

Amazon SageMaker Canvas

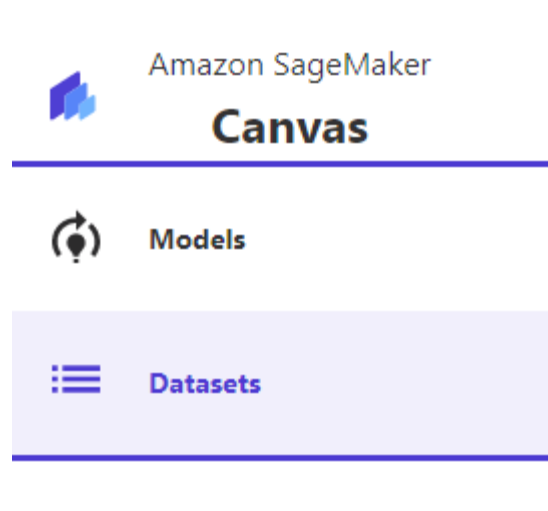
Use AWS Academy in Virginia

What is SageMaker Canvas

- Another tool to automate the ML workflow to reduce manual efforts
- SageMaker Canvas uses Autopilot (next lecture) service under the hood
- It is a fully managed no code ML solution with interactive UI
- You can finish the entire ML workflow with a few point-and-click

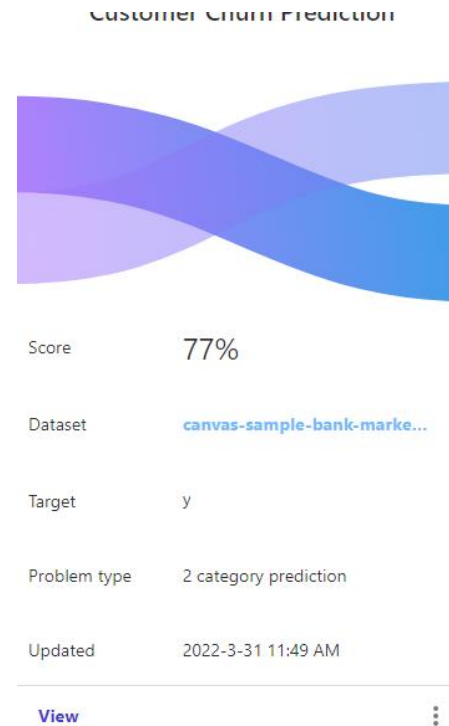
Canvas Interface

- There are two main pages in Canvas: the **Models** page and the **Datasets** page.



Model Page

- The **Models** page shows the models you've created in Canvas.



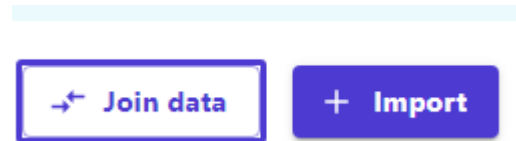
Datasets page

- The **Datasets** page shows the datasets you've imported

Source	Columns	Rows	Cells	Created	Status
S3	21	41,190	864,990	12/21/2021 8:58 AM	Ready

Steps to build a model

- First, **join** or **import** new datasets into Canvas.



- After importing new datasets, **choose** a dataset you want to build a model with.

Datasets

Name	Source	Columns
<input checked="" type="checkbox"/> canvas-sample-bank-marketing.csv	S3	21

- Then click on Create a model



Prepare to build

- In the **Build** tab, before building a model, review and prepare your chosen dataset at the bottom of the page.

Then, select the **Target column** in the top left section.

Select

Build

Select a column to predict

Choose the target column. The model that you build will predict the values of that column.

Target column

y

Value distribution

no

yes

canvas-sample-bank-marketing.csv

Random sample: 20.0k rows

Column name ↓

y

Target

previous

✓

poutcome

✓

pdays

✓

nr.employed

✓

month

✓

marital

✓

loan

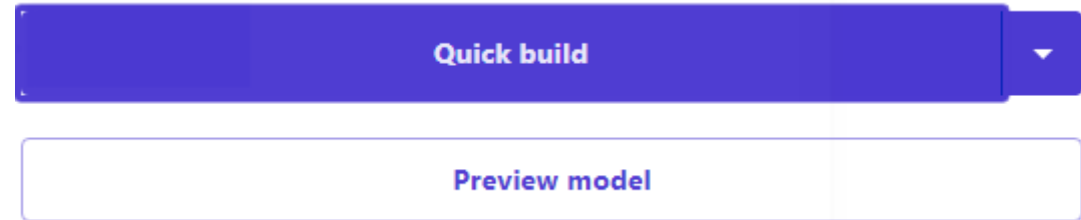
✓

inh

✓

Click to build your model

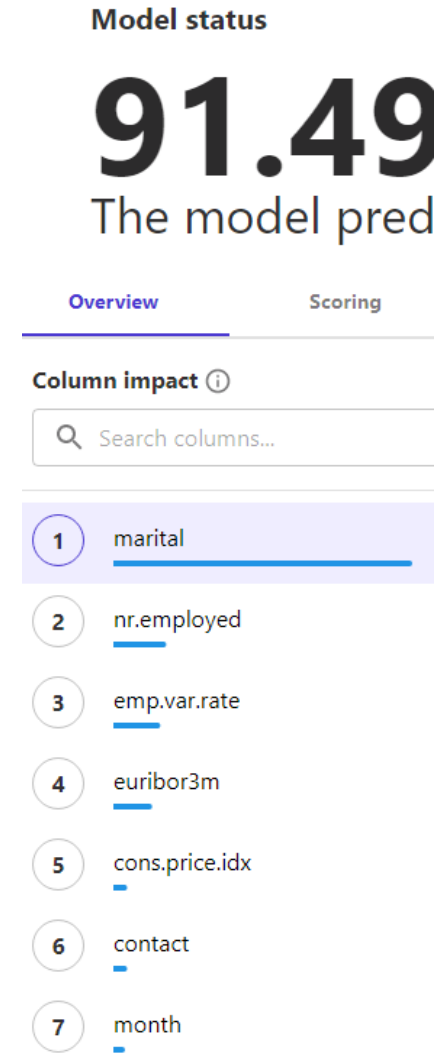
- After selecting the **Target column**, start building your model.



