el módulo **Execute R Script** (Ejecutar script R) y escriba el comentario "Establecer ajuste de costos".
7. En el panel **Propiedades**, elimine el texto predeterminado del parámetro **Script R** y escriba este script:



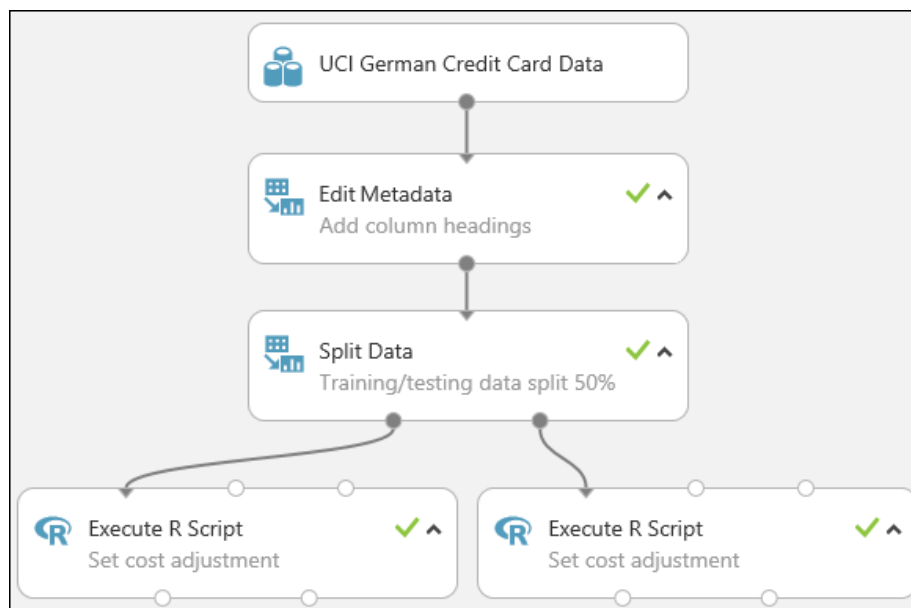
```
1 dataset1 <- mam1.mapInputPort(1)
2 data.set<-dataset1[dataset1[,21]==1,]
3 pos<-dataset1[dataset1[,21]==2,]
4 for (i in 1:5) data.set<-rbind(data.set,pos)
5 mam1.mapOutputPort("data.set")
```



Hay que hacer esta misma operación de replicación para cada salida del módulo **Split Data** (Dividir datos) para que los datos de entrenamiento y prueba tengan el mismo ajuste con relación al costo. La forma más sencilla de hacerlo consiste en duplicar el módulo **Execute R Script** (Ejecutar script R) que se acaba de crear y conectarlo al otro puerto de salida del módulo **Split Data** (Dividir datos).

8. Haga clic con el botón derecho en el módulo **Execute R Script** (Ejecutar script R) y seleccione **Copiar**.
9. Haga clic con el botón derecho en el lienzo del experimento y seleccione **Pegar**.
10. Arrastre el nuevo módulo a la posición correspondiente y luego conecte el puerto de salida de la derecha del módulo **Split Data** (Dividir datos) al primer puerto de entrada de este nuevo módulo **Execute R Script** (Ejecutar script R).
11. En la parte inferior del lienzo, haga clic en **Ejecutar**.

Nuestro experimento tiene ahora un aspecto similar al siguiente:



Paso 8: Entrenamiento de varios modelos

Una de las ventajas del uso de Azure Machine Learning Studio para crear modelos de aprendizaje automático es la capacidad para probar más de un tipo de modelo



a la vez en un solo experimento y comparar los resultados. Este tipo de experimentación ayuda a encontrar la mejor solución al problema.

En el experimento que vamos a crear en este laboratorio, crearemos dos tipos diferentes de modelos y después compararemos los resultados de su puntuación para decidir qué algoritmo usar en nuestro experimento final.

