

Cloudera Machine Learning Workshop

- 1) Open CDP, using the “admin” user within the Test Drive link.

Your link should look something like (remember click the link in your email not the link below)

http://login.trycdp.com/auth/realms/trycdp-trialxx/protocol/saml/clients/samlclient?tn=trialxx_admin@trycdp.com&p=X

X *xx represents the trial user #

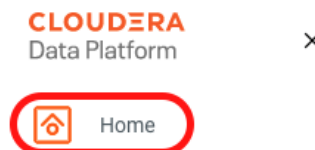
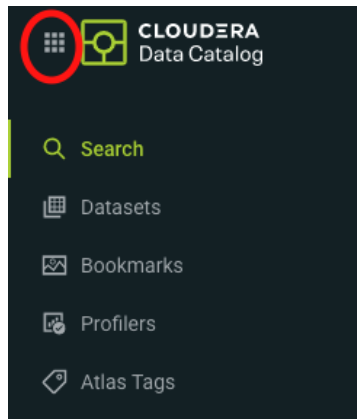
*X represents the password

- 2) Click the “Machine Learning” tile within the CDP Home Screen



How do you get to the CDP Home Screen?

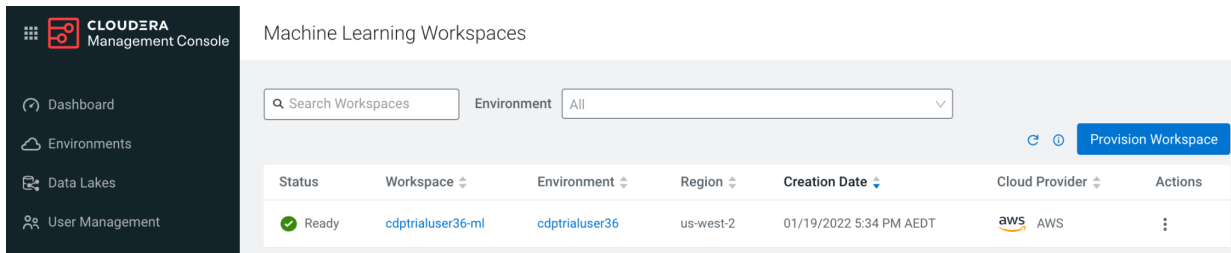
- From any experience such as “Data Catalog”, click the 9 square at the top left and then click “Home”



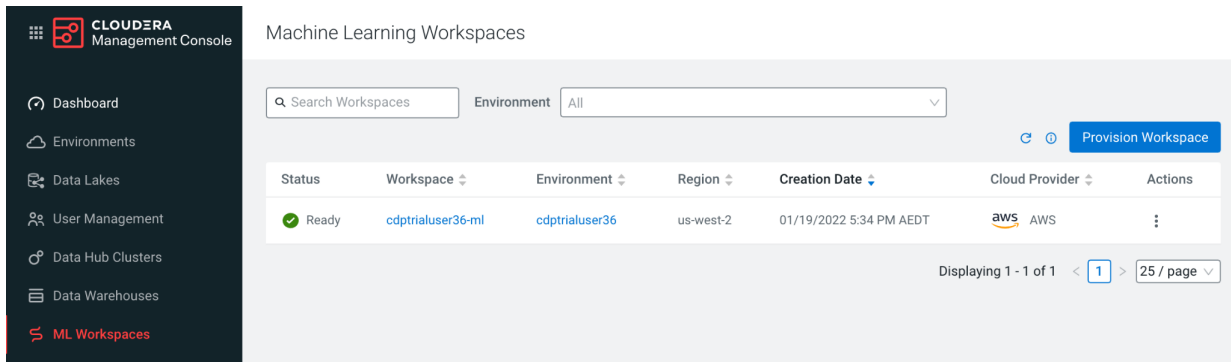
Part 1 - REQUIRED IN ORDER FOR LAB TO WORK PROPERLY: Create a CML Project

(A Workspace should already be provisioned for you)

Workspaces are the heart of the Cloudera Machine Learning (CML)



A Workspace is a small cluster that runs on a kubernetes service to provide teams of data scientists to develop, test, train, and ultimately deploy machine learning models. **Click into the Workspace by clicking the Workspace name.**



You can visualize all of the Projects and Resources that are part of the Projects page. Next we will create a Project where we will develop and deploy models along with other CML features. **Click on “New Project”**