- You can check out the information on the building time and progress of your model by staying in the **Build** tab

Analyzing your model

- After the model has been built, in the **Analyze** tab, check which column has the most impact on your predictions in the **Overview** section, and review how well your model has been built in the **Scoring** section.



Generate predictions

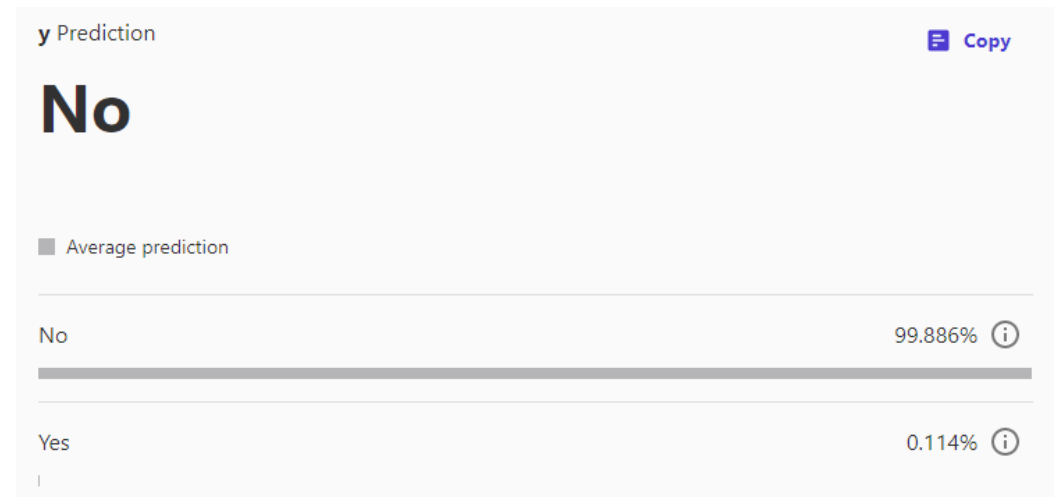
- In the **Prediction** tab, try to change the dataset values to see how the prediction results change in **Single** prediction. You can also import new datasets and generate predictions in **Batch** prediction.

Predict target values




Batch prediction

Single prediction

Modify values to predict y in real time.



When you start the following datasets are available to use

 Amazon SageMaker Canvas		Datasets 🔍 Data					
 Models							
 Datasets							
Name		Source	Columns	Rows	Cells	Created	Status
<input type="checkbox"/>	canvas-sample-loans-part-2.csv	S3	5	1,000	5,000	11/13/2022 9:04 AM	Ready
<input type="checkbox"/>	canvas-sample-housing.csv	S3	10	1,000	10,000	11/13/2022 9:04 AM	Ready
<input type="checkbox"/>	canvas-sample-sales-forecasting.csv	S3	5	1,000	5,000	11/13/2022 9:04 AM	Ready
<input type="checkbox"/>	canvas-sample-loans-part-1.csv	S3	19	1,000	19,000	11/13/2022 9:04 AM	Ready
<input type="checkbox"/>	canvas-sample-maintenance.csv	S3	9	1,000	9,000	11/13/2022 9:04 AM	Ready
<input type="checkbox"/>	canvas-sample-shipping-logs.csv	S3	12	1,000	12,000	11/13/2022 9:04 AM	Ready
<input type="checkbox"/>	canvas-sample-product-descriptions.csv	S3	5	120	600	11/13/2022 9:04 AM	Ready
<input type="checkbox"/>	canvas-sample-diabetic-readmission.csv	S3	16	1,000	16,000	11/13/2022 9:04 AM	Ready

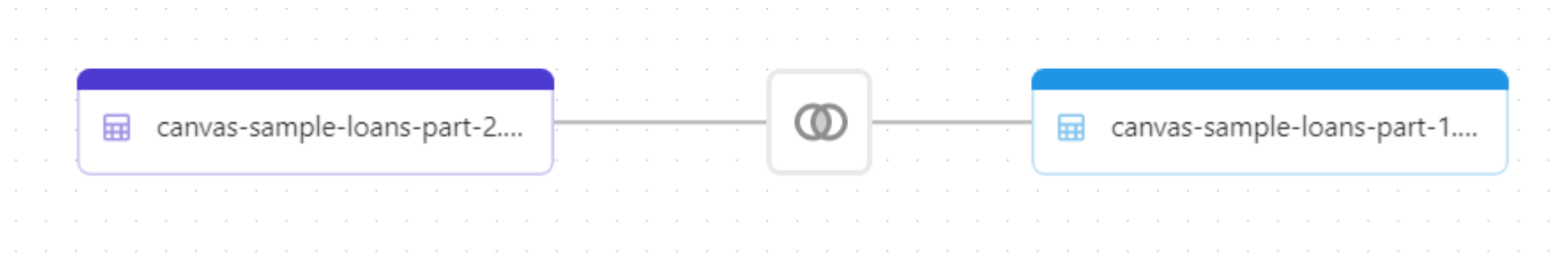
Click on Join data set and import sample loans part 1 and 2

- You can use loan sample data set to predict whether a customer will repay a loan. Use the **loan_status** column as the target column
- The data source is <https://www.kaggle.com/datasets/wordsforthewise/lending-club>

