

Руководство 1. Прогнозирование кредитного риска — Студия машинного обучения (классическая)

Статья • 11.02.2019

ОБЛАСТЬ ПРИМЕНЕНИЯ:  Студия машинного обучения (классическая) 

[Машинное обучение Azure](#)

Важно!

Поддержка Студии машинного обучения (классической) будет прекращена 31 августа 2024 г. До этой даты рекомендуется перейти на [Машинное обучение Azure](#).

Начиная с 1 декабря 2021 года вы не сможете создавать новые ресурсы Студии машинного обучения (классической). Существующие ресурсы Студии машинного обучения (классическая версия) можно будет использовать до 31 августа 2024 г.

- См. [сведения о переносе проектов машинного обучения из Студии машинного обучения \(классическая версия\) в Машинное обучение Azure](#).
- См. дополнительные сведения о [Машинном обучении Azure](#).

Прекращается поддержка документации по Студии машинного обучения (классическая версия). В будущем она может не обновляться.

В этом руководстве подробно описывается процесс разработки решения прогнозной аналитики. Мы создадим в Студии машинного обучения (классической) простую модель, а затем развернем ее в качестве веб-службы машинного обучения. Развернутая модель позволяет создавать прогнозы на основе новых данных. Это **первое руководство в серии, состоящей из трех частей**.

Предположим, требуется спрогнозировать кредитный риск претендентов на кредит на основе сведений, которые они предоставляют в заявке на кредит.

Оценка кредитных рисков — это сложная задача. Мы немного упростим ее для этого руководства. На примере этой задачи вы создадите решение прогнозной аналитики с помощью Студии машинного обучения (классической). Для этого вы используете Студию машинного обучения (классическую) и веб-службу машинного обучения.

В этом руководстве из трех частей описание начинается с общедоступных данных по кредитным рискам. Затем необходимо разработать и обучить модель прогнозирования. и, наконец, развернете модель в качестве веб-службы.

Эта часть руководства состоит из таких этапов:

- ✓ создание рабочей области Студии машинного обучения (классической);
- ✓ отправка существующих данных;
- ✓ Создание эксперимента

Затем вы сможете использовать этот эксперимент для [обучения модулей в части 2](#), чтобы [развернуть их в части 3](#).

Предварительные требования

В этом учебнике предполагается, что вы уже работали со Студией машинного обучения (классической) и имеете некоторое представление о машинном обучении. При этом предполагается, что вы не являетесь специалистом в этой области.

Если вы не использовали **Студию машинного обучения (классическую)**, прежде чем приступить к работе с этим учебником, [создайте свой первый эксперимент по обработке и анализу данных в Студии машинного обучения \(классической\)](#). При ознакомлении со Студией машинного обучения (классической) вы научитесь выполнять базовые операции. переносить модули в эксперимент, соединять их друг с другом, запускать эксперимент и просматривать результаты.

Совет

Рабочую копию эксперимента, разработанного в этом руководстве, можно найти в **Коллекции решений ИИ Azure**. Перейдите по ссылке **Tutorial - Predict credit risk** (Учебник. Прогнозирование кредитных рисков) и нажмите кнопку **Open in Studio** (Открыть в Студии), чтобы скачать в рабочую область

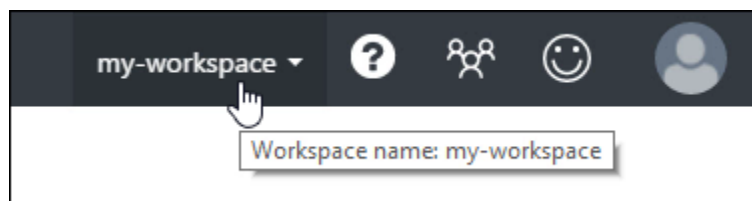
Студии машинного обучения (классической) копию этого эксперимента.

создание рабочей области Студии машинного обучения (классической);

Чтобы использовать Студию машинного обучения (классическую), требуется рабочая область Студии машинного обучения (классической). Такая рабочая область содержит инструменты, необходимые для создания, публикации экспериментов и управления ими.

Инструкции по созданию рабочей области Студии машинного обучения (классической) и предоставлению к ней общего доступа см. в [этой статье](#).

Создав рабочую область, откройте Студию машинного обучения (классическую) (<https://studio.azureml.net/Home>). Если у вас несколько рабочих областей, можно выбрать рабочую область на панели инструментов в правом верхнем углу окна.



💡 Совет

Владелец рабочей области может предоставить общий доступ к экспериментам, над которыми он работает, пригласив других пользователей в свою рабочую область. Это можно сделать в Студии машинного обучения (классической) на странице **Параметры**. Для каждого пользователя необходима учетная запись Майкрософт или учетная запись организации.

На странице **Параметры** щелкните **Пользователи**, а затем — **Invite more users** (Пригласить пользователей) в нижней части окна.