Existen varios modelos entre los que se puede elegir. Para ver cuáles están disponibles, expanda el nodo **Machine Learning** de la paleta de módulos y luego expanda **Initialize Model** (Inicializar modelo) y los nodos que incluye. Teniendo en cuenta el objetivo de este experimento, seleccione los módulos **Two-Class Support Vector Machine** (Máquina de vectores de soporte de dos clases, SVM) y **Two-Class Boosted Decision Tree** (Árbol de decisión ampliado de dos clases).

Agregará tanto el módulo **Two-Class Boosted Decision Tree** (Árbol de decisión ampliado de dos clases) como el módulo **Two-Class Support Vector Machine** (Máquina de vectores de soporte de dos clases) en este experimento.

Paso 9: Two-Class Boosted Decision Tree (Árbol de decisión ampliado de dos clases).

En primer lugar, configure el modelo del árbol de decisión ampliado.

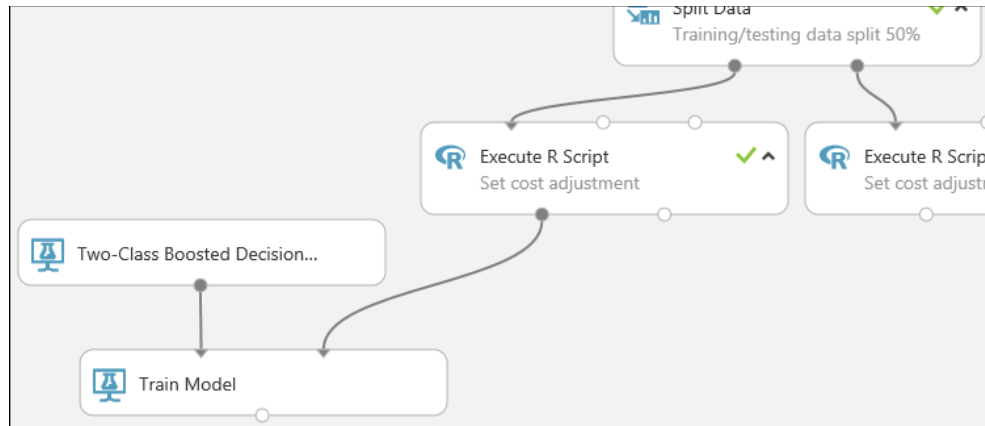
1. Busque el módulo **Two-Class Boosted Decision Tree** (Árbol de decisión ampliado de dos clases) en la paleta de módulos y arrástrelo al lienzo.
2. Busque el módulo **Train Model** (Entrenar modelo), arrástrelo al lienzo y conecte la salida del módulo **Two-Class Boosted Decision Tree** (Árbol de decisión ampliado de dos clases) al puerto de entrada izquierdo del módulo **Train Model** (Entrenar modelo).

El módulo **Two-Class Boosted Decision Tree** (Árbol de decisión ampliado de dos clases) inicializa el modelo genérico, y **Train Model** (Entrenar modelo) usa los datos de entrenamiento para entrenar el modelo.

3. Conecte la salida izquierda del módulo **Execute R Script (Ejecutar script R)** izquierdo al puerto de entrada de la derecha del módulo **Train Model (Entrenar modelo)** (en este laboratorio usó los datos procedentes del lado izquierdo del módulo Split Data [Dividir datos] para el entrenamiento).

