Machine Learning Zoomcamp FAQ

The purpose of this document is to capture frequently asked technical questions.

We did this for our data engineering course and it worked quite well. Check this document for inspiration on how to structure your questions and answers:

Data Engineering Zoomcamp FAQ

General course-related questions

How do I sign up?

In the course GitHub repository there's a link. Here it is: https://airtable.com/shryxwLd0COOEaqXo

Is it going to be live? When?

The course videos are pre-recorded, you can start watching the course right now.

We will also occasionally have office hours - live sessions where we will answer your questions. The office hours sessions are recorded too.
You can see the office hours as well as the pre-recorded course videos in the course playlist on YouTube.

What if I miss a session?

Everything is recorded, so you won't miss anything. You will be able to ask your questions for office hours in advance and we will cover them during the live stream. Also, you can always ask questions in Slack.

How much theory will you cover?

The bare minimum. The focus is more on practice, and we'll cover the theory only on the intuitive level.: https://mlbookcamp.com/article/python

For example, we won't derive the gradient update rule for logistic regression (there are other great courses for that), but we'll cover how to use logistic regression and make sense of the results.

I don't know math. Can I take the course?

Yes! We'll cover some linear algebra in the course, but in general, there will be very few formulas, mostly code.

I filled the form, but haven't received a confirmation email. Is it normal?

The process is automated now, so you should receive the email eventually. If you haven’t, check your promotions tab in Gmail as well as spam.

If you unsubscribed from our newsletter, you won't get course related updates too.

But don't worry, it’s not a problem. To make sure you don’t miss anything, join the #course-ml-zoomcamp channel in Slack and our telegram channel with announcements. This is enough to follow the course.

How long is the course?

Approximately 4 months, but may take more if you want to do some extra activities (an extra project, an article, etc)

How much time do I need for this course?

Around ~10 hours per week. Timur Kamaliev did a detailed analysis of how much time students of the previous cohort needed to spend on different modules and projects. Full article

Will I get a certificate?

Yes, if you finish at least 2 out of 3 projects, you will get a certificate. This is what it looks like: link. There’s also a version without a robot: link.

How much Python should I know?

Check this article. If you know everything in this article, you know enough. If you don’t, read the article and join the course too :)

I’m new to Slack and can’t find the course channel. Where is it?

Here’s how you join a channel in Slack:
https://slack.com/help/articles/205239967-Join-a-channel

- Click “All channels” at the top of your left sidebar. If you don't see this option, click “More” to find it.
- Browse the list of public channels in your workspace, or use the search bar to search by channel name or description.
- Select a channel from the list to view it.
- Click Join Channel.

Do we need to provide the GitHub link to only our code corresponding to the homework questions?
Yes. You are required to provide the URL to your repo in order to receive a grade

The course has already started. Can I still join it?

Yes, you can. You won’t be able to submit some of the homeworks, but you can still take part in the course.

In order to get a certificate, you need to submit 2 out of 3 course projects. It means that if you join the course at the end of November and manage to work on two projects, you will still be eligible for a certificate.

When does the next iteration start?

The course is available in the self-paced mode too, so you can go through the materials at any time. But if you want to do it as a cohort with other students, the next iteration will happen in September 2023.

Can I submit the homework after the due date?

No, it’s not possible. The form is closed after the due date. But don’t worry, homework is not mandatory for finishing the course.

I just joined. What should I do next? How can I access course materials?

Welcome to the course! Go to the course page (http://mlzoomcamp.com/), scroll down and start going through the course materials.

Click on the links and start watching the videos.

Or you can just use this link: http://mlzoomcamp.com/#syllabus

What are the deadlines in this course?

You can see them here (it's taken from the 2023 cohort page)

What's the difference between the previous iteration of the course (2022) and this one (2023)?

There's not much difference. There was one special module (BentoML) in the previous iteration of the course, but the rest of the modules are the same as in 2022. The homework this year is different.

The course videos are from the previous iteration. Will you release new ones or we'll use the videos from 2021?

We won't re-record the course videos. The focus of the course and the skills we want to teach remained the same, and the videos are still up-to-date.

If you haven't taken part in the previous iteration, you can start watching the videos. It'll be useful for you and you will learn new things. However, we recommend using Python 3.10 now instead of Python 3.8.

Submitting learning in public links

When you post about what you learned from the course on your social media pages, use the tag #mlzoomcamp. When you submit your homework, there's a section in the form for putting the links there. Separate multiple links by any whitespace character (linebreak, space, tab, etc).

For posting the learning in public links, you get extra scores. But the number of scores is limited to 7 points: if you put more than 7 links in your homework form, you'll get only 7 points.

Adding community notes

You can create your own github repository for the course with your notes, homework, projects, etc.
Then fork the original course repo and add a link under the 'Community Notes' section to the notes that are in your own repo.
After that's done, create a pull request to sync your fork with the original course repo.

(By Wesley Barreto)

Computing the hash for the leaderboard and project review

Leaderboard Link:
https://docs.google.com/spreadsheets/d/e/2PACX-1vQzLGpva63gb2rIilFnpZMRSb-buyr5oGh8jmDtIb8DANo4n6hDalra_WRCl4EZwO1JvaC4UIS62n5h/pubhtml

Python Code:

```
from hashlib import sha1

def compute_hash(email):
    return sha1(email.lower().encode('utf-8')).hexdigest()
```

You need to call the function as follows:

```
print(compute_hash('YOUR_EMAIL_HERE'))
```

The quotes are required to denote that your email is a string.

(By Wesley Barreto)

1. Introduction to Machine Learning

wget is not recognized as an internal or external command

If you get “wget is not recognized as an internal or external command”, you need to install it.

On Ubuntu, run

```
sudo apt-get install wget
```

On Windows, the easiest way to install wget is to use Chocolatey:

```
choco install wget
```

Or you can download a binary from here and put it to any location in your PATH (e.g. C:/tools/)

On Mac, the easiest way to install wget is to use brew.

Brew install wget

Alternatively, you can use a Python wget library, but instead of simply using “wget” you’ll need to use

```
python -m wget
```

You need to install it with pip first:

```
pip install wget
```

Using Anaconda:

```
pip install wget
```

And then in your python code, for example in your jupyter notebook, use:

```
import wget
wget.download(“URL”)
```

This should download whatever is at the URL in the same directory as your code.

(Memoona Tahira)

```
url =
"https://raw.githubusercontent.com/alexeygrigorev/datasets/master/housing.csv
"
df = pd.read_csv(url)
```

Valid URL schemes include http, ftp, s3, gs, and file.

Retrieving csv inside notebook

You can use

!wget https://raw.githubusercontent.com/alexeygrigorev/datasets/master/housing.csv

To download the data too. The exclamation mark !, lets you execute shell commands inside your notebooks.

Windows WSL and VS Code

If you have a Windows 11 device and would like to use the built in WSL to access linux you can use the Microsoft Learn link Set up a WSL development environment | Microsoft Learn. To connect this to VS Code download the Microsoft verified VS Code extension ‘WSL’ this will allow you to remotely connect to your WSL Ubuntu instance as if it was a virtual machine.

(Tyler Simpson)

Uploading the homework to Github

This is my first time using Github to upload a code. I was getting the below error message when I type

```
git push -u origin master:
```

```
error: src refspec master does not match any
error: failed to push some refs to
'https://github.com/XXXXXX/1st-Homework.git'
```

Solution:

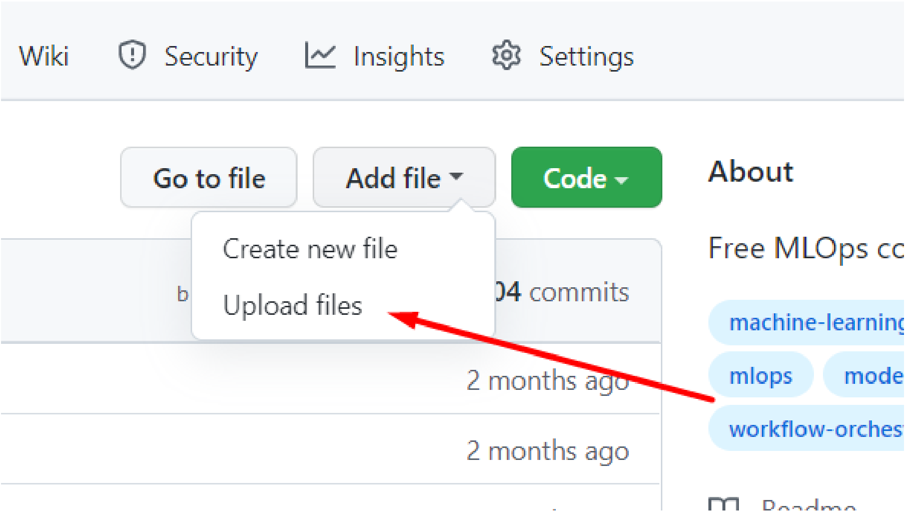
The error message got fixed by running below commands:

```
git commit -m "initial commit"
git push origin main
```

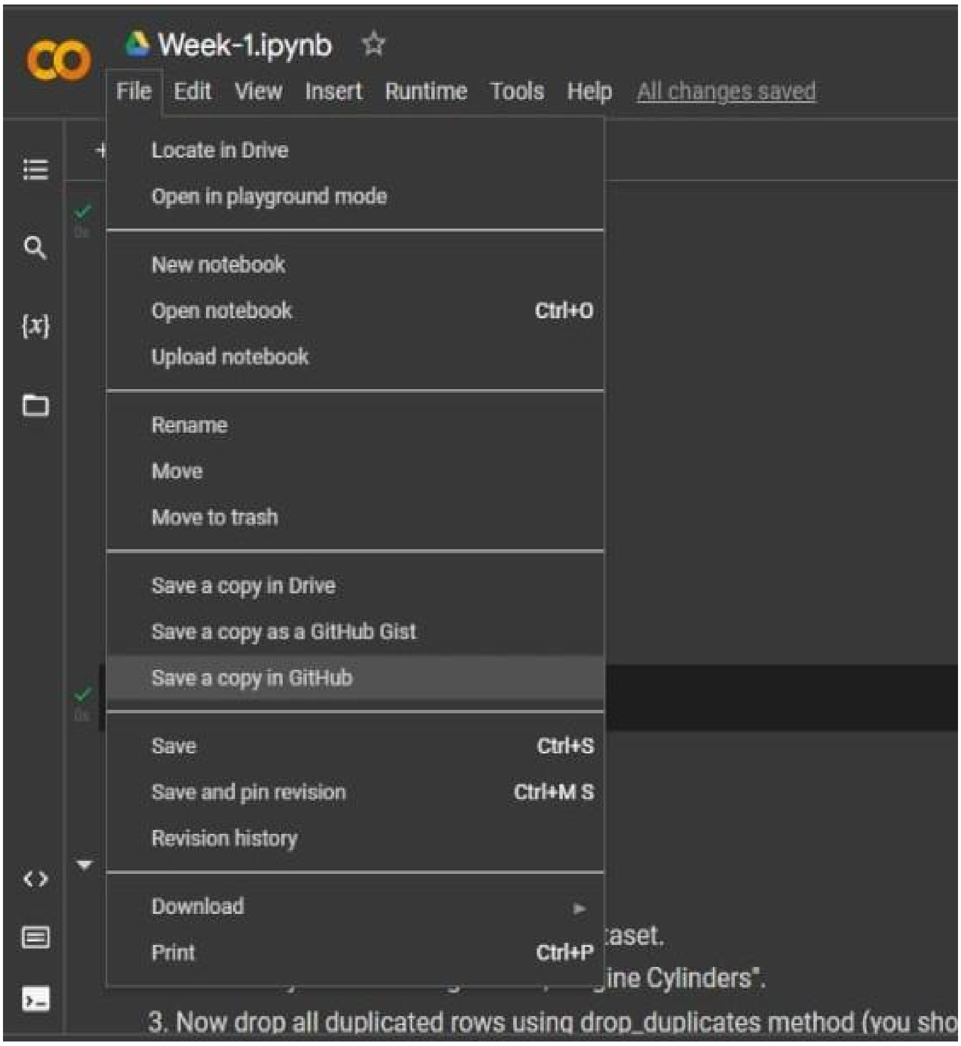
If this is your first time to use Github, you will find a great & straightforward tutorial in this link https://dennisivy.com/github-quickstart

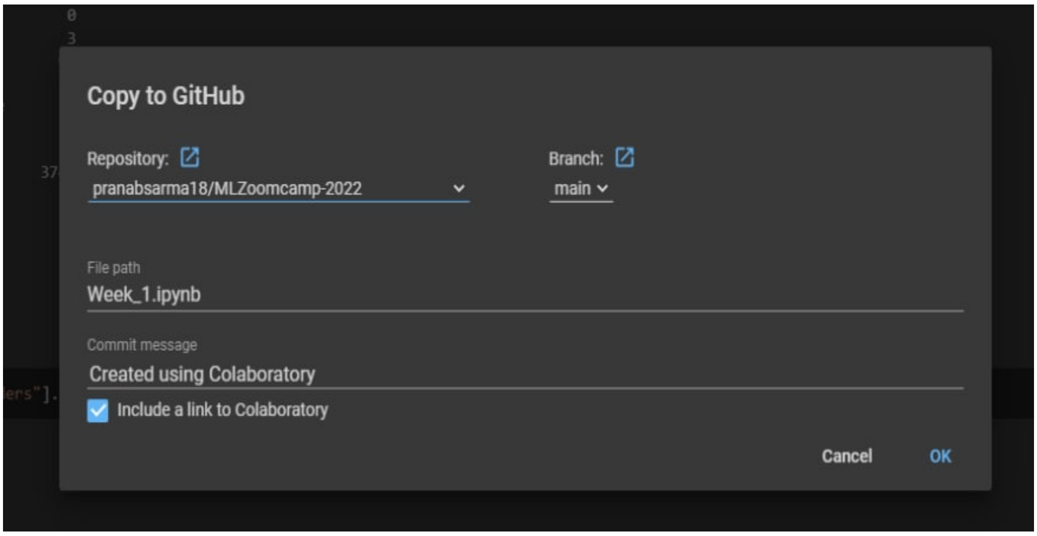
(Asia Saeed)

You can also use the “upload file” functionality from GitHub for that



If you write your code on Google colab you can also directly share it on your Github.