отправка существующих данных;

Чтобы разработать модель прогнозирования кредитного риска, потребуются

данные, которые можно использовать для обучения и последующей проверки модели. В рамках этого руководства вы используете набор данных "UCI Statlog (German Credit Data) Data Set" из репозитория машинного обучения UC Irvine Machine Learning Repository. Его можно найти по следующей ссылке:
[https://archive.ics.uci.edu/ml/datasets/Statlog+\(German+Credit+Data\)](https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data))

Вы будете использовать файл с именем **german.data**. Загрузите этот файл на свой жесткий диск.

Набор данных **german.data** содержит строки из 20 переменных для 1000 претендентов на получение кредита. Эти 20 переменных представляют набор свойств набора данных (*характеристический вектор*), обеспечивающих определяющие характеристики каждого претендента на кредит. Дополнительный столбец в каждой строке представляет расчетный кредитный риск претендента, причем 700 претендентов идентифицированы с низким уровнем кредитного риска, а 300 — с высоким уровнем риска.

На веб-сайте UCI представлено описание атрибутов вектора свойств для этих данных. В них входят финансовые сведения, кредитная история, данные о занятости и личные сведения. Для каждого претендента приведен двоичный рейтинг, указывающий низкий или высокий кредитный риск.

Вы будете использовать эти данные для обучения модели прогнозирующей аналитики. По завершении ваша модель должна уметь принимать вектор свойств для новых претендентов и прогнозировать для них степень кредитного риска.

Есть один интересный момент.

В описании набора данных на веб-сайте UCI упоминается, какой ценой обходится ошибка классификации кредитного риска человека. Если модель прогнозирует высокий кредитный риск для тех, у кого он на самом деле низкий, это означает, что модель допустила ошибку классификации.

Но для финансового учреждения в пять раз дороже обходится обратная ошибка классификации: когда модель прогнозирует низкий кредитный риск для тех, у кого он на самом деле высокий.

Поэтому необходимо обучить модель таким образом, чтобы стоимость ошибки классификации второго типа была в пять раз выше, чем прочих ошибок классификации.

Один простой способ сделать это при обучении модели в вашем эксперименте — 5 раз повторять записи, представляющие претендента с высоким кредитным риском.

Если модель ошибочно отнесет к категории низкого кредитного риска кого-либо, чей кредитный риск на самом деле высокий, то модель допустит эту ошибку классификации пять раз, по одному разу для каждой повторяющейся записи. Это увеличит стоимость ошибки в результатах подготовки.

Преобразование формата набора данных

В исходном наборе данных используется формат с разделителями-пробелами. Студия машинного обучения (классическая) лучше работает с файлами, в которых используются разделители-запятые (CSV-файлами), поэтому вы преобразуете набор данных, заменив пробелы запятыми.

Преобразовать эти данные можно разными способами. Один из них — с помощью следующей команды Windows PowerShell:

PowerShell

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

Другой — с помощью команды `sed` Unix:

Консоль


```
sed 's/ /,/g' german.data > german.csv
```

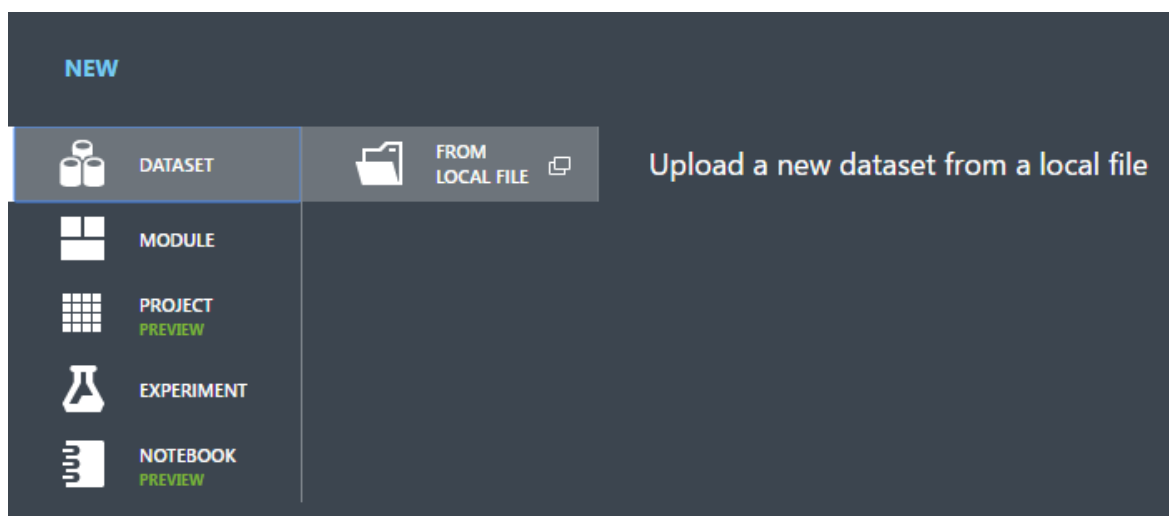
В любом случае вы создаете версию разделенных запятыми данных в файле **german.csv**, который будет использоваться в эксперименте.