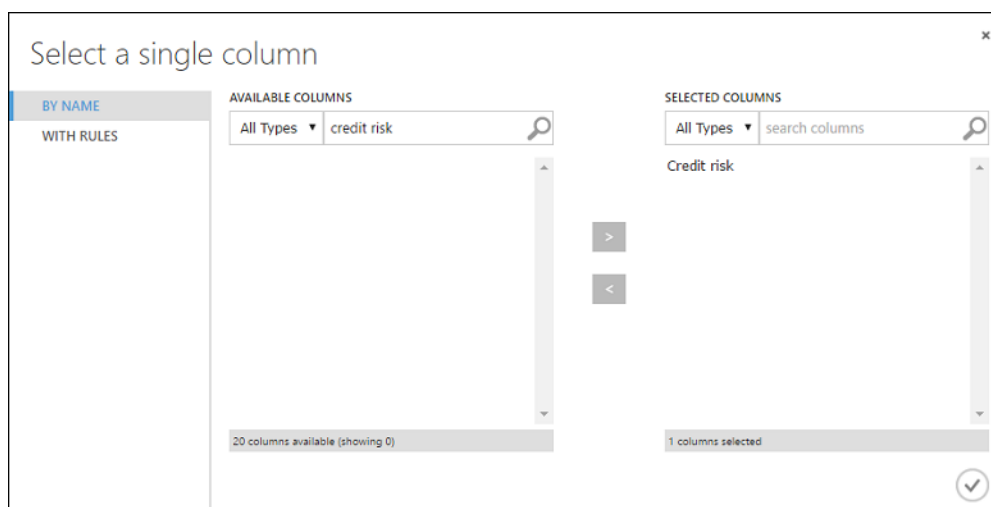


Esta parte del experimento tiene ahora un aspecto similar al siguiente:



Ahora es necesario indicar al módulo **Train Model** (Entrenar modelo) que desea que el modelo prediga el valor del riesgo crediticio.

4. Seleccione el módulo **Train Model** (Entrenar modelo). En el panel **Propiedades**, haga clic en **Launch column selector** (Iniciar el selector de columnas).
5. En el cuadro de diálogo **Select a single column** (Seleccionar una sola columna), escriba "riesgo de crédito" en el campo de búsqueda en **Columnas disponibles**, seleccione "Riesgo de crédito" a continuación y haga clic en el botón de la flecha derecha (>) para mover "Riesgo de crédito" a **Columnas seleccionadas**.



6. Haga clic en la marca de verificación **Aceptar**



Paso 10: Máquina de vectores de soporte de dos clases

A continuación, configure el modelo SVM.

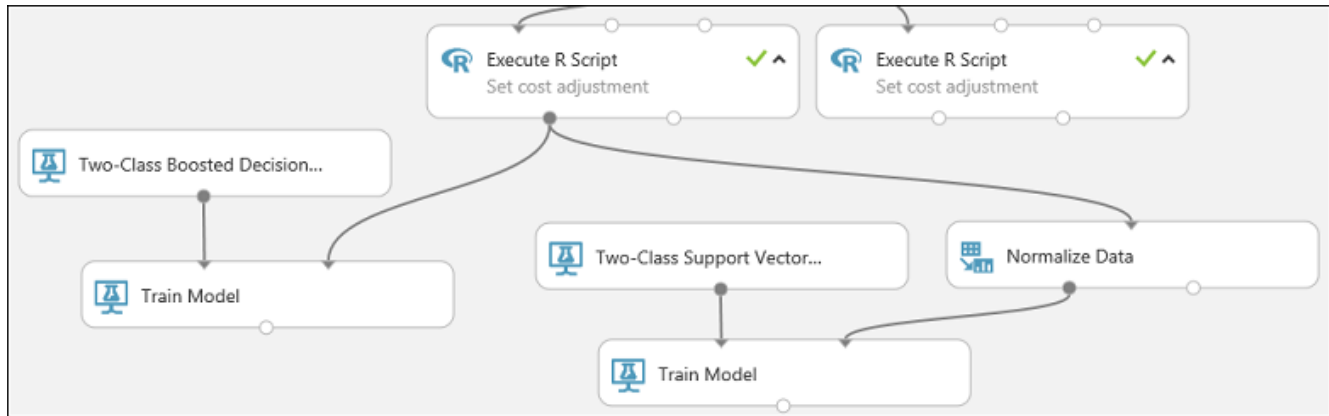
En primer lugar, una breve explicación sobre SVM. Los árboles de decisión ampliados funcionan bien con características de todo tipo. Sin embargo, dado que el módulo SVM genera un clasificador lineal, el modelo que genera tiene el mejor error de prueba cuando todas las características numéricas tienen la misma escala. Para convertir todas las características numéricas a la misma escala, utilice una transformación "Tanh", con el módulo **Normalize Data** (Normalizar datos). Esto transforma los números en el intervalo [0,1]. El módulo SVM convierte las características de cadena en características categóricas y luego en características binarias 0/1. Por lo tanto, no hace falta transformar manualmente las características de cadena. Además, no queremos transformar la columna Credit Risk (Riesgo crediticio, columna 21): es numérica, pero es el valor sobre cuya predicción estamos entrenando al modelo; por tanto, es necesario dejarla tal cual.

