



Microsoft Machine Learning in a Day Workshop

Student Guide

March 2019

Information in this document, including URL and other Internet Web site references, is subject to change without notice. Unless otherwise noted, the example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted herein are fictitious, and no association with any real company, organization, product, domain name, e-mail address, logo, person, place or event is intended or should be inferred. Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

The names of manufacturers, products, or URLs are provided for informational purposes only and Microsoft makes no representations and warranties, either expressed, implied, or statutory, regarding these manufacturers or the use of the products with any Microsoft technologies. The inclusion of a manufacturer or product does not imply endorsement of Microsoft of the manufacturer or product. Links may be provided to third party sites. Such sites are not under the control of Microsoft and Microsoft is not responsible for the contents of any linked site or any link contained in a linked site, or any changes or updates to such sites. Microsoft is not responsible for webcasting or any other form of transmission received from any linked site. Microsoft is providing these links to you only as a convenience, and the inclusion of any link does not imply endorsement of Microsoft of the site or the products contained therein.

© 2018 Microsoft Corporation. All rights reserved.

Microsoft and the trademarks listed at <https://www.microsoft.com/en-us/legal/intellectualproperty/Trademarks/Usage/General.aspx> are trademarks of the Microsoft group of companies. All other trademarks are property of their respective owners.

Contents

Machine Learning in a Day student guide.....	1
Abstract and learning objectives.....	1
Architecture	2
Section 1: Set up a Data Science Virtual Machine (DSVM) on Azure Portal.....	3
Section 2: Download Telco dataset and create Blob Storage Account and Container access	9
Section 3: Upload CSV file to Blob Storage Container Manually.....	15
Section 4: Build a clustering unsupervised model in Azure Machine Learning Studio	18
Section 5: Create a logistic regression model in Azure ML Studio	31
Section 6: Export scored dataset with prediction scores from Blob to Power BI Desktop	39

Machine Learning in a Day student guide

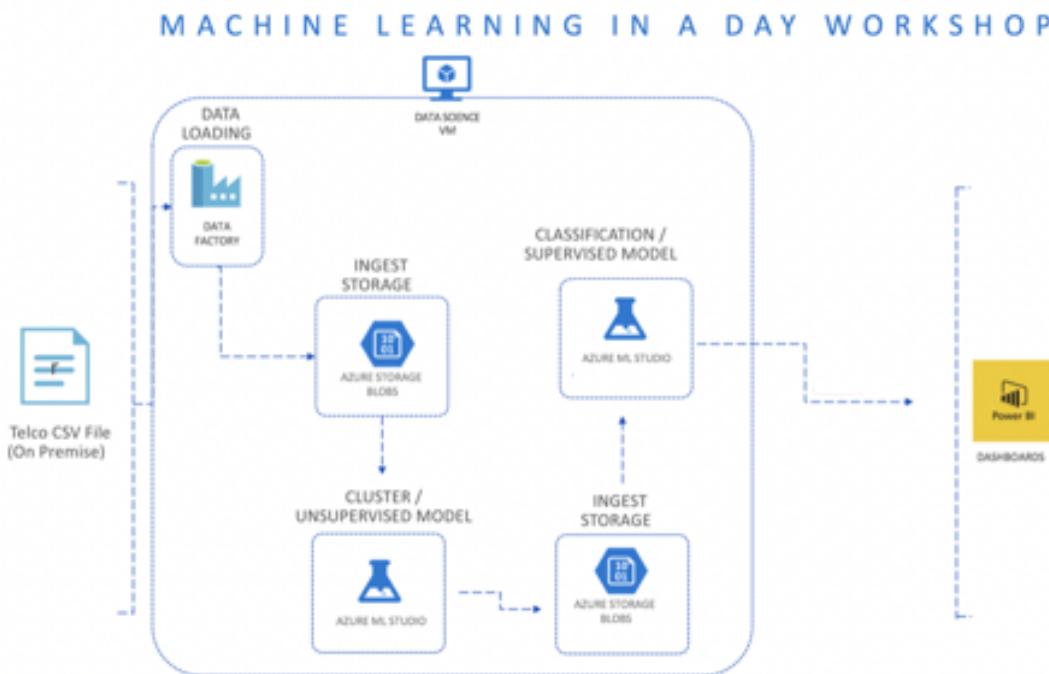
Abstract and learning objectives

In this workshop, you will build a Machine Learning model that will be used to predict whether customers remain with business or move onto other services based on their behavior, demographics, and spending. You will leverage several Microsoft products and services on Microsoft Azure to deliver the Machine Learning Model.

The dataset that will be used in the workshop contains the following information

1. Customers who left within the last month – the column is called **Churn**
2. Services that each customer has signed up for
 - a. Phone
 - b. Multiple lines
 - c. Internet
 - d. Online security
 - e. Online backup
 - f. Device protection
 - g. Tech support
 - h. Streaming TV and movies
3. Customer account information
 - a. How long they've been a customer
 - b. Contract
 - c. Payment method
 - d. Paperless billing
 - e. Monthly charges
 - f. Total charges
4. Demographic information about customers
 - a. Gender
 - b. Age range
 - c. If they have partners and dependents

Architecture



Section 1: Set up a Data Science Virtual Machine (DSVM) on Azure Portal

1. The use of this documentation requires the subscription of a Microsoft Azure account either from a personal email account or a work email account.
2. Visit portal.azure.com and click on **Create a resource** on the left-hand side and type in **Data Science Virtual Machine – Windows 2012** and select the **Create** button as seen in the following screenshot:

The screenshot shows the Azure portal interface. On the left, there's a sidebar with various service icons and a 'Create a resource' button highlighted with a yellow box. The main pane has a search bar containing 'data science virtual machine'. Below it, a table lists several VM options, with the first one, 'Data Science Virtual Machine - Windows 2012', also highlighted with a yellow box. To the right, a detailed description of the selected VM is shown, along with links to documentation and a 'Create' button. At the very bottom of the page, another 'Create' button is visible.

3. Assign the following basic configuration settings for your Data Science Virtual Machine and the select the **OK** button as seen in the following screenshot.

Create a virtual machine

[Basics](#) [Disks](#) [Networking](#) [Management](#) [Guest config](#) [Tags](#) [Review + create](#)

Create a virtual machine that runs Linux or Windows. Select an image from Azure marketplace or use your own customized image. Complete the Basics tab then Review + create to provision a virtual machine with default parameters or review each tab for full customization.

Looking for classic VMs? [Create VM from Azure Marketplace](#)

PROJECT DETAILS

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

* Subscription [CCG Training Sandbox](#)
 * Resource group [DS-Practice](#) [Create new](#)

INSTANCE DETAILS

* Virtual machine name [dsvmMLinaday](#)
 * Region [East US](#)
 Availability options [No infrastructure redundancy required](#)
 * Image [Data Science Virtual Machine - Windows 2012](#) [Browse all images and disks](#)
 * Size [Standard DS3 v2](#)
 4 vcpus, 14 GB memory [Change size](#)

ADMINISTRATOR ACCOUNT

* Username [demouser](#)
 * Password [#MLinaDay2019](#)
 * Confirm password [#MLinaDay2019](#)

Please note that you can keep your Names and passwords different from this document as long as you remember what you used for future use. In this section, we will assign a username of **demouser** and a password of **#MLinaDay2019** (Please keep track of both the Username and Password as you will need both of them several times as we go through future sections of this workshop).

- After clicking “Change size,” choose the following size VM (**D2S_v3**) for the purposes of this workshop as seen in the following screenshot:

Select a VM size

Browse available virtual machine sizes and their features

Search by VM size... [Clear all filters](#)

Size : Small (0-4) Generation : Current Family : General purpose Premium disk : Supported [Add filter](#)

Showing 13 of 198 VM sizes. | Subscription: CCG Training Sandbox | Region: East US | Current size: Standard_DS3_v2

VM SIZE	OFFERING	FAMILY	VCPUS	RAM (GB)	DATA DISKS	MAX IOPS	TEMPORARY STORA...	Premium disk sup...	COST/MONTH (ESTI...)
B1ms	Standard	General purpose	1	2	2	800	4 GB	Yes	\$18.30
B1s	Standard	General purpose	1	1	2	400	4 GB	Yes	\$10.42
B2ms	Standard	General purpose	2	8	4	2400	16 GB	Yes	\$67.85
B2s	Standard	General purpose	2	4	4	1600	8 GB	Yes	\$36.90
B4ms	Standard	General purpose	4	16	8	3600	32 GB	Yes	\$135.41
D2S_v3	Standard	General purpose	2	8	4	3200	16 GB	Yes	\$139.87
D4s_v3	Standard	General purpose	4	16	8	6400	32 GB	Yes	\$279.74
DS1_v2	Standard	General purpose	1	3.5	4	3200	7 GB	Yes	\$93.74

- For settings, the only configuration that is truly needed is to navigate to the Management tab and enable **auto-shutdown** every evening at **7pm EST** (your manager will thank you later on !!!!) and then select **Review + create** as seen in the following screenshot:

Create a virtual machine

Basics • Disks Networking Management Guest config Tags Review + create

Configure monitoring and management options for your VM.

MONITORING

Boot diagnostics On Off

OS guest diagnostics On Off

* Diagnostics storage account

IDENTITY

System assigned managed identity On Off

AUTO-SHUTDOWN

Enable auto-shutdown On Off

Shutdown time

Time zone

Notification before shutdown On Off

Review + create **Previous** **Next : Guest config >**

6. In the summary section and confirm everything that was selected is accurate and let the provisioning process of the DSVM begin by clicking on **Create** as seen in the following screenshot:

Create a virtual machine

 Validation passed

Basics Disks Networking Management Guest config Tags Review + create

PRODUCT DETAILS

Data Science Virtual Machine - Windows 2012 by Microsoft	Not covered by credits ⓘ 0.0000 USD/hr
Standard D2s v3 by Microsoft	Subscription credits apply ⓘ 0.1880 USD/hr Pricing for other VM sizes

TERMS

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. See the [Azure Marketplace Terms](#) for additional details.

BASICS

Subscription	CCG Training Sandbox
Resource group	DS-Practice
Virtual machine name	mliadDSVMbees
Region	East US
Availability options	No infrastructure redundancy required
Username	demouser
Already have a Windows license?	No

DISKS

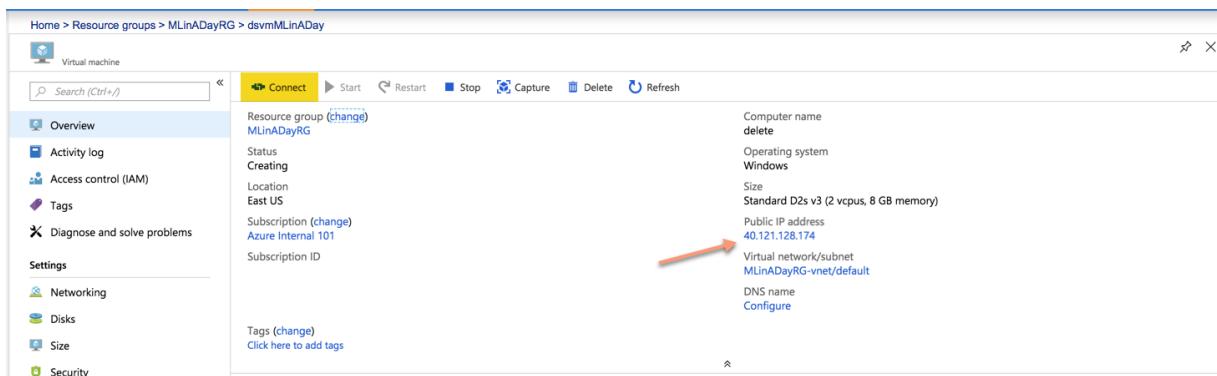
OS disk type	Premium SSD
Use managed disks	Yes

NETWORKING

Virtual network	CCG-INTERNAL
-----------------	--------------

Buttons: Create, Previous, Next, Download a template for automation

- Once the provisioning is complete, you can go to your resource and view your DSVM overview and obtain your **Public IP address** as seen in the following screenshot:



Home > Resource groups > MLinADayRG > dsvmMLinADay

Virtual machine

Search (Ctrl+F)

Connect Start Restart Stop Capture Delete Refresh

Overview

Activity log Access control (IAM) Tags Diagnose and solve problems

Settings

Networking Disks Size Security

Resource group (changed) MLinADayRG

Status Creating

Location East US

Subscription (change) Azure Internal 101

Subscription ID

Tags (change) Click here to add tags

Computer name delete

Operating system Windows

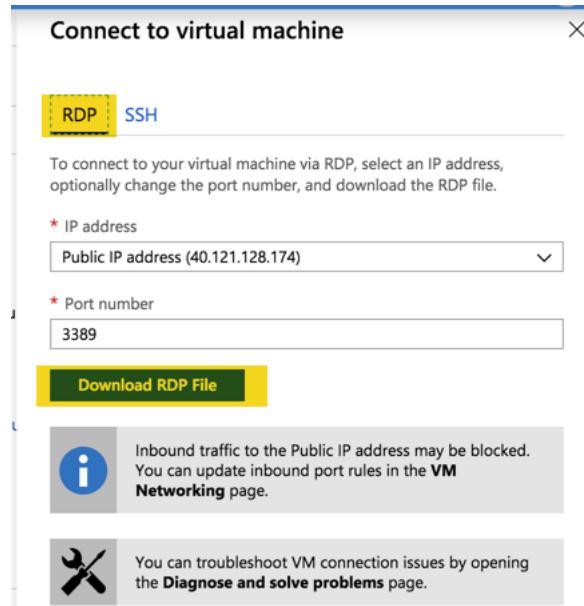
Size Standard D2s v3 (2 vcpus, 8 GB memory)

Public IP address 40.121.128.174

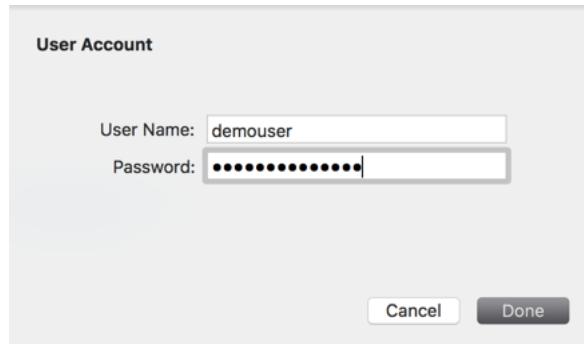
Virtual network/subnet MLinADayRG-vnet/default

DNS name Configure

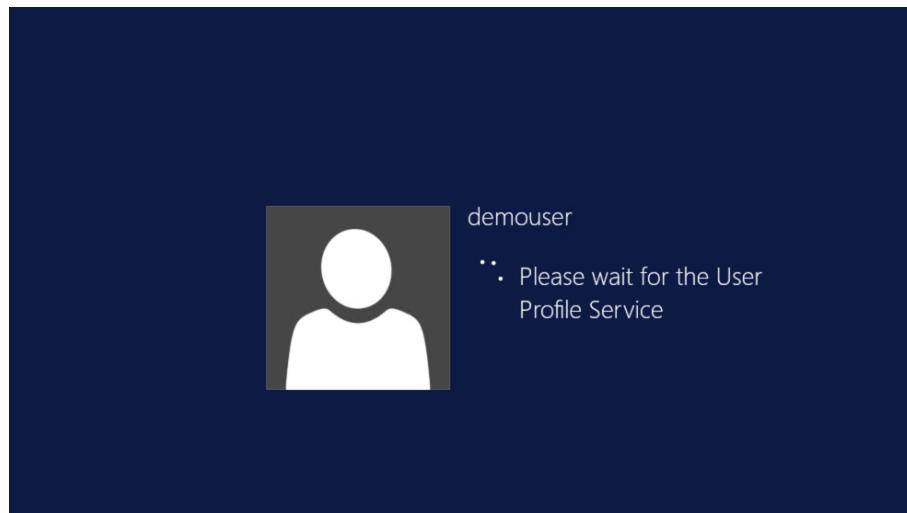
- Make sure that your DSVM is currently running and click on the **Connect** button to enter your virtual machine using **RDP** as seen in the following screenshot:



9. Enter your login credentials that were created back in Step # 3 as seen in the following screenshot:



10. It may take a few extra moments while your profile is being created and you will see the following window:

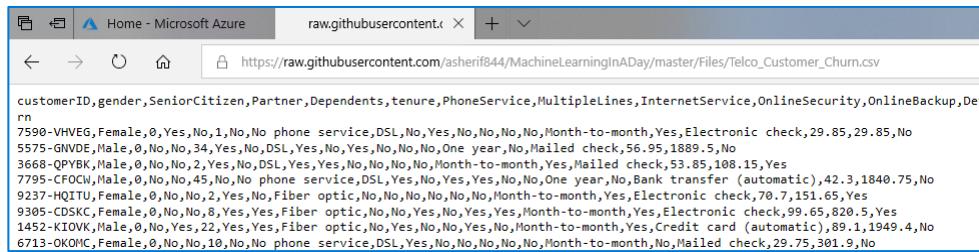


11. Once you log in using your credentials using the RDP approach, you should see the following remote server for the data science virtual machine as seen in the following screenshot:



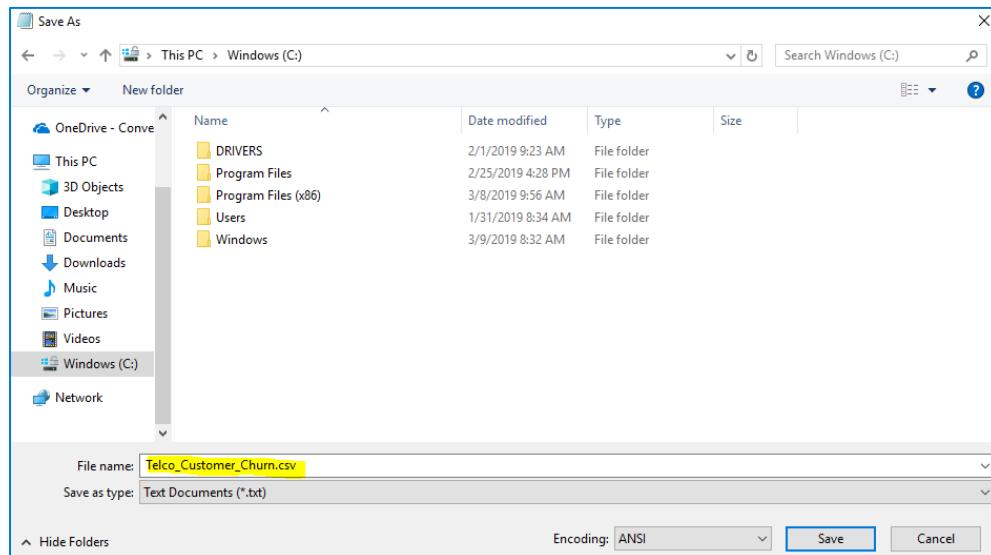
Section 2: Download Telco dataset and create Blob Storage Account and Container access

- The Telco Churn Dataset ([TELCO_CUSTOMER_CHURN.csv](https://raw.githubusercontent.com/asherif844/MachineLearningInADay/master/Files/Telco_Customer_Churn.csv)) can be accessed through the following link:
https://raw.githubusercontent.com/asherif844/MachineLearningInADay/master/Files/Telco_Customer_Churn.csv



The screenshot shows a Microsoft Edge browser window with the URL https://raw.githubusercontent.com/asherif844/MachineLearningInADay/master/Files/Telco_Customer_Churn.csv. The page displays a large block of comma-separated values (CSV) data, which is the Telco Customer Churn dataset. The data includes columns such as customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, Dev, and various service usage and payment details.

- Place your cursor in the screen and press Ctrl + A to select all the text. Next press Ctrl + C to copy all the text.
- Open a new notepad editor by clicking your windows menu, typing Notepad, and selecting the Notepad application when it appears in the search results.
- In your new Notepad window, press Ctrl + V to paste the data.
- Select **File** and **Save As...**
- For now, the dataset can be saved in the C:\ drive as seen in the following screenshot. Be sure to end the file name with **.csv** to specify what type of data the file contains.



- Next, we will return to our Azure portal (<https://portal.azure.com>) and create a Blob storage account to house our data. We can do this by selecting **Create A Resource**, typing **Blob** in the search, and then selecting **Storage Account** as seen in the following screenshot:

The screenshot shows the Azure Marketplace interface. On the left, there's a sidebar with a 'Create a resource' button highlighted with a red box. Below it are sections for 'All services', 'FAVORITES' (containing 'Dashboard', 'All resources', 'Resource groups', 'App Services', 'Function Apps', and 'SQL databases'), and a 'My Saved List' section with 2 items. The main area is titled 'Marketplace' and shows a search bar with 'blob' typed in, also highlighted with a red box. A red arrow points from the search bar to the search icon. To the right, the results are listed under 'Everything', including 'Compute', 'Networking', 'Storage', and 'Web'. The 'Storage' section is expanded, showing a list of storage account types. One item, 'Storage account - blob, file, table, queue', is highlighted with a red box.

- Once you create a **Blob** account, you can name and customize the properties of the account as the following, for optimal performance in this workshop (Please note than any existing resource that is created from this point forward in the workshop should ideally use the same existing **Resource Group** that we created in Section 1):

The screenshot shows the 'Create storage account' wizard in the 'Basics' step. At the top, there are tabs for 'Basics' (highlighted with a yellow box), 'Advanced', 'Tags', and 'Review + create'. Below the tabs, a descriptive text states: 'Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below.' A 'Learn more' link is provided.

PROJECT DETAILS

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

* Subscription: Azure Internal 101

* Resource group: MLinADayRG (with a 'Create new' link)

INSTANCE DETAILS

The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. [Choose classic deployment model](#)

* Storage account name: blobminaday

* Location: East US

Performance: Standard (highlighted with a yellow box)

Account kind: StorageV2 (general purpose v2) (highlighted with a yellow box)

Replication: Locally-redundant storage (LRS) (highlighted with a yellow box)

Access tier (default): Cool (highlighted with a yellow box)

At the bottom, there are navigation buttons: 'Review + create' (highlighted with a blue box), 'Previous', and 'Next : Advanced >'.

- Ensure that the Basics setting includes **Standard** Performance, **StorageV2** Account Kind, **Locally-redundant storage**, and **Cool** Access Tier.
- Continue on to the Advanced Section of the settings and ensure that **secure transfer required** has been disabled as seen in the following screenshot:

The screenshot shows the 'Create storage account' wizard with the 'Advanced' tab selected. Under 'SECURITY', 'Secure transfer required' is set to 'Disabled'. Under 'VIRTUAL NETWORKS', 'Allow access from' is set to 'All networks'. Under 'DATA LAKE STORAGE GEN2 (PREVIEW)', 'Hierarchical namespace' is set to 'Disabled'.

11. Next, we can just click on **Review and Create** and ensure that our validation has passed as seen in the following screenshot:

The screenshot shows the 'Create storage account' wizard with the 'Review + create' tab selected. A green bar at the top indicates 'Validation passed'. The 'BASICS' section shows the following configuration:

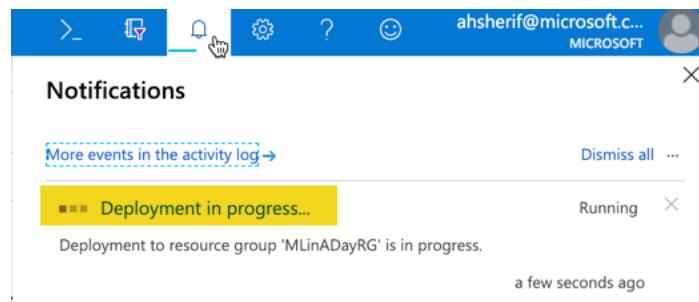
- Subscription: Azure Internal 101
- Resource group: MLinADayRG
- Location: East US
- Storage account name: blobmlinaday
- Deployment model: Resource manager
- Account kind: StorageV2 (general purpose v2)
- Replication: Locally-redundant storage (LRS)
- Performance: Standard
- Access tier (default): Cool

The 'ADVANCED' section shows:

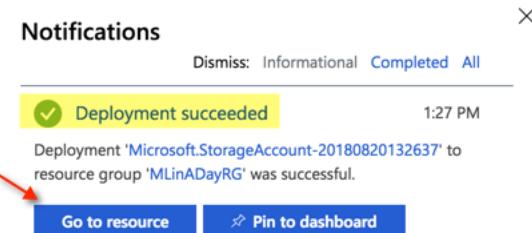
- Secure transfer required: Disabled
- Allow access from: All networks
- Hierarchical namespace: Disabled

At the bottom, there are 'Create', 'Previous', and 'Next' buttons, along with a link to 'Download a template for automation'.

12. We can then just click on **Create** and begin the deployment process of the blob storage.
13. While the deployment is underway, we can check the progress as seen in the following screenshot:



14. Once the Blob account has been successfully deployed, we can go to the storage account by clicking on **Go to Resource**, as seen in the following screenshot:



15. ****Important Step ***** Ensure that your blob account has the following setting configured to ensure that **Secure Transfer Required** is **Disabled** by selecting Configuration as seen in the following screenshot:

This screenshot shows the 'blobmlinaday - Configuration' page in the Azure portal. The left sidebar lists navigation options: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Storage Explorer (preview), Settings, Access keys, CORS, Configuration (which is selected and highlighted in blue), Encryption, Shared access signature, Firewalls and virtual networks, Advanced Threat Protection (pr...), Properties, and Locks. The main content area shows general account information: 'Storage account' and 'Save' and 'Discard' buttons. It notes that the cost depends on usage and options chosen. The 'Account kind' is set to 'Storage (general purpose v1)'. A warning message states: 'This account can be upgraded to a General Purpose v2 account with additional features. Upgrading is permanent and will result in billing changes.' with a 'Learn more' link and an 'Upgrade' button. Under 'Settings', the 'Secure transfer required' setting is shown as 'Disabled' (selected) and 'Enabled' (disabled). Other settings include 'Performance' (Standard selected), 'Replication' (Locally-redundant storage (LRS)), 'Azure Active Directory authentication for Azure Files (Preview)' (Enabled), and 'Data Lake Storage Gen2 (preview)' (Hierarchical namespace (Enabled)).

16. Back in the Overview section, select **Blobs** under the **Services** section as seen in the following screenshot:

blobmlinaday
Storage account

Search (Ctrl+ /) | Open in Explorer | Move | Delete | Refresh

Overview

- Resource group (change) MLinADayRG
- Status Primary: Available
- Location East US
- Subscription (change) Azure Internal 101
- Subscription ID ba3edd26-e2b1-4cc5-a19e-5bd21d7e9f5d
- Tags (change) Click here to add tags

Services

- Blobs** REST-based object storage for unstructured data Learn more

Access keys | CORS | Configuration | Encryption | Shared access signature

17. We should now be in the Containers section; however, if this is the first time, we are entering the Container, we should see a message that says **You don't have any containers yet. Click '+ Container' to get started.**
18. Go ahead and create a new Container called **Churn** and set **Public access level** to **Container (anonymous read access for containers and blobs)**, as seen in the following screenshot:

blobmlinaday - Blobs
Storage account

Search (Ctrl+ /) | + Container | Refresh | Delete

New container

* Name: churn

Public access level: Container (anonymous read access for containers and blobs)

OK | Cancel

19. The only thing that we will need right now is an **Access Keys** that will be used in future sections of this workshop. We can copy **key1** and paste it somewhere handy in a notepad editor as seen in the following screenshot:

The screenshot shows the 'Access keys' section of the Azure Storage Account settings. The storage account name is 'blobmlinaday'. It displays two access keys: 'key1' and 'key2'. An orange arrow points to the 'key1' key, which is highlighted with a yellow background. The 'key1' key value is '2Le22d38roB6lFqAeds/dFXRfc6VoKUDBXqcsjhzRv8r3YDaB1xI7FdmyzLcaH04NAFrL+OQmXsNlo4NzICg=='. The 'key2' key value is 'e3rNW07ji+s37Xy7/v84m8mCYSYDQSH+p9lxUmnfpgrwta02/YyaOEpp9JhYcReHuqZLv0s/qYGbm6CPxaE8yA=='. Both key values have copy icons to their right.

20. Once the container has been successfully created, we can move onto the next section of adding data to the container. We'll do it manually to keep this workshop simple.

Section 3: Upload CSV file to Blob Storage Container Manually

- Access your blob storage container once again by selecting **Storage Accounts** on the left-hand side of the Portal menu, selecting your Blob account, and finally selecting **Blobs** under **Blob Service** as seen in the following screenshot:

The screenshot shows the Microsoft Azure portal interface. On the left, the navigation menu is open, with 'Storage accounts' selected. In the main content area, the 'blobmlinaday' storage account is displayed. Under the 'Blob service' section, the 'Blobs' option is highlighted with a yellow box. The right pane provides detailed information about the storage account, including its name, location, and various configuration settings like encryption and replication.

- Select your **container** which should be called churn as seen in the following screenshot:

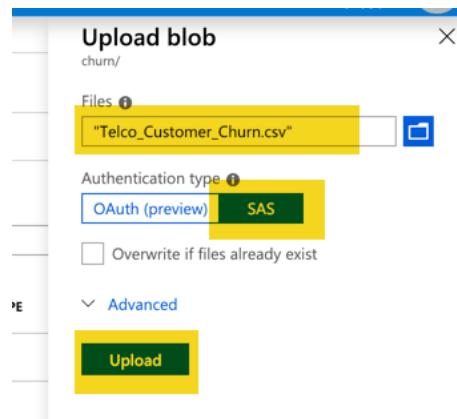
The screenshot shows the 'blobmlinaday - Blobs' blade. On the left, the 'Blobs' option under the 'Blob service' section is highlighted with a yellow box. The main area displays a table of containers. One row, containing the word 'churn', is highlighted with a yellow box. The table includes columns for NAME, LAST MODIFIED, PUBLIC ACCESS ..., and LEASE STATE.

NAME	LAST MODIFIED	PUBLIC ACCESS ...	LEASE STATE
churn	11/25/2018, 11:20:31 AM	Container	Available

- Once inside your container, select the upload icon to begin the uploading of your csv spreadsheet as seen in the following screenshot:

The screenshot shows the Azure Storage Accounts interface for the 'churn' container. On the left, there's a sidebar with options like Overview, Access Control (IAM), Settings, Access policy, Properties, Metadata, and Editor (preview). The main area has tabs for Upload, Refresh, Delete, Acquire lease, Break lease, View snapshots, and Create snapshot. A search bar at the top says 'Search (Ctrl+ /)'. Below it, a 'Location: churn' section has a search input 'Search blobs by prefix (case-sensitive)'. A table lists blobs with columns for NAME, MODIFIED, ACCESS TIER, and BLOB TYPE. The table shows 'No blobs found.'

4. Select your file from your local hard drive and select Upload as seen in the following screenshot. The Authentication type option no longer appears in the latest version:



5. You should now see your CSV file in your container as seen in the following screenshot:

The screenshot shows the 'churn' container page again. The table now lists the 'Telco_Customer_Churn.csv' file. The columns are NAME, MODIFIED, ACCESS TIER, BLOB TYPE, SIZE, and LEASE STATE. The file details are: NAME 'Telco_Customer_Churn.csv', MODIFIED '11/25/2018, 11:28:06 AM', ACCESS TIER 'Cool (Inferred)', BLOB TYPE 'Block blob', SIZE '954.59 KiB', and LEASE STATE 'Available'.