Передача набора данных в Студию машинного обучения (классическую)

После преобразования данных в формат CSV вам необходимо отправить их в Студию машинного обучения (классическую).

1. Откройте домашнюю страницу Студии машинного обучения (классической) (<https://studio.azureml.net>).

- Щелкните меню  в верхнем левом углу окна, щелкните **Машинное обучение Azure**, выберите **Студия** и выполните вход.
- В нижней части окна щелкните **+СОЗДАТЬ**.
- Выберите **НАБОР ДАННЫХ**.
- Выберите **ИЗ ЛОКАЛЬНОГО ФАЙЛА**.



- В диалоговом окне **Upload a new dataset** (Отправка нового набора данных) нажмите кнопку "Обзор" и найдите созданный ранее файл **german.csv**.
- Введите имя для набора данных. В этом руководстве назовем его "UCI German Credit Card Data".
- Для типа данных выберите **Универсальный CSV-файл без заголовка (.nh.csv)**.
- При желании добавьте описание.
- Нажмите кнопку с флажком **ОК**.

×

Upload a new dataset

SELECT THE DATA TO UPLOAD:

Choose File

german.csv

☐ This is the new version of an existing dataset

ENTER A NAME FOR THE NEW DATASET:

UCI German Credit Card Data

SELECT A TYPE FOR THE NEW DATASET:

Generic CSV File With no header (.nh.csv) ▼

PROVIDE AN OPTIONAL DESCRIPTION:

Downloaded from
<http://archive.ics.uci.edu/ml/datasets>

✓

Данные будут отправлены в модуль набора данных, который можно использовать в эксперименте.

Для управления отправленными в Студию (классическую) наборами данных откройте вкладку **Наборы данных** в левой части окна Студии (классической).

Microsoft Azure Machine Learning Studio

PROJECTS

EXPERIMENTS

WEB SERVICES

NOTEBOOKS

DATASETS

TRAINED MODELS

SETTINGS

datasets

MY DATASETS SAMPLES

	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	

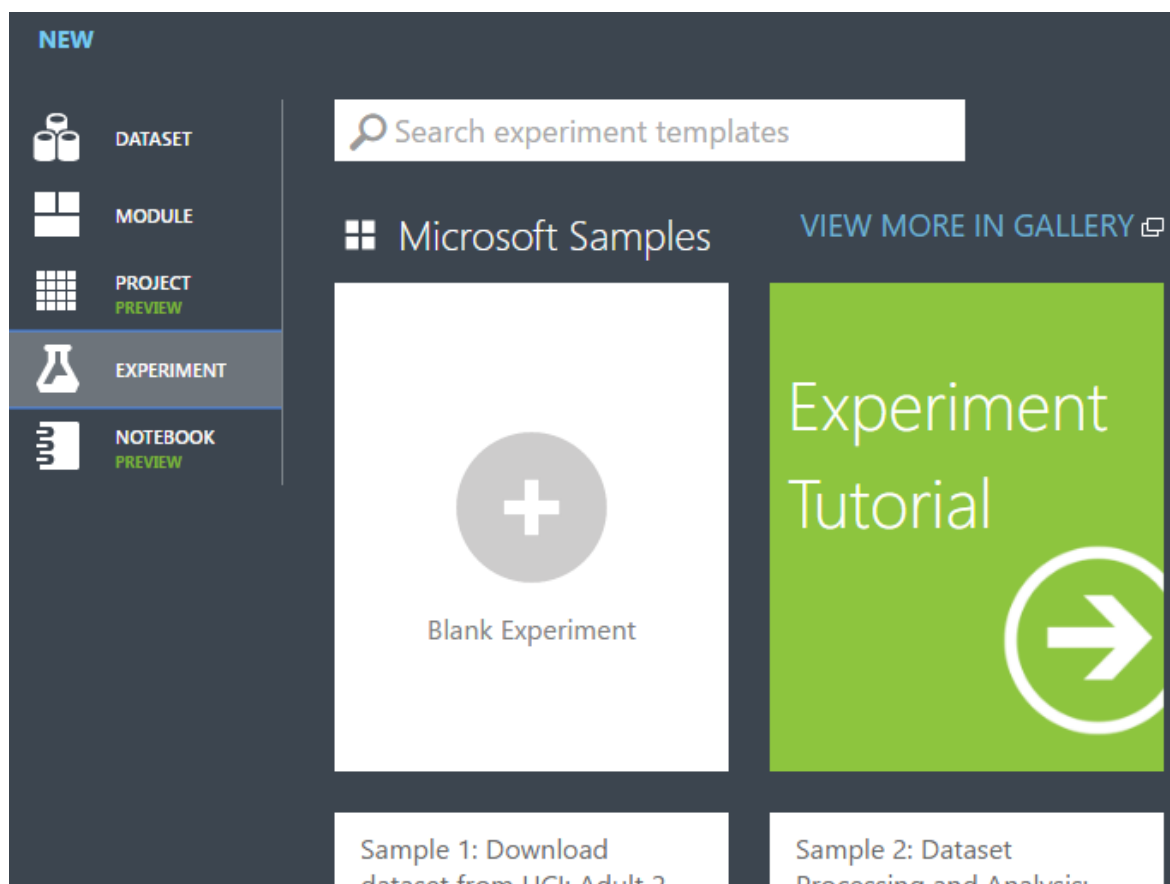
Дополнительные сведения об импорте других типов данных в эксперимент см. в