Projects

Project quick find + trial36_admin

View Resource Usage Details

Active Workloads

SESSIONS	EXPERIMENTS	MODELS	JOB	APPLICATIONS
0	0	0	0	0

User Resources | Workspace Resources

CPU: 0.0 vCPU 3.5 available
Memory: 0.0 GiB 17.5 available

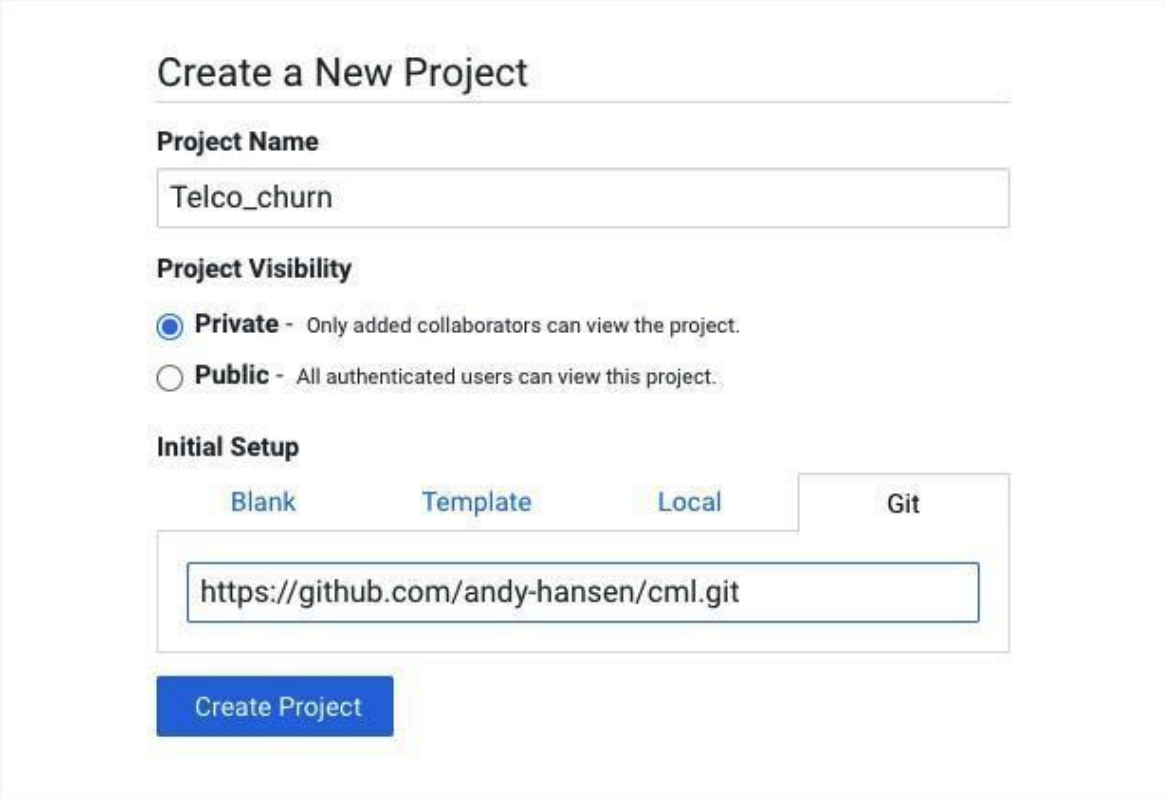
Search Projects Scope: My Projects Creator: All

Telco_churn1

Created by: trial36_admin Last worked on: a minute ago

When creating a new project give a Name, Visibility, and initial configuration.

Project Name: Telco_churn **Visibility:** Private **Initial Setup:** Git ->
<https://github.com/andy-hansen/cml.git>



The screenshot shows the 'Create a New Project' interface. It has three main sections: 'Project Name', 'Project Visibility', and 'Initial Setup'. The 'Project Name' field contains 'Telco_churn'. The 'Project Visibility' section has two radio buttons: 'Private' (selected) and 'Public'. The 'Initial Setup' section has four tabs: 'Blank', 'Template', 'Local', and 'Git' (selected). Below the 'Git' tab is a text field containing the URL 'https://github.com/andy-hansen/cml.git'. At the bottom is a blue 'Create Project' button.

Create a New Project

Project Name

Telco_churn

Project Visibility

☒ **Private** - Only added collaborators can view the project.

☐ **Public** - All authenticated users can view this project.

Initial Setup

Blank Template Local Git

<https://github.com/andy-hansen/cml.git>

Create Project

Part 2: CML Project Overview

Overview gives you access to all the features of a CML project. We will only have files copied from the Github repo currently. Initially it is good to start on the management components of a project.