Para configurar el modelo SVM, realice lo siguiente:

1. Busque el módulo **Two-Class Support Vector Machine** (Máquina de vectores de soporte de dos clases) en la paleta de módulos y arrástrelo al lienzo.
2. Haga clic con el botón derecho en el módulo **Train Model** (Entrenar modelo), seleccione **Copiar**, haga clic con el botón derecho en el lienzo y seleccione **Pegar**. La copia del módulo **Train Model** (Entrenar modelo) tiene la misma selección de columnas que el original.
3. Conecte la salida del módulo **Máquina de vectores de soporte de dos clases** al puerto de entrada izquierdo del módulo **Entrenar modelo**.
4. Busque el módulo **Normalizar datos** y arrástrelo al lienzo.
5. Conecte la salida de la izquierda del módulo **Ejecutar script R** de la izquierda a la entrada de este módulo (tenga en cuenta que el puerto de salida de un módulo puede estar conectado a más de un módulo distinto).
6. Conecte el puerto de salida izquierdo del módulo **Normalize Data (Normalizar datos)** al puerto de entrada derecho del segundo módulo **Train Model (Entrenar modelo)**.



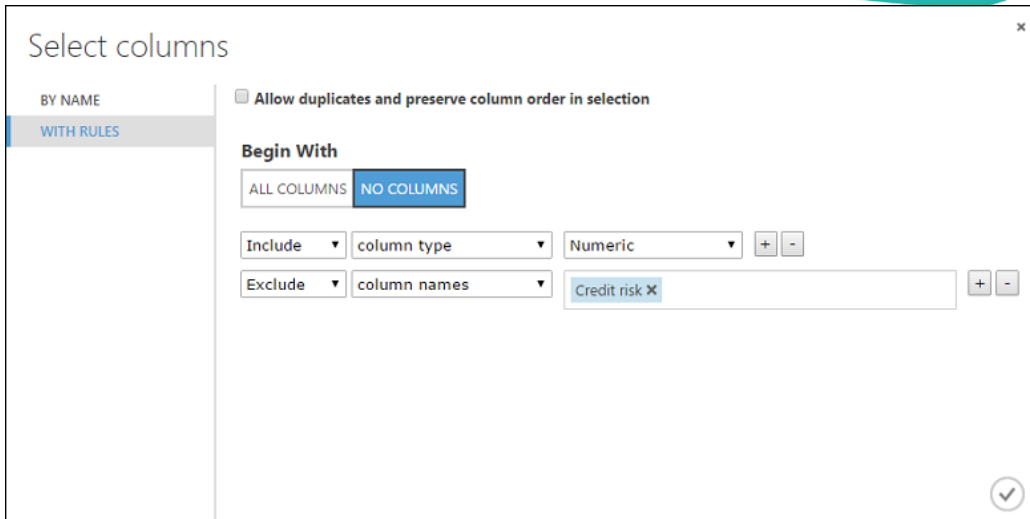
Esta parte de nuestro experimento debería tener ahora un aspecto similar al siguiente:



Configure ahora el módulo **Normalize Data** (Normalizar datos):