статье [Импорт обучающих данных в Студию машинного обучения \(классическую\)](#).

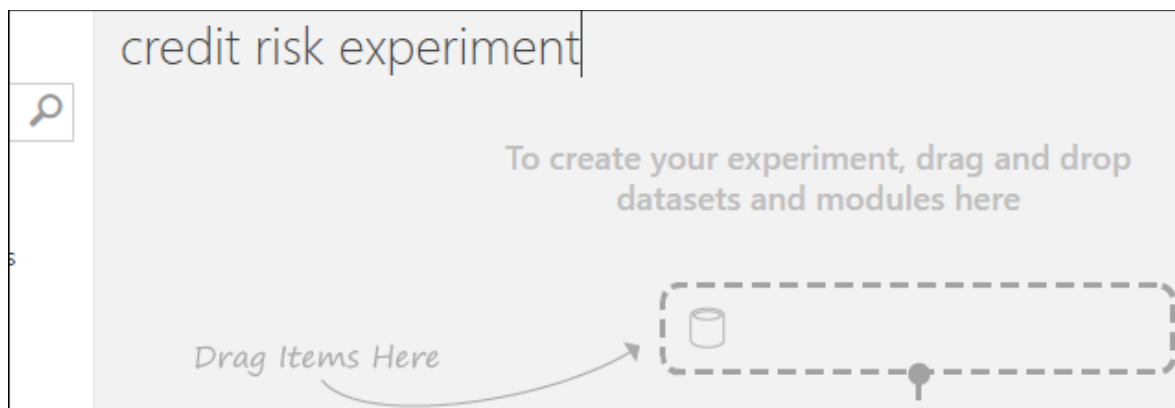
Создание эксперимента

Следующий этап в этом учебнике — создание эксперимента в Студии машинного обучения (классическая), в котором используется отправленный вами набор данных.

1. В Студии машинного обучения (классической) щелкните **+СОЗДАТЬ** в нижней части окна.
2. Выберите **ЭКСПЕРИМЕНТА**, а затем выберите "Пустой эксперимент".

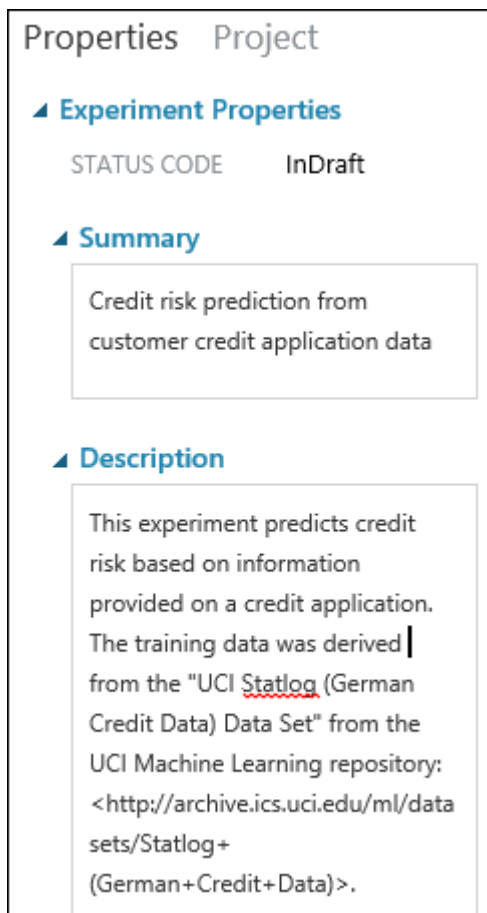


3. В верхней части холста выберите имя эксперимента по умолчанию и присвойте ему понятное имя.