Copy to GitHub

Repository: pranabsarma18/MLZoomcamp-2022

Branch: main

File path: Week_1.ipynb

Commit message: Created using Colaboratory

☑ Include a link to Colaboratory

(By Pranab Sarma)

Singular Matrix Error

I'm trying to invert the matrix but I got error that the matrix is singular matrix

The singular matrix error is caused by the fact that not every matrix can be inverted. In particular, in the homework it happens because you have to pay close attention when dealing with multiplication (the method .dot) since multiplication is not commutative! X.dot(Y) is not necessarily equal to Y.dot(X), so respect the order otherwise you get the wrong matrix.

Conda is not an internal command

I have a problem with my terminal. Command

```
conda create -n ml-zoomcamp python=3.9
```

doesn’t work.

If you’re on Windows and just installed Anaconda, you can use Anaconda’s own terminal called “Anaconda Prompt”.

If you don’t have Anaconda or Miniconda, you should install it first

(Tatyana Mardvilko)

Reading csv files with pandas straight from the url

We can read the csv files without the need to download it manually and load it straight to pandas.
The downside is if the source link is removed, we lose access to the file

```
url =
“https://raw.githubusercontent.com/alexeygrigorev/datasets/mas
ter/housing.csv”
df = pd.read_csv(url)
```

(Nikki Satmaka)

Downloading and reading csv files with pandas

How to download the assignment file?

Solution:

- Download it manually and save in your local directory
- Read it with `df = pd.read_csv('data.csv')`
- Then follow as in the video with the title "ML Zoomcamp 1.9 - Introduction to Pandas".

(Bhaskar Sarma)

Read-in the File in Windows OS

How do I read the dataset with Pandas in Windows?

I used the code below but not working

```
df = pd.read_csv('C:\Users\username\Downloads\data.csv')
```

Unlike Linux/Mac OS, Windows uses the backslash (\) to navigate the files that cause the conflict with Python. The problem with using the backslash is that in Python, the '\' has a purpose known as an escape sequence. Escape sequences allow us to include special characters in strings, for example, "\n" to add a new line or "\t" to add spaces, etc. To avoid the issue we just need to add "r" before the file path and Python will treat it as a literal string (not an escape sequence).

Here's how we should be loading the file instead:

```
df = pd.read_csv(r'C:\Users\username\Downloads\data.csv')
```

(Muhammad Awon)

'403 Forbidden' error message when you try to push to a GitHub repository

Type the following command:

```
git config -l | grep url
```

The output should look like this:

```
remote.origin.url=https://github.com/github-username/github-repository-name.git
```

Change this to the following format and make sure the change is reflected using command in step 1:

```
git remote set-url origin
"https://github-username@github.com/github-username/github-repository-name.git"
```

(Added by

Dheeraj Karra)

Fatal: Authentication failed for 'https://github.com/username

I had a problem when I tried to push my code from Git Bash:

```
remote: Support for password authentication was removed on
August 13, 2021.
remote: Please see
https://docs.github.com/en/get-started/getting-started-with-git/about-remote-repositories#cloning-with-https-urls for
information on currently recommended modes of authentication.
fatal: Authentication failed for 'https://github.com/username
```

Solution:
Create a personal access token from your github account and use it when you make a push of your last changes.

https://docs.github.com/en/authentication/connecting-to-github-with-ssh/generating-a-new-ssh-key-and-adding-it-to-the-ssh-agent

Bruno Bedón

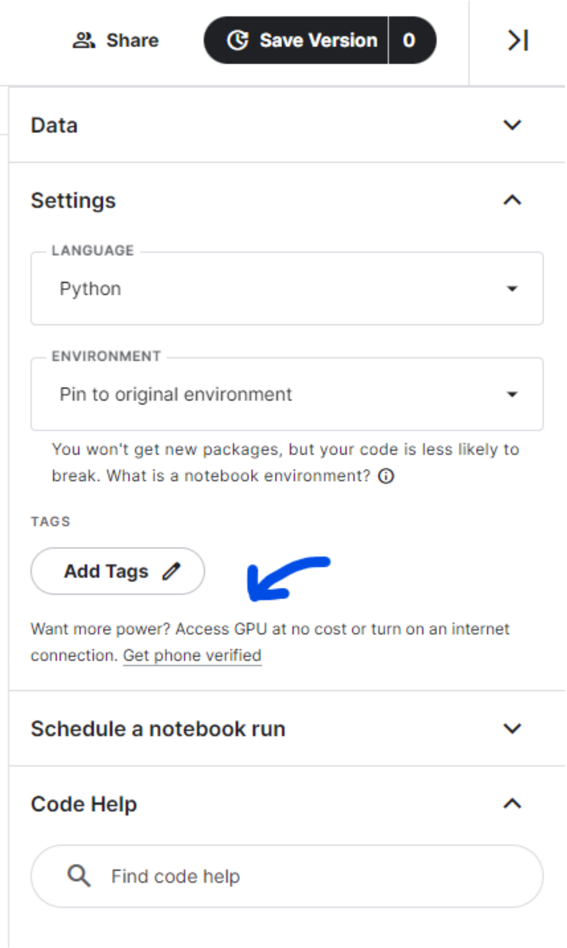
wget: unable to resolve host address 'raw.githubusercontent.com'

In Kaggle, when you are trying to !wget a dataset from github (or any other public repository/location), you get the following error:

```
Getting  this error while trying to import data- !wget
https://raw.githubusercontent.com/alexeygrigorev/datasets/master/housing.csv
--2022-09-17 16:55:24--
https://raw.githubusercontent.com/alexeygrigorev/datasets/master/housing.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... failed:
Temporary failure in name resolution.
wget: unable to resolve host address 'raw.githubusercontent.com'
```

Solution:

In your Kaggle notebook settings, turn on the Internet for your session. It's on the settings panel, on the right hand side of the Kaggle screen. You'll be asked to verify your phone number so Kaggle knows you are not a bot.



LANGUAGE: Python

ENVIRONMENT: Pin to original environment

You won't get new packages, but your code is less likely to break. What is a notebook environment?

TAGS

Want more power? Access GPU at no cost or turn on an internet connection. Get phone verified

Setting up an environment using VS Code

I found this video quite helpful:
Creating Virtual Environment for Python from VS Code

[Native Jupiter Notebooks support in VS Code] In VS Code you can also have a native Jupiter Notebooks support, i.e. you do not need to open a web browser to code in a Notebook. If you have port forwarding enabled + run a 'jupyter notebook ' command from a remote machine + have a remote connection configured in .ssh/config (as Alexey's video suggests) - VS Code can execute remote Jupyter Notebooks files on a remote server from your local machine: https://code.visualstudio.com/docs/datascience/jupyter-notebooks .

[Git support from VS Code] You can work with Github from VSCode - staging and commits are easy from the VS Code's UI: https://code.visualstudio.com/docs/sourcecontrol/overview

(Added by Ivan Brigida)

Conda Environment Setup

With regards to creating an environment for the project, do we need to run the command "conda create -n ...." and "conda activate ml-zoomcamp" everytime we open vs code to work on the project?

Answer:

"conda create -n ...." is just run the first time to create the environment. Once created, you just need to run "conda activate ml-zoomcamp" whenever you want to use it.

(Added by Wesley Barreto)

conda env export > environment.yml will also allow you to reproduce your existing environment in a YAML file. You can then recreate it with ***conda env create -f environment.yml***

Floating Point Precision

I was doing Question 7 from Week1 Homework and with step6: Invert XTX, I created the inverse. Now, an inverse when multiplied by the original matrix should return in an Identity matrix. But when I multiplied the inverse with the original matrix, it gave a matrix like this:

```
Inverse * Original:
[[ 1.00000000e+00 -1.38777878e-16]
 [ 3.16968674e-13  1.00000000e+00]]
```

Solution:

It's because floating point math doesn't work well on computers as shown here:

https://stackoverflow.com/questions/588004/is-floating-point-math-broken

(Added by Wesley Barreto)

What does pandas.DataFrame.info() do?

Answer:
It prints the information about the dataset like:

- Index datatype
- No. of entries
- Column information with not-null count and datatype
- Memory usage by dataset

We use it as:

```
df.info()
```

(Added by Aadarsha Shrestha & Emoghena Itakpe)

NameError: name 'np' is not defined

Pandas and numpy libraries are not being imported

```
NameError: name 'np' is not defined
NameError: name 'pd' is not defined
```

If you're using numpy or pandas, make sure you use the first few lines before anything else.

```
import pandas as pd
import numpy as np
```

Added by Manuel Alejandro Aponte

How to select column by dtype

What if there were hundreds of columns? How do you get the columns only with numeric or object data in a more concise way?

```
df.select_dtypes(include=np.number).columns.tolist()
df.select_dtypes(include='object').columns.tolist()
```

Added by Gregory Morris

How to identify the shape of dataset in Pandas

There are many ways to identify the shape of dataset, one of them is using .shape attribute!

```
df.shape
df.shape[0] # for identify the number of rows
df.shape[1 # for identify the number of columns
```

Added by Radikal Lukafiardi

How to avoid Value errors with array shapes in homework?

First of all use np.dot for matrix multiplication. When you compute matrix-matrix multiplication you should understand that order of multiplying is crucial and affects the result of the multiplication!

Dimension Mismatch
To perform matrix multiplication, the number of columns in the 1st matrix should match the number of rows in the 2nd matrix. You can rearrange the order to make sure that this satisfies the condition.

Added by Leah Gotladera

Question 5: Select average median_house_valve based on the houses located near the bay.Faced some difficulty to get the result.

One of the way to get the average median_house_valve for the houses located in the bay is to use the group by command:

```
dfh.groupby('ocean_proximity').median_house_value.mean()
```

Added by Krishna Anand

Question 5: Select average median_house_value based on the houses located near the bay.

Using group by command will actually give you additional information, as it will give average median_house_value prices for all the categories in ocean_proximity column. One precise way to get the required information would be to filter the data for Near Bay area and then take the mean.

Added by Hrithik Advani

Question 7: Select all the options located on islands. Confused me a little at first.

This sentence actually means you have to focus on the column "ocean_proximity". Check out the values of this column.

Added by Piyush Sonewar

Question 7: Mathematical formula for linear regression

In Question 7 we are asked to calculate

- $XTX = X^T \cdot X$
- $XTX_{inv} = (X^T \cdot X)^{-1}$
- $w = (X^T \cdot X)^{-1} X^T \cdot y$

The initial problem $w = X^{-1} \cdot y$ can be solved by this, where a Matrix X is multiplied by some unknown weights w resulting in the target y.

Added by Sylvia Schmitt

Question 7: FINAL MULTIPLICATION not having 5 column

This is most likely that you interchanged the first step of the multiplication

You used $XTX = X.\ X^T$ instead of $XTX = X^T \cdot X$

Added by Emmanuel Ikpesu

Question 7: Multiplication operators.

Note, that matrix multiplication (matrix-matrix, matrix-vector multiplication) can be written as * operator in some sources, but performed as @ operator or np.matmul() via numpy. * operator performs element-wise multiplication (Hadamard product).
numpy.dot() or ndarray.dot() can be used, but for matrix-matrix multiplication @ or np.matmul() is preferred (as per numpy doc).
If multiplying by a scalar numpy.multiply() or * is preferred.