1. Haga clic para seleccionar el módulo **Normalize Data** (Normalizar datos). En el panel **Propiedades**, seleccione **Tanh** para el parámetro **Transformation method** (Método de transformación).
2. Haga clic en **Launch column selector** (Iniciar el selector de columnas), seleccione "No columns" (Sin columnas) en **Comenzar con**, seleccione **Incluir** en el primer menú desplegable, **Tipo de columna** en el segundo y **Númerica** en el tercero. Esto especifica que todas las columnas numéricas (y solo numéricas) se deben transformar.
3. Haga clic en el signo más (+) a la derecha de esta fila; de esta forma, se crea una fila de menús desplegables. Seleccione **Excluir** en la primera lista desplegable y **Nombres de columna** en la segunda, y escriba "Riesgo de crédito" en el campo de texto. Esto especifica que se debe ignorar la columna Credit Risk (Riesgo crediticio) (debemos hacerlo porque se trata de una columna numérica y, de lo contrario, se transformaría).
4. Haga clic en la marca de verificación **Aceptar**.





El módulo **Normalize Data** (Normalizar datos) está configurado ahora para realizar una transformación Tanh en todas las columnas numéricas excepto en la columna de riesgo de crédito.

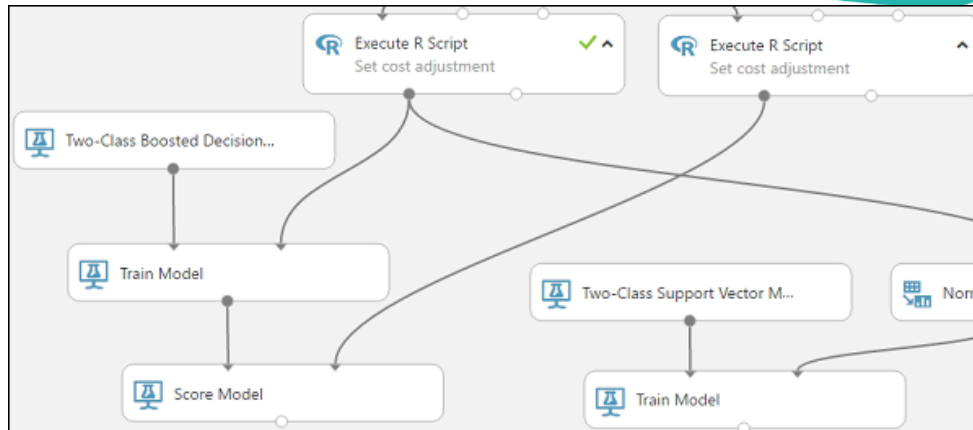
Paso 11: Puntuación y evaluación de modelos

Se utilizan los datos de prueba que se separaron mediante el módulo **Split Data** (Dividir datos) para puntuar los modelos entrenados. A continuación podremos comparar los resultados de los dos modelos para ver cuál de ellos generó mejores resultados.

Agregar los módulos Score Model (Puntuar modelo)

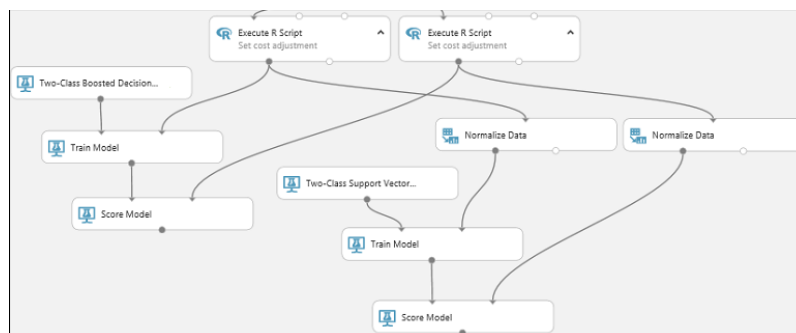
1. Busque el módulo **Score Model** (Puntuar modelo) y arrástrelo al lienzo.
2. Conecte el módulo **Train Model** (Entrenar modelo) que está conectado al módulo **Two-Class Boosted Decision Tree** (Árbol de decisión ampliado de dos clases) al puerto de entrada izquierdo del módulo **Score Model** (Puntuar modelo).
3. Conecte el módulo derecho **Execute R Script** (Ejecutar script R) (los datos de prueba) al puerto de entrada derecho del módulo **Score Model** (Puntuar modelo).