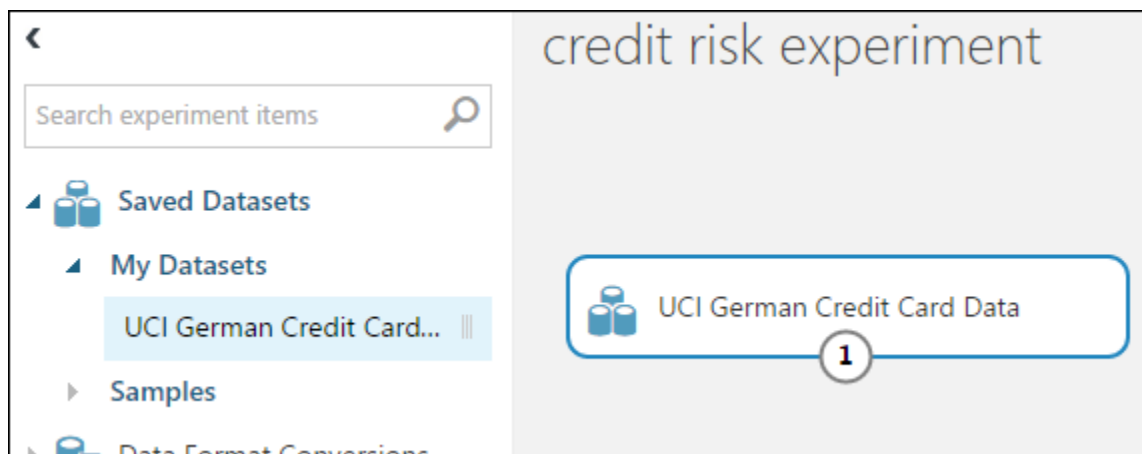
💡 Совет

Рекомендуем заполнить поля **Сводка** и **Описание** для эксперимента в области **Свойства**. Эти свойства позволят вам задокументировать эксперимент, чтобы любой, кто откроет его позднее, смог понять ваши цели и методы.



4. На палитре модулей слева от полотна эксперимента разверните **Saved Datasets**(Сохраненные наборы данных).

5. Найдите набор данных, созданный в разделе **Мои наборы данных** , и перетащите его на холст. Набор данных можно также отыскать, введя имя в поле **Поиск** над палитрой.



Подготовка данных

Чтобы просмотреть первые 100 строк данных и некоторые статистические сведения для всего набора данных, щелкните выходной порт набора данных (маленький кружок в нижней части) правой кнопкой мыши и выберите команду **Визуализировать**.

Так как файл данных поступил без заголовков столбцов, Студия (классическая) предоставляет универсальные заголовки (Col1, Col2 и т. д.). Понятные заголовки столбцов не имеют существенного значения для создания модели, однако они облегчат работу с данными в эксперименте. Кроме того, при последующей публикации этой модели в веб-службе заголовки помогают определять столбцы пользователям службы.

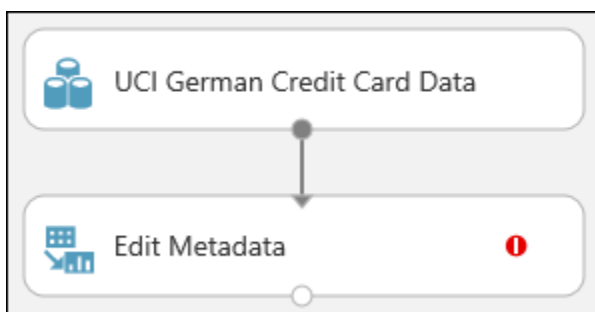
Вы можете добавить заголовки столбцов с помощью модуля [Edit Metadata](#) (Изменение метаданных).

Модуль [Edit Metadata](#) (Изменение метаданных) используется для изменения метаданных, связанных с набором данных. В этом случае вы используете его, чтобы предоставить более понятные имена заголовков столбцов.

Для использования модуля [Edit Metadata](#) (Изменение метаданных) сначала необходимо указать, какие столбцы будут изменены (в нашем случае все). Далее следует указать действие, которое будет выполнено с этими столбцами (в нашем случае изменение заголовков столбцов).

1. В палитре модуля введите "metadata" в поле **Поиск** . Модуль **Edit Metadata** (Изменение метаданных) появляется в списке модулей.
2. Щелкните и перетащите модуль **Edit Metadata** (Изменение метаданных) на холст и вставьте его под набором данных, добавленным ранее.
3. Подключите набор данных к модулю **Edit Metadata** (Изменение метаданных). Щелкните порт вывода набора данных (маленький кружок в нижней части набора данных), перетащите в порт ввода модуля **Edit Metadata** (Изменение метаданных) (маленький кружок в верхней части модуля) и отпустите кнопку мыши. Набор данных и модуль остаются подключенными даже при их перемещении на холсте.