The screenshot shows the Cloudera Machine Learning interface for a project named 'Telco_churn1'. The left sidebar contains navigation links: All Projects, Overview (selected), Sessions, Experiments, Models, Jobs, Applications, Files, Collaborators, and Project Settings. The main content area displays the project name 'Telco_churn1' with a lock icon and the source 'Telco_churn from github'. It includes buttons for 'Fork' and 'New Session'. Below this, there are sections for 'Models' (no models yet), 'Jobs' (no jobs yet), and 'Files'. The 'Files' section shows a table of files and folders.

	Name ^	Size	Last Modified
<input type="checkbox"/>	flask	-	a few seconds ago
<input type="checkbox"/>	images	-	a few seconds ago
<input type="checkbox"/>	models	-	a few seconds ago
<input type="checkbox"/>	raw	-	a few seconds ago
<input type="checkbox"/>	0_bootstrap.py	2.22 kiB	a few seconds ago Edit
<input type="checkbox"/>	1_data_ingest.py	7.41 kiB	a few seconds ago Edit
<input type="checkbox"/>	2_data_exploration.ipynb	602.63 kiB	a few seconds ago Edit
<input type="checkbox"/>	3_model_building.ipynb	71.53 kiB	a few seconds ago Edit
<input type="checkbox"/>	4_train_models.py	9.56 kiB	a few seconds ago Edit
<input type="checkbox"/>	5_model_serve_explainer.py	8.10 kiB	a few seconds ago Edit
<input type="checkbox"/>	6_application.py	7.74 kiB	a few seconds ago Edit
<input type="checkbox"/>	7a_ml_ops_simulation.py	7.96 kiB	a few seconds ago Edit

Collaborators:

For our demo we aren't adding additional collaborators.

You can give access to other users with certain permissions for the encompassing project so teams of users can collaborate together. You can set up Admins, Contributor, Operator, and Viewer permissions.

The screenshot shows the 'Collaborators' page for the 'Telco_churn1' project. The left sidebar is the same as the previous screenshot. The main content area shows the project name 'Telco_churn1' followed by 'Collaborators'. It includes a search bar and a list of collaborators. A message states: 'This project is private. Only collaborators can view and edit this project. [Change Settings.](#)' Below this is a section 'Add Collaborator' with a red error message: 'Email is not configured. Please contact your administrator.' There is a search bar for adding collaborators and a table showing the current collaborators.

Collaborator	Permission
trial36_admin	Owner

Granting Admin or Contributor permission to other users may have security impact since it gives them full access to your project files and running sessions.

Project Settings:

Taking a look at Project Settings, this is where you can define several options for the current project. You have the ability to define different engines where your code in CML will run. There are project variables that can be defined and used throughout your code. SSH tunnels can also be configured to connect to other services as needed. More details can be found in our docs [here](#).

The screenshot shows the Cloudera Machine Learning interface. On the left is a dark sidebar with navigation links: All Projects, Overview, Sessions, Experiments, Models, Jobs, Applications, Files, Collaborators, and Project Settings (highlighted). The main content area is titled 'Project Settings' and includes tabs for Options, Runtime/Engine, Advanced, SSH Tunnels, and Delete Project. The 'Options' tab is active, showing a form with the following fields:

- Project Name:** A text input field containing 'Telco_churn1'.
- Project Description:** A text input field containing 'Telco_churn from github'.
- Visibility:** Two radio button options:
 - Private:** Selected. Description: 'Only Collaborators can view or edit the project.'
 - Public:** Unselected. Description: 'All authenticated users can view this project. Collaborators can also edit the project.'

At the bottom of the form is a blue button labeled 'Update Project'.

-> Change the Runtime/Engine to Legacy Engine -> Click Save Engine

This screenshot shows the same 'Project Settings' page as before, but with a modal dialog box open in the center. The dialog box is titled 'Legacy Engine' and contains the following text:

Warning: By selecting this option, runtimes will not be available for new sessions and only legacy engines will be available.

Are you sure you want to use this option?

At the bottom of the dialog box are two buttons: 'Cancel' and 'Use Legacy Engine'.

[trial36_admin](#) / [telco_churn1](#) / [Project Settings](#) / [Runtime/Engine](#)

[+](#) [trial36_admin](#)

[← All Projects](#)

[Overview](#)

[Sessions](#)

[Experiments](#)

[Models](#)

[Jobs](#)

[Applications](#)

[Files](#)

[Collaborators](#)

[Project Settings](#)

Project Settings

[Options](#)
[Runtime/Engine](#)
[Advanced](#)
[SSH Tunnels](#)
[Delete Project](#)

Default Engine: ☐ ML Runtime [?](#) ☒ Legacy Engine [?](#)

Engine Image

Select the Docker image that Cloudera Machine Learning should use to run sessions, jobs and applications in this project. If you'd like to use a different image, contact your site administrator.