Added by Andrii Larkin

Error launching Jupyter notebook

If you face an error kind of ImportError: cannot import name 'contextfilter' from 'jinja2' (anaconda\lib\site-packages\jinja2\__init__.py) when launching a new notebook for a brand new environment.

Switch to the main environment and run "pip install nbconvert --upgrade".

Added by George Chizhmak

wget https://raw.githubusercontent.com/alexeygrigorev/datasets/master/housing.csv hangs on MacOS Ventura M1

If you face this situation and see IPv6 addresses in the terminal, go to your System Settings > Network > *your network connection* > Details > **Configure IPv6** > set to **Manually** > OK. Then try again

In case you are using mac os and having trouble with WGET

Wget doesn't ship with macOS, so there are other alternatives to use.
No worries, we got curl:

1. example:

```
curl -o ./housing.csv
https://raw.githubusercontent.com/alexeygrigorev/datasets/master/housing.csv
```

2. Explanations:
   a. curl: a utility for retrieving information from the internet.
   b. -o: Tell it to store the result as a file.
   c. filename: You choose the file's name.
   d. Links: Put the web address (URL) here, and cURL will extract data from it and save it under the name you provide.

More about it at:
Curl Documentation

Added by David Espejo

cHow to output only a certain number of decimal places

You can use round() function or f-strings

```
 round(number, 4)  - this will round number up to 4 decimal
places
 print(f'Average mark for the Homework is {avg:.3f}') - using
F string
```

Also there is pandas.Series. round idf you need to round values in the whole Series
Please check the documentation
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.round.html#pandas.Series.round

Added by Olga Rudakova

Question 4 : How many unique values does the ocean proximity column have?

I found a very easy way to answer this question. Thought others may find it useful.

data = pd.read_csv('housing.csv')
len(data["ocean_proximity"].value_counts())

Added by Kailash

Question 4 : How many unique values does the ocean proximity column have?

This is an add on to the above answer. I find the method below more easier. Others may find this helpful.

```
data = pd.read_csv('housing.csv')
data["ocean_proximity"].nunique()
```

Added by Hrithik Advani

2. Machine Learning for Regression

How do I get started with Week 2?

Here are the crucial links for this Week 2 that starts September 18, 2023

- Ask questions for Live Sessions: https://app.sli.do/event/vsUpjYsayZ8A875Hq8dpUa/live/questions
- Calendar for weekly meetings: https://calendar.google.com/calendar/u/0/r?cid=cGtjZ2tkbGc1OG9yb2lxa2Vwc2g4YXMzMmNAZ3JvdXAuY2FsZW5kYXIuZ29vZ2xlLmNvbQ&pli=1
- Week 2 HW: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/02-regression/homework.md
- Submit HW Week 2: https://docs.google.com/forms/d/e/1FAIpQLSf8eMtnErPFqzzFsEdLap_GZ2sMih-H-Y7F_IuPGqt4fOmOJw/viewform (also available at the bottom of the above link)
- All HWs: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/
- GitHub for theory: https://github.com/alexeygrigorev/mlbookcamp-code/tree/master/course-zoomcamp
- Youtube Link: 2.X --- https://www.youtube.com/watch?v=vM3SqPNlStE&list=PL3MmuxUbc_hIhxl5Ji8t4O6lPAOpHaCLR&index=12

- FAQs: https://docs.google.com/document/d/1LpPanc33QJJ6BSsyxVg-pWNMplal84TdZtq10naIhD8/edit#heading=h.lpz96zg7l47j

~~Nukta Bhatia~~

LinAlgError: Singular matrix

It’s possible that when you follow the videos, you’ll get a Singular Matrix error. We will explain why it happens in the Regularization video. Don’t worry, it’s normal that you have it.

California housing dataset

You can find a detailed description of the dataset ere https://inria.github.io/scikit-learn-mooc/python_scripts/datasets_california_housing.html

KS

Getting NaNs after applying .mean()

I was using for loops to apply rmse to list of y_val and y_pred. But the resulting rmse is all nan.

I found out that the problem was when my data reached the mean step after squaring the error in the rmse function. Turned out there were nan in the array, then I traced the problem back to where I first started to split the data: I had only use fillna(0) on the train data, not on the validation and test data. So the problem was fixed after I applied fillna(0) to all the dataset (train, val, test). Voila, my for loops to get rmse from all the seed values work now.

Added by Sasmito Yudha Husada

Target variable transformation

Why should we transform the target variable to logarithm distribution? Do we do this for all machine learning projects?

Only if you see that your target is highly skewed. The easiest way to evaluate this is by plotting the distribution of the target variable.

This can help to understand skewness and how it can be applied to the distribution of your data set.

https://en.wikipedia.org/wiki/Skewness

Pastor Soto

Null column is appearing even if I applied .fillna()

When creating a duplicate of your dataframe by doing the following:

```
X_train = df_train
X_val = df_val
```

You’re still referencing the original variable, this is called a shallow copy. You can make sure that no references are attaching both variables and still keep the copy of the data do the following to create a deep copy:

```
X_train = df_train.copy()
X_val = df_val.copy()
```

Added by Ixchel García

Can I use Scikit-Learn’s train_test_split for this week?

Yes, you can. Here we implement it ourselves to better understand how it works, but later we will only rely on Scikit-Learn’s functions. If you want to start using it earlier — feel free to do it

Can I use LinearRegression from Scikit-Learn for this week?

Yes, you can. We will also do that next week, so don’t worry, you will learn how to do it.

Corresponding Scikit-Learn functions

I wanted to know which Scikit-Learn functions are the equivalents for the linear regression implemented

Random seed 42

One of the questions on the homework calls for using a random seed of 42. When using 42, all my missing values ended up in my training dataframe and not my validation or test dataframes, why is that?

The purpose of the seed value is to randomly generate the proportion split. Using a seed of 42 ensures that all learners are on the same page by getting the same behavior (in this case, all missing values ending up in the training dataframe). If using a different seed value (e.g. 9), missing values will then appear in all other dataframes.

The answer I get for one of the homework questions doesn't match any of the options. What should I do?

That’s normal. We all have different environments: our computers have different versions of OS and different versions of libraries — even different versions of Python.

If it’s the case, just select the option that’s closest to your answer

Meaning of mean in homework 2, question 3

In question 3 of HW02 it is mentioned: ‘For computing the mean, use the training only’. What does that mean?

It means that you should use only the training data set for computing the mean, not validation or test data set. This is how you can calculate the mean

```
df_train['column_name'].mean( )
```

Another option:

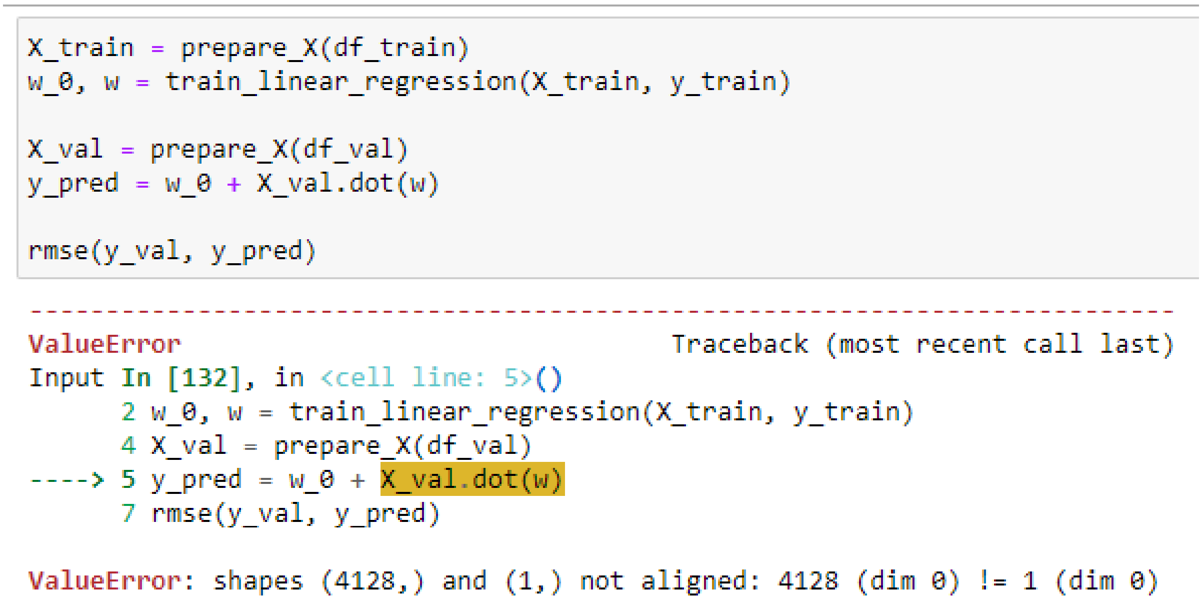
```
df_train[‘column_name’].describe()
```

(Bhaskar Sarma)

When should we transform the target variable to logarithm distribution?

When the target variable has a long tail distribution, like in prices, with a wide range, you can transform the target variable with `np.log1p()` method, but be aware if your target variable has negative values, this method will not work

ValueError: shapes not aligned

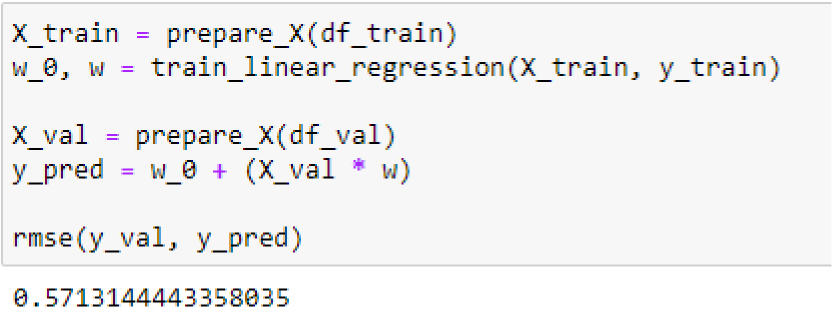


```
X_train = prepare_X(df_train)
w_0, w = train_linear_regression(X_train, y_train)

X_val = prepare_X(df_val)
y_pred = w_0 + X_val.dot(w)

rmse(y_val, y_pred)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Input In [132], in <cell line: 5>()
      2 w_0, w = train_linear_regression(X_train, y_train)
      4 X_val = prepare_X(df_val)
----> 5 y_pred = w_0 + X_val.dot(w)
      7 rmse(y_val, y_pred)

ValueError: shapes (4128,) and (1,) not aligned: 4128 (dim 0) != 1 (dim 0)
```

If we try to perform an arithmetic operation between 2 arrays of different shapes or different dimensions, it throws an error like *operands could not be broadcast together with shapes.* There are some scenarios when broadcasting can occur and when it fails.

If this happens sometimes we can use * operator instead of dot() method to solve the issue. So that the error is solved and also we get the dot product.



```
X_train = prepare_X(df_train)
w_0, w = train_linear_regression(X_train, y_train)

X_val = prepare_X(df_val)
y_pred = w_0 + (X_val * w)

rmse(y_val, y_pred)
```

```
0.5713144443358035
```

(Santhosh Kumar)

How to copy a dataframe without changing the original dataframe?

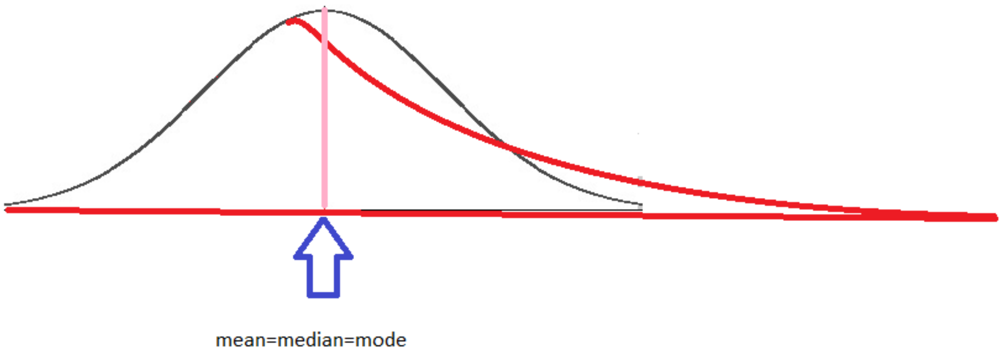
Copy of a dataframe is made with `X_copy = X.copy()`.

This is called creating a deep copy. Otherwise it will keep changing the original dataframe if used like this: `X_copy = X`.

Any changes to X_copy will reflect back to X. This is not a real copy, instead it is a “view”.

(Memoona Tahira)

What does ‘long tail’ mean?