<input type="checkbox"/>	canvas-sample-sales-forecasting.csv	S3	5	1,000	5,000	03/12/2023 2:32 PM	Ready	⋮
<input checked="" type="checkbox"/>	canvas-sample-loans-part-1.csv	S3	19	1,000	19,000	03/12/2023 2:32 PM	Ready	
<input type="checkbox"/>	canvas-sample-housing.csv	S3	10	1,000	10,000	03/12/2023 2:32 PM	Ready	
<input type="checkbox"/>	canvas-sample-product-descriptions.csv	S3	5	120	600	03/12/2023 2:32 PM	Ready	
<input type="checkbox"/>	canvas-sample-shipping-logs.csv	S3	12	1,000	12,000	03/12/2023 2:32 PM	Ready	
<input checked="" type="checkbox"/>	canvas-sample-loans-part-2.csv	S3	5	1,000	5,000	03/12/2023 2:32 PM	Ready	
<input type="checkbox"/>	canvas-sample-diabetic-readmission.csv	S3	16	1,000	16,000	03/12/2023 2:32 PM	Ready	
<input type="checkbox"/>	canvas-sample-maintenance.csv	S3	9	1,000	9,000	03/12/2023 2:32 PM	Ready	

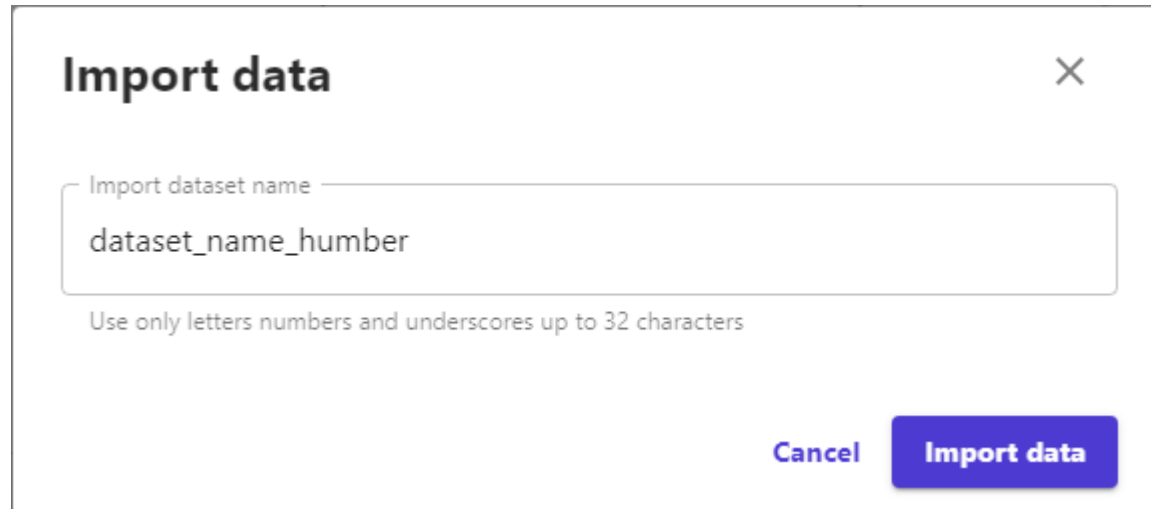
Join them based on id

- After join you will see that the resulting table has all the columns



Import data and name it

- After clicking on import, you need to set a **name** and click on **Import data**

A dialog box titled "Import data" with a close button (X) in the top right corner. It contains a text input field with the placeholder "Import dataset name" and the text "dataset_name_humber" entered. Below the input field is a note: "Use only letters numbers and underscores up to 32 characters". At the bottom right are two buttons: "Cancel" and "Import data".

Import data ×

Import dataset name

dataset_name_humber

Use only letters numbers and underscores up to 32 characters

[Cancel](#) [Import data](#)

Create model from that imported data

- Select the data
- Click on **create** model
- Set a **name** for the new model

Datasets

Search Dataset

Join data Import

1 selected Create a model Delete

Name	Source	Columns	Rows	Cells	Created	Status
<input checked="" type="checkbox"/> dataset_name_humber	Joined	23	1,000	23,000	03/12/2023 3:07 PM	Ready
<input type="checkbox"/> canvas-sample-sales-forecasting.csv	S3	5	1,000	5,000	03/12/2023 2:32 PM	Ready
<input type="checkbox"/> canvas-sample-loans-part-1.csv	S3	19	1,000	19,000	03/12/2023 2:32 PM	Ready
<input type="checkbox"/> canvas-sample-housing-new	S3	10	1,000	10,000	03/12/2023 2:32 PM	Ready

Create new model

Model name

humber_model

Use only letters, numbers, and underscores up to 32 characters.


Cancel Create

Select a column to predict

- Canvas will automatically detect that this is a **3+ category prediction** problem (also known as **multi-class classification**). If the wrong model type is detected, you can change it manually with the **Change type** link at the center of the screen.

Select a column to predict

Choose the target column. The model that you build predicts values for the column that you select.


 Target column
loan_status

Value distribution

fully paid	
charged off	
current	

Model type

SageMaker Canvas automatically recommends the appropriate model type for your analysis.

 3+ category prediction

Your model classifies loan_status into 3 or more categories.

[Change type](#)