El módulo **Score Model** (Puntuar modelo) ahora puede utilizar la información de crédito de los datos de prueba, ejecutarla a través del modelo y comparar las predicciones que el modelo genera con la columna de riesgo de crédito real de los datos de prueba.

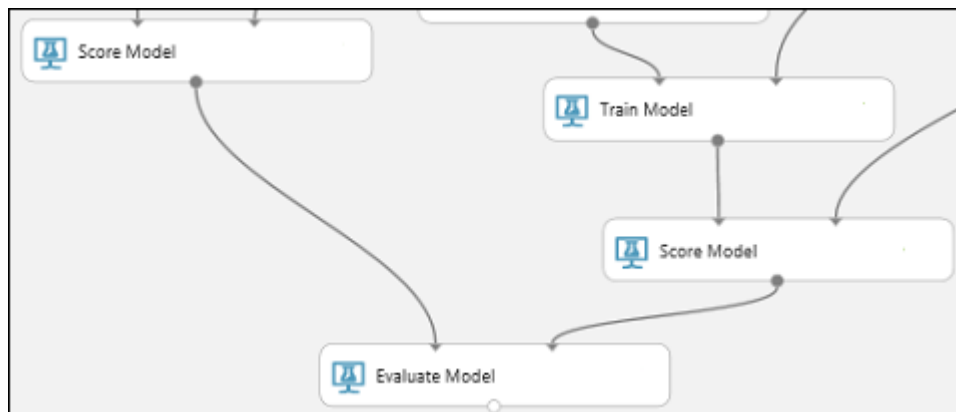
4. Copie y pegue el módulo **Score Model** (Puntuar modelo) para crear una segunda copia.
5. Conecte la salida del modelo SVM; es decir, el puerto de salida del módulo **Train Model** (Entrenar modelo) que está conectado al módulo **Two-Class Support Vector Machine** (Máquina de vectores de soporte de dos clases), al puerto de entrada del segundo módulo **Score Model** (Puntuar modelo).
6. En cuanto al modelo SVM, tiene que realizar la misma transformación en los datos de prueba que la que realizó con los datos de entrenamiento. Así pues, copie y pegue el módulo **Normalize Data** (Normalizar datos) para crear una segunda copia y conéctelo al módulo derecho **Execute R Script** (Ejecutar script R).
7. Conecte la salida izquierda del segundo módulo **Normalize Data** (Normalizar datos) al puerto de salida derecho del segundo módulo **Score Model** (Puntuar modelo).



Agregar el módulo Evaluate Model (Evaluar modelo)

Para evaluar los dos resultados de puntuación y compararlos, use un módulo Evaluate Model (Evaluar modelo).

1. Busque el módulo **Evaluate Model** (Evaluar modelo) y arrástrelo al lienzo.
2. Conecte el puerto de salida del módulo **Score Model** (Puntuar modelo) asociado al modelo del árbol de decisión ampliado al puerto de entrada izquierdo del módulo **Evaluate Model** (Evaluar modelo).
3. Conecte el otro módulo **Score Model** (Puntuar modelo) al puerto de entrada derecho.