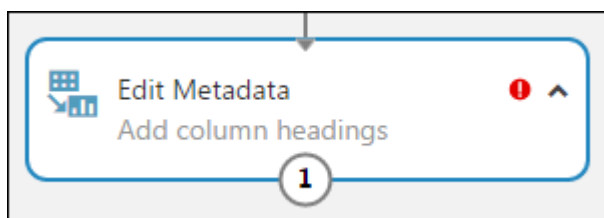
Теперь наш эксперимент должен выглядеть следующим образом:



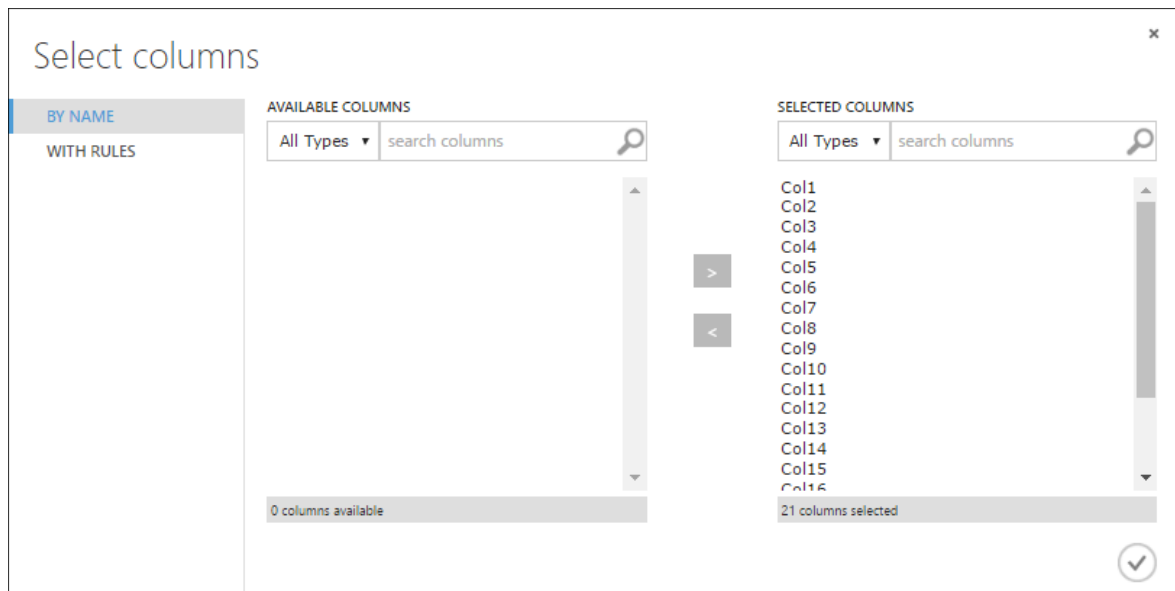
Красный восклицательный знак указывает на то, что для этого модуля еще не заданы свойства. Поэтому вы сделаете это сейчас.

💡 Совет

Дважды щелкните модуль и введите текст, чтобы добавить комментарий. Это поможет вам увидеть описание модуля и его действие в рамках эксперимента. В этом случае дважды щелкните модуль **Edit Metadata** (Изменение метаданных) и введите комментарий "Добавить заголовки столбцов". Щелкните в любом месте холста, чтобы закрыть текстовое поле. Чтобы отобразить комментарий, щелкните стрелку вниз в модуле.



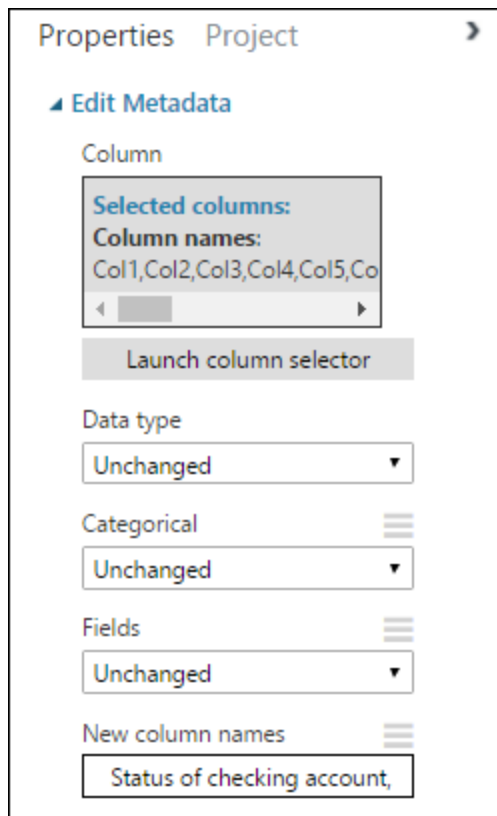
4. Выберите **Edit Metadata** (Изменение метаданных), а затем в области **Свойства** справа от холста щелкните **Launch column selector** (Запустить средство выбора столбцов).
5. В диалоговом окне **Select columns** (Выбор столбцов) выберите все строки в разделе **Available Columns** (Доступные столбцы) и нажмите кнопку **>**, чтобы переместить их в раздел **Selected Columns** (Выбранные столбцы). Диалоговое окно должно выглядеть следующим образом:



6. Нажмите кнопку с флажком **OK**.
7. На панели **Свойства** найдите параметр **New column names** (Новые имена столбцов). В этом поле введите список имен для 21 столбца набора данных с разделителями-запятыми и в порядке расположения столбцов. Вы можете получить имена столбцов из документации по наборам данных на веб-сайте UCI или скопировать и вставить следующий список.

```
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
```

Панель свойств выглядит следующим образом:



💡 Совет

Если вы хотите проверить заголовки столбцов, запустите эксперимент (щелкните **RUN** (Выполнить) под холстом эксперимента). После его завершения (в модуле **Edit Metadata** (Изменение метаданных) появится зеленая галочка) щелкните порт вывода модуля **Edit Metadata** (Изменение метаданных), а затем выберите **Визуализировать**. Аналогичным образом можно просматривать выход любого модуля, чтобы следить за ходом использования данных в эксперименте.

Создание обучающих и тестовых наборов данных

Вам необходимы данные для обучения модели и данные для ее тестирования. Поэтому на следующем шаге эксперимента вы разделите набор данных на два отдельных набора: один будет использоваться для обучения модели, а другой — для ее тестирования.

Для этого используйте модуль **Split Data** (Разделение данных).

1. Найдите модуль **Split Data** (Разделение данных), перетащите его на холст и

подключите к модулю [Edit Metadata](#) (Изменение метаданных).

- По умолчанию коэффициент разделения равен 0,5 и задан параметр **Randomized split** (Случайное разделение). Это означает, что случайно выбранная половина данных будет выводиться через один порт модуля [Split Data](#) (Разделение данных), а другая половина — через другой. Эти параметры можно отрегулировать, а также использовать параметр **Случайное начальное значение**, чтобы изменить соотношение между данными для обучения и тестирования. Для этого примера оставьте их как есть.

Совет

Свойство **Fraction of rows in the first output dataset** (Доля строк в первом выходном наборе данных) определяет, какой объем данных будет выводиться через *левый* порт вывода. Например, если коэффициент установлен на 0,7, то 70% данных выходит через левый порт, а 30% — через правый.

- Дважды щелкните модуль [Split Data](#) (Разделение данных) и введите комментарий "Соотношение данных обучения и тестирования составляет 50 %".

Вы можете использовать выходные данные модуля [Split Data](#) (Разделение данных) по своему усмотрению. Но в этом случае предлагаем использовать левый выход для обучения данных, а правый — для их оценки.

Как отмечалось на [предыдущем шаге](#), стоимость ошибочной классификации высокого кредитного риска как низкого в пять раз выше, чем стоимость ошибочной классификации низкого кредитного риска как высокого. Чтобы это учесть, создайте новый набор данных, отражающий эту функцию стоимости. В новом наборе данных каждый пример высокого риска реплицируется пять раз, а каждый пример низкого риска не реплицируется.

Вы можете выполнить эту репликацию с помощью кода R:

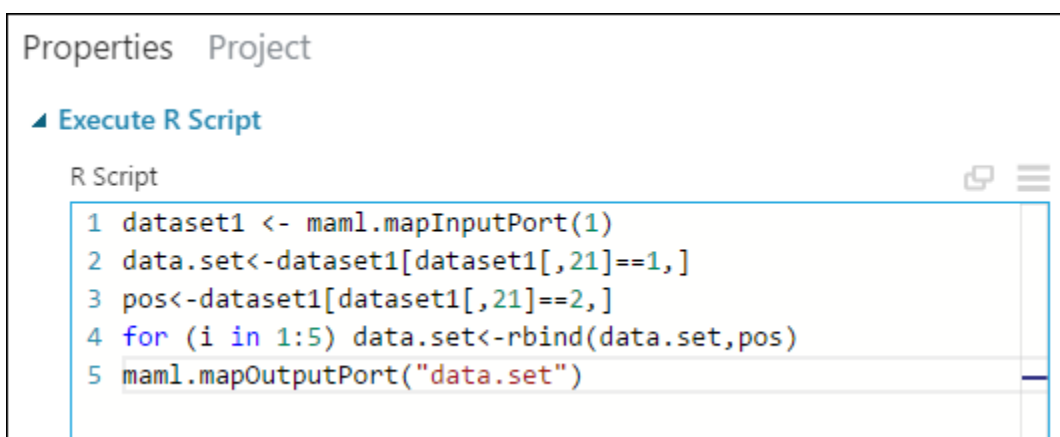
- Найдите и перетащите модуль [Execute R Script](#) (Выполнение скрипта R) на холст эксперимента.
- Подключите левый порт вывода модуля [Split Data](#) (Разделение данных) к

первому порту ввода (Dataset1) модуля [Execute R Script](#) (Выполнение скрипта R).