One of the most important characteristics of the normal distribution is that mean=median=mode, this means that the most popular value, the mean of the distribution and 50% of the sample are under the same value, this is equivalent to say that the area under the curve (black) is the same on the left and on the right. The long tail (red curve) is the result of having a few observations with high values, now the behaviour of the distribution changes, first of all, the area is different on each side and now the mean, median and mode are different. As a consequence, the mean is no longer representative, the range is larger than before and the probability of being on the left or on the right is not the same.

(Tatiana Dávila)

What is standard deviation?submit

In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range. *[Wikipedia]* The formula to calculate standard deviation is:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

(Aadarsha Shrestha)

3. Machine Learning for Classification

How do I get started with Week 3?

- Ask questions for Live Sessions: https://app.sli.do/event/vsUpjYsayZ8A875Hq8dpUa/live/questions

- Calendar for weekly meetings: https://calendar.google.com/calendar/u/0/r?cid=cGtjZ2tkbGc1OG9yb2lxa2Vwc2g4YXMzMmNAZ3JvdXAuY2FsZW5kYXIuZ29vZ2xlLmNvbQ&pli=1
- Week 3 HW: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/03-classification/homework.md
- Submit HW Week 3: https://docs.google.com/forms/d/e/1FAIpQLSeXS3pqsv_smRkYmVx-7g6KIZDnG29g2s7pdHo-ASKNqtfRFQ/viewform
- All HWs: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/
- Evaluation Matrix: https://docs.google.com/spreadsheets/d/e/2PACX-1vQCwqAtkjl07MTW-SxWUK9GUvMQ3Pv_fF8UadcuIYLgHa0PlNu9BRWtfLgivI8xSCncQs82HDwGXSm3/pubhtml
- GitHub for theory: https://github.com/alexeygrigorev/mlbookcamp-code/tree/master/course-zoomcamp
- Youtube Link: 3.X --- https://www.youtube.com/watch?v=0Zw04wdeTQo&list=PL3MmuxUbc_hIhxl5Ji8t4O6lPAOpHaCLR&index=29

~~Nukta Bhatia~~

Why did we change the targets to binary format when calculating mutual information score in the homework?

Solution: Mutual Information score calculates the relationship between categorical variables or discrete variables. So in the homework, because the target which is median_house_value is continuous, we had to change it to binary format which in other words, makes its values discrete as either 0 or 1. If we allowed it to remain in the continuous variable format, the mutual information score could be calculated, but the algorithm would have to divide the continuous variables into bins and that would be highly subjective. That is why continuous variables are not used for mutual information score calculation.

—Odimegwu David—-

What data should we use for correlation matrix

Q2 asks about correlation matrix and converting median_house_value from numeric to binary. Just to make sure here we are only dealing with df_train not df_train_full, right? As the question explicitly mentions the train dataset.

Yes. I think it is only on df_train. The reason behind this is that df_train_full also contains the validation dataset, so at this stage we don't want to make conclusions based on the validation data, since we want to test how we did without using that portion of the data.

Pastor Soto

What data should be used for EDA?

Should we perform EDA on the base of train or train+validation or train+validation+test dataset?

It's indeed good practice to only rely on the train dataset for EDA. Including validation might be okay. But we aren't supposed to touch the test dataset, even just looking at it isn't a good idea. We indeed pretend that this is the future unseen data

Alena Kniazeva

Fitting DictVectorizer on validation

Validation dataset helps to validate models and prediction on unseen data. This helps get an estimate on its performance on fresh data. It helps optimize the model.

Edidiong Esu

Below is an extract of Alexey's book explaining this point. Hope is useful

When we apply the fit method, this method is looking at the content of the df_train dictionaries we are passing to the DictVectorizer instance, and fit is figuring out (training) how to map the values of these dictionaries. If categorical, applies one-hot encoding, if numerical it will leave it as it is.

With this context, if we apply the fit to the validation model, we are "giving the answers" and we are not letting the "fit" do its job for data that we haven't seen. By not applying the fit to the validation model we can know how well it was trained.

Below is an extract of Alexey's book explaining this point.

Humberto Rodriguez

There is no need to initialize another instance of dictvectorizer after fitting it on the train set as it will overwrite what it learnt from being fit on the train data.

The correct way is to fit_transform the train set, and only transform the validation and test sets.

Memoona Tahira

Feature elimination

For Q5 in homework, should we calculate the smallest difference in accuracy in real values (i.e. -0.001 is less than -0.0002) or in absolute values (i.e. 0.0002 is less than 0.001)?

We should select the “smallest” difference, and not the “lowest”, meaning we should reason in absolute values.

If the difference is negative, it means that the model actually became better when we removed the feature.

FutureWarning: Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will be removed in 1.2

Instead use the method “.get_feature_names_out()” from DictVectorizer function and the warning will be resolved , but we need not worry about the waning as there won't be any warning

Santhosh Kumar

Logistic regression crashing Jupyter kernel

Fitting the logistic regression takes a long time / kernel crashes when calling predict() with the fitted model.

Make sure that the target variable for the logistic regression is binary.

Konrad Muehlberg

How to select the alpha parameter in Q6

Question: Regarding RMSE, how do we decide on the correct score to choose? In the study group discussion about week two homework, all of us got it wrong and one person had the lowest score selected as well.

Answer: You need to find RMSE for each alpha. If RMSE scores are equal, you will select the lowest alpha which is 0.

Asia Saeed

Second variable that we need to use to calculate the mutual information score

Question: Could you please help me with HW3 Q3: "Calculate the mutual information score with the (binarized) price for the categorical variable that we have. Use the training set only." What is the second variable that we need to use to calculate the mutual information score?

Answer: You need to calculate the mutual info score between the binarized price (above_average) variable & ocean_proximity, the only original categorical variable in the dataset.

Asia Saeed

Features for homework Q5

Do we need to train the model only with the features: total_rooms, total_bedrooms, population and households? or with all the available features and then pop once at a time each of the previous features and train the model to make the accuracy comparison?

You need to create a list of all features in this question and evaluate the model one time to obtain the accuracy, this will be the original accuracy, and then remove one feature each time, and in each time, train the model, find the accuracy and the difference between the original accuracy and the found accuracy. Finally, find out which feature has the smallest absolute accuracy difference.

While calculating differences between accuracy scores while training on the whole model, versus dropping one feature at a time and comparing its accuracy to the model to judge impact of the feature on the accuracy of the model, do we take the smallest difference or smallest absolute difference?

Since order of subtraction between the two accuracy scores can result in a negative number, we will take its absolute value as we are interested in the **smallest** value difference, not the **lowest** difference value. Case in point, if difference is -4 and -2, the smallest difference is abs(-2), and not abs(-4)

What is the difference between OneHotEncoder and DictVectorizer?

Both work in similar ways. That is, to convert categorical features to numerical variables for use in training the model. But the difference lies in the input. OneHotEncoder uses an array as input while DictVectorizer uses a dictionary.

Both will produce the same result. But when we use OneHotEncoder, features are sorted alphabetically. When you use DictVectorizer you stack features that you want.

Tanya Mard

Use of random seed in HW3

For the test_train_split question on week 3's homework, are we supposed to use 42 as the random_state in both splits or only the 1st one?

Answer: for both splits random_state = 42 should be used

(Bhaskar Sarma)

Correlation before or after splitting the data

Should correlation be calculated after splitting or before splitting. And lastly I know how to find the correlation but how do i find the two most correlated features.

Answer: Correlation matrix of your train dataset. Thus, after splitting. Two most correlated features are the ones having the highest correlation coefficient in terms of absolute values.

4. Evaluation Metrics for Classification

How do I get started with Week 4?

- Ask questions for Live Sessions: https://app.sli.do/event/vsUpjYsayZ8A875Hq8dpUa/live/questions
- Calendar for weekly meetings: https://calendar.google.com/calendar/u/0/r?cid=cGtjZ2tkbGc1OG9yb2lxa2Vwc2g4YXMzMmNAZ3JvdXAuY2FsZW5kYXIuZ29vZ2xlLmNvbQ&pli=1
- Week 4 HW: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/04-evaluation/homework.md
- Submit HW Week 3: https://docs.google.com/forms/d/e/1FAIpQLSemEBRhhN1RNiXW-dSwO4b7AGBg5x2kor4-UiCHthzMbo6q9g/viewform
- All HWs: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/
- Evaluation Matrix: https://docs.google.com/spreadsheets/d/e/2PACX-1vQCwqAtkjl07MTW-SxWUK9GUvMQ3Pv_fF8UadcuIYLgHa0PlNu9BRWtfLgivI8xSCncQs82HDwGXSm3/pubhtml
- GitHub for theory: https://github.com/alexeygrigorev/mlbookcamp-code/tree/master/course-zoomcamp
- YouTube Link: 4.X --- https://www.youtube.com/watch?v=gmg5jw1bM8A&list=PL3MmuxUbc_hIhxl5Ji8t4O6lPAOpHaCLR&index=40

~~Nukta Bhatia~~

How to get all classification metrics?

How to get classification metrics - precision, recall, f1 score, accuracy simultaneously

Use classification_report from sklearn. For more info check here.

Abhishek N

Multiple thresholds for Q4

I am getting multiple thresholds with the same F1 score, does this indicate I am doing something wrong or is there a method for choosing? I would assume just pick the lowest?

Choose the one closest to any of the options

Added by Azeez Enitan Edunwale

ValueError: This solver needs samples of at least 2 classes in the data, but the data contains only one class: 0

Solution description: duplicating the

```
df.churn = (df.churn == 'yes').astype(int)
```

This is causing you to have only 0's in your churn column. In fact, match with the error you are getting: ValueError: This solver needs samples of at least 2 classes in the data, but the data contains only one class: 0.
It is telling us that it only contains 0's.
Delete one of the below cells and you will get the accuracy

Humberto Rodriguez

I’m not getting the exact result in homework

That’s fine, use the closest option

Use AUC to evaluate feature importance of numerical variables

Check the solutions from the 2021 iteration of the course. You should use roc_auc_score.

What does KFold do?

What does this line do?

```
KFold(n_splits=n_splits, shuffle=True, random_state=1)
```

If I do it inside the loop [0.01, 0.1, 1, 10] or outside the loop in Q6, HW04 it doesn't make any difference to my answers. I am wondering why and what is the right way, although it doesn't make a difference!

Did you try using a different random_state? From my understanding, KFold just makes N (which is equal to n_splits) separate pairs of datasets (train+val).

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

In my case changing random state changed results

(Arthur Minakhmetov)

Changing the random state makes a difference in my case too, but not whether it is inside or outside the for loop. I think I have got the answer. kFold = KFold(n_splits=n_splits, shuffle = True, random_state = 1) is just a generator object and it contains only the information n_splits, shuffle and random_state. The k-fold splitting actually happens in the next for loop for train_idx, val_idx in kFold.split(df_full_train): . So it doesn't matter where we generate the object, before or after the first loop. It will generate the same information. But from the programming point of view, it is better to do it before the loop. No point doing it again and again inside the loop

(Bhaskar Sarma)

ValueError: multi_class must be in ('ovo', 'ovr')

I’m getting “ValueError: multi_class must be in ('ovo', 'ovr')” when using roc_auc_score to evaluate feature importance of numerical variables in question 1.

I was getting this error because I was passing the parameters to roc_auc_score incorrectly (df_train[col] , y_train) . The correct way is to pass the parameters in this way: roc_auc_score(y_train, df_train[col])

Asia Saeed

Difference between predict(X) and predict_proba(X)[:, 1]

In case of using predict(X) for this task we are getting the binary classification predictions which are 0 and 1. This may lead to incorrect evaluation values.

The solution is to use predict_proba(X)[:,1], where we get the probability that the value belongs to one of the classes.

Vladimir Yesipov

Why are FPR and TPR equal to 0.0, when threshold = 1.0?

For churn/not churn predictions, I need help to interpret the following scenario please, what is happening when:

The threshold is 1.0
FPR is 0.0
And TPR is 0.0

When the threshold is 1.0, the condition for belonging to the positive class (churn class) is g(x)>=1.0 But g(x) is a sigmoid function for a binary classification problem. It has values between 0 and 1. This function never becomes equal to outermost values, i.e. 0 and 1.

That is why there is no object, for which churn-condition could be satisfied. And that is why there is no any positive (churn) predicted value (neither true positive, nor false positive), if threshold is equal to 1.0

Alena Kniazeva

I didn’t fully understand the ROC curve. Can I move on?

It's a complex and abstract topic and it requires some time to understand. You can move on without fully understanding the concept.

Nonetheless, it might be useful for you to rewatch the video, or even watch videos/lectures/notes by other people on this topic, as the ROC AUC is one of the most important metrics used in Binary Classification models.