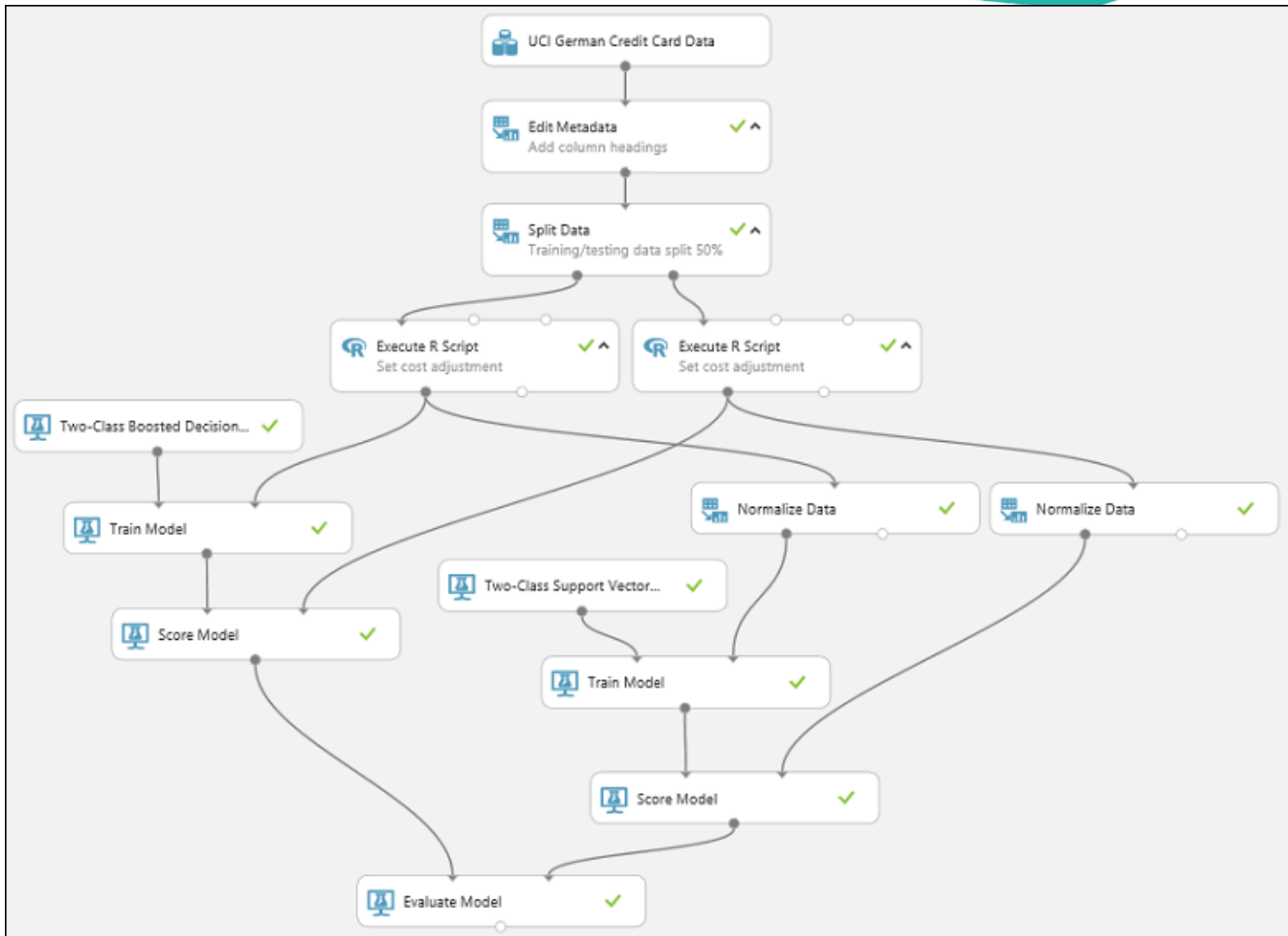


Ejecutar el experimento y comprobar los resultados

Para ejecutar el experimento, haga clic en el botón **EJECUTAR** bajo el lienzo. Esto puede tardar unos minutos. Aparece un indicador giratorio en cada módulo para indicar que está en ejecución y, cuando el módulo acaba, aparece una marca de verificación de color verde. Cuando todos los módulos tengan una marca de verificación, habrá finalizado la ejecución del experimento.