3. Дважды щелкните модуль [Execute R Script](#) (Выполнение скрипта R) и введите комментарий "Задать корректировку стоимости".
4. На панели **Свойства** удалите текст по умолчанию в параметре R-скрипт и введите следующий сценарий:

R

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



Вам необходимо проделать одну и ту же операцию репликации для каждого выхода модуля [Split Data](#) (Разделение данных), чтобы данные для обучения и тестирования получили одинаковую корректировку стоимости. Проще всего это сделать, скопировав созданный вами модуль [Execute R Script](#) (Выполнение сценария R) и подключив его к другому порту вывода модуля [Split Data](#) (Разделение данных).

1. Щелкните правой кнопкой мыши модуль [Execute R Script](#) (Выполнение скрипта R) и выберите **Копировать**.
2. Щелкните правой кнопкой мыши полотно эксперимента и выберите **Вставить**.
3. Перетащите модуль в позицию и подключите правый порт вывода модуля [Split Data](#) (Разделение данных) к первому порту ввода нового модуля

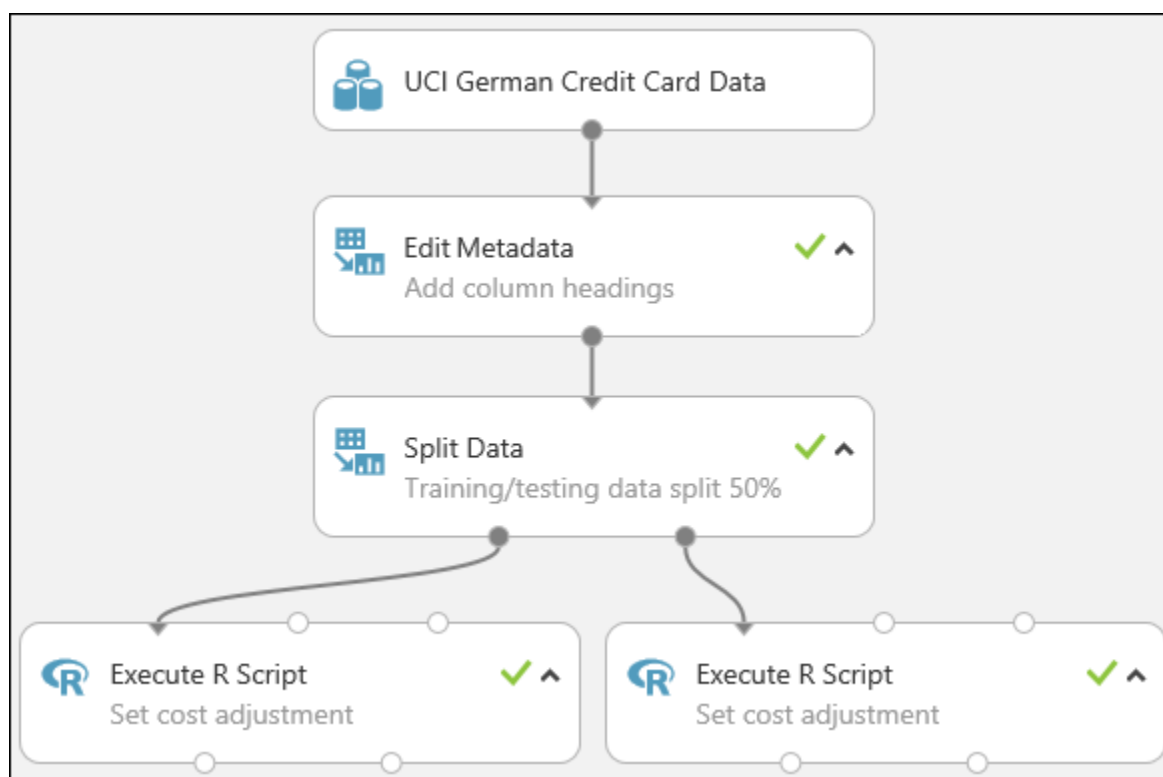
Выполнить сценарий R.

4. В нижней части холста нажмите кнопку **Run** (Выполнить).

💡 Совет

Копия модуля "Выполнение скрипта R" содержит тот же скрипт, что и исходный модуль. При копировании и вставке модуля на полотно копия сохраняет все свойства оригинала.

На эксперимент теперь выглядит приблизительно так:



Подробнее об использовании сценариев R в экспериментах см. в разделе [Расширение возможностей эксперимента с помощью R](#).

Очистка ресурсов

Если вы не планируете дальше использовать ресурсы, созданные при работе с этой статьей, удалите их, чтобы плата не взималась. О том, как это сделать, см. в статье [Экспорт и удаление встроенных в продукт данных пользователей из Студии машинного обучения Azure](#).

Дальнейшие действия

В рамках этого руководства вы выполнили следующие действия:

- ✓ создание рабочей области Студии машинного обучения (классической);
- ✓ отправка существующих данных в рабочую область;
- ✓ Создание эксперимента

Теперь вы готовы начать обучение и анализ моделей для этих данных.

Руководство 2. Обучение и анализ моделей