5. Deploying Machine Learning Models

How do I get started with Week 5?

- Ask questions for Live Sessions: https://app.sli.do/event/vsUpjYsayZ8A875Hq8dpUa/live/questions
- Calendar for weekly meetings: https://calendar.google.com/calendar/u/0/r?cid=cGtjZ2tkbGc1OG9yb2lxa2Vwc2g4YXMzMmNAZ3JvdXAuY2FsZW5kYXIuZ29vZ2xlLmNvbQ&pli=1
- Week 5 HW: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/05-deployment/homework.md
- Submit HW Week 5: https://docs.google.com/forms/d/e/1FAIpQLScptn4Z0Ls62g0GcjJIYYi0PqEOsbbsShMJOAQouft915A8Kg/viewform
- All HWs: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/
- HW 3 Solution: https://github.com/alexeygrigorev/mlbookcamp-code/blob/master/course-zoomcamp/cohorts/2022/03-classification/homework_3.ipynb
- Evaluation Matrix: https://docs.google.com/spreadsheets/d/e/2PACX-1vQCwqAtkjl07MTW-SxWUK9GUvMQ3Pv_fF8UadcuIYLgHa0PlNu9BRWtfLgivI8xSCncQs82HDwGXSm3/pubhtml
- GitHub for theory: https://github.com/alexeygrigorev/mlbookcamp-code/tree/master/course-zoomcamp
- YouTube Link: 5.X --- https://www.youtube.com/watch?v=agIFak9A3m8&list=PL3MmuxUbc_hIhxl5Ji8t4O6lPAOpHaCLR&index=49

~~~ Nukta Bhatia ~~~

Errors related to the default environment: WSL, Ubuntu, proper Python version, installing pipenv etc.

While weeks 1-4 can relatively easily be followed and the associated homework completed with just about any default environment / local setup, week 5 introduces several layers of abstraction and dependencies.

It is advised to prepare your “homework environment” with a cloud provider of your choice. A thorough step-by-step guide for doing so for an AWS EC2 instance is provided in an introductory video taken from the MLOPS course here:

https://www.youtube.com/watch?v=IXSiYkP23zo

Note that (only) small AWS instances can be run for free, and that larger ones will be billed hourly based on usage (but can and should be stopped when not in use).

Alternative ways are sketched here:
https://github.com/alexeygrigorev/mlbookcamp-code/blob/master/course-zoomcamp/01-intro/06-environment.md

Error building Docker images on Mac with M1 silicon

Do you get errors building the Docker image on the Mac M1 chipset?

The error I was getting was:

```
Could not open '/lib64/ld-linux-x86-64.so.2': No such file or
directory
```

The fix (from here): vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv

Open mlbookcamp-code/course-zoomcamp/01-intro/environment/Dockerfile
Replace line 1 with

```
FROM --platform=linux/amd64 ubuntu:latest
```

Now build the image as specified. In the end it took over 2 hours to build the image but it did complete in the end.
~~~

David Colton

Cannot connect to the docker daemon. Is the Docker daemon running?

Working on getting Docker installed - when I try running hello-world I am getting the error.

```
Docker: Cannot connect to the docker daemon at
unix:///var/run/docker.sock. Is the Docker daemon running ?
```

Solution description

If you’re getting this error on WSL, re-install your docker: remove the docker installation from WSL and install Docker Desktop on your host machine (Windows).

On Linux, start the docker daemon with either of these commands:

- sudo dockerd
- sudo service docker start

Added by Ugochukwu Onyebuchi

The command '/bin/sh -c pipenv install --deploy --system && rm -rf /root/.cache' returned a non-zero code: 1

After using the command “docker build -t churn-prediction .” to build the Docker image, the above error is raised and the image is not created.

In your Dockerfile, change the Python version in the first line the Python version installed in your system:

```
FROM python:3.7.5-slim
```

To find your python version, use the command `python --version`. For example:

```
python --version
```

```
>> Python 3.9.7
```

Then, change it on your Dockerfile:

```
FROM python:3.9.7-slim
```

Added by Filipe Melo

Running “pipenv install sklearn==1.0.2” gives errors. What should I do?

When the facilitator was adding sklearn to the virtual environment in the lectures, he used sklearn==0.24.1 and it ran smoothly. But while doing the homework and you are asked to use the 1.0.2 version of sklearn, it gives errors.

The solution is to use the full name of sklearn. That is, run it as “pipenv install scikit-learn==1.0.2” and the error will go away, allowing you to install sklearn for the version in your virtual environment.

Odimegwu David

Why do we need the --rm flag

What is the reason we don’t want to keep the docker image in our system and why do we need to run docker containers with `--rm` flag?

1. For best practice, you don’t want to have a lot of abandoned docker images in your system. You just update it in your folder and trigger the build one more time.
2. They consume extra space on your disk. Unless you don’t want to re-run the previously existing containers, it is better to use the `--rm` option.
3. The right way to say: “Why do we remove the docker container in our system?”. Well the docker image is still kept; it is the container that is not kept. Upon execution, images are not modified; only containers are.
4. The option `--rm` is for removing containers. The images remain until you remove them manually. If you don’t specify a version when building an image,

it will always rebuild and replace the latest tag. `docker images` shows you all the image you have pulled or build so far.

5. During development and testing you usually specify `--rm` to get the containers auto removed upon exit. Otherwise they get accumulated in a stopped state, taking up space. `docker ps -a` shows you all the containers you have in your host. Each time you change Pipfile (or any file you baked into the container), you rebuild the image under the same tag or a new tag. It's important to understand the difference between the term "docker image" and "docker container". Image is what we build with all the resources baked in. You can move it around, maintain it in a repository, share it. Then we use the image to spin up instances of it and they are called containers.

Added by Muhammad Awon

Failed to read Dockerfile

When you create the dockerfile the name should be dockerfile and needs to be without extension. One of the problems we can get at this point is to create the dockerfile as a dockerfile extension Dockerfile.dockerfile which creates an error when we build the docker image. Instead we just need to create the file without extension: Dockerfile and will run perfectly.

Added by Pastor Soto

Install docker on MacOS

Refer to the page https://docs.docker.com/desktop/install/mac-install/ remember to check if you have apple chip or intel chip.

I cannot pull the image with docker pull command

Problem: When I am trying to pull the image with the docker pull svizor/zoomcamp-model command I am getting an error:

```
Using default tag: latest
Error response from daemon: manifest for svizor/zoomcamp-model:latest not found: manifest unknown: manifest unknown
```

Solution: The docker by default uses the latest tag to avoid this use the correct tag from image description. In our case use command:

docker pull svizor/zoomcamp-model:3.9.12-slim

Added by Vladimir Yesipov

Where does pipenv create environments and how does it name them?

It creates them in **~/.local/share/virtualenvs/*environment_name*.**

The environment name is the name of the last folder in the folder directory where we used the pipenv install command (or any other pipenv command). E.g. If you run any pipenv command in folder path **~/home/user/Churn-Flask-app**, it will create an environment named Churn-Flask-app-some_random_characters, and it's path will be like this: **/home/user/.local/share/virtualenvs/churn-flask-app-i_mzGMjX.**

All libraries of this environment will be installed inside this folder. To activate this environment, I will need to cd into the project folder again, and type `pipenv shell`. In short, **the location of the project folder acts as an identifier for an environment, in place of any name.**

(Memoona Tahira)

How do I debug a docker container?

Launch the container image in interactive mode and overriding the entrypoint, so that it starts a bash command.

```
docker run -it --entrypoint bash <image>
```

If the container is already running, execute a command in the specific container:

```
docker ps (find the container-id)
docker exec -it <container-id> bash
```

(Marcos MJD)

The input device is not a TTY when running docker in interactive mode (Running Docker on Windows in GitBash)

```
$ docker exec -it 1e5a1b663052 bash
```

```
the input device is not a TTY.  If you are using mintty, try prefixing the command
with 'winpty'
```

Fix:

```
winpty docker exec -it 1e5a1b663052 bash
```

A TTY is a terminal interface that supports escape sequences, moving the cursor around, etc.
Winpty is a Windows software package providing an interface similar to a Unix pty-master for communicating with Windows console programs.

More info on terminal, shell, console applications hi and so on:
https://conemu.github.io/en/TerminalVsShell.html

(Marcos MJD)

Error: failed to compute cache key: "/model2.bin" not found: not found

Initially, I did not assume there was a model2. I copied the original model1.bin and dv.bin. Then when I tried to load using

```
COPY ["model2.bin", "dv.bin", "./"]
```

then I got the error above in MINGW64 (git bash) on Windows.

The temporary solution I found was to use

```
COPY ["*", "./"]
```

which I assume combines all the files from the original docker image and the files in your working directory.

Added by Muhammed Tan

f-strings

f-String not properly keyed in: does anyone knows why i am getting error after import pickle?

The first error showed up because your f-string is using () instead of {} around C. So, should be: f’model_C={C}.bin’
The second error as noticed by Sriniketh, your are missing one parenthesis it should be pickle.dump((dv, model), f_out)

(Humberto R.)

'pipenv' is not recognized as an internal or external command, operable program or batch file.

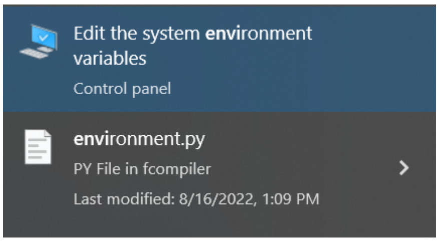
This error happens because pipenv is already installed but you can't access it from the path.

This error comes out if you run.

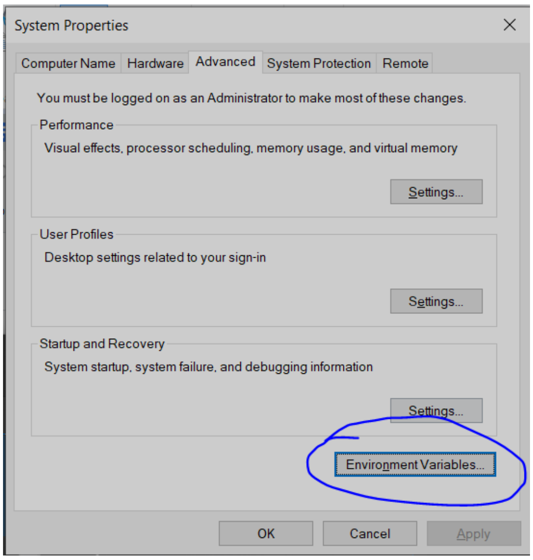
```
pipenv  --version
pipenv shell
```

Solution for Windows

1) Open this option


Edit the system environment variables
Control panel
environment.py
PY File in fcompiler
Last modified: 8/16/2022, 1:09 PM

2) Click here



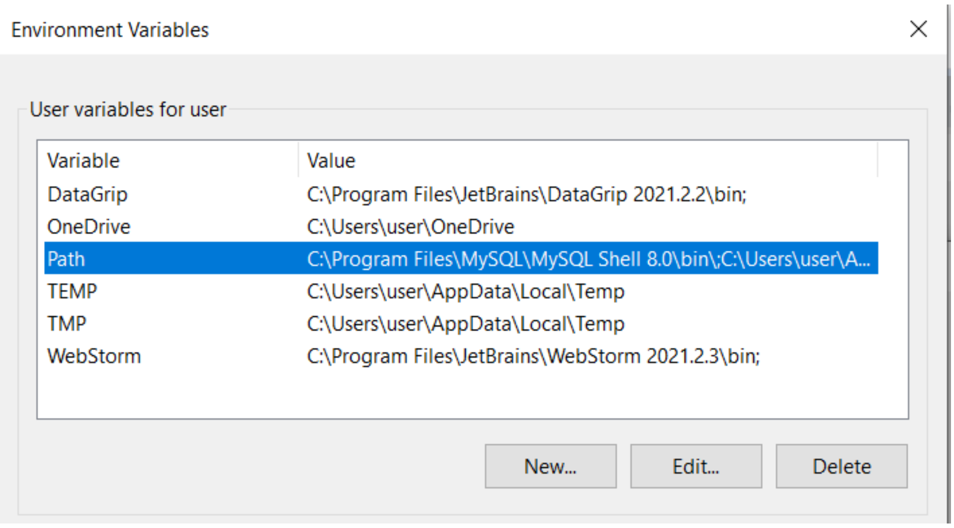
You must be logged on as an Administrator to make most of these changes.