Changing the project's Docker image will change the Docker image of any job set to "Always use this project's default engine".

[Save Engine](#)

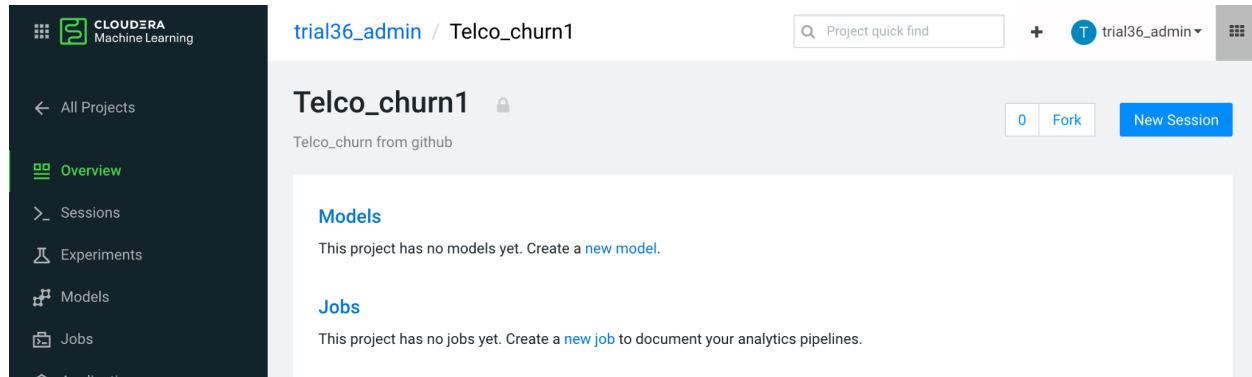
Third-party editors

Cloudera Machine Learning allows you to launch sessions with third-party, web-based editors. To add an editor to the Start New Session menu, first launch a session with the built-in Workbench editor and install the third-party editor of your choice. Then, come back to this page and provide a name for the editor and the command to start the editor server. Ensure that you start the server on the port specified by the `CDSW_APP_PORT` environment variable.

[+ New Editor](#)

Part 3: CML Sessions and Workbench


Sessions allow you to perform actions such as run R or Python code. They also provide access to an interactive command prompt and terminal. Sessions will be built on a specified Engine Image, which is a docker container that is deployed onto the Workspace. In addition you can specify how many resources are used per session. **From the Overview page click on New Session**



Session Name: telco_churn_session_1 Editor: Workbench Kernel: Python 3 Engine Image: Default Resource Profile: 1vCPU/2 GiB Memory

Then select Start Session

Start A New Session

 This Project is configured to run on Legacy Engines.
 You can try the new ML Runtimes by switching over on the [Project Settings](#) page.
 To learn more visit the [documentation](#).

Session Name

Engine

Editor ⓘ

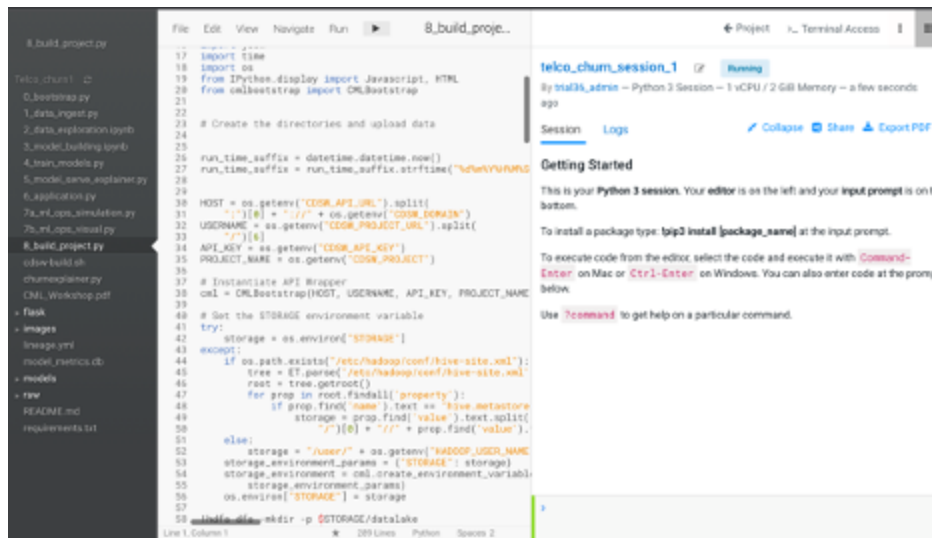
Kernel ⓘ

Engine Image - [Configure](#)

Default engine image - docker.repository.cloudera.com/cloudera/cdsw/engine:15-cml-2021.09-2

Resource Profile

The Workbench is now starting up and deploying a container onto the workspace at this point. Going from left to right you will see the project files, editor pane, and session pane. **Once you see the flashing red line on the bottom of the session pane turn steady green the container has been successfully started.**



Open up the script `8_build_project.py` from the left pane. In the editor pane we are going to select and run the script in several parts. Throughout the script you will see breaks in the code defined by Part 1, Part 2, Part 3, etc.

Part 1: Will install any required packages to be used. As an example, flask is installed as part of the project. A variable is also set as part of the project. Last but not least we are loading a file from the project as a test dataset and moving that to a s3 location.

Select and highlight the contents of part 1, then Right click -> Run Line(s)

```

12 import subprocess
13 import datetime
14 import xml.etree.ElementTree as ET
15 import requests
16 import json
17 import time
18 import os
19 from IPython.display import Javascript, HTML
20 from cmlbootstrap import CMLBootstrap
21
22 # Create the directories and upload data
23
24 run_time_suffix = datetime.datetime.now()
25 run_time_suffix = run_time_suffix.strftime("%d%k%SY%h%M%S")
26
27 HOST = os.getenv("CDSW_API_URL").split(
28     "://")[0] + "://" + os.getenv("CDSW_DOMAIN")
29 USERNAME = os.getenv("CDSW_PROJECT_URL").split(
30     "://")[0]
31 API_KEY = os.getenv("CDSW_API_KEY")
32 PROJECT_NAME = os.getenv("CDSW_PROJECT")
33
34 # Instantiate API Wrapper
35 cml = CMLBootstrap(HOST, USERNAME, API_KEY, PROJECT_NAME)
36
37 # Set the STORAGE environment variable
38 try:
39     storage = os.environ["STORAGE"]
40 except:
41     if os.path.exists("/etc/hadoop/conf/hive-site.xml"):
42         tree = ET.parse("/etc/hadoop/conf/hive-site.xml")
43         root = tree.getroot()
44         for prop in root.findall('property'):
45             if prop.find('name').text == 'hive.metastore.warehouse.dir':
46                 storage = prop.find('value').text.split("/")[2]
47             else:
48                 storage = "/user/" + os.getenv("HADOOP_USER_NAME")
49                 storage_environment_params = {"STORAGE": storage}
50                 storage_environment = cml.create_environment_variable(
51                     storage_environment_params)
52                 os.environ["STORAGE"] = storage
53
54 hdfs dfs -mkdir -p $STORAGE/datalake
55 hdfs dfs -mkdir -p $STORAGE/datalake/data
56 hdfs dfs -mkdir -p $STORAGE/datalake/data/churn
57 hdfs dfs -copyFromLocal /home/cdsw/raw/NA_Fn-UseC_-Telco-Customer-Churn-...csv $STORAGE

```

Part 2: The telco churn dataset is ingested from s3 and a hive table is created using Spark.

Select and highlight the contents of part 2, then Right click -> Run Line(s)

Part 3: Create a CML Job and start the job. A job automates the action of launching an engine, running a script, and tracking the results, all in one batch process. Jobs are created within the purview of a single project and can be configured to run on a recurring schedule. You can customize the engine environment for a job, set up email alerts for successful or failed job runs, and email the output of the job to yourself or a colleague.

Select and highlight the contents of part 3, then Right click -> Run Line(s)

Once the Job is started we will look at what was created. Click the Project button (top right corner of screen).

← Project
>_ Terminal Access
⋮

telco_churn_session_1
Running

By [trial36_admin](#) — Python 3 Session — 1 vCPU / 2 GiB Memory — a few seconds ago

[Session](#)
[Logs](#)
[Spark UI](#)
[Collapse](#)
[Share](#)
[Export PDF](#)

```

"notifications": [
  {
    "user_id": user_obj["id"],
    "user": user_obj,
    "success": False, "failure": False, "ti
  }
],
"recipients": {},
"attachments": [],
"include_logs": True,
"report_attachments": [],
"success_recipients": [],
"failure_recipients": [],
"timeout_recipients": [],
"stopped_recipients": []
}
    
```

Click on **Jobs** to explore the job that was created and details can be found by **Clicking on the Job Name**.

[trial36_admin / Telco_churn1 / Jobs](#)

+
T trial36_admin
⋮

[← All Projects](#)

[Overview](#)

[Sessions](#)

[Experiments](#)

[Models](#)

[Jobs](#)

[Applications](#)

[Files](#)

[Collaborators](#)

[Project Settings](#)

Jobs
New Job

Job Dependencies for Train Model 19012022090853

○
Train Model 19012022090853

+ Add Job Dependency

Name
Runs / Failures
Duration
Status
Latest Run
Actions

Train Model 19012022090853
1 / 0
00:12
Success
3 minutes ago
Run

Creator

Part 4: Using CML, you can create any function within a script and deploy it to a REST API. In a machine learning project, this will typically be a predict function that will accept an input and return a prediction based on the model's parameters.

Select and highlight the contents of part 4, then Right click -> Run Line(s)

When getting to deploying the Model this can take a little time, and is a good spot to stretch the legs and refill your coffee.

The screenshot shows the Cloudera Machine Learning (CML) interface. On the left is a file explorer with a tree view containing files like `8_build_project.py`, `telco_churn1`, `0_bootstrap.py`, `1_data_ingest.py`, `2_data_exploration.ipynb`, `3_model_building.ipynb`, `4_train_models.py`, `5_model_serve_explainer.py`, `6_application.py`, `7a_ml_ops_simulation.py`, `7b_ml_ops_visual.py`, `8_build_project.py`, `__pycache__`, `cdsw-build.sh`, `churnexplainer.py`, `CML_Workshop.pdf`, `flask`, `images`, `lineage.yml`, `model_metrics.db`, `models`, `raw`, `README.md`, and `requirements.txt`.

The main editor displays the `8_build_project.py` script. The code defines a `Model Explainer` function and a `new_model_details` dictionary. It then creates a new model, sets authentication, and enters a loop to wait for the model to be deployed. A context menu is open over the deployment loop, with options: `Run Line(s)`, `Cut`, `Copy`, `Paste`, and `Select All`. The `Run Line(s)` option is selected.

On the right, the `telco_churn_session_1` terminal window is active, showing the output of the script. It displays the model details, including the name `Model Explainer 19012022090853`, namespace `mlx-user-1`, and project `cdp:ml:us-west-1:a5c7fda-39bb-4604-a4f7-49fd919becfd:workspace:8f8aaba8-ff43-4b31-bc3a-c7939926e69e/42879631-20a1-4f19-bb30-2d291a7fe7d5`. The terminal also shows the command `> is_deployed = False` and the start of a `while` loop.

Part 5: Applications give data scientists a way to create ML web applications/dashboards and easily share them with other business stakeholders. Applications can range from single visualizations embedded in reports, to rich dashboard solutions such as Tableau. They can be interactive or non-interactive.

Applications stand alongside other existing forms of workloads in CML (sessions, jobs, experiments, models). Like all other workloads, applications must be created within the scope of a project. Each application is launched within its own isolated engine. Additionally, like models, engines launched for applications do not time out automatically. They will run as long as the web application needs to be accessible by any users and must be stopped manually when needed.

Select and highlight the contents of part 5, then Right click -> Run Line(s)

When deploying the CML Flask Application you will be provided with a URL to follow at the end of your session pane output. **Click to Open Application UI**

The screenshot shows a CloudERA IDE interface. On the left, a file explorer lists various files including 'telco_churn1', '0_bootstrap.py', '1_data_ingest.py', '2_data_exploration.ipynb', '3_model_building.ipynb', '4_train_models.py', '5_model_serve_explainer.py', '6_application.py', '7a_ml_ops_simulation.py', '7b_ml_ops_visual.py', '8_build_project.py', 'requirements.txt', 'CML_Workshop.pdf', 'flask', 'images', 'lineage.yml', 'model_metrics.db', 'models', 'raw', 'README.md', and 'requirements.txt'. The main editor displays the file '8_build_project.py'. The code in this file is a Python script that uses the CloudERA API to create and manage applications. It includes comments and code for creating an application, waiting for it to be deployed, and then opening the application UI. The script is currently at line 289. On the right, the 'Terminal Access' pane shows the output of the script. It indicates that the application was created successfully and provides a URL to open the application UI: 'https://19012822184580.ml-d09f3fad-5b2.cdptrial.g76q-bstg.cloudera.a.site'. The output also shows the application is deployed and the URL to open the application UI is provided.

Within the Flask Application UI you can experiment with some of the parameters that will run against the model to predict how likely a customer is to churn:

Refractor																				
Id	Probability	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges
5540	0.691	Female	No	No	No	3	Yes	No	Fiber	No	Yes	No	Yes	Yes	Yes	Month	Yes	Elect	97.9	315.3
2168	0.661	Male	No	No	No	4	Yes	No	Fiber	No	Yes	No	Yes	No	Yes	Month	No	Elect	86.05	308.1
3973	0.642	Male	No	No	No	28	Yes	Yes	Fiber	No	Yes	Yes	No	Yes	Yes	Month	Yes	Elect	105.7	2979.
53	0.640	Female	Yes	Yes	No	8	Yes	Yes	Fiber	No	Yes	No	No	No	No	Month	Yes	Credi	80.65	633.3
6338	0.549	Male	No	Yes	No	15	Yes	No	Fiber	No	No	No	No	No	No	Month	Yes	Bank	68.6	1108.
4247	0.407	Male	No	Yes	No	29	Yes	Yes	Fiber	Yes	Yes	No	No	No	No	Month	Yes	Elect	84.45	2467.
6100	0.378	Female	No	No	No	2	Yes	Yes	DSL	No	Yes	No	No	No	No	Month	No	Male	54.15	101.6
6294	0.094	Female	No	Yes	No	71	Yes	Yes	Fiber	No	Yes	Yes	Yes	Yes	Yes	Two y	No	Elect	109.2	7707.
5020	0.053	Female	No	No	No	72	Yes	No	Fiber	Yes	Yes	Yes	Yes	Yes	Yes	Two y	Yes	Elect	109.9	7624.
5887	0.002	Male	Yes	Yes	No	71	Yes	Yes	No	No in	No in	No in	No in	No in	No in	Two y	No	Credi	25.45	1789.

Single Prediction View

Churn Probability 0.692

Contract	Month-to-month	0.12	Month-to-month	One year	Two year	
Dependents	No	0	No	Yes		
DeviceProtection	No	0	No	No internet service	Yes	
InternetService	Fiber optic	0.20	DSL	Fiber optic	No	
MonthlyCharges	97.9	-0.24	mean 64.80	min 18.25	max 118.75	<input type="text"/> <input type="button" value="Submit"/>
MultipleLines	No	-0.04	No	No phone service	Yes	
OnlineBackup	Yes	0	No	No internet service	Yes	
OnlineSecurity	No	0	No	No internet service	Yes	
PaperlessBilling	Yes	0	No	Yes		
Partner	No	0	No	Yes		
PaymentMethod	Electronic check	0.04	Bank transfer (automatic)	Credit card (automatic)	Electronic check	Mailed check
PhoneService	Yes	0	No	Yes		
SeniorCitizen	No	-0.04	No	Yes		
StreamingMovies	Yes	0.08	No	No internet service	Yes	
StreamingTV	Yes	0.08	No	No internet service	Yes	
TechSupport	Yes	0	No	No internet service	Yes	
TotalCharges	315.3	-0.12	mean 2283.30	min 18.80	max 8684.80	<input type="text"/> <input type="button" value="Submit"/>
gender	Female	0	Female	Male		
tenure	3	0.29	mean 32.42	min 1.00	max 72.00	<input type="text"/> <input type="button" value="Submit"/>

Part 6: Run the last code snippet in the workbench to complete the project build script (8_build_project.py). This goes through a process of simulating a model that drifts over 1000 calls to the model. The file contains comments with details of how this is done. **Select and highlight the contents of part 6, then Right click -> Run Line(s)**



```

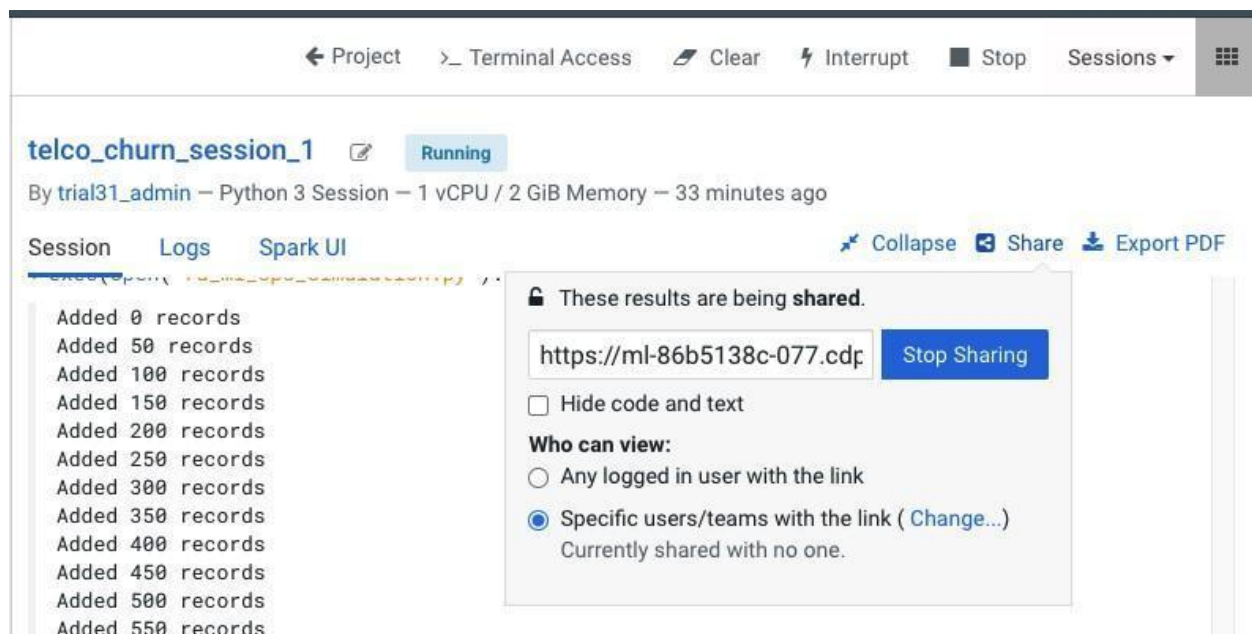
284
285 #####
286 #####
287 #####
288 #####
289 #####
290
291 # This will run the model operations section that makes calls to t
292 # mertics and track metric aggregations
293
294 exec(open("7a_ml_ops_simulation.py").read())
295

```

Line 291, Column 1 ★ 295 Lines Python Spaces 2

You can also share the workbench session with other users if they would like to view results of the code. A URL can be shared out if users outside of CML would like to view code/results of work.

Click on share at the top of the Session pane



← Project >_ Terminal Access ✎ Clear ⚡ Interrupt ■ Stop Sessions ▾

telco_churn_session_1 Running

By trial31_admin — Python 3 Session — 1 vCPU / 2 GiB Memory — 33 minutes ago

Session Logs Spark UI Collapse Share Export PDF

Added 0 records

Added 50 records

Added 100 records

Added 150 records

Added 200 records

Added 250 records

Added 300 records

Added 350 records

Added 400 records

Added 450 records

Added 500 records

Added 550 records

🔒 These results are being shared.

Stop Sharing

☐ Hide code and text

Who can view:

☐ Any logged in user with the link

☒ Specific users/teams with the link ([Change...](#))

Currently shared with no one.

A full CML project should now be running with a Job, Model, and Application deployed! Go back to the project and take a look at the Model and Applications page.

Going to the **Model** page will show you the deployed model. Clicking on the Model will give various tabs of monitoring and statistics in addition to previous deployments of the same model.

On the **Applications** page we can see our running flask app we looked at previously.

Models and Applications will continue to run even if the workbench session is stopped or timeout occurs. These will run on engines as part of the CML workspace

trial31_admin / Telco_churn

Project quick find

trial31_admin

Telco_churn

2 running

0 Fork

New Session

Overview

Sessions

Experiments

Models

Jobs

Applications

Files

Collaborators

Project Settings

Models

Model	Status	Replicas	CPU	Memory	Last Deployed	Actions
Model Explainer 02022021210720	Deployed	1 / 1	1	2.00 GiB	Feb 2, 2021, 03:08 PM	Stop

Jobs

Name	Runs / Failures	Duration	Status	Latest Run	Actions
Train Model 02022021210720	1 / 0	00:12	Success	27 minutes ago	Run

Files

Name	Size	Last Modified
__pycache__	-	27 minutes ago
flask	-	26 minutes ago
images	-	38 minutes ago
models	-	38 minutes ago
raw	-	27 minutes ago
0_bootstrap.py	2.22 kiB	38 minutes ago
1_data_ingest.py	7.41 kiB	38 minutes ago
2_data_exploration.ipynb	602.63 kiB	38 minutes ago
3_model_building.ipynb	71.53 kiB	38 minutes ago
4_train_models.py	9.56 kiB	38 minutes ago
5_model_serve_explainer.py	8.10 kiB	38 minutes ago
6_application.py	7.74 kiB	38 minutes ago
7a_ml_ops_simulation.py	7.96 kiB	26 minutes ago