6. Your file should be ~ 955 kB. If you click on your CSV file and select **Edit Blob**, you can preview your actual data that is in the CSV file within the **Blob** container as seen in the following screenshot:

The screenshot shows the Azure Storage Blob service interface. On the left, there are navigation buttons for Upload, Refresh, More, Save, Discard, Refresh, Download, and Delete. Below these are sections for Location ('churn') and NAME, where 'Telco_Customer_Churn.csv' is listed. The main area is titled 'Overview' and displays the contents of the CSV file. The first few lines of the data are:

```

1 CustomerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,Online
2 7590-VHVEG,Female,0,Yes,No,1,No,No phone service,DSL,No,Yes,No,No,Month-to-month,Yes,Electronic che
3 5575-GNVDL,Male,0,No,No,34,Yes,No,DSL,Yes,No,Yes,No,No,One year,No,Mailed check,56.95,1889.5,No
4 3668-QPYBK,Male,0,No,No,2,Yes,No,DSL,Yes,NO,NO,NO,Month-to-month,YES,Mailed check,53.85,108.15,YES
5 7795-CFOCW,Male,0,No,No,45,NO,NO phone service,DSL,YES,NO,YES,YES,NO,NO,One year,NO,Bank transfer (automat
6 9237-HQITU,Female,0,No,No,2,Yes,No,Fiber optic,NO,NO,NO,NO,Month-to-month,YES,Electronic check,70.7
7 9305-CDSKC,Female,0,No,No,8,Yes,Fiber optic,NO,NO,YES,NO,YES,Month-to-month,YES,Electronic check,'
8 1452-KIOVK,Male,0,No,YES,22,YES,Fiber optic,NO,YES,NO,NO,NO,Month-to-month,YES,Credit card (automat
9 6713-OKOMC,Female,0,No,No,10,NO,NO phone service,DSL,YES,NO,NO,NO,NO,Month-to-month,NO,Mailed check,29
10 7892-POOKP,Female,0,YES,NO,28,YES,Fiber optic,NO,YES,NO,YES,YES,Month-to-month,YES,Electronic che
11 6388-TABGU,Male,0,No,YES,62,YES,NO,DSL,YES,YES,NO,NO,NO,One year,NO,Bank transfer (automatic),56.15,34
12 9763-GRSKD,Male,0,YES,YES,13,YES,NO,DSL,YES,NO,NO,NO,NO,Month-to-month,YES,Mailed check,49.95,587.45,N
13 7469-LKBCI,Male,0,No,NO,16,YES,NO,NO,NO,internet service,NO,internet service,NO,internet service,NO,int
14 8091-TTVAX,Male,0,YES,NO,58,YES,Fiber optic,NO,NO,YES,NO,YES,One year,NO,Credit card (automatic),
15 0280-XJGEX,Male,0,No,NO,49,YES,Fiber optic,NO,YES,NO,YES,YES,Month-to-month,YES,Bank transfer (au
16 5129-JLPIS,Male,0,No,NO,25,YES,Fiber optic,YES,NO,YES,YES,YES,Month-to-month,YES,Electronic check,
17 3655-SNOYZ,Female,0,YES,YES,69,YES,Fiber optic,YES,YES,YES,YES,YES,Two year,NO,Credit card (auto
18 8191-KWSZG,Female,0,NO,NO,52,YES,NO,NO,NO,internet service,NO,internet service,NO,internet service,NO,int
19 9959-WOFGT,Male,0,No,YES,71,YES,Fiber optic,YES,NO,YES,YES,Two year,NO,Bank transfer (automatic),
20 4190-MFLWU,Female,0,YES,YES,10,YES,NO,DSL,NO,NO,YES,NO,Month-to-month,NO,Credit card (automatic),5
21 4183-MYFRB,Female,0,NO,NO,21,YES,NO,Fiber optic,NO,YES,NO,NO,YES,Month-to-month,YES,Electronic check,
22 8779-QRMV,Male,1,NO,NO,1,NO,NO phone service,DSL,NO,NO,NO,NO,Month-to-month,YES,Electronic check
23 1680-VDCCW,Male,0,YES,NO,12,YES,NO,NO,NO,internet service,NO,internet service,NO,internet service,NO,int
24 1066-JKSGK,Male,0,NO,NO,1,YES,NO,NO,NO,internet service,NO,internet service,NO,internet service,NO,int
25 3638-WEABW,Female,0,YES,NO,58,YES,YES,DSL,NO,YES,NO,NO,Two year,YES,Credit card (automatic),59.9,3
26 6322-HRPFA,Male,0,YES,YES,49,YES,NO,DSL,YES,YES,NO,YES,NO,NO,Month-to-month,NO,Credit card (automatic),59
27 6865-JZNKO,Female,0,NO,NO,30,YES,NO,DSL,YES,YES,NO,NO,NO,Month-to-month,YES,Bank transfer (automatic),

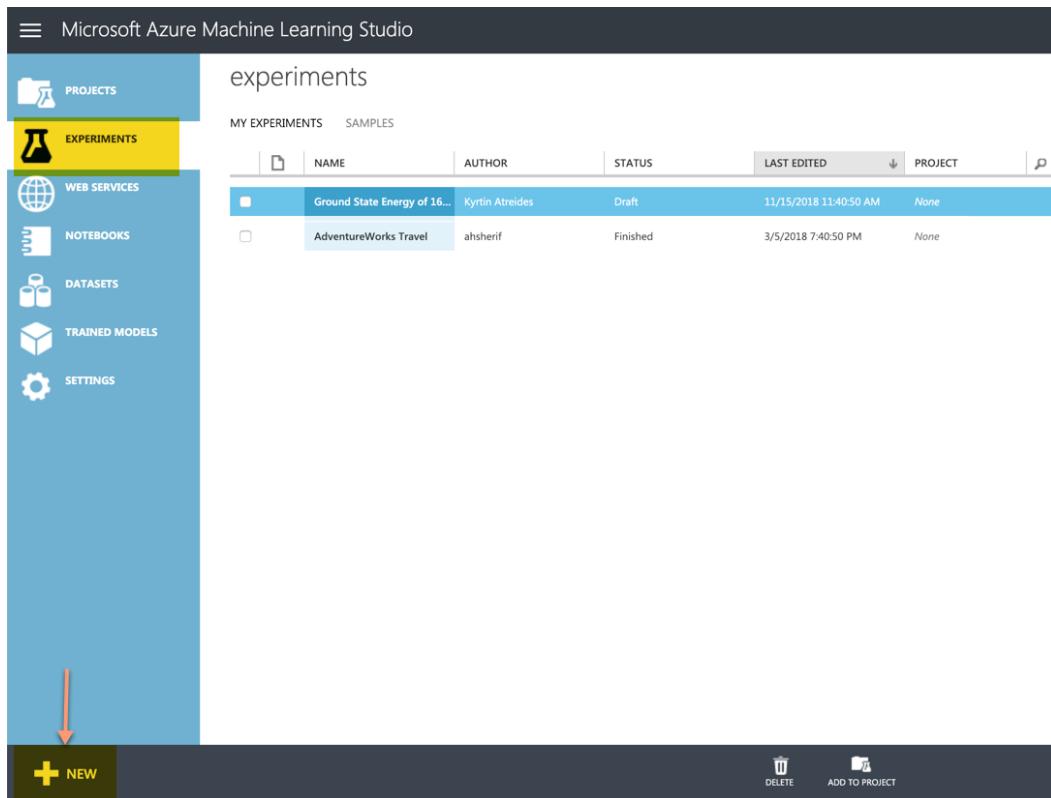
```

At the bottom right, there are buttons for Edit, Csv, and Download.

7. We have now confirmed our file is in our Blob account and we can continue onto Section 4.

Section 4: Build a clustering unsupervised model in Azure Machine Learning Studio

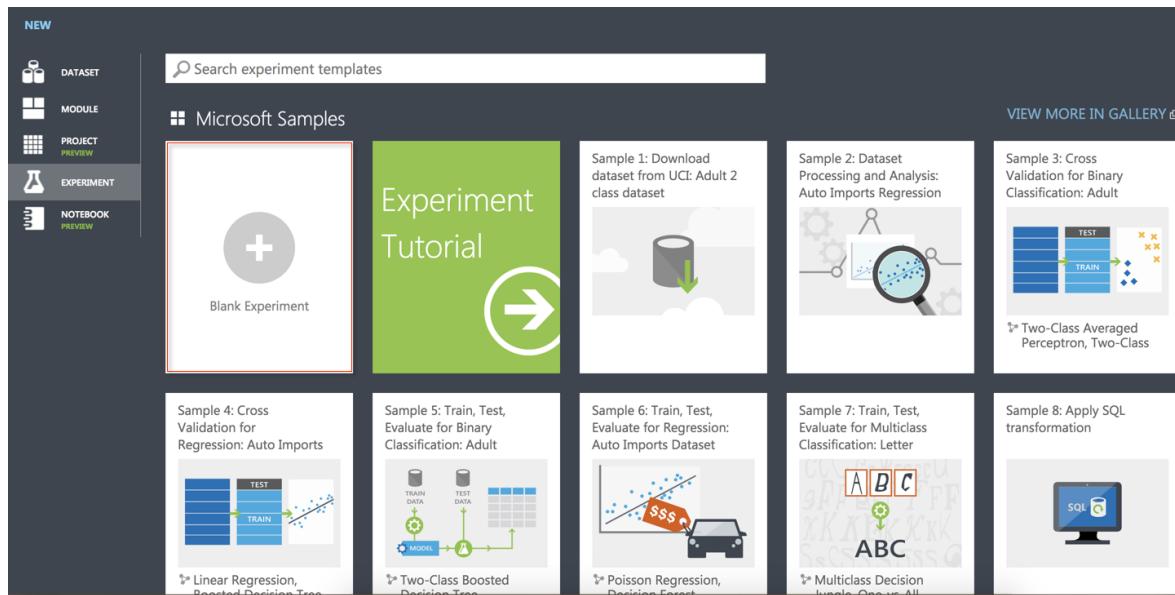
1. Clustering is the process of grouping similar entities together. The goal of this unsupervised machine learning technique is to find similarities in the fields that we have about our customers and group them into 10 distinct clusters.
2. Sign into Azure Machine Learning Studio using your Azure login Credentials in the following website:
<https://studio.azureml.net/>
3. If you are not already directed to create a new experiment, go ahead and select the Experiments tab on the menu on the left-hand side as seen in the following screenshot. All current and future experiments will reside in this section.



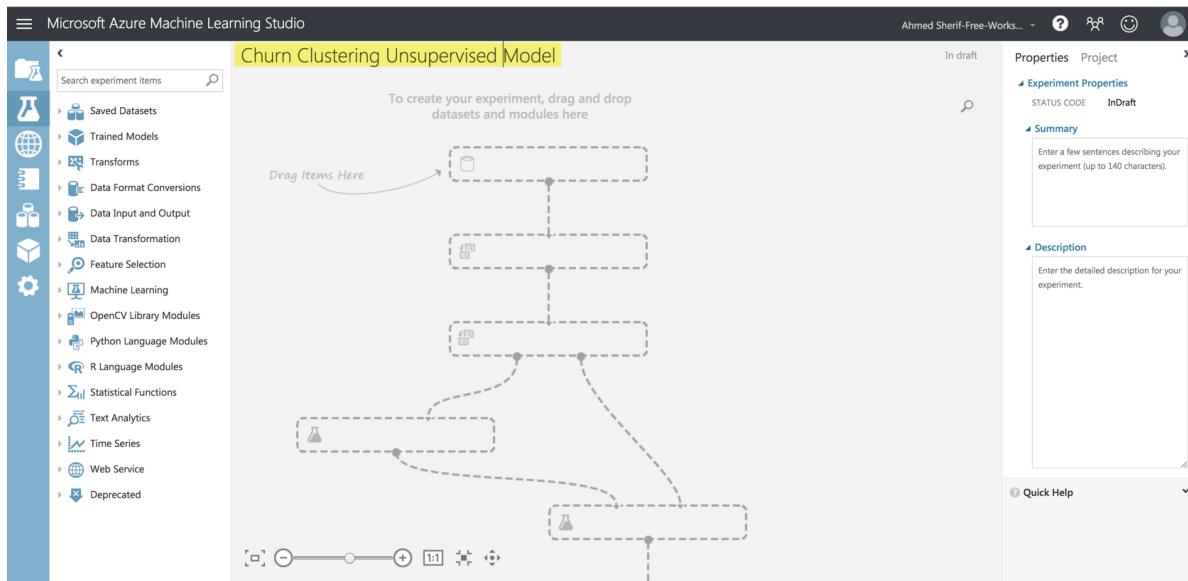
The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, there is a vertical sidebar with icons for Projects, Experiments (which is highlighted in yellow), Web Services, Notebooks, Datasets, Trained Models, and Settings. The main area is titled "experiments" and contains a table titled "MY EXPERIMENTS". The table has columns for NAME, AUTHOR, STATUS, LAST EDITED, and PROJECT. It lists two experiments: "Ground State Energy of 16..." by Kyrrin Atreides (Draft, 11/15/2018 11:40:50 AM, None) and "AdventureWorks Travel" by ahsherif (Finished, 3/5/2018 7:40:50 PM, None). At the bottom of the sidebar, there is a green button with a white plus sign and the word "NEW", with a red arrow pointing to it. Below the table, there are "DELETE" and "ADD TO PROJECT" buttons.

	NAME	AUTHOR	STATUS	LAST EDITED	PROJECT
<input type="checkbox"/>	Ground State Energy of 16...	Kyrrin Atreides	Draft	11/15/2018 11:40:50 AM	None
<input type="checkbox"/>	AdventureWorks Travel	ahsherif	Finished	3/5/2018 7:40:50 PM	None

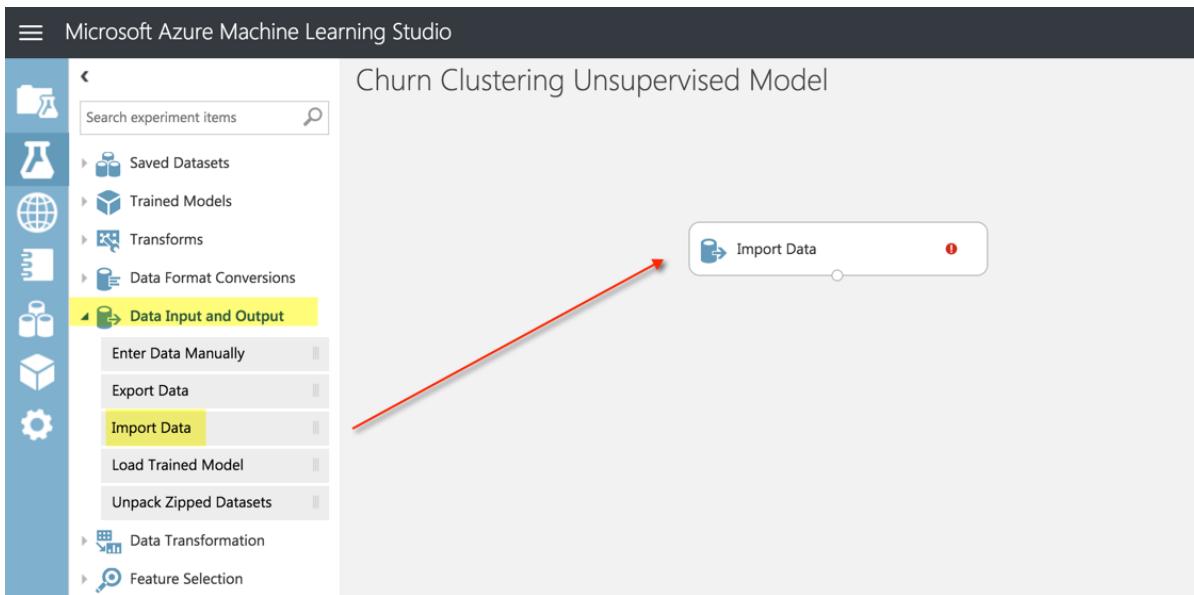
4. Next click on +NEW on the button of the screen and select **Blank Experiment** as seen in the following screenshot:



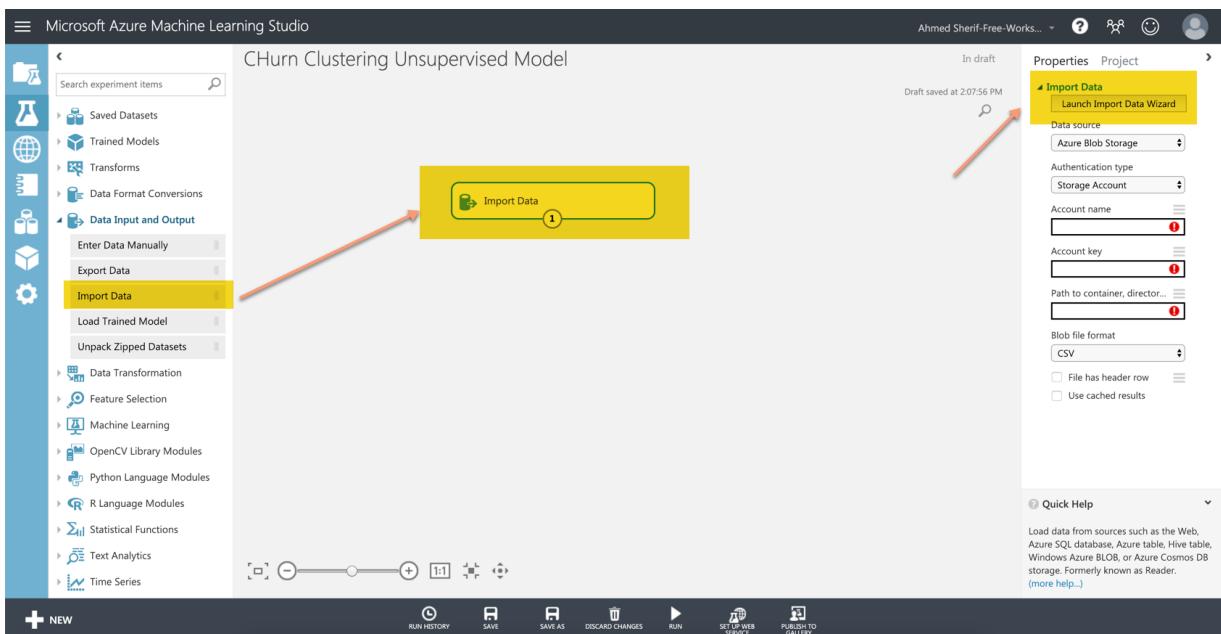
- Click on the title of the experiment and change the name to **Churn Clustering Unsupervised Model** as seen in the following screenshot:



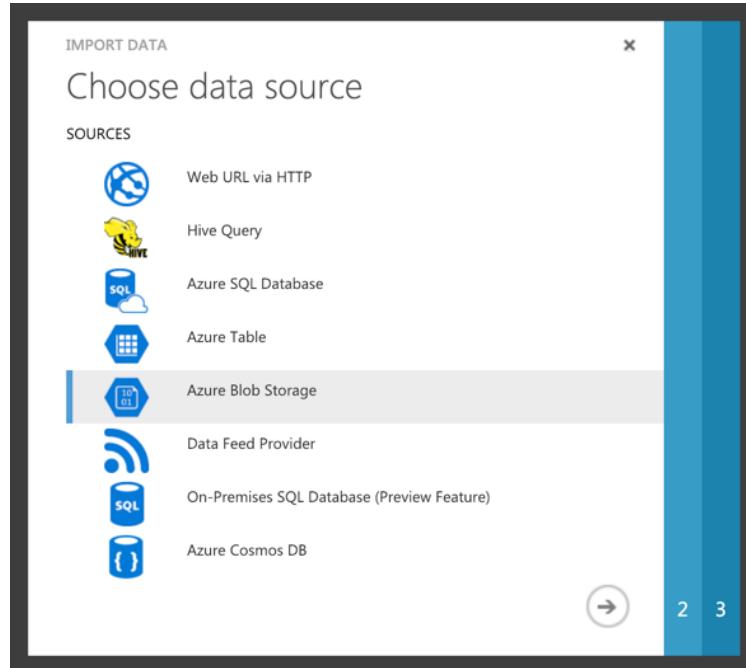
- Expand the **Data Input and Output** category and drag the **Import Data** into the Workflow as seen in the following screenshot:



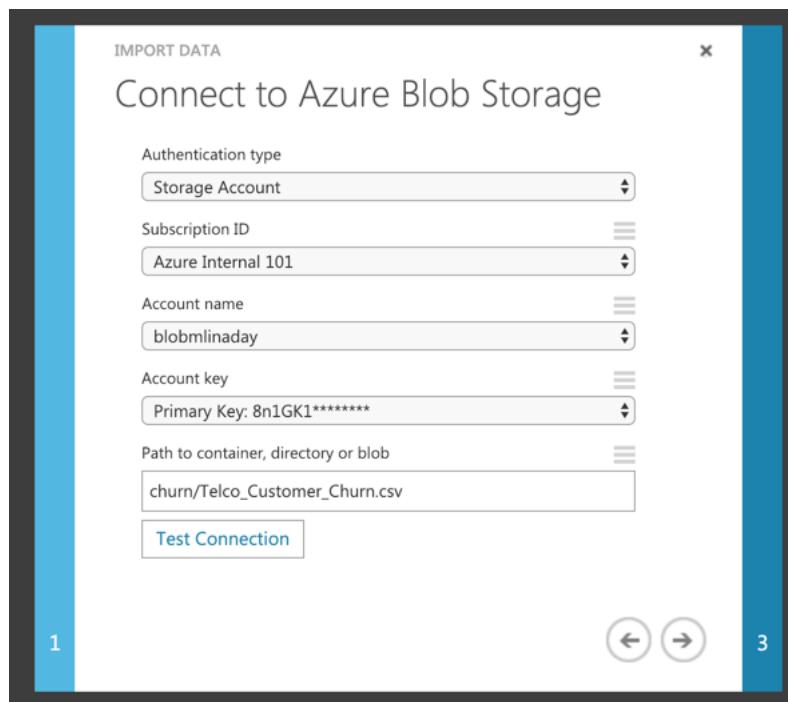
- As you click on the **Import Data** icon in the workflow, you will have the option to configure your data to pull from the Azure Blob Container you created a few sections ago using the Launch Import Wizard as seen in the following screenshot:



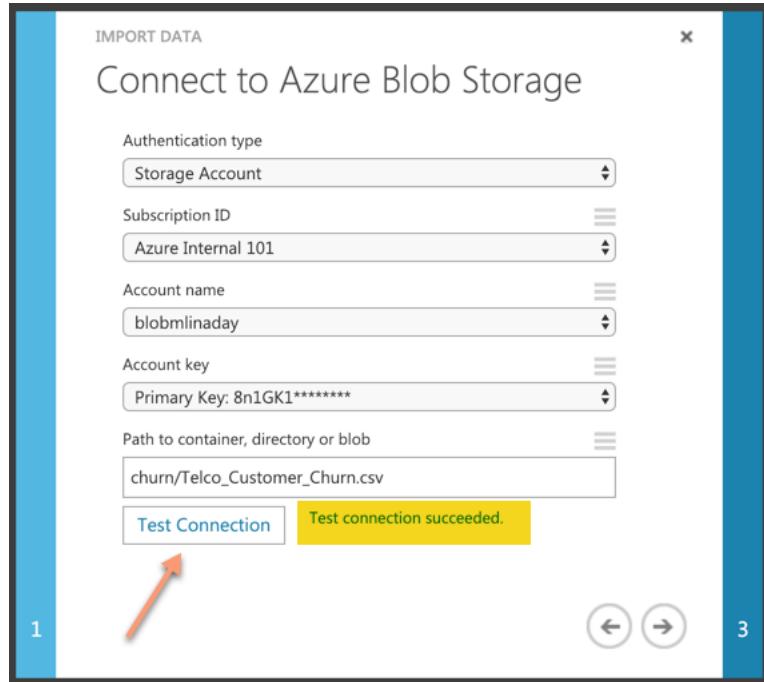
- Select Azure Blob Storage



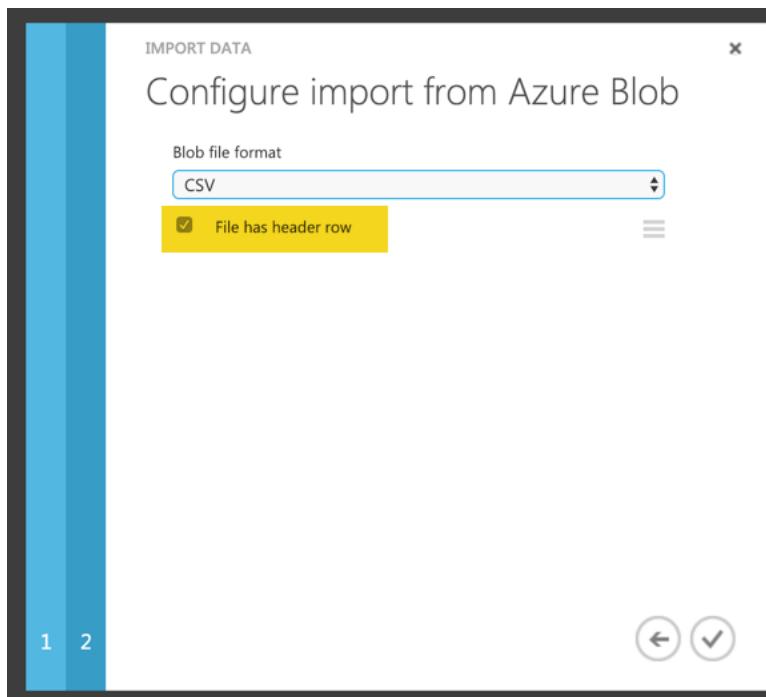
9. Specify the following credentials from the drop-down boxes for your accounts:



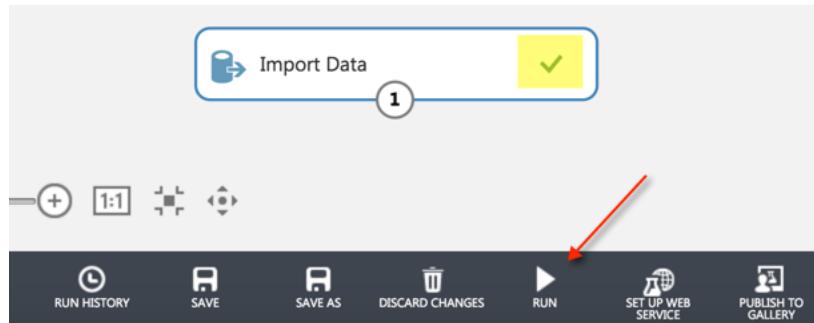
10. Test the connection to make sure the credentials work. Once it is validated you should see a green checkmark as seen in the following screenshot:



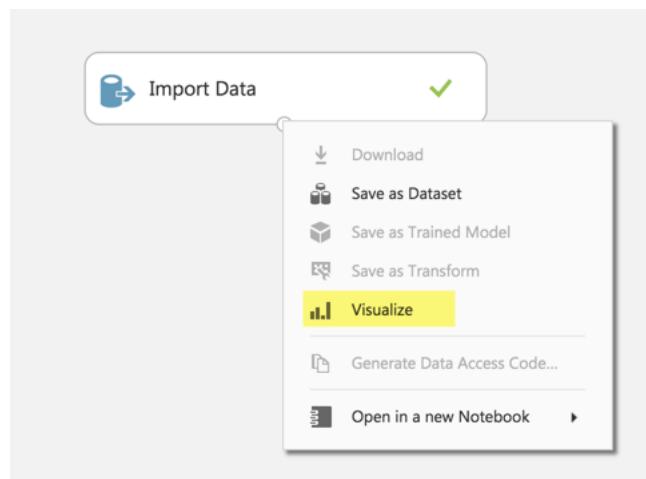
11. Check to include a **file header row** for the data and maintain a **CSV** file structure as seen in the following screenshot:



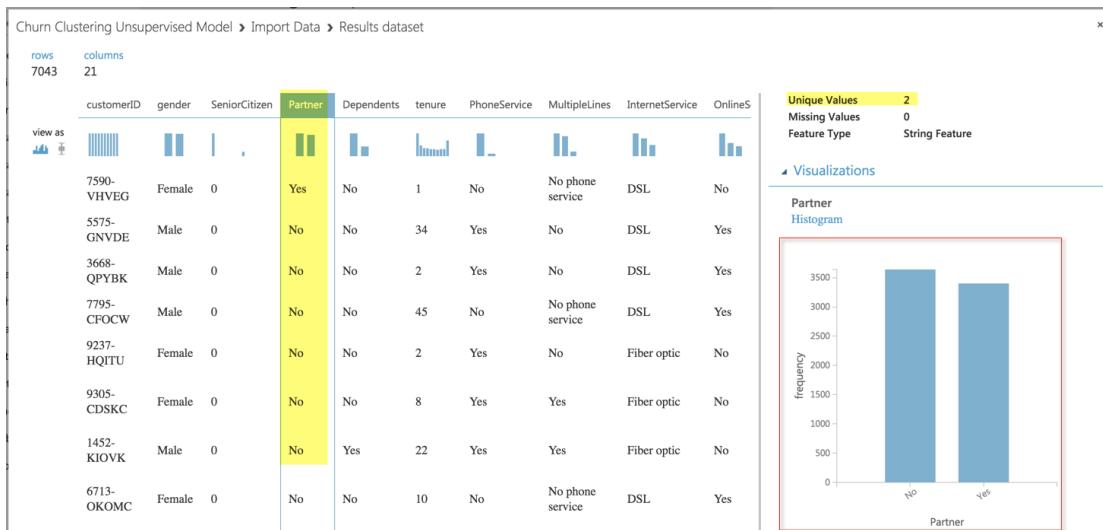
12. Back in the Azure ML Studio workflow, the credentials can be confirmed by clicking on the **RUN** icon at the bottom of the screen and seeing the green check mark next to **Import Data**.



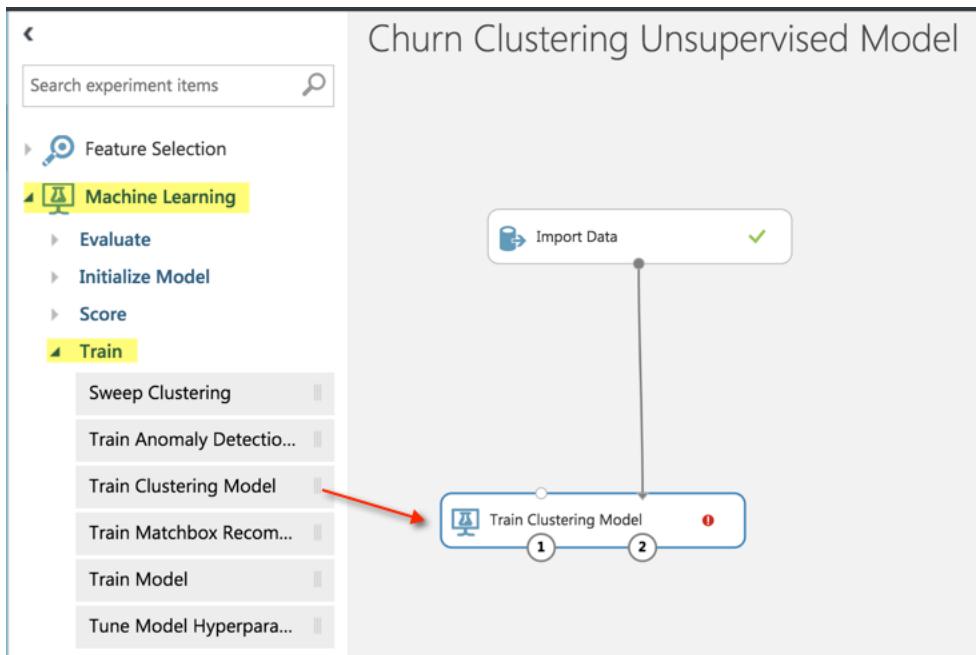
13. ***Every time a new node is added and attached to an existing node, you will need to select the RUN icon to see the new results***
14. Once we have confirmed the connection is working, we can right-click on the number 1 on the first workflow and **Visualize** the dataset as seen in the following screenshot:



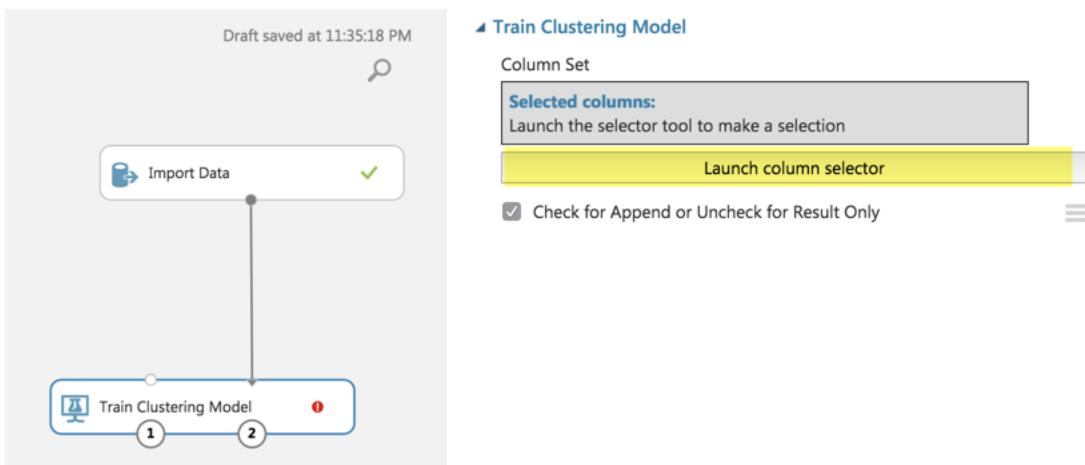
15. Take some time to click on the columns to identify the summary statistics for both the numeric fields and the categorical fields as seen in the following screenshot:



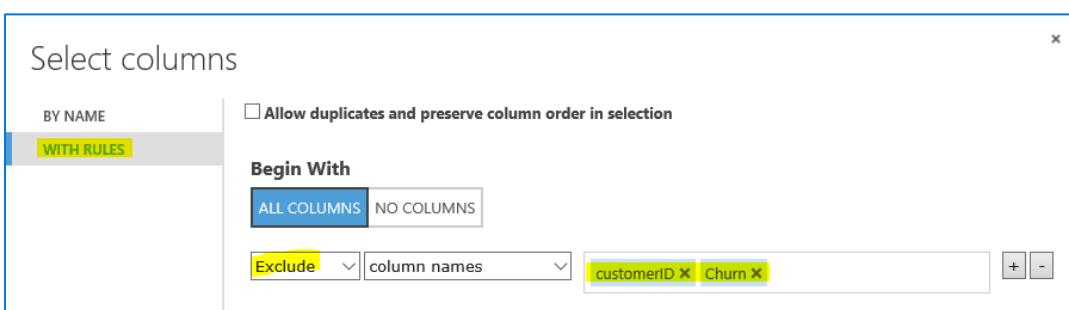
16. Import **Train Clustering Model** from the **Machine Learning** Node and connect it to the **Import Data** module as seen in the following screenshot:



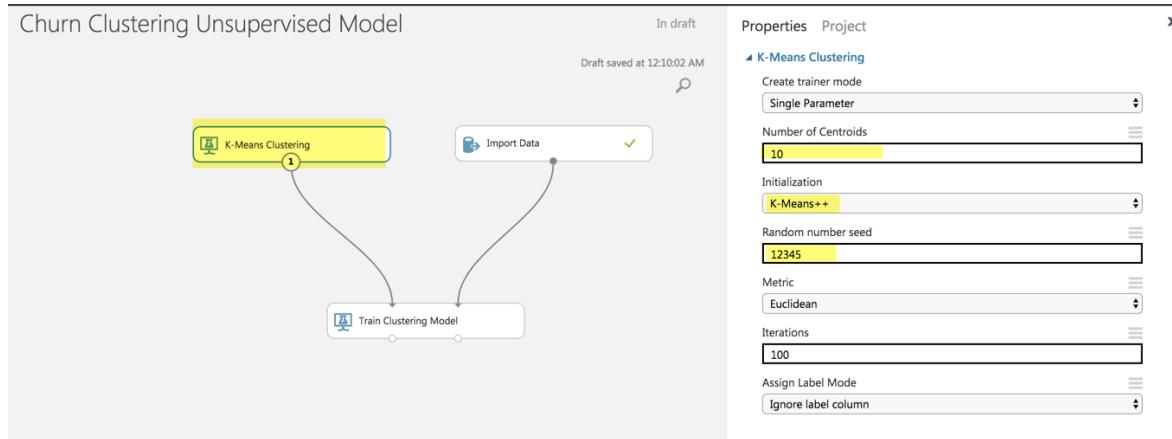
17. Next, click on the **Train Clustering Model** module, **Launch Column Selector**, connect the node back to the **Import Data** node, and configure the columns to be selected as seen in the following screenshot:



18. **CustomerID** is not a column that will help us out in clustering customers into 1 of 10 categories; therefore, we will remove it from the cluster algorithm. **Churn** is also not helpful, because customers who have churned are no longer customers.
 19. So, you will need to configure the **clustering model** and **Select All columns** from the dataset except for the column called **customerID** and **Churn** as seen in the following screenshot:

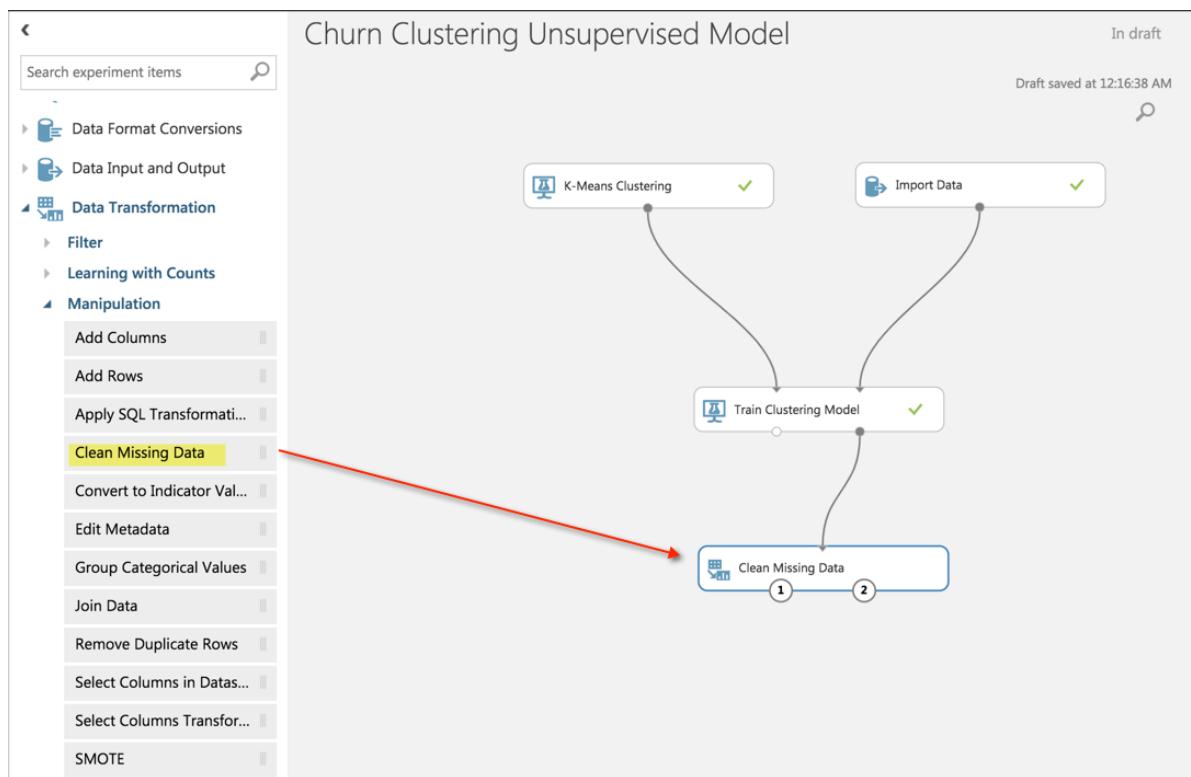


20. Next, we will add a **K-Means Clustering** model from the Machine Learning → Initialize Model → Clustering section. Attach it to the **Train Clustering Model** and assign the following parameters to the clustering model, as seen in the following screenshot:

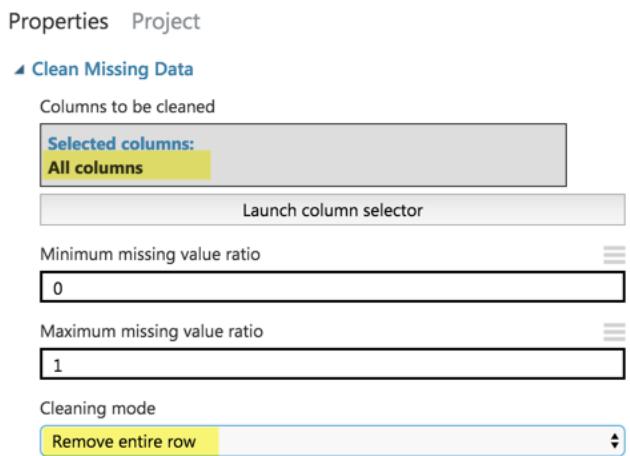


21. Select **RUN**

22. Next, we will do some data cleanup and remove any rows that have missing values by pulling in the **Clean Missing Data** module and connecting it to the **Train Clustering Model** as seen in the following screenshot:



23. We can configure the **Clean Missing Data** module by clicking on it and selecting the following configuration to remove any rows if they meet the following criteria:

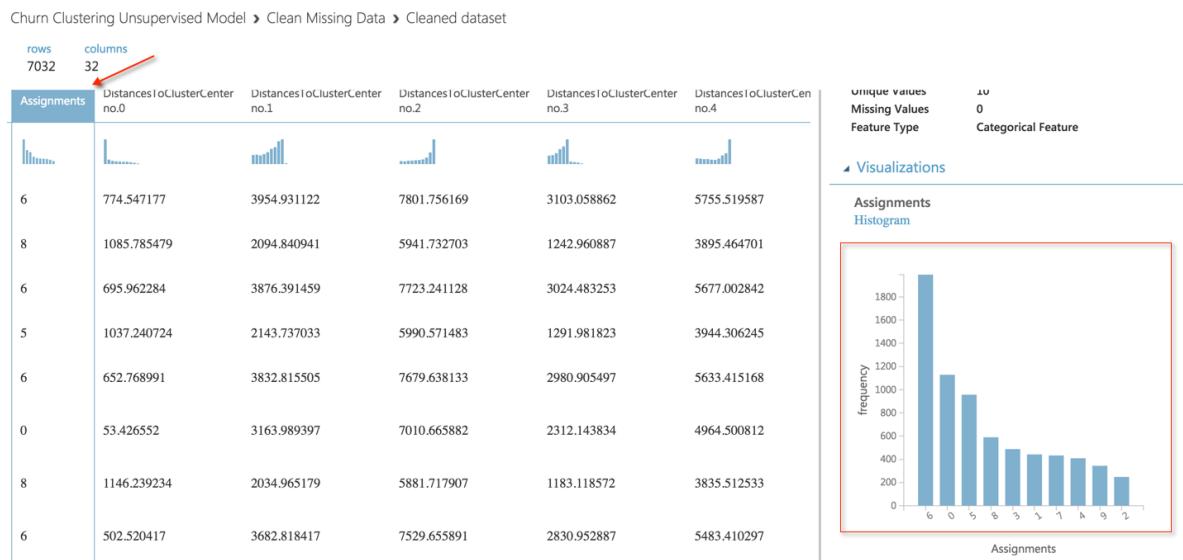


24. Select RUN

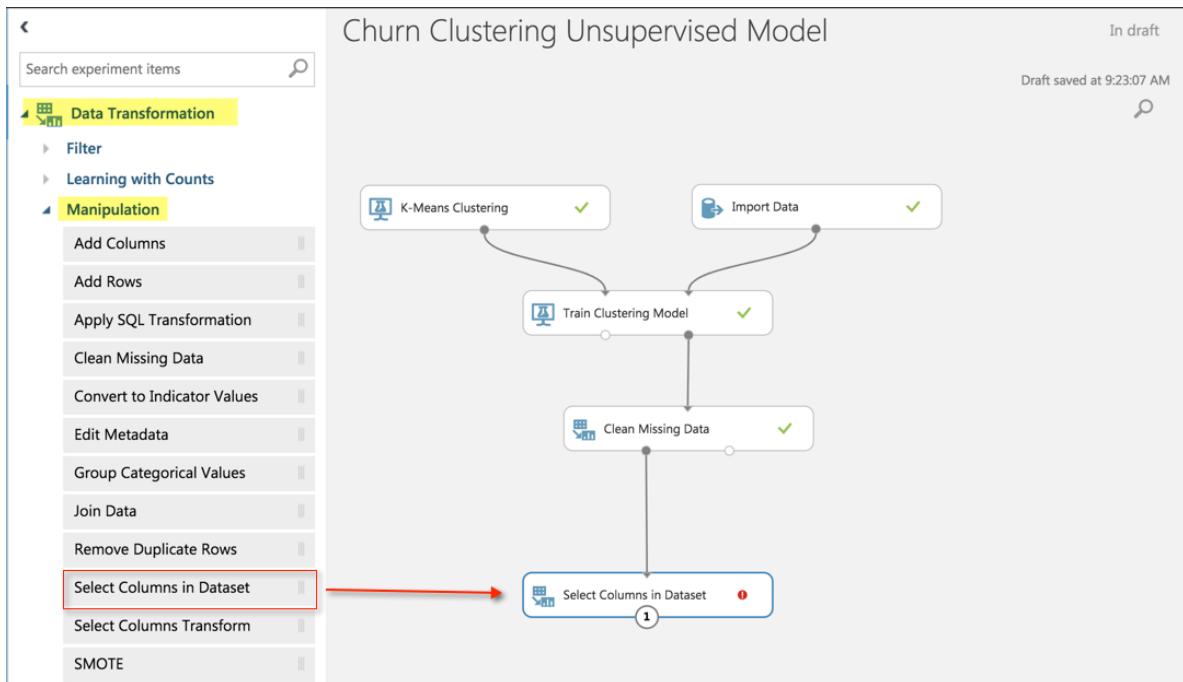
25. The clustering outcome produces several columns that are not necessary for our purposes for this workshop as can be seen by visualizing the **Clean Missing Data** module as seen in the following screenshot:

Churn Clustering Unsupervised Model > Clean Missing Data > Cleaned dataset							
rows	columns						
Churn	Assignments	DistancesToClusterCenter no.0	DistancesToClusterCenter no.1	DistancesToClusterCenter no.2	DistancesToClusterCenter no.3	DistancesToClusterCenter no.4	DistancesToC no.4
No	6	774.547177	3954.931122	7801.756169	3103.058862	5755.519587	
No	8	1085.785479	2094.840941	5941.732703	1242.960887	3895.464701	
Yes	6	695.962284	3876.391459	7723.241128	3024.483253	5677.002842	
No	5	1037.240724	2143.737033	5990.571483	1291.981823	3944.306245	
Yes	6	652.768991	3832.815505	7679.638133	2980.905497	5633.415168	
Yes	0	53.426552	3163.989397	7010.665882	2312.143834	4964.500812	
No	8	1146.239234	2034.965179	5881.717907	1183.118572	3835.512532	
No	6	502.520417	3682.818417	7529.655891	2830.952887	5483.410297	

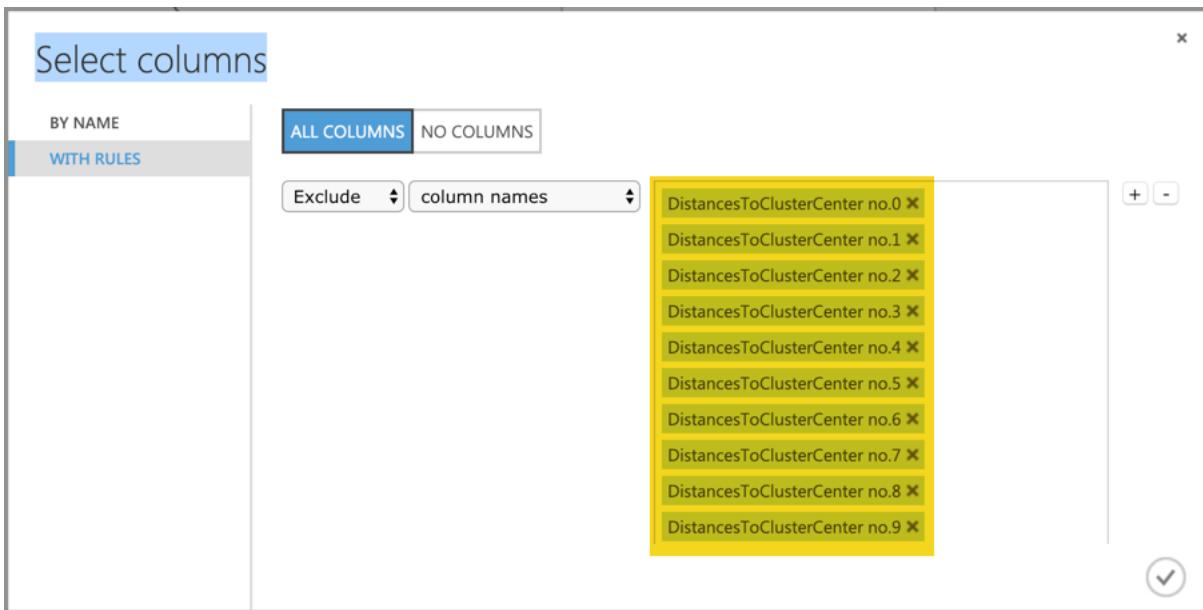
26. We are truly only interested in the assignments, which is grouping each customer into a category of **0** through **9** as seen when we visualize a histogram of the unique values as seen in the following screenshot:



27. All other columns that start with **DistancesToClusterCenter** are not relevant for our purposes and can be removed. The easiest way to do this is to drag into our workflow **Select Columns in Dataset** as seen in the following screenshot:

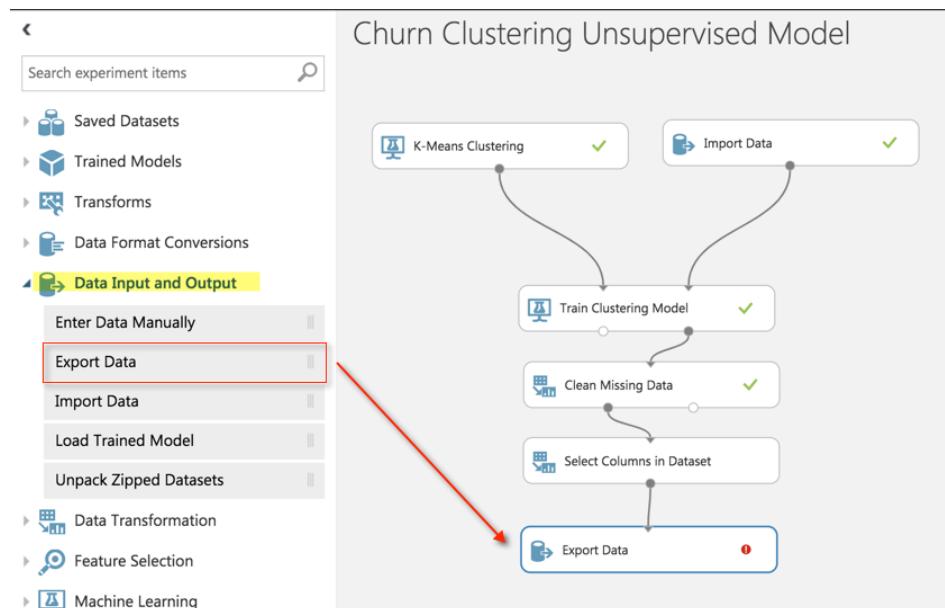


28. We can then configure the module for **Select Columns in Dataset** by **Launching Column Selector**, select **WITH RULES**, and exclude column names DistancetoClusterCenter no.0 through DistancetoClusterCenter no.9, as seen in the following screenshot:



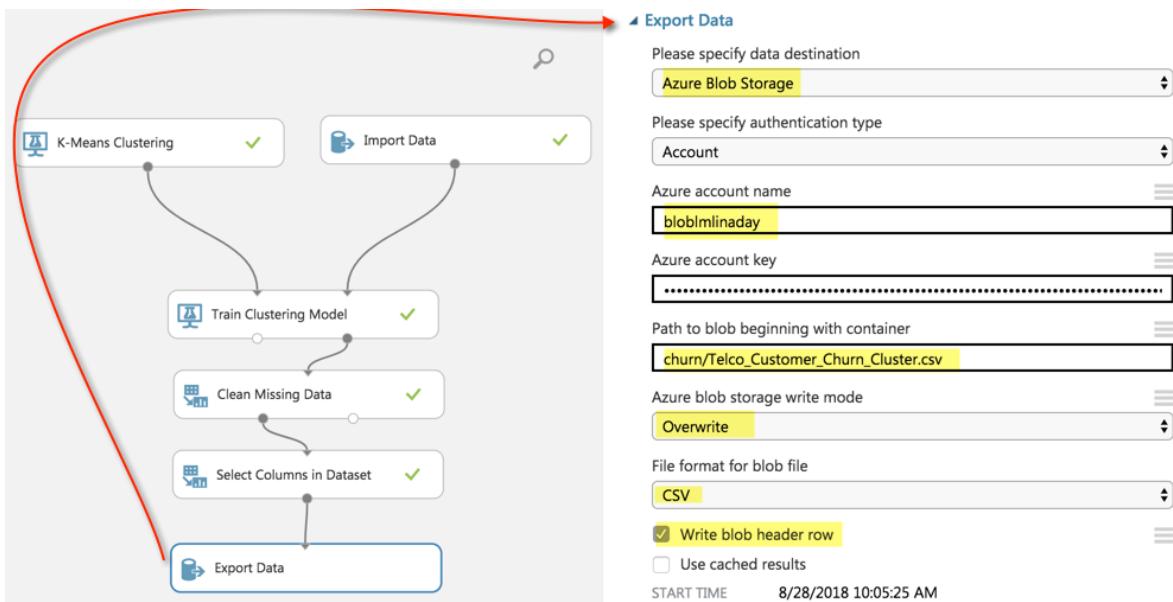
29. Select **RUN**

30. We are finally ready to send out revised dataset with only the columns that we need back to blob storage using the **Export Data** module as seen in the following screenshot:

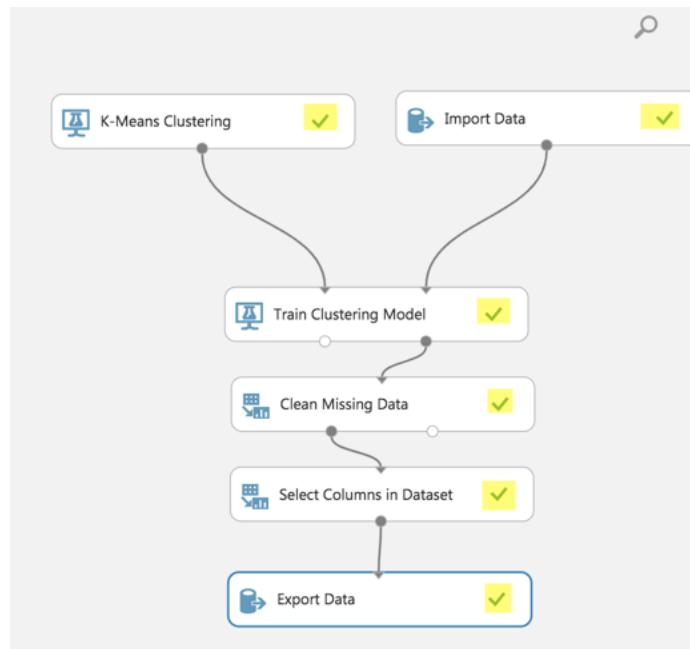


31. Unlike the Import Data module, we do not have a wizard to help us out. So, we will have to input our credentials manually.

32. We will need to configure the connection to export the new dataset, **Telco_Customer_Churn_Cluster.csv**, back to the same blob container, **churn**, using the following configuration as seen in the following screenshot (*please note that you should continue to use the same Key1 from Blob storage that was saved to a notepad from an earlier section*):



33. Once we are done, we can execute our workflow one last time by clicking on the **RUN** button.
34. Once completed, we should hopefully see all green checkmarks next to each step of our workflow as seen in the following screenshot:



35. We can revisit our Blob container in the Azure Portal and view our newly added dataset alongside our existing dataset as seen in the following screenshot:

Home > Resource groups > MLinADayRG > blobmlinaday - Blobs > churn

churn
Container

Search (Ctrl+/
Upload Refresh Delete Acquire lease Break lease View snapshots Create snapshot

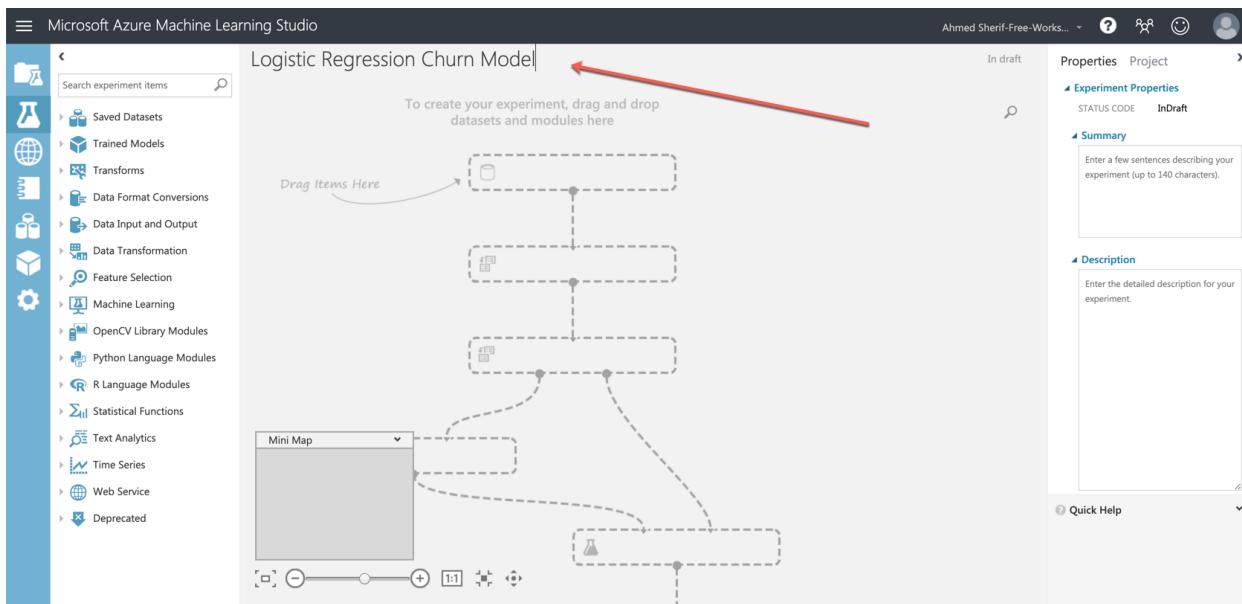
Location: churn

Search blobs by prefix (case-sensitive)

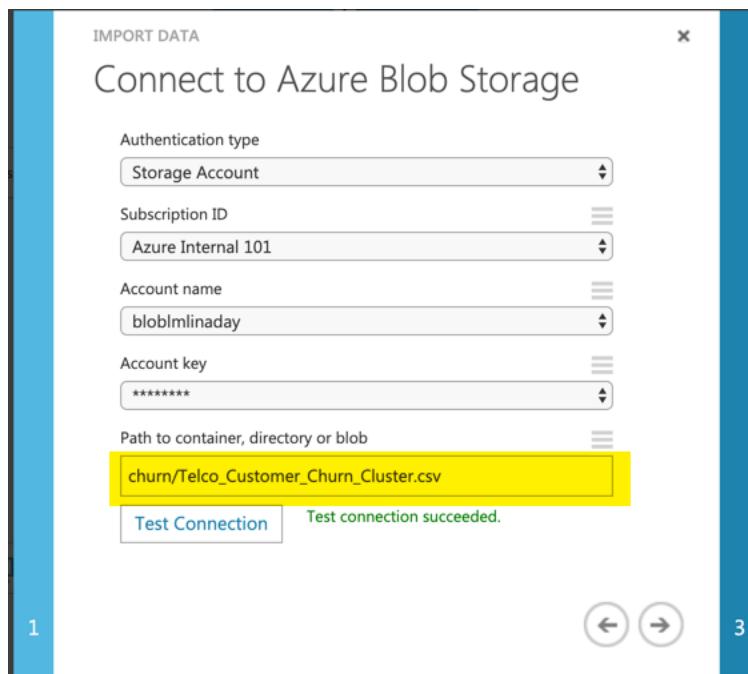
NAME	MODIFIED	BLOB TYPE
Telco_Customer_Churn_Cluster.csv		Block blob
Telco_Customer_Churn.csv		Block blob

Section 5: Create a logistic regression model in Azure ML Studio

- Just as we did in the previous section, we will go ahead and create a new **Blank Experiment** in Azure ML Studio that is named Logistic Regression Churn Model as seen in the following screenshot:



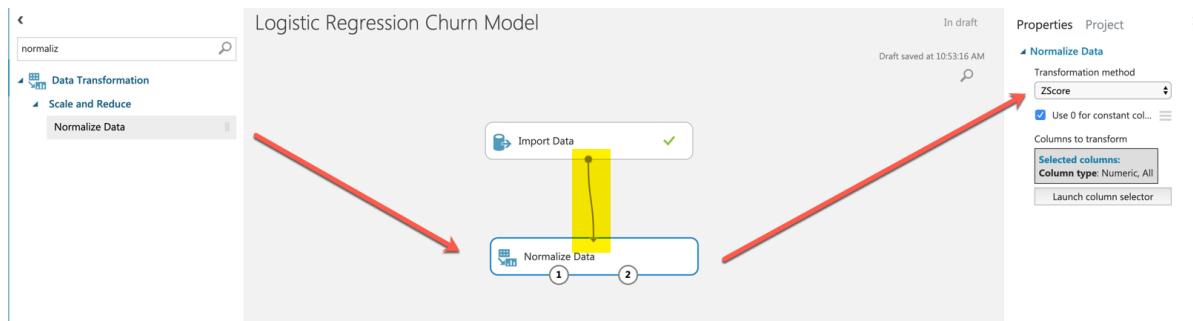
- Once again, launch the import wizard but this time but this import the **Telco_Customer_Churn_Cluster.csv** file as seen in the following screenshot:



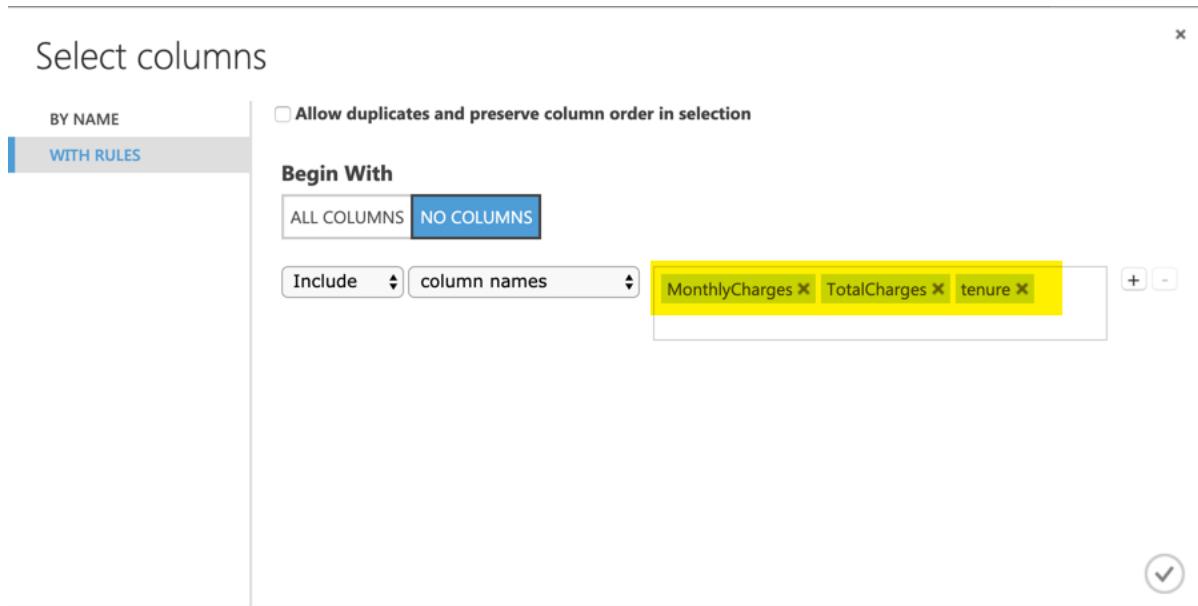
- Ensure the File has header row and then run the experiment to confirm that the data has exported as seen in the following screenshot:



4. We must do some modifications to our numeric fields to normalize them. We only have three numeric fields: **Tenure**, **TotalCharges** and **MonthlyCharges**.
5. Drag the **Normalize Data** node onto the workflow and select **ZScore** transformation to normalize data as seen in the following screenshot:



6. Launch the column selector and select Begin with NO COLUMNS and Include only **Tenure**, **MonthlyCharges** and **TotalCharges** as seen in the following screenshot:

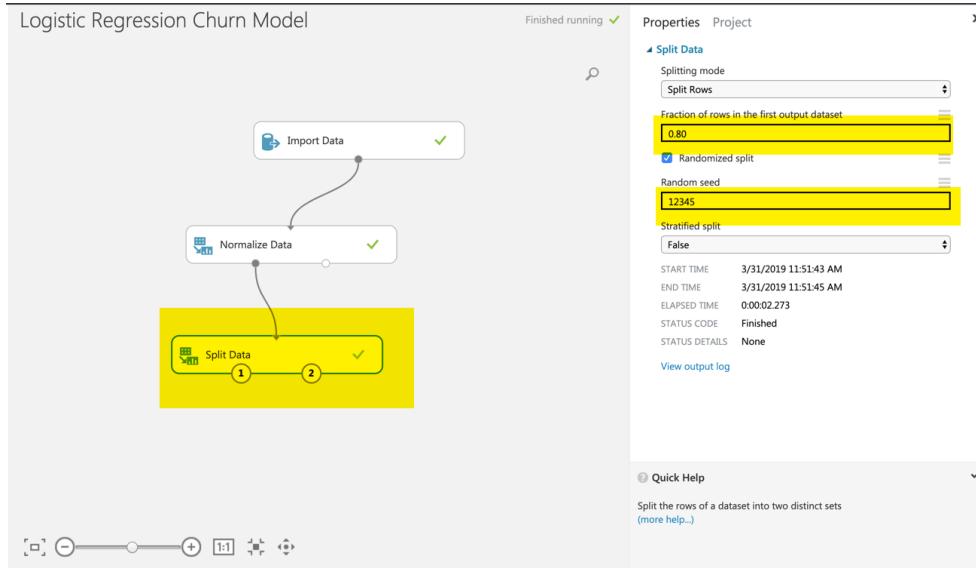


7. Select the check mark and **RUN** the experiment
8. Once the experiment has finished, the values for **Tenure**, **Monthly Charges** and **Total Charges** have now been normalized as seen in the following screenshot:

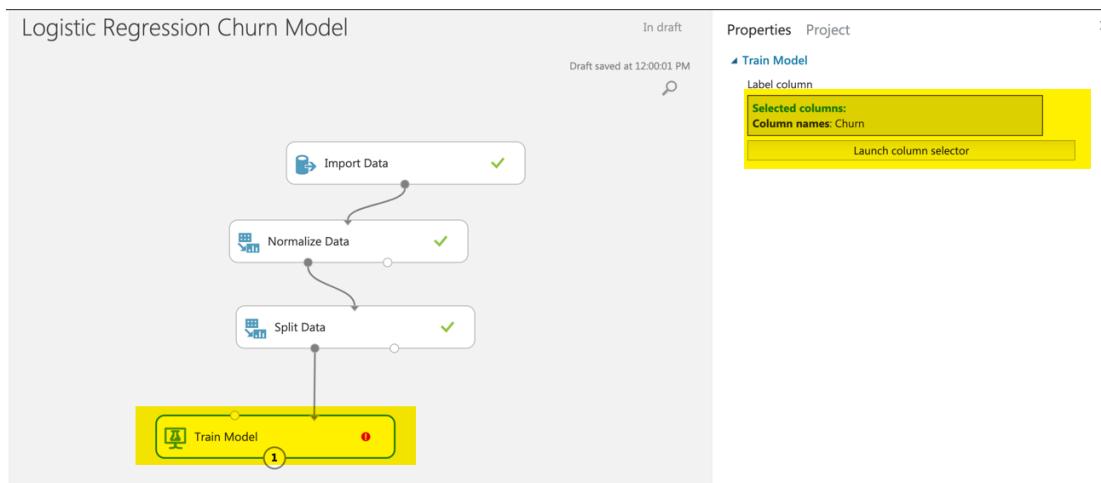
Logistic Regression Churn Model ➤ Normalize Data ➤ Transformed dataset

rows	columns								
7043	21								
Churn	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	No	No	Month-to-month	Yes	Electronic check	-1.160323	-0.994194	No
No	No	No	No	One year	No	Mailed check	-0.259629	-0.17374	No
No	No	No	No	Month-to-month	Yes	Mailed check	-0.36266	-0.959649	Yes
Yes	No	No	No	One year	No	Bank transfer (automatic)	-0.746535	-0.195248	No

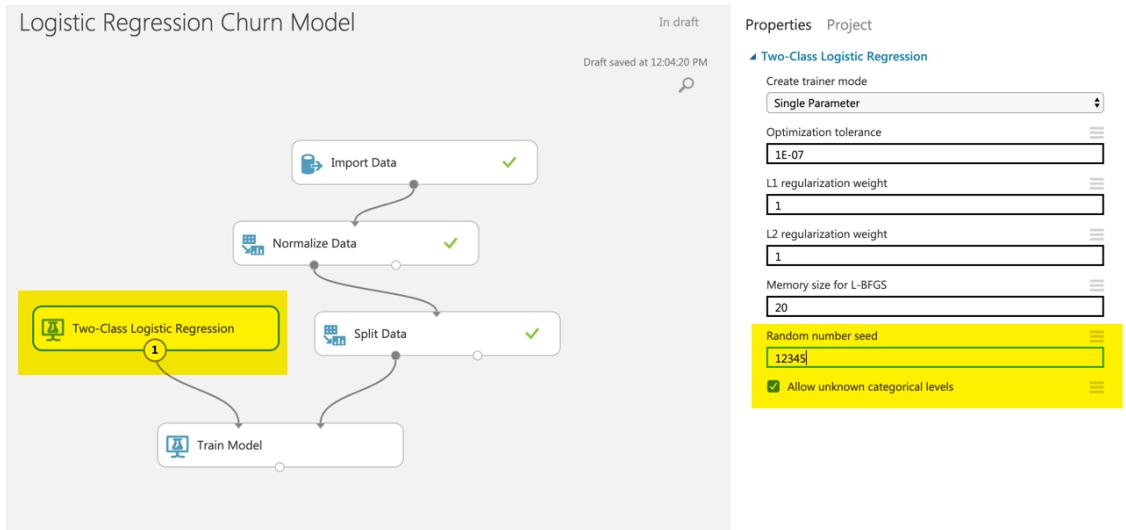
9. **What is a Z-Score.** Basically, a Z-Score is a way to compare results to a normal distribution or normal population. It is a way to measure how many standard deviations something is away from the mean, positively or negatively.
10. We are now ready to split our data into a training and testing split by pulling the **split data** node onto the workspace with the following parameters as seen in the following screenshot:



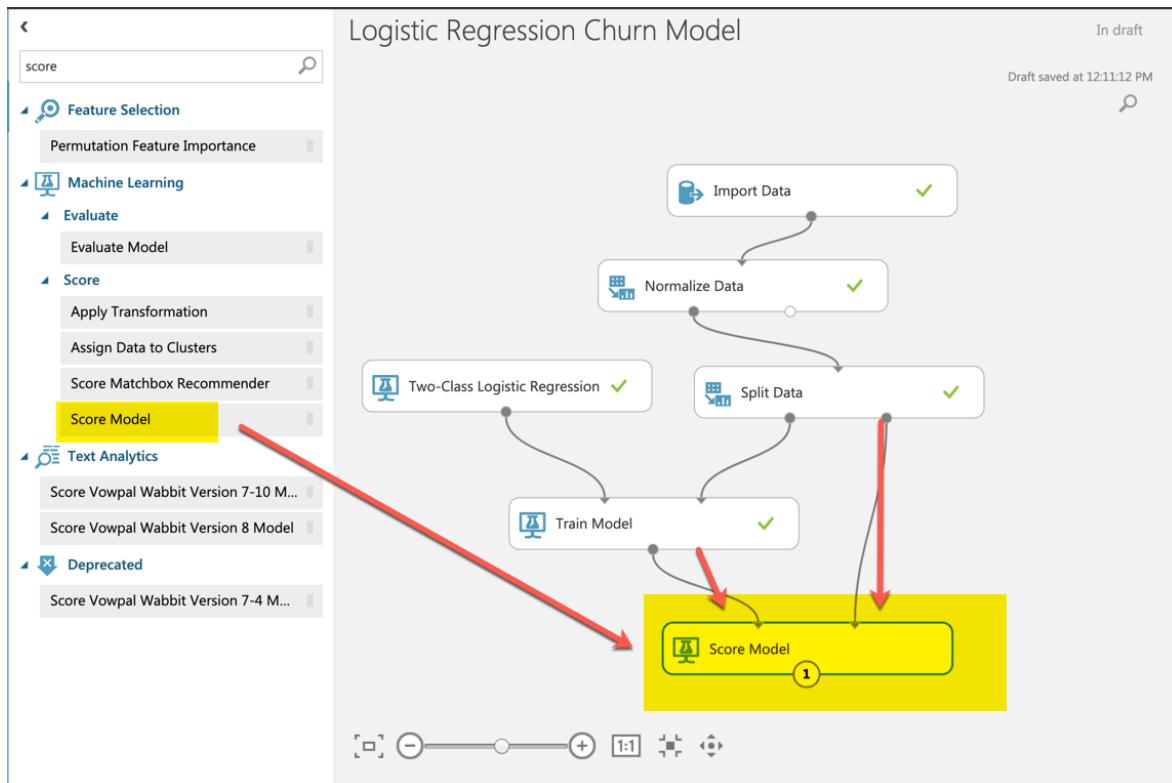
11. Run the experiment
12. Drag a **Train Model** node and connect it to the canvas and select **Churn** as the selected column as seen in the following screenshot:



13. Connect the **two-class logistic regression** node to the top of the **Train Model** node and specify a seed as seen in the following screenshot:



14. We now want to score our model against the test data set that we split earlier on in the section as seen in the following screenshot:

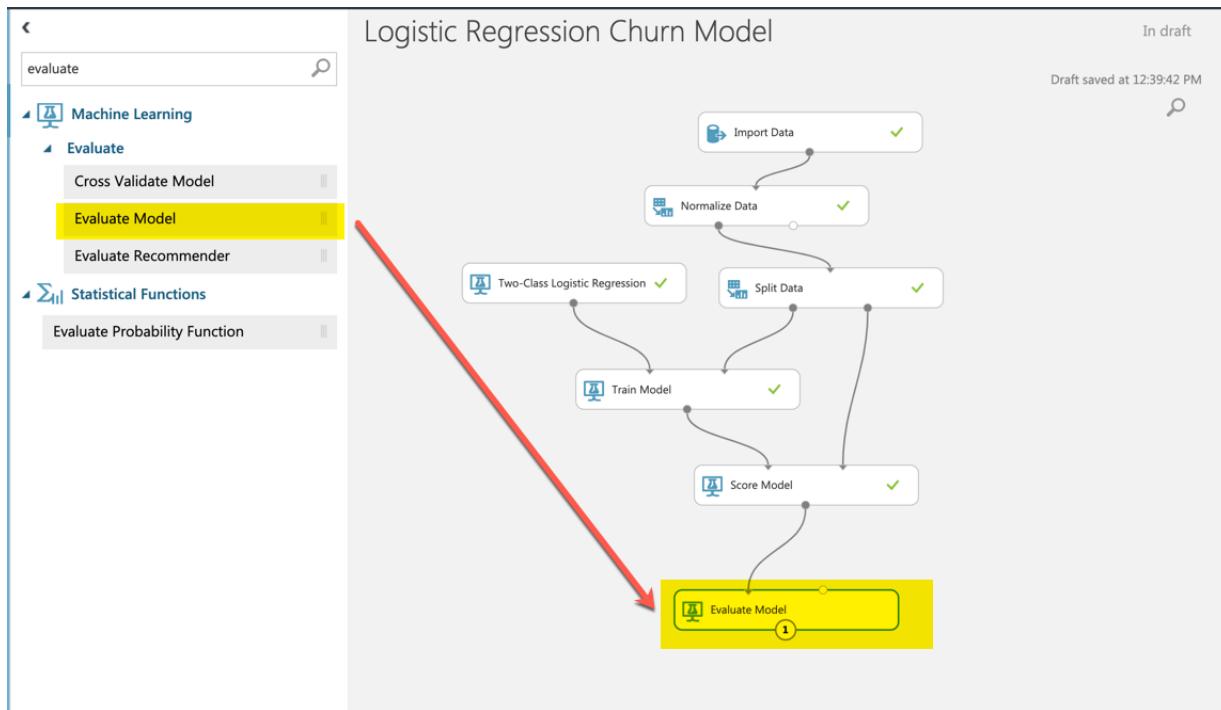


15. Run the experiment

16. If we view the scored model results, we can see how we did compare to what the actual prediction outcome was as seen in the following screenshot:

Churn	Scored Labels	Scored Probabilities
Yes	Yes	0.714217
No	No	0.330289
No	No	0.492767
No	No	0.453898
No	No	0.030879
No	No	0.270078
Yes	No	0.184071
Yes	Yes	0.513516

17. Rather than looking at each individual dataset from our dataset, we can evaluate our model as a whole using the **evaluate** node as seen in the following screenshot:



18. Run the experiment

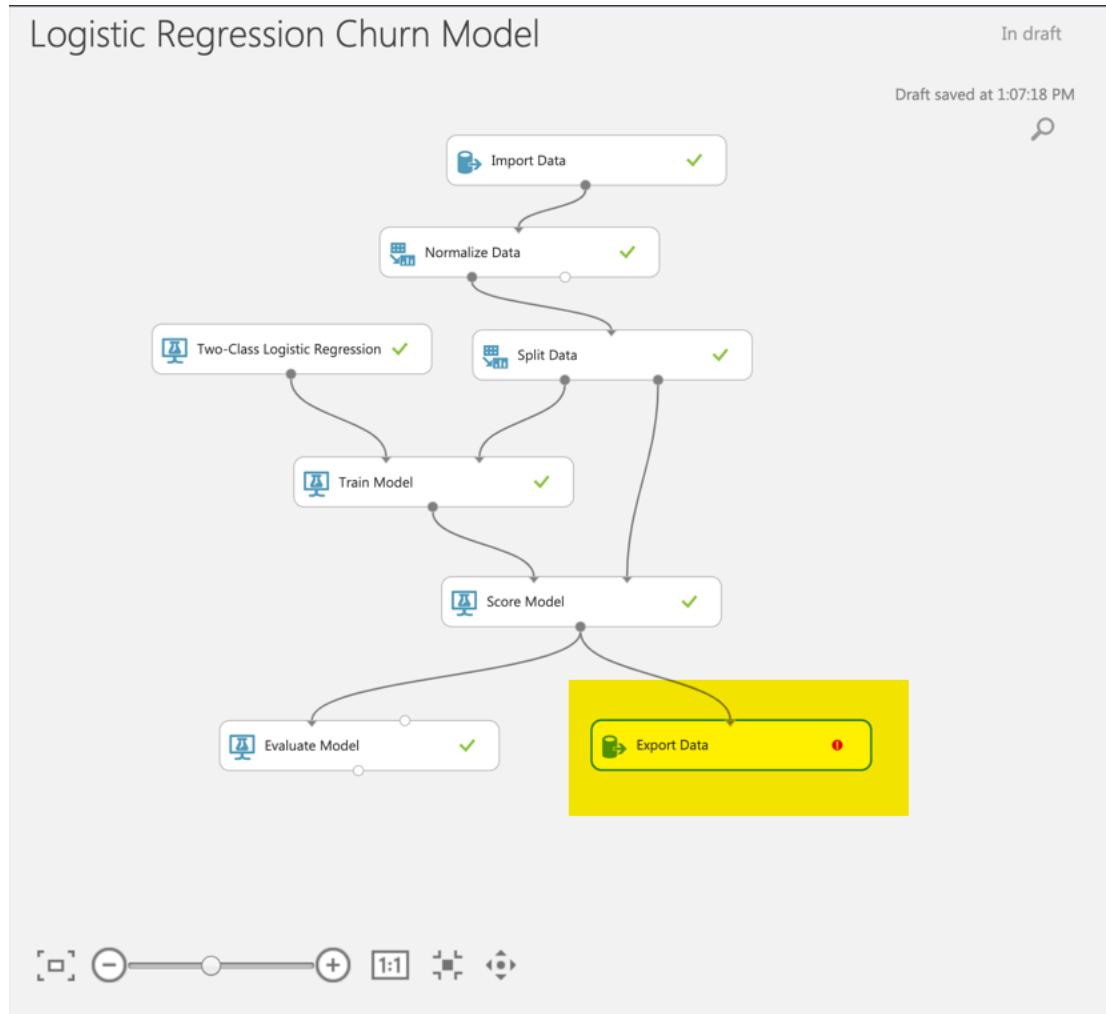
19. Right-click on the Evaluate model to view the performance of the model as seen in the following screenshot:



20. We see a few of the following metrics to evaluate our model

- [ROC Curve](#)
- [AUC Score](#)
- [Confusion Matrix](#)

21. We are now ready to export our Scored Dataset back to blob storage using the **Export Data** Node as seen in the following screenshot:



22. Enter the same credentials as you did in the previous section when exporting the scored results as a CSV file. However, this time we will name the file path **churn/Telco_Customer_Churn_Final.csv** as seen in the following screenshot:

Properties Project >

Export Data

Please specify data destination
Azure Blob Storage

Please specify authentication type
Account

Azure account name
blobmlinaday

Azure account key
.....