Performance
Visual effects, processor scheduling, memory usage, and virtual memory

User Profiles
Desktop settings related to your sign-in

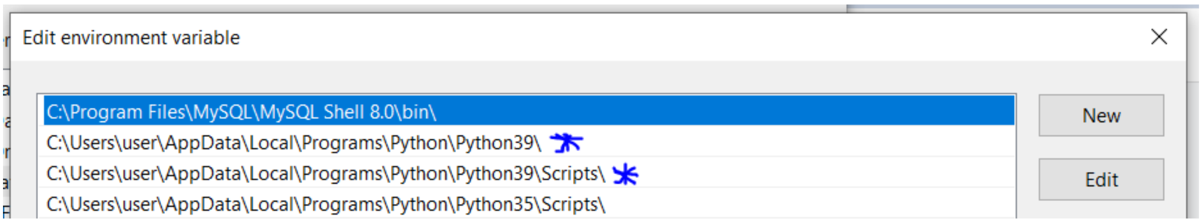
Startup and Recovery
System startup, system failure, and debugging information

3) Click in Edit Button



Variable	Value
DataGrip	C:\Program Files\JetBrains\DataGrip 2021.2.2\bin;
OneDrive	C:\Users\user\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\user\A...
TEMP	C:\Users\user\AppData\Local\Temp
TMP	C:\Users\user\AppData\Local\Temp
WebStorm	C:\Program Files\JetBrains\WebStorm 2021.2.3\bin;

4) Make sure the next two locations are on the PATH, otherwise, add it.
C:\Users\AppData\....\Python\PythonXX\
C:\Users\AppData\....\Python\PythonXX\Scripts\



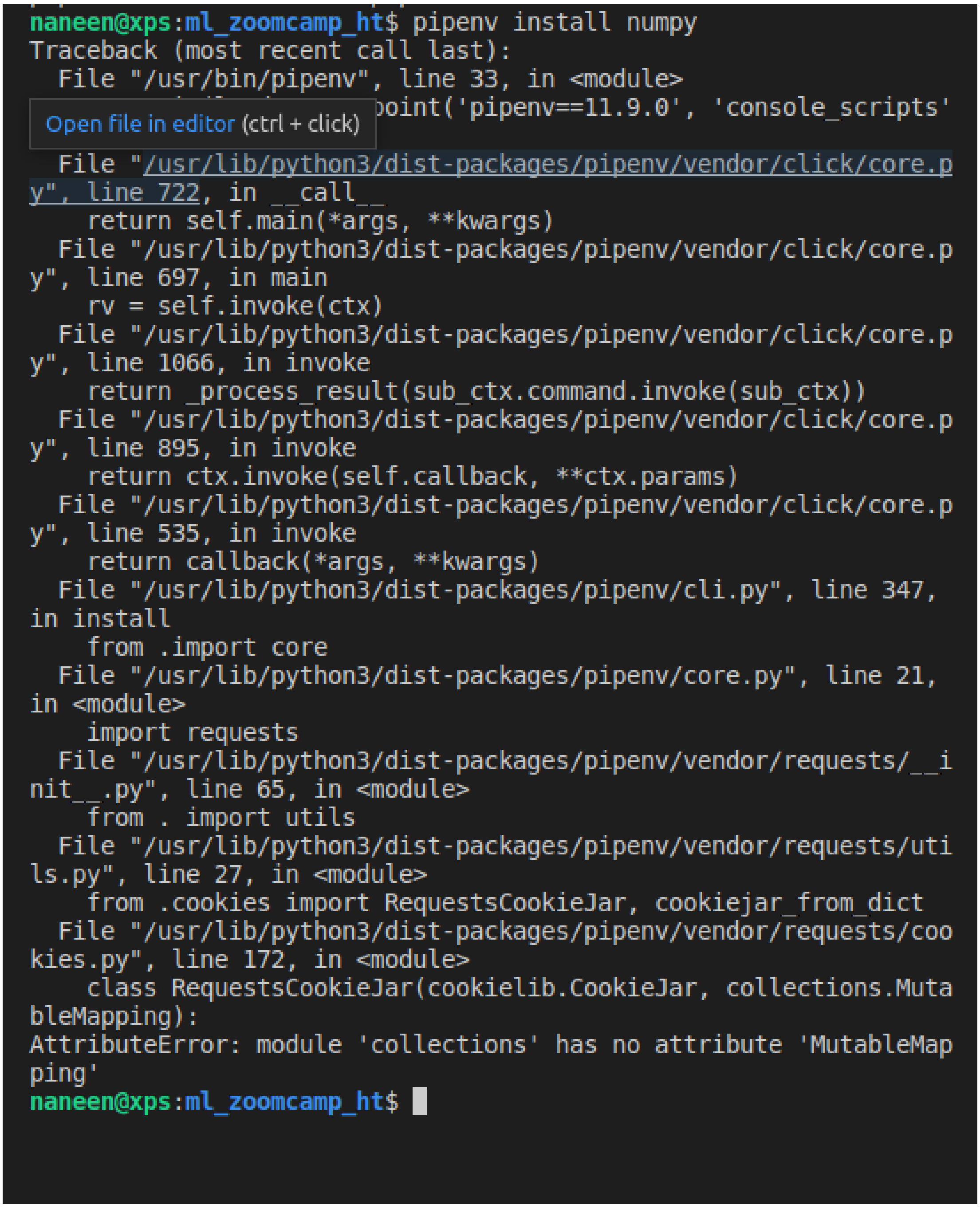
```
C:\Program Files\MySQL\MySQL Shell 8.0\bin\
C:\Users\user\AppData\Local\Programs\Python\Python39\ *
C:\Users\user\AppData\Local\Programs\Python\Python39\Scripts\ *
C:\Users\user\AppData\Local\Programs\Python\Python35\Scripts\
```

Added by Alejandro Aponte

Note: this answer assumes you don't use Anaconda. For Windows, using Anaconda would be a better choice and less prone to errors.

AttributeError: module 'collections' has no attribute 'MutableMapping'

Following the instruction from video week-5.6, using pipenv to install python libraries throws below error



```
naneen@xps:ml_zoomcamp_ht$ pipenv install numpy
Traceback (most recent call last):
  File "/usr/bin/pipenv", line 33, in <module>
                                      oint('pipenv==11.9.0', 'console_scripts'

  File "/usr/lib/python3/dist-packages/pipenv/vendor/click/core.p
y", line 722, in __call__
    return self.main(*args, **kwargs)
  File "/usr/lib/python3/dist-packages/pipenv/vendor/click/core.p
y", line 697, in main
    rv = self.invoke(ctx)
  File "/usr/lib/python3/dist-packages/pipenv/vendor/click/core.p
y", line 1066, in invoke
    return _process_result(sub_ctx.command.invoke(sub_ctx))
  File "/usr/lib/python3/dist-packages/pipenv/vendor/click/core.p
y", line 895, in invoke
    return ctx.invoke(self.callback, **ctx.params)
  File "/usr/lib/python3/dist-packages/pipenv/vendor/click/core.p
y", line 535, in invoke
    return callback(*args, **kwargs)
  File "/usr/lib/python3/dist-packages/pipenv/cli.py", line 347,
in install
    from .import core
  File "/usr/lib/python3/dist-packages/pipenv/core.py", line 21,
in <module>
    import requests
  File "/usr/lib/python3/dist-packages/pipenv/vendor/requests/__i
nit__.py", line 65, in <module>
    from . import utils
  File "/usr/lib/python3/dist-packages/pipenv/vendor/requests/uti
ls.py", line 27, in <module>
    from .cookies import RequestsCookieJar, cookiejar_from_dict
  File "/usr/lib/python3/dist-packages/pipenv/vendor/requests/coo
kies.py", line 172, in <module>
    class RequestsCookieJar(cookielib.CookieJar, collections.Muta
bleMapping):
AttributeError: module 'collections' has no attribute 'MutableMap
ping'
naneen@xps:ml_zoomcamp_ht$ 
```

Solution to this error is to make sure that you are working with python==3.9 (as informed in the very first lesson of the zoomcamp) and not python==3.10.

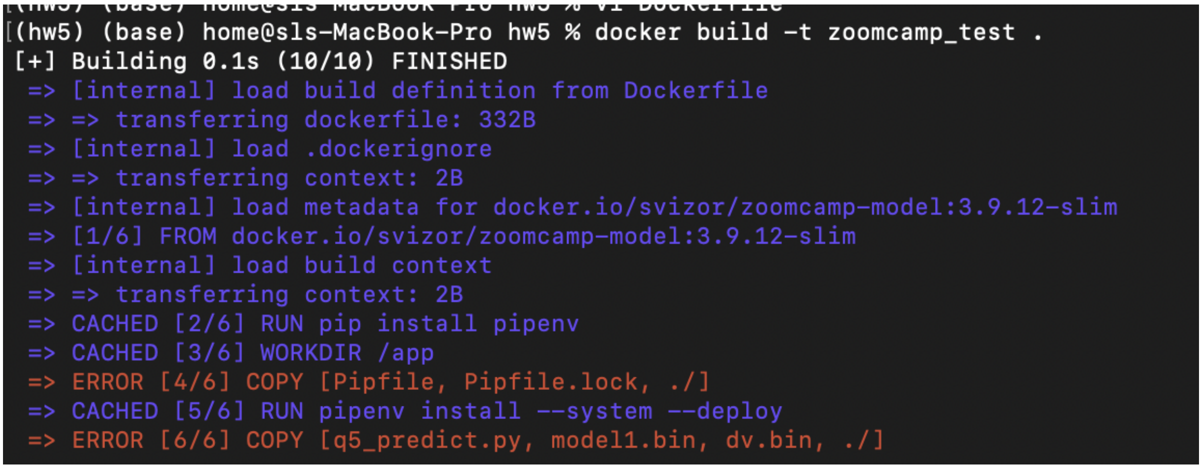
Added by Hareesh Tummala

ConnectionError: ('Connection aborted.', RemoteDisconnected('Remote end closed connection without response'))

Set the host to ‘0.0.0.0’ on the flask app and dockerfile then RUN the url using localhost.

(Theresa S.)

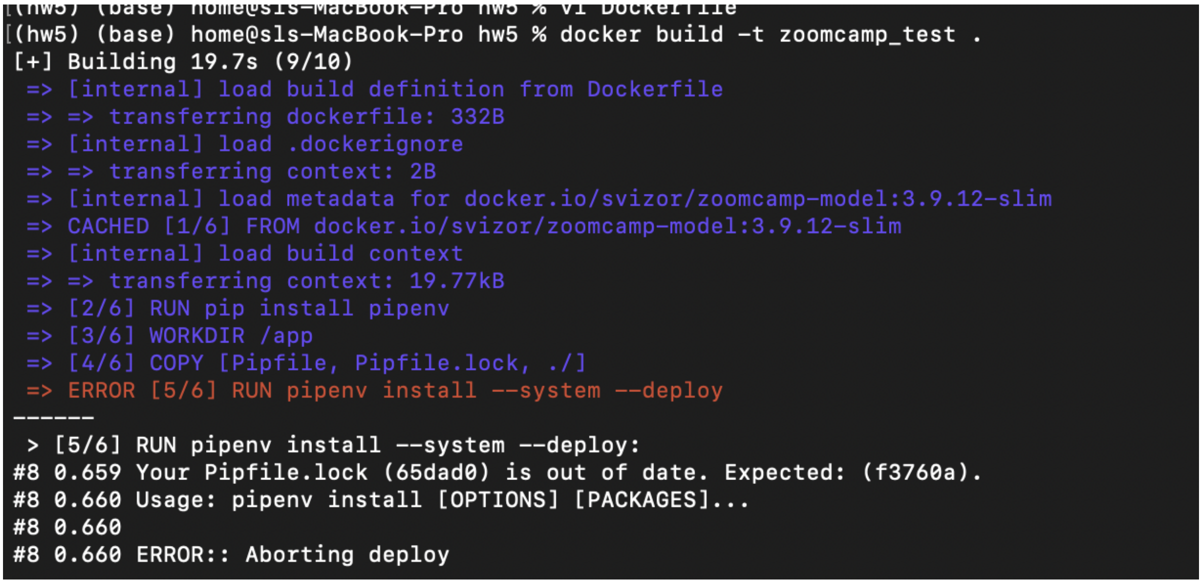
ERROR COPY …



```
[(hw5) (base) home@sls-MacBook-Pro hw5 % vi Dockerfile
[(hw5) (base) home@sls-MacBook-Pro hw5 % docker build -t zoomcamp_test .
 [+] Building 0.1s (10/10) FINISHED
  => [internal] load build definition from Dockerfile
  => => transferring dockerfile: 332B
  => [internal] load .dockerignore
  => => transferring context: 2B
  => [internal] load metadata for docker.io/svizor/zoomcamp-model:3.9.12-slim
  => [1/6] FROM docker.io/svizor/zoomcamp-model:3.9.12-slim
  => [internal] load build context
  => => transferring context: 2B
  => CACHED [2/6] RUN pip install pipenv
  => CACHED [3/6] WORKDIR /app
  => ERROR [4/6] COPY [Pipfile, Pipfile.lock, ./]
  => CACHED [5/6] RUN pipenv install --system --deploy
  => ERROR [6/6] COPY [q5_predict.py, model1.bin, dv.bin, ./]
```