Statistics about data

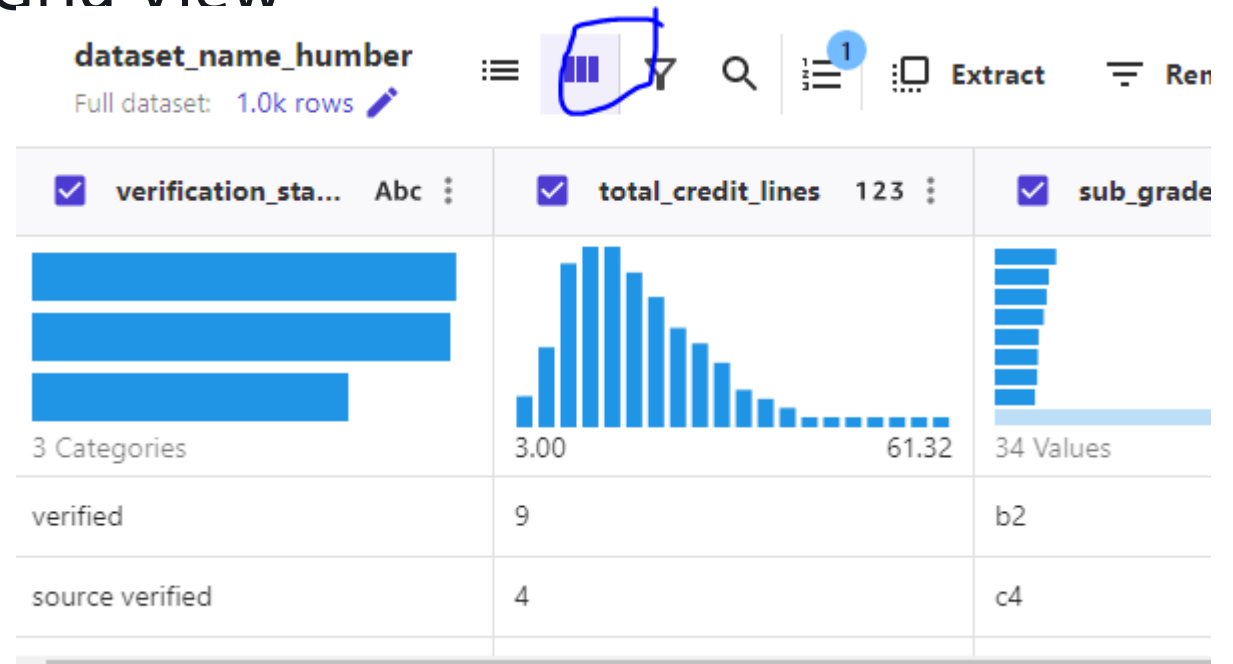
- At the bottom half of the screen, you can take a look at some of the statistics of the dataset, including missing and mismatched values, unique vales, mean and median values.
- You can see Column view and Grid View

dataset_name_humber
Full dataset: 1.0k rows

Column view icon (circled in blue)

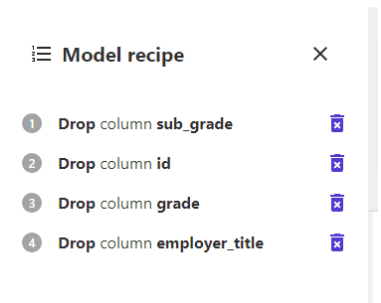
<input type="checkbox"/>	Column name ↓	Data type	Missing ⓘ
<input checked="" type="checkbox"/>	verification_status	Categorical	0.00% (0)
<input checked="" type="checkbox"/>	total_credit_lines	Numeric	0.00% (0)
<input checked="" type="checkbox"/>	sub_grade	Text	0.00% (0)
<input checked="" type="checkbox"/>	revolving_line_utilization_rate	Numeric	0.00% (0)
<input checked="" type="checkbox"/>	purpose	Text	0.00% (0)
<input checked="" type="checkbox"/>	open_credit_lines	Numeric	0.00% (0)

■ Total columns: 22 ■ Total rows: 1,000 ■ Total cells: 22,000 ☒ Show



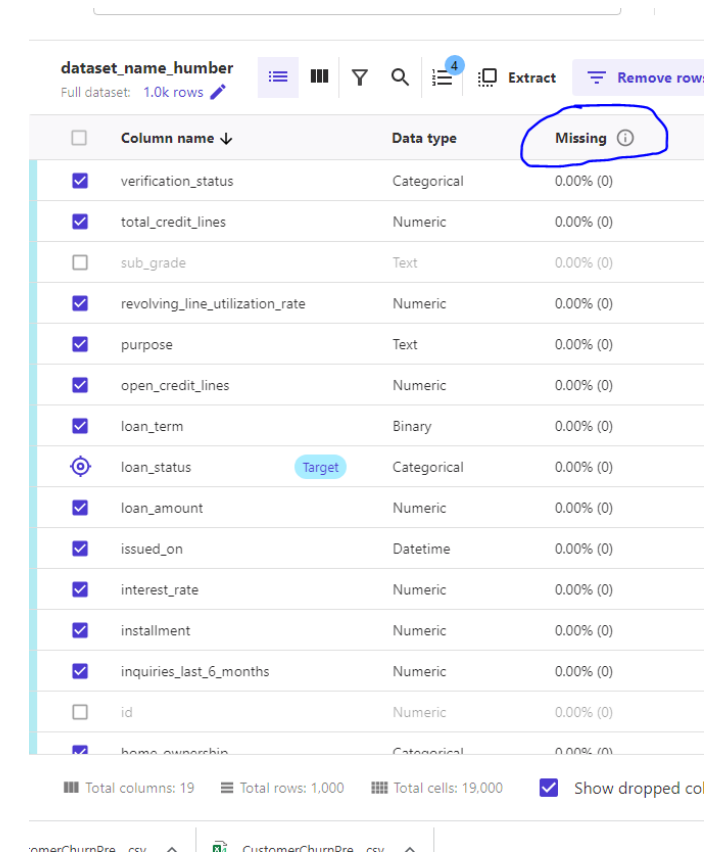
Drop columns

- You can also drop some of the columns, if we don't want to use them for the prediction, by simply **un-checking** them with the left checkbox.
- We deselect the columns that do not add value to the model training process:
- **id**, since it's a primary key, it does not have valuable information;
- **employer_title**,
- **Grade**: This number is something specific to the company that has shared this data set in Kaggle so we do not use it in this case
- **sub_grade**: Same as above



Missing values

- Dataset does not have a lot of missing values



The screenshot shows a data table interface with the following components:

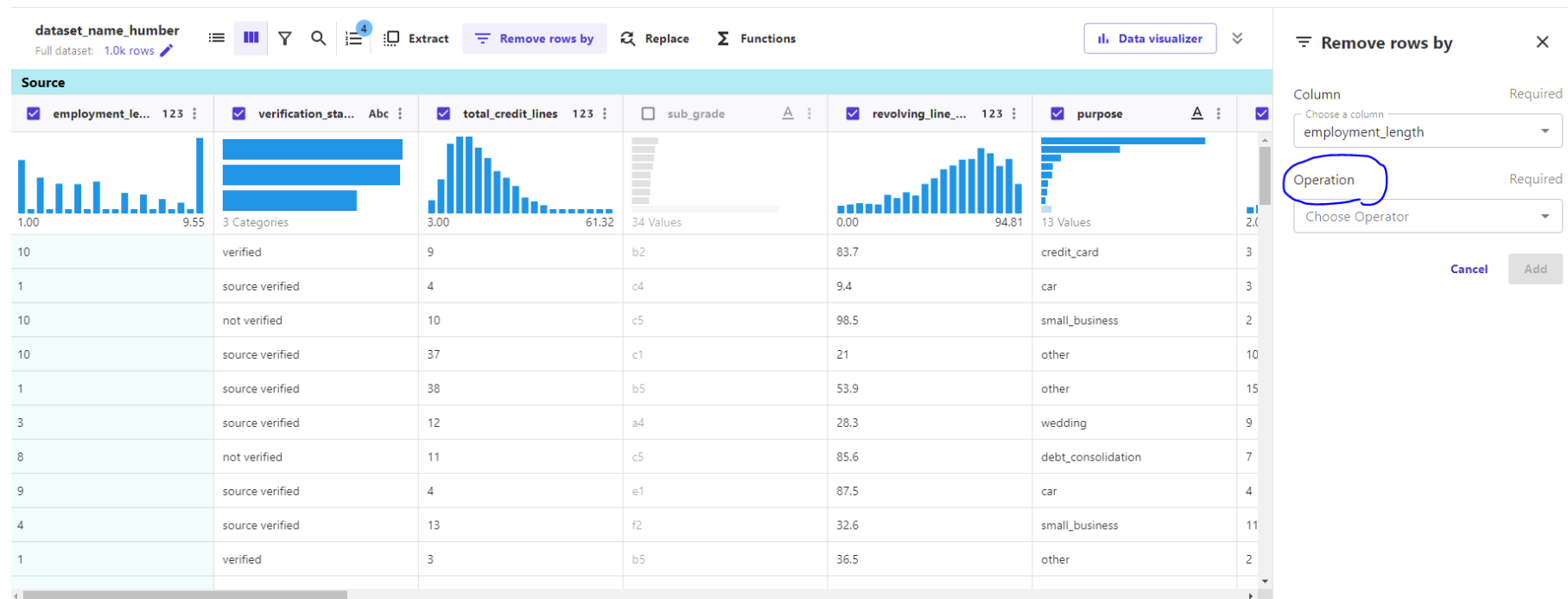
- Header:** "dataset_name_humber" with a sub-header "Full dataset: 1.0k rows".
- Toolbar:** Includes icons for list, columns, filter, search, and a blue circle with the number "4". Buttons for "Extract" and "Remove rows" are also present.
- Table:** A table with 4 columns: "Column name", "Data type", and "Missing". The "Missing" column shows "0.00% (0)" for all rows. The "loan_status" column is highlighted with a blue "Target" label.
- Footer:** Summary statistics: "Total columns: 19", "Total rows: 1,000", "Total cells: 19,000". A checkbox for "Show dropped columns" is checked.

<input type="checkbox"/>	Column name ↓	Data type	Missing ⓘ
<input checked="" type="checkbox"/>	verification_status	Categorical	0.00% (0)
<input checked="" type="checkbox"/>	total_credit_lines	Numeric	0.00% (0)
<input type="checkbox"/>	sub_grade	Text	0.00% (0)
<input checked="" type="checkbox"/>	revolving_line_utilization_rate	Numeric	0.00% (0)
<input checked="" type="checkbox"/>	purpose	Text	0.00% (0)
<input checked="" type="checkbox"/>	open_credit_lines	Numeric	0.00% (0)
<input checked="" type="checkbox"/>	loan_term	Binary	0.00% (0)
<input checked="" type="checkbox"/>	loan_status	Categorical	0.00% (0)
<input checked="" type="checkbox"/>	loan_amount	Numeric	0.00% (0)
<input checked="" type="checkbox"/>	issued_on	Datetime	0.00% (0)
<input checked="" type="checkbox"/>	interest_rate	Numeric	0.00% (0)
<input checked="" type="checkbox"/>	installment	Numeric	0.00% (0)
<input checked="" type="checkbox"/>	inquiries_last_6_months	Numeric	0.00% (0)
<input type="checkbox"/>	id	Numeric	0.00% (0)
<input checked="" type="checkbox"/>	home_ownership	Categorical	0.00% (0)

Total columns: 19 Total rows: 1,000 Total cells: 19,000 ☒ Show dropped columns

We can remove the rows with missing values

- We do not have to handle missing values since our data does not have one.
- But if you needed to handle those, you can filter the missing values



Quick Build Features

- Once you've explored this section, it's time to finally train the model!
- Before building a complete model, it is a good practice to have a general idea about the performances that our model will have by training a **Quick Build**.
- A quick model trains **fewer combinations of models** and **hyper-parameters** in order to prioritize speed over accuracy
- Note that quick build is not available for models bigger than 50k rows. Let's go ahead and click **Quick build**.

Select Quick Build

- It takes between 2-15 mins

Change type

Quick build ▲

Standard build

Choose accuracy over speed. Building usually takes between 2–4 hours.

Quick build

Choose speed over accuracy. Building usually takes 2–15 minutes. You can't share quick build models.

Model overview

Your model is being created. Quick build usually takes 2–15 minutes. You can now leave this view.