El experimento debería tener ahora un aspecto similar al siguiente:





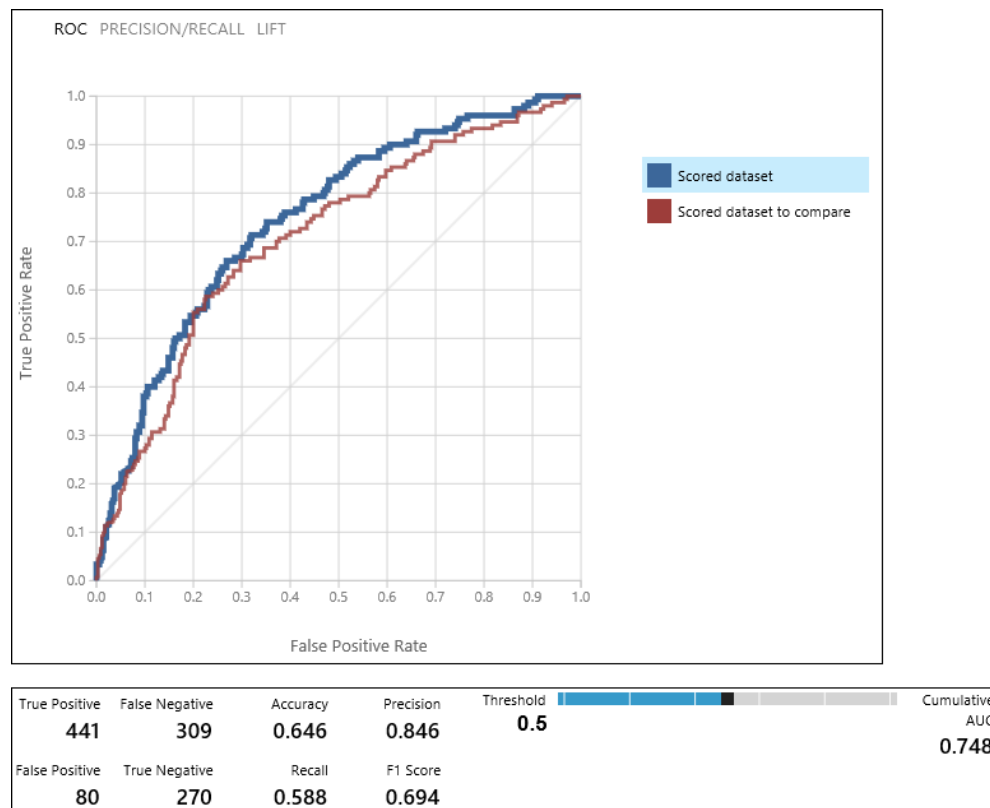
Para comprobar los resultados, haga clic en el puerto de salida del módulo Evaluate Model (Evaluar modelo) y seleccione Visualizar.

El módulo Evaluate Model (Evaluar modelo) produce un par de curvas y métricas que permiten comparar los resultados de los dos modelos de puntuación. Puede ver los resultados como curvas de características operativas del receptor (ROC), curvas de precisión/sensibilidad o curvas de elevación. También se muestran otros datos como la matriz de confusión y los valores del área bajo la curva (AUC) acumulados, entre otras métricas. También puede cambiar el valor del umbral moviendo el control deslizante a la izquierda o a la derecha, y comprobar cómo afecta esta acción al conjunto de métricas.

A la derecha del gráfico, haga clic en **Scored dataset** (Conjunto de datos puntuados) o en **Scored dataset to compare** (Conjunto de datos puntuados para



comparar) con el fin de resaltar la curva asociada y mostrar debajo las métricas asociadas. En la leyenda de las curvas, "Conjunto de datos puntuados" corresponde al puerto de entrada izquierdo del módulo **Evaluate Model** (Evaluar modelo); en este caso, se trata del modelo del árbol de decisión ampliado. "Conjunto de datos puntuados para comparar" corresponde al puerto de entrada derecho (el modelo SVM en nuestro caso). Al hacer clic en una de estas etiquetas, la curva del modelo correspondiente se resalta y muestra las métricas correspondientes tal y como se muestra en el gráfico siguiente.



Si examina estos valores, podrá decidir cuál es el modelo que más se acerca a ofrecerle los resultados que busca.