Solution:
This error occurred because I used single quotes around the filenames. Stick to double quotes

Fix error during installation of Pipfile inside Docker container

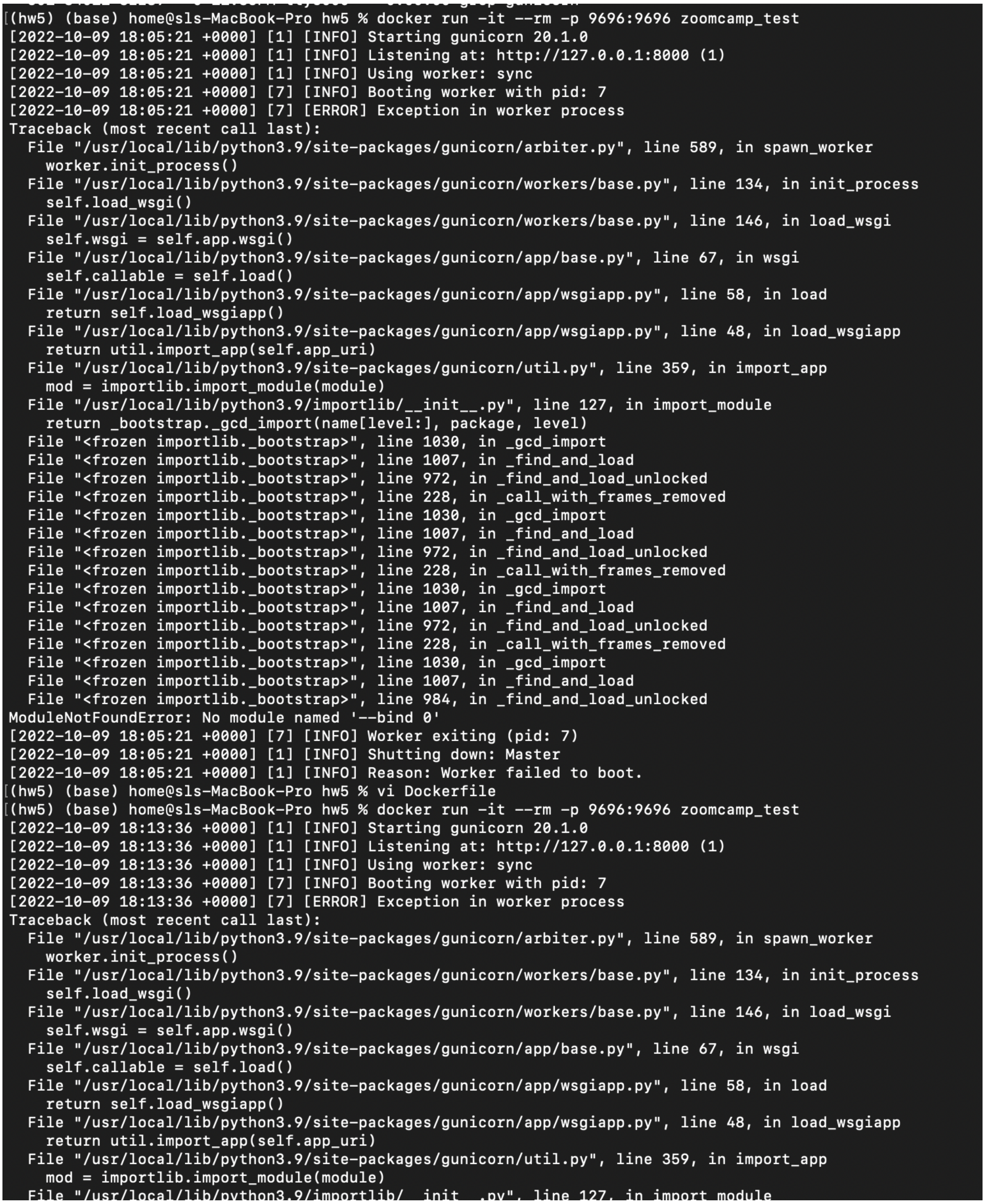


```
[(hw5) (base) home@sls-MacBook-Pro hw5 % vi Dockerfile
[(hw5) (base) home@sls-MacBook-Pro hw5 % docker build -t zoomcamp_test .
[+] Building 19.7s (9/10)
 => [internal] load build definition from Dockerfile
 => => transferring dockerfile: 332B
 => [internal] load .dockerignore
 => => transferring context: 2B
 => [internal] load metadata for docker.io/svizor/zoomcamp-model:3.9.12-slim
 => CACHED [1/6] FROM docker.io/svizor/zoomcamp-model:3.9.12-slim
 => [internal] load build context
 => => transferring context: 19.77kB
 => [2/6] RUN pip install pipenv
 => [3/6] WORKDIR /app
 => [4/6] COPY [Pipfile, Pipfile.lock, ./]
 => ERROR [5/6] RUN pipenv install --system --deploy
------
 > [5/6] RUN pipenv install --system --deploy:
#8 0.659 Your Pipfile.lock (65dad0) is out of date. Expected: (f3760a).
#8 0.660 Usage: pipenv install [OPTIONS] [PACKAGES]...
#8 0.660
#8 0.660 ERROR:: Aborting deploy
```

I tried the first solution on Stackoverflow which recommended running `pipenv lock` to update the Pipfile.lock. However, this didn't resolve it. But the following switch to the pipenv installation worked

```
RUN pipenv install --system --deploy --ignore-pipfile
```

How to fix error after running the Docker run command



```
[(hw5) (base) home@sls-MacBook-Pro hw5 % docker run -it --rm -p 9696:9696 zoomcamp_test
[2022-10-09 18:05:21 +0000] [1] [INFO] Starting gunicorn 20.1.0
[2022-10-09 18:05:21 +0000] [1] [INFO] Listening at: http://127.0.0.1:8000 (1)
[2022-10-09 18:05:21 +0000] [1] [INFO] Using worker: sync
[2022-10-09 18:05:21 +0000] [7] [INFO] Booting worker with pid: 7
[2022-10-09 18:05:21 +0000] [7] [ERROR] Exception in worker process
Traceback (most recent call last):
  File "/usr/local/lib/python3.9/site-packages/gunicorn/arbiter.py", line 589, in spawn_worker
    worker.init_process()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/workers/base.py", line 134, in init_process
    self.load_wsgi()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/workers/base.py", line 146, in load_wsgi
    self.wsgi = self.app.wsgi()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/app/base.py", line 67, in wsgi
    self.callable = self.load()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/app/wsgiapp.py", line 58, in load
    return self.load_wsgiapp()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/app/wsgiapp.py", line 48, in load_wsgiapp
    return util.import_app(self.app_uri)
  File "/usr/local/lib/python3.9/site-packages/gunicorn/util.py", line 359, in import_app
    mod = importlib.import_module(module)
  File "/usr/local/lib/python3.9/importlib/__init__.py", line 127, in import_module
    return _bootstrap._gcd_import(name[level:], package, level)
  File "<frozen importlib._bootstrap>", line 1030, in _gcd_import
  File "<frozen importlib._bootstrap>", line 1007, in _find_and_load
  File "<frozen importlib._bootstrap>", line 972, in _find_and_load_unlocked
  File "<frozen importlib._bootstrap>", line 228, in _call_with_frames_removed
  File "<frozen importlib._bootstrap>", line 1030, in _gcd_import
  File "<frozen importlib._bootstrap>", line 1007, in _find_and_load
  File "<frozen importlib._bootstrap>", line 972, in _find_and_load_unlocked
  File "<frozen importlib._bootstrap>", line 228, in _call_with_frames_removed
  File "<frozen importlib._bootstrap>", line 1030, in _gcd_import
  File "<frozen importlib._bootstrap>", line 1007, in _find_and_load
  File "<frozen importlib._bootstrap>", line 972, in _find_and_load_unlocked
  File "<frozen importlib._bootstrap>", line 228, in _call_with_frames_removed
  File "<frozen importlib._bootstrap>", line 1030, in _gcd_import
  File "<frozen importlib._bootstrap>", line 1007, in _find_and_load
  File "<frozen importlib._bootstrap>", line 984, in _find_and_load_unlocked
ModuleNotFoundError: No module named '--bind 0'
[2022-10-09 18:05:21 +0000] [7] [INFO] Worker exiting (pid: 7)
[2022-10-09 18:05:21 +0000] [1] [INFO] Shutting down: Master
[2022-10-09 18:05:21 +0000] [1] [INFO] Reason: Worker failed to boot.
[(hw5) (base) home@sls-MacBook-Pro hw5 % vi Dockerfile
[(hw5) (base) home@sls-MacBook-Pro hw5 % docker run -it --rm -p 9696:9696 zoomcamp_test
[2022-10-09 18:13:36 +0000] [1] [INFO] Starting gunicorn 20.1.0
[2022-10-09 18:13:36 +0000] [1] [INFO] Listening at: http://127.0.0.1:8000 (1)
[2022-10-09 18:13:36 +0000] [1] [INFO] Using worker: sync
[2022-10-09 18:13:36 +0000] [7] [INFO] Booting worker with pid: 7
[2022-10-09 18:13:36 +0000] [7] [ERROR] Exception in worker process
Traceback (most recent call last):
  File "/usr/local/lib/python3.9/site-packages/gunicorn/arbiter.py", line 589, in spawn_worker
    worker.init_process()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/workers/base.py", line 134, in init_process
    self.load_wsgi()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/workers/base.py", line 146, in load_wsgi
    self.wsgi = self.app.wsgi()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/app/base.py", line 67, in wsgi
    self.callable = self.load()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/app/wsgiapp.py", line 58, in load
    return self.load_wsgiapp()
  File "/usr/local/lib/python3.9/site-packages/gunicorn/app/wsgiapp.py", line 48, in load_wsgiapp
    return util.import_app(self.app_uri)
  File "/usr/local/lib/python3.9/site-packages/gunicorn/util.py", line 359, in import_app
    mod = importlib.import_module(module)
  File "/usr/local/lib/python3.9/importlib/__init__.py", line 127, in import_module
```

Solution
This error was because there was another instance of guincorn running. So I thought of removing this along with the zoomcamp_test image. However, it didn’t let me remove the orphan container. So I did the following
Running the following commands

```
docker ps -a <to list all docker containers>
docker images <to list images>
docker stop <container ID>
docker rm <container ID>
docker rmi image
```

I rebuilt the Docker image, and ran it once again; this time it worked correctly and I was able to serve the test script to the endpoint.

Bind for 0.0.0.0:9696 failed: port is already allocated

I was getting the below error when I rebuilt the docker image although the port was not allocated, and it was working fine.
Error message:

```
Error response from daemon: driver failed programming external
connectivity on endpoint beautiful_tharp
(875be95c7027cebb853a62fc4463d46e23df99e0175be73641269c3d180f7
796): Bind for 0.0.0.0:9696 failed: port is already allocated.
```

Solution description

Issue has been resolved running the following command:

```
docker kill $(docker ps -q)
```

https://github.com/docker/for-win/issues/2722

Asia Saeed

Installing md5sum on Macos

Install it by using command

```
% brew install md5sha1sum
```

Then run command to check hash for file to check if they the same with the provided

```
% md5sum model1.bin dv.bin
```

Olga Rudakova

How to run a script while a web-server is working?

Problem description:
I started a web-server in terminal (command window, powershell, etc.). How can I run another python script, which makes a request to this server?

Solution description:
Just open another terminal (command window, powershell, etc.) and run a python script.

Alena Kniazeva

Version-conflict in pipenv

Problem description:
In video 5.5 when I do `pipenv shell` and then `pipenv run gunicorn --bind 0.0.0.0:9696 predict:app`, I get the following warning:

```
UserWarning: Trying to unpickle estimator DictVectorizer from
version 1.1.1 when using version 0.24.2. This might lead to
breaking code or invalid results. Use at your own risk.
```

Solution description:
When you create a virtual env, you should use the same version of Scikit-Learn that you used for training the model on this case it's 1.1.1. There is version conflicts so we need to make sure our model and dv files are created from the version we are using for the project.