Remove

imn

Expected build time

2–15 minutes

Build type

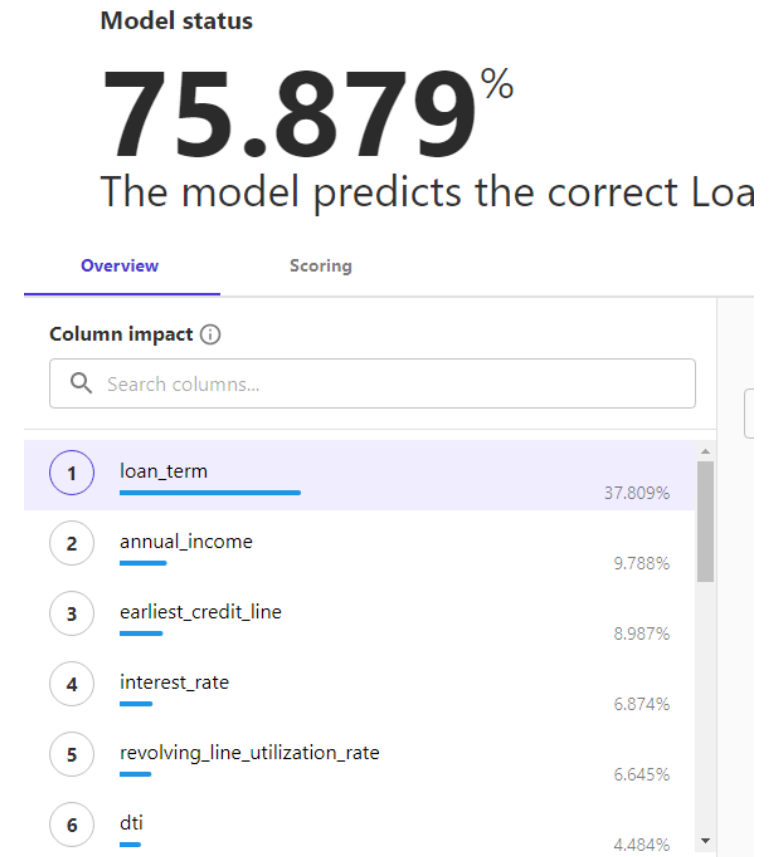
Quick build

Detailed progress

Generating column impact

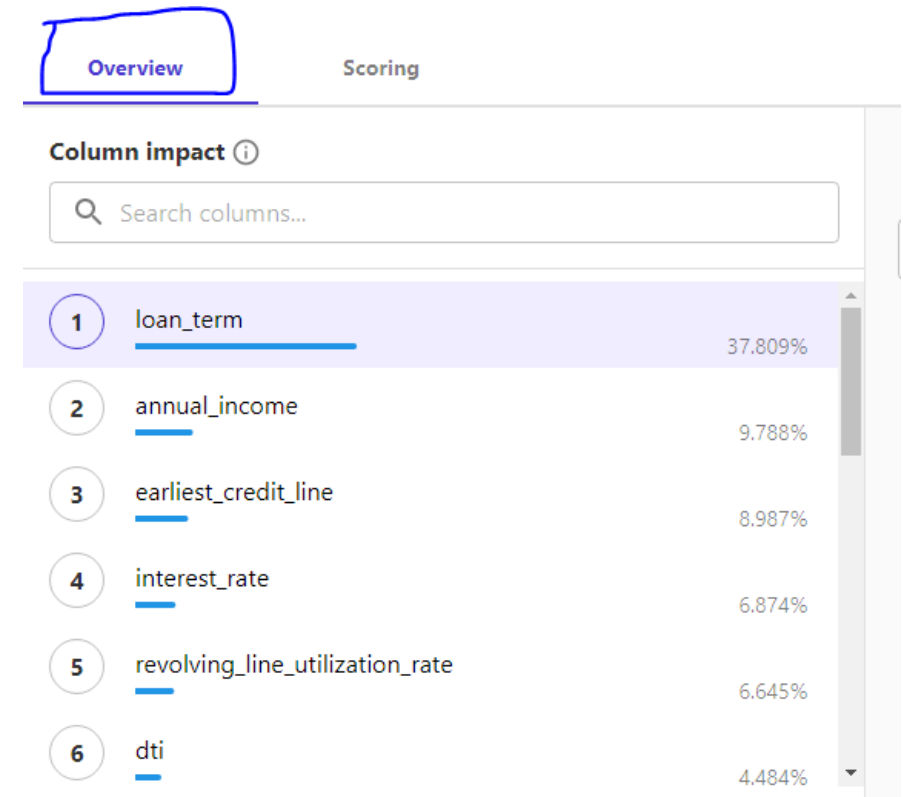
Analyze the model

- Once done, Canvas will automatically move to the **Analyze** tab, to show us the results of our quick training:



Column importance

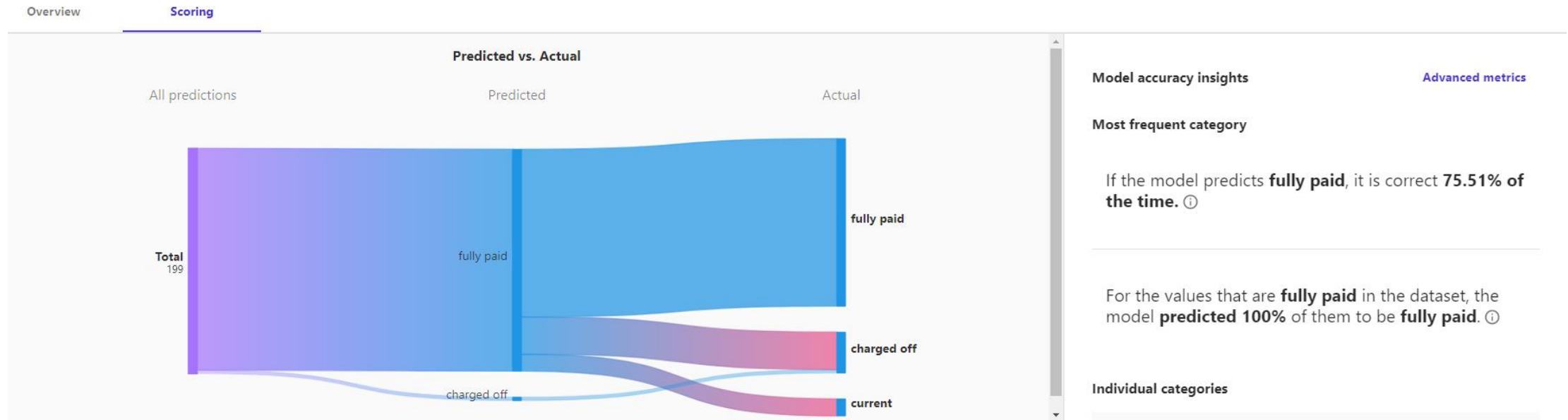
- In the overview tab you see the column importance