Path to blob beginning with container
churn/Telco_Customer_Churn_Final.csv

Azure blob storage write mode
Overwrite

File format for blob file
CSV

Write blob header row

Use cached results

23. Run the Experiment

- If you go to your blob storage account, you should now see your newly exported dataset with the scored results as seen in the following screenshot:

Home > Resource groups > MLInADayRG > blobmlinaday - Blobs > churn

churn
Container

Upload Refresh Change access level Delete Acquire lease Break lease View snapshots Create snapshot

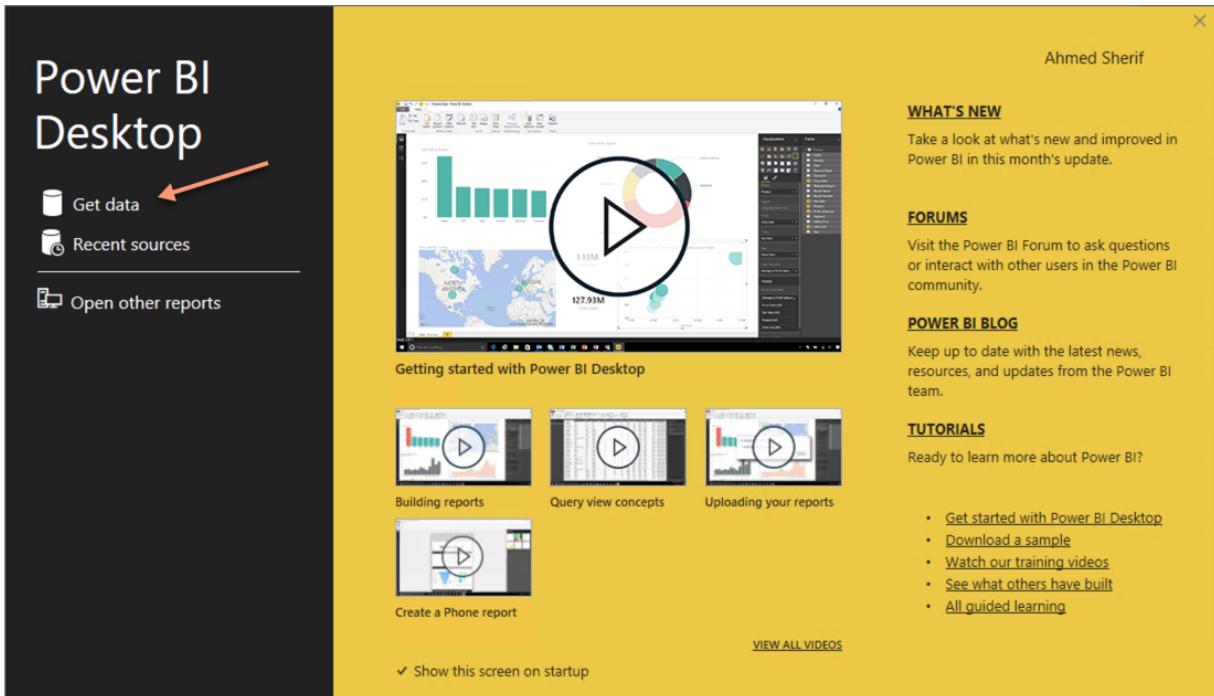
Search blobs by prefix (case-sensitive) Show deleted blobs

Authentication method: Access key (Switch to Azure AD User Account)
Location: churn

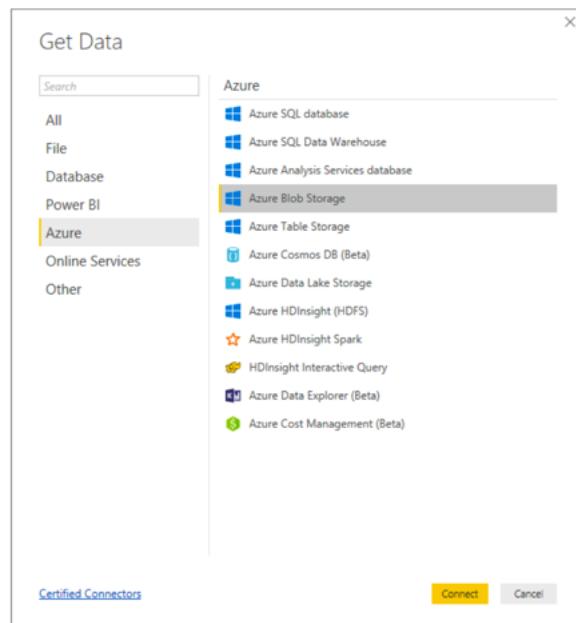
NAME	MODIFIED	ACCESS TIER	BLOB TYPE	SIZE	LEASE STATE
Telco_Customer_Churn_Cluster.csv	11/25/2018, 2:37:12 PM	Cool (Inferred)	Block blob	966.64 KiB	Available
Telco_Customer_Churn_Final.csv	3/31/2019, 2:37:46 PM	Cool (Inferred)	Block blob	275.49 KiB	Available
Telco_Customer_Churn.csv	11/25/2018, 11:28:06 AM	Cool (Inferred)	Block blob	954.59 KiB	Available

Section 6: Export scored dataset with prediction scores from Blob to Power BI Desktop

1. Start up Power BI Desktop and select Get Data as seen in the following screenshot:



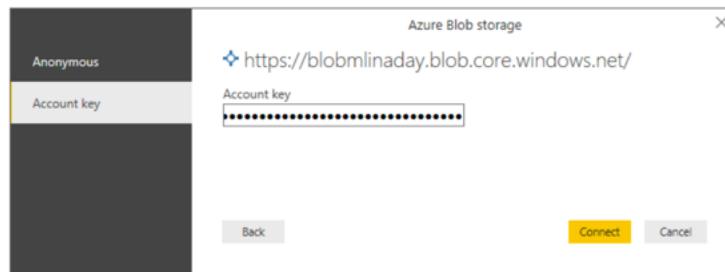
2. Select **Azure** and then **Azure Blob Storage** as seen in the following screenshot:



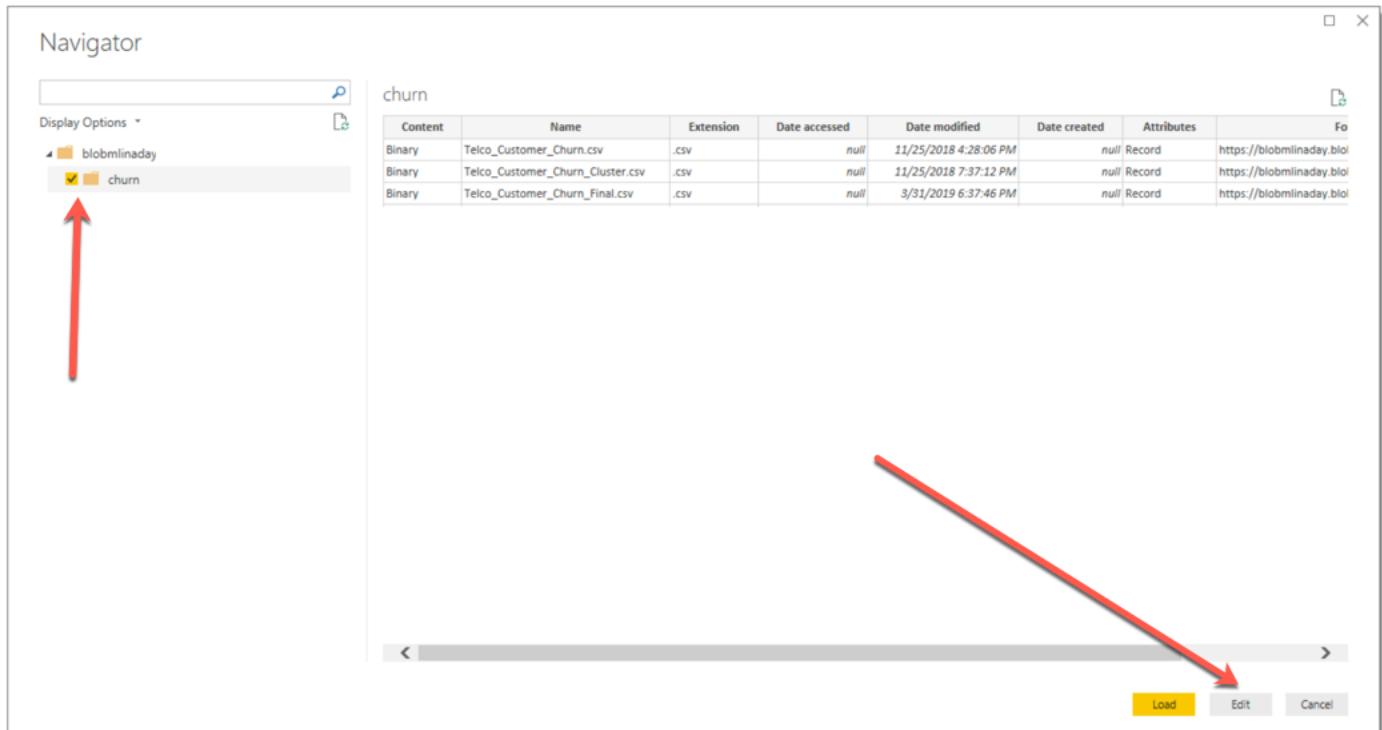
3. Enter your Blob Account Name and select OK



4. Enter your blob account key and select OK



5. You will see all of your files in the container you've selected as seen in the following screenshot:



Content	Name	Extension	Date accessed	Date modified	Date created	Attributes	File
Binary	Telco_Customer_Churn.csv	.CSV	null	11/25/2018 4:28:06 PM	null	Record	https://blobmlinaday.blob.core.windows.net/churn/Telco_Customer_Churn.csv
Binary	Telco_Customer_Churn_Cluster.csv	.CSV	null	11/25/2018 7:37:12 PM	null	Record	https://blobmlinaday.blob.core.windows.net/churn/Telco_Customer_Churn_Cluster.csv
Binary	Telco_Customer_Churn_Final.csv	.CSV	null	3/31/2019 6:37:46 PM	null	Record	https://blobmlinaday.blob.core.windows.net/churn/Telco_Customer_Churn_Final.csv

6. Do not select Load, but instead select EDIT and then select the double arrows next to content as seen in the following screenshot:

Name	Extension	Date accessed	Date modified	Date created	Attributes	Folder Path
Telco_Customer_Churn.csv	.csv	null	11/25/2018 4:28:06 PM	null	Record	https://blobmlinaday.blob.core.windows.net/churn/
Telco_Customer_Churn_Cluster.csv	.csv	null	11/25/2018 7:37:12 PM	null	Record	https://blobmlinaday.blob.core.windows.net/churn/
Telco_Customer_Churn_Final.csv	.csv	null	3/31/2019 7:33:21 PM	null	Record	https://blobmlinaday.blob.core.windows.net/churn/

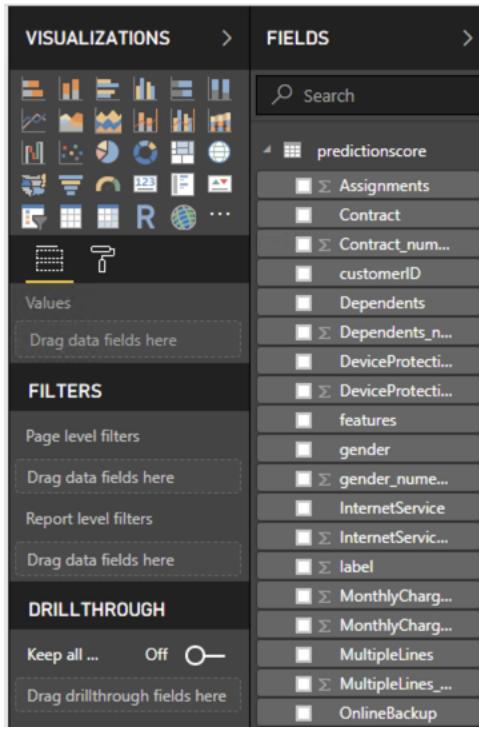
7. Select **Telco_Customer_Churn_Final.csv** as the Example File as seen in the following screenshot:

8. Select OK

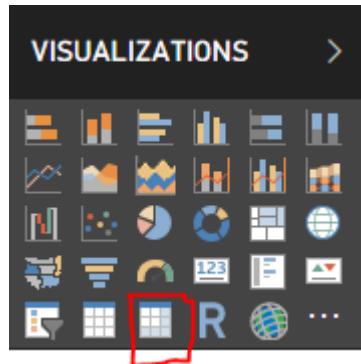
9. Scroll all the way to the end of the last column and select on the **Scored Probabilities** column. Right-click on it and select **Change Type to Decimal Number** as seen in the following screenshot:

10. Select Close and Apply and return to the main Power BI canvas

11. Once all of that has been accomplished, the dataset with all of the columns will be available for us to visualize inside of Power BI as seen in the following screenshot:



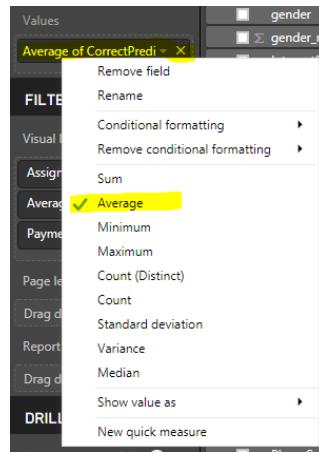
12. We can build a heatmap showing how accurate our model is within different segments of the population. Select the Matrix visualization from the Visualizations bar.



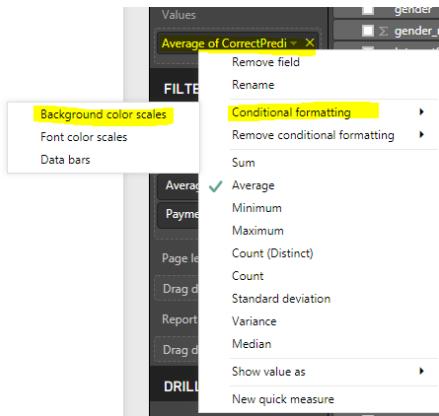
13. Drag **Assignments** into the Rows area. These **Assignments** represent the 10 customer segments we created in Azure Machine Learning Studio. Next, Drag **PaymentMethod** into the Columns area. Drag **Scored Probabilities** into the Values area.

The screenshot shows the Power BI Fields pane on the right side of the interface. In the 'Values' section, the 'Scored Probabilities' field is selected. Three red arrows point from the text '14. Right click on Scored Probabilities in the Values area and select Average.' to the 'Scored Probabilities' field in the Fields pane.

14. Right click on **Scored Probabilities** in the Values area and select **Average**.



15. Right click on **Average of Scored Probabilities**, which now appears in the Values area, but this time select Conditional formatting and Background color scales.



16. We now have a heatmap showing how accurate our predictions are within each customer segment and for customers of each payment type.

Assignments	Bank transfer (automatic)	Credit card (automatic)	Electronic check	Mailed check	Total
0	0.19	0.23	0.51	0.13	0.31
1	0.18	0.08	0.25	0.16	0.18
2	0.05	0.05	0.18	0.10	0.08
3	0.22	0.13	0.35	0.10	0.23
4	0.11	0.08	0.23	0.12	0.13
5	0.21	0.09	0.47	0.08	0.22
6	0.25	0.24	0.59	0.25	0.37
7	0.15	0.12	0.34	0.12	0.21
8	0.28	0.18	0.49	0.19	0.30
9	0.09	0.07	0.22	0.02	0.12
Total	0.19	0.14	0.45	0.19	0.27

17. We have successfully completed this workshop!! Feel free to play with other visualizations in Power BI while fellow participants finish up.