Bhaskar Sarma

Python_version and Python_full_version error after running pipenv install:

If you install packages via pipenv install, and get an error that ends like this:

```
pipenv.vendor.plette.models.base.ValidationError:
{'python_version': '3.9', 'python_full_version': '3.9.13'}
python_full_version: 'python_version' must not be present with
'python_full_version'
python_version: 'python_full_version' must not be present with
'python_version'
```

Do this:

1. open Pipfile in nano editor, and remove either the `python_version` or `python_full_version` line, press CTRL+X, type Y and click Enter to save changed
2. Type pipenv lock to create the Pipfile.lock.
3. Done. Continue what you were doing

Your Pipfile.lock (221d14) is out of date (during Docker build)

If during running the docker build command, you get an error like this:

```
Your Pipfile.lock (221d14) is out of date. Expected: (939fe0).
Usage: pipenv install [OPTIONS] [PACKAGES]...

ERROR:: Aborting deploy
```

Option 1: Delete the pipfile.lock via rm Pipfile, and then rebuild the lock via pipenv lock from the terminal before retrying the docker build command.

Option 2: If it still doesn’t work, remove the pipenv environment, Pipfile and Pipfile.lock, and create a new one before building docker again. Commands to remove pipenv environment and removing pipfiles:

```
pipenv  --rm
rm Pipfile*
```

You are using windows. Conda environment. You then use waitress instead of gunicorn. After a few runs, suddenly mlflow server fails to run.

Ans: Pip uninstall waitress mflow. Then reinstall just mlflow. By this time you should have successfully built your docker image so you dont need to reinstall waitress. All good. Happy learning.

Added by BLAQ

Completed creating the environment locally but could not find the environment on AWS.

Ans: so you have created the env. You need to make sure you're in eu-west-1 (ireland) when you check the EB environments. Maybe you're in a different region in your console.

Added by Edidiong Esu

Installing waitress on Windows via GitBash: “waitress-serve” command not found

Running **'pip install waitress'** as a command on GitBash was not downloading the executable file **'waitress-serve.exe'**. You need this file to be able to run commands with waitress in Git Bash. To solve this:

- open a Jupyter notebook and run the same command ' **pip install waitress'**. This way the executable file will be downloaded. The notebook may give you this warning : **'WARNING: The script waitress-serve.exe is installed in 'c:\Users\....\anaconda3\Scripts' which is not on PATH. Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.'**
- Add the path where 'waitress-serve.exe' is installed into gitbash's PATH as such:
    - enter the following command in gitbash: **nano ~/.bashrc**
    - add the path to 'waitress-serve.exe' to PATH using this command**: export PATH="/path/to/waitress:$PATH"**

    - close gitbash and open it again and you should be good to go

Added by Bachar Kabalan

6. Decision Trees and Ensemble Learning

How to get started with Week 6?

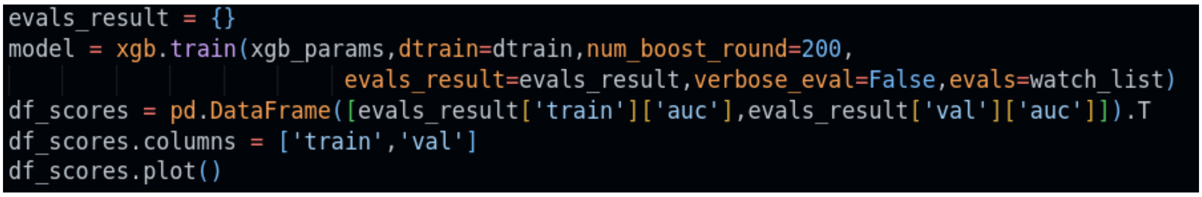
- Ask questions for Live Sessions: https://app.sli.do/event/vsUpjYsayZ8A875Hq8dpUa/live/questions
- Calendar for weekly meetings: https://calendar.google.com/calendar/u/0/r?cid=cGtjZ2tkbGc1OG9yb2lxa2Vwc2g4YXMzMmNAZ3JvdXAuY2FsZW5kYXIuZ29vZ2xlLmNvbQ&pli=1
- Week 6 HW: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/06-trees/homework.md
- Submit HW Week 6: https://docs.google.com/forms/d/e/1FAIpQLSdgFwQSLTAI_4wOAGnT2sMiWX8dindkTUwAV29Rlg0a1m67Ng/viewform
- All HWs: https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/cohorts/2023/
- HW 4 Solution: https://github.com/alexeygrigorev/mlbookcamp-code/blob/master/course-zoomcamp/cohorts/2022/04-evaluation/homework_4.ipynb
- Evaluation Matrix: https://docs.google.com/spreadsheets/d/e/2PACX-1vQCwqAtkjl07MTW-SxWUK9GUvMQ3Pv_fF8UadcuIYLgHa0PlNu9BRWtfLgivI8xSCncQs82HDwGXSm3/pubhtml
- GitHub for theory: https://github.com/alexeygrigorev/mlbookcamp-code/tree/master/course-zoomcamp
- YouTube Link: 6.X --- https://www.youtube.com/watch?v=GJGmlfZoCoU&list=PL3MmuxUbc_hIhxl5Ji8t4O6lPAOpHaCLR&index=57
- FAQs: https://docs.google.com/document/d/1LpPanc33QJJ6BSsyxVg-pWNMplal84TdZtq10naIhD8/edit#heading=h.lpz96zg7l47j

~~~Nukta Bhatia~~~
~~~

How to get the training and validation metrics from XGBoost?

During the XGBoost lesson, we created a parser to extract the training and validation auc from the standard output. However, we can accomplish that in a more straightforward way.

We can use the evals_result parameters, which takes an empty dictionary and updates it for each tree. Additionally, you can store the data in a dataframe and plot it in an easier manner.



```
evals_result = {}
model = xgb.train(xgb_params,dtrain=dtrain,num_boost_round=200,
                         evals_result=evals_result,verbose_eval=False,evals=watch_list)
df_scores = pd.DataFrame([evals_result['train']['auc'],evals_result['val']['auc']]).T
df_scores.columns = ['train','val']
df_scores.plot()
```

Added by Daniel Coronel

How to solve regression problems with random forest in scikit-learn?

You should create sklearn.ensemble.RandomForestRegressor object. It’s rather similar to sklearn.ensemble.RandomForestClassificator for classification problems. Check https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html for more information.

Alena Kniazeva

ValueError: feature_names must be string, and may not contain [,] or <

In question 6, I was getting **ValueError: feature_names must be string, and may not contain [,] or <** when I was creating DMatrix for train and validation

Solution description
The cause of this error is that some of the features names contain special characters like = and <, and I fixed the error by removing them as follows:
features= [i.replace("=<", "_").replace("=","_") for i in features]

Asia Saeed

What is eta in XGBoost

Sometimes someone might wonder what eta means in the tunable hyperparameters of XGBoost and how it helps the model.

ETA is the learning rate of the model. XGBoost uses gradient descent to calculate and update the model. In gradient descent, we are looking for the minimum weights that help the model to learn the data very well. This minimum weights for the features is updated each time the model passes through the features and learns the features during training. Tuning the learning rate helps you tell the model what speed it would use in deriving the minimum for the weights.

ValueError: continuous format is not supported

Calling `roc_auc_score()` to get auc is throwing the above error.

Solution to this issue is to make sure that you pass **y_actuals** as 1st argument and **y_pred** as 2nd argument.

```
roc_auc_score(y_train, y_pred)
```

Hareesh Tummala

Question 3 of homework 6 if i see that rmse goes up at a certain number of n_estimators but then goes back down lower than it was befofe, should the answer be the number of n_estimators after which rmse initially went up, or the number after which it was its overall lowest value?

When rmse stops improving means, when it stops to decrease or remains almost similar.

Pastor Soto

ValueError: Unknown label type: 'continuous'

Solution: This problem happens because you use DecisionTreeClassifier instead of DecisionTreeRegressor. You should check if you want to use a Decision tree for classification or regression.

Alejandro Aponte

Different values of auc, each time code is re-run

When I run `dt = DecisionTreeClassifier()` in jupyter in same laptop, each time I re-run it or do (restart kernel + run) I get different values of auc. Some of them are 0.674, 0.652, 0.642, 0.669 and so on. Anyone knows why it could be? I am referring to 7:40-7:45 of video 6.3.

Solution: try setting the random seed e.g

```
dt = DecisionTreeClassifier(random_state=22)
```

Bhaskar Sarma

Does it matter if we let the Python file create the server or if we run gunicorn directly?

They both do the same, it's just less typing from the script.

Asked by Andrew Katoch, Added by Edidiong Esu

Difference between xgb.XGBClassifier and xgb.train?

Essentially everything that is done of XGB classifier can be done using xgb.train
Detail discussion about this can be found here:
https://stackoverflow.com/questions/47152610/what-is-the-difference-between-xgb-train-and-xgb-xgbregressor-or-xgb-xgbclassif

Pastor Soto

7. Production-Ready Machine Learning (Bento ML)

This section was covered in the previous iteration of the course (2022). For more information, refer to this separate Document

Midterm projects

This section is moved to Projects

8. Neural Networks and Deep Learning

How to get started with Week 8?

How do I push from Saturn Cloud to Github?

Connecting your GPU on Saturn Cloud to Github repository is not compulsory, since you can just download the notebook and copy it to the Github folder. But if you like technology to do things for you, then follow the solution description below:

Solution description: Follow the instructions in these github docs to create an SSH private and public key:

1. https://docs.github.com/en/authentication/connecting-to-github-with-ssh/generating-a-new-ssh-key-and-adding-it-to-the-ssh-agent
2. https://docs.github.com/en/authentication/connecting-to-github-with-ssh/adding-a-new-ssh-key-to-your-github-account?tool=webui
3. Then the second video on this module about saturn cloud would show you how to add the ssh keys to secrets and authenticate through a terminal.

Or alternatively, you could just use the public keys provided by Saturn Cloud by default. To do so, follow these steps:

1. Click on your username and on manage
2. Down below you will see the Git SSH keys section.
3. Copy the default public key provided by Saturn Cloud
4. Paste these key into the SSH keys section of your github repo
5. Open a terminal on Saturn Cloud and run this command "ssh -T git@github.com"
6. You will receive a successful authentication notice.

Odimegwu David

How to upload kaggle data to Saturn Cloud?

Problem description: Uploading the data to saturn cloud from kaggle can be time saving, specially if the dataset is large.

You can just download to your local machine and then upload to a folder on saturn cloud, but there is a better solution that needs to be set once and you have access to all kaggle datasets in saturn cloud.

On your notebook run:

```
!pip install -q kaggle
```

Go to Kaggle website (you need to have an account for this):

- Click on your profile image -> Account
- Scroll down to the API box
- Click on Create New API token

It will download a json file with the name kaggle.json store on your local computer. We need to upload this file in the .kaggle folder

- On the notebook click on folder icon on the left upper corner
- This will take you to the root folder
- Click on the .kaggle folder
- Once inside of the .kaggle folder upload the kaggle.json file that you downloaded

Run this command on your notebook:

```
!chmod 600 /home/jovyan/.kaggle/kaggle.json
```

Download the data using this command:

```
!kaggle datasets download -d agrigorev/dino-or-dragon
```

Create a folder to unzip your files:

```
!mkdir data
```

Unzip your files inside that folder

```
!unzip dino-or-dragon.zip -d data
```

Pastor Soto

Error: (ValueError: Unable to load weights saved in HDF5 format into a subclassed Model which has not created its variables yet. Call the Model first, then load the weights.) when loading model.

Problem description:
When loading saved model getting error: ValueError: Unable to load weights saved in HDF5 format into a subclassed Model which has not created its variables yet. Call the Model first, then load the weights.

Solution description:

Before loading model need to evaluate the model on input data: model.evaluate(train_ds)

Added by Vladimir Yesipov

Host key verification failed.

Problem description:
Getting an error using <git clone git@github.com:alexeygrigorev/clothing-dataset-small.git>
The error:
Cloning into 'clothing-dataset'...
Host key verification failed.
fatal: Could not read from remote repository.
Please make sure you have the correct access rights
and the repository exists.

Solution description:
when cloning the repo, you can also chose https - then it should work. This happens when you don't have your ssh key configured.
<git clone https://github.com/alexeygrigorev/clothing-dataset-small.git>
Added by Gregory Morris

The same accuracy on epochs

Problem description
The accuracy and the loss are both still the same or nearly the same while training.

Solution description
In the homework, you should set `class_mode='binary'` while reading the data.
Also, problem occurs when you choose the wrong optimizer, batch size, or learning rate

Added by Ekaterina Kutovaia

Model breaking after augmentation – high loss + bad accuracy

Problem:
When resuming training after augmentation, the loss skyrockets (1000+ during first epoch) and accuracy settles around 0.5 – i.e. the model becomes as good as a random coin flip.

Solution:
Check that the augmented ImageDataGenerator still includes the option “rescale” as specified in the preceding step.

Added by Konrad Mühlberg

Missing channel value error while reloading model:

While doing:

```
import tensorflow as tf
from tensorflow import keras
model = tf.keras.models.load_model('model_saved.h5')
```

If you get an error message like this:

```
ValueError: The channel dimension of the inputs should be defined. The
input_shape received is (None, None, None, None), where axis -1