Overview		Scoring
Column impact ⓘ		
<input type="text" value="Search columns..."/>		
1	loan_term	37.809%
2	annual_income	9.788%
3	earliest_credit_line	8.987%
4	interest_rate	6.874%
5	revolving_line_utilization_rate	6.645%
6	dti	4.484%

Scoring

- On the scoring tab you see model metrics



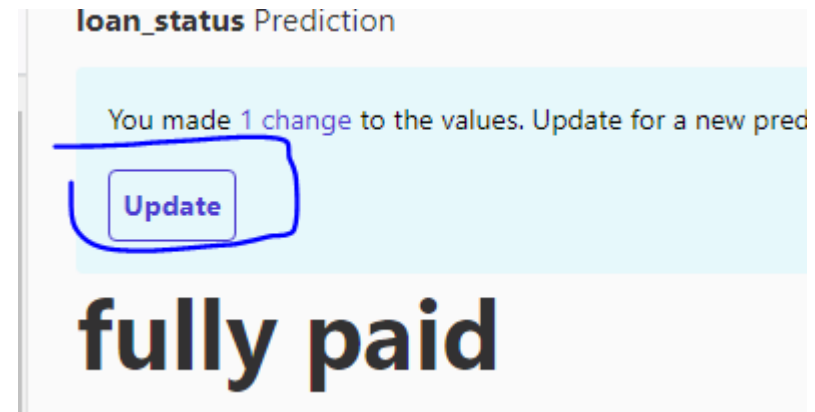
Advanced Metrics

- You can get deeper quality metrics by going to the Advanced Metrics

Advanced metrics								
Average f1 ⓘ	Average accuracy ⓘ	Average precision ⓘ	Average recall ⓘ	Average AUC ⓘ				
33.81%	75.879%	58.503%	36.111%	Not available				
					Predicted values		Class	
					fully paid	charged off	current	fully paid
Actual values	fully paid	148	0	0	Precision ⓘ		75.51%	
	charged off	33	3	0	Recall ⓘ		100%	
	current	15	0	0	Accuracy ⓘ		75.879%	
					F1 ⓘ		0.86	
					Auc ⓘ		Not available	

Predict

- Now that the model is trained, let's use for some predictions.
- Select **Predict** at the bottom of the **Analyze** page, or choose the **Predict** tab.
- Click on **Single Prediction**
- Put some numbers there just to simulate the prediction
- This is good for **what-if** scenarios
- Click on update to see the result



Add Version

- You can add a new version of this model and this time you train with Standard method



- The console says it take 2-4 hours to create the model but for me it took around **15** minutes

Compare scoring with quick build

Advanced metrics



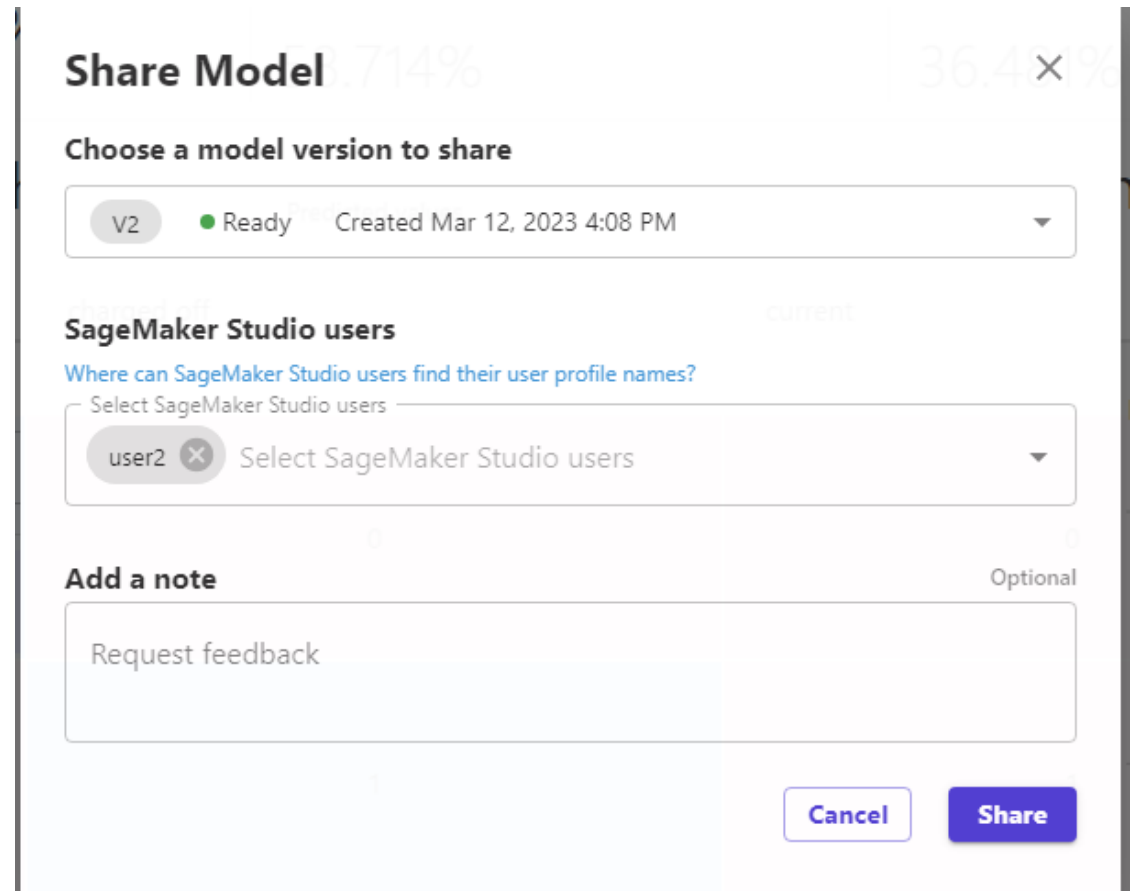
Average f1 ⓘ	Average accuracy ⓘ	Average precision ⓘ	Average recall ⓘ	Average AUC ⓘ
34.494%	75.622%	58.714%	36.481%	Not available

		Predicted values			Class
		fully paid	charged off	current	fully paid ▼
Actual values	fully paid	150	0	0	Precision ⓘ 76.142%
	charged off	34	1	1	Recall ⓘ 100%
	current	13	1	1	Accuracy ⓘ 75.622%
					F1 ⓘ 0.865
					Auc ⓘ Not available

[Close](#)[Download](#)

You can share that model

- Assume you have user 2 in that domain

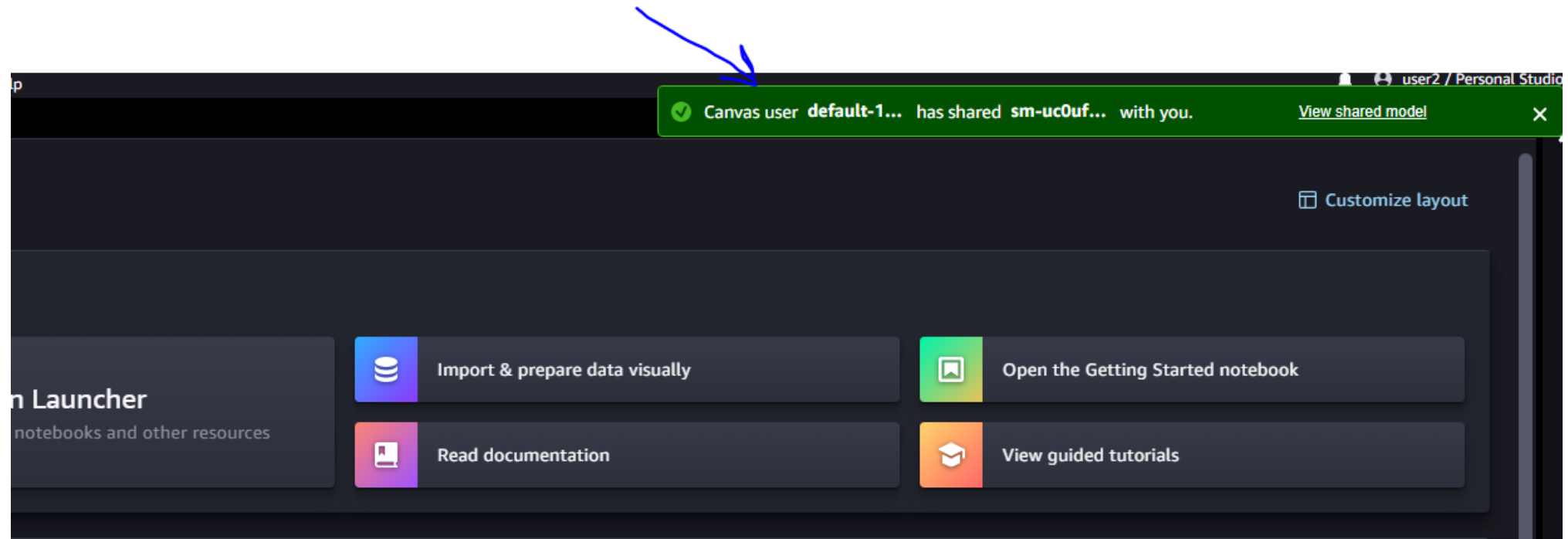


The screenshot shows a 'Share Model' dialog box with the following sections:

- Share Model**: The title of the dialog.
- Choose a model version to share**: A dropdown menu showing 'V2' as the selected version, with a status of 'Ready' and a creation time of 'Created Mar 12, 2023 4:08 PM'.
- SageMaker Studio users**: A section with a link 'Where can SageMaker Studio users find their user profile names?' and a dropdown menu labeled 'Select SageMaker Studio users' showing 'user2' as the selected user.
- Add a note**: A text input field with the placeholder text 'Request feedback'.
- Buttons**: 'Cancel' and 'Share' buttons at the bottom right.

What user2 sees

- When user2 login to Studio a message is shown
- Click on view shared models
- You can deploy it by this user session



Important

- Billing in SageMaker Canvas consists of the following components:
 - **Session charges** – You are charged for the number of hours that you are logged in to or using SageMaker Canvas.
 - **Training charges** – You are charged for training models based on the size of the dataset you use.
- Make sure you follow below instructions to avoid unnecessary cost:
 - If you're not using Amazon SageMaker Canvas, you can log out of your session. A *session* starts as soon as you launch SageMaker Canvas from the console. Logging out ends the session. You are only billed for the duration of the session.
 - When you log out, your models and datasets aren't affected, but SageMaker Canvas cancels any **Quick build** tasks. If you log out of SageMaker Canvas while running a **Quick build**, your build might be interrupted until you log back in. When you log back in, SageMaker Canvas automatically restarts the build.
 - Click on logout button now



Log out

Check if you have active session

- SageMaker console → Select domains → Select domain details → under user profile → select user profile name
- Under the Apps → find canvas → see App type column
- See the status → if it shows ready, that means there is an active session and you get charged. Make sure it is in the deleted mode

Amazon SageMaker > Domains > Domain: day05 > User Details: user2

User Details

General details about this user profile.

Apps			
App name	Status	App type	Created
default	✓ Ready	JupyterServer	Sun Mar 12 2023 16:50
default	✗ Deleted	Canvas	Sun Mar 12 2023 16:44

Assignment

- I have uploaded White Wine quality data set to BB in **archive.zip** file (source: <https://www.kaggle.com/datasets/piyushagni5/white-wine-quality>)
- Create a report and explain the following actions:
 - Divide the data set to training and test set
 - Add the data set to SageMaker Canvas
 - Select the columns that you think add value in prediction
 - Build the model using quick build
 - Check accuracy when the columns are added or removed
 - Predict values using test data set
 - Create different versions
 - Try standard build and share with other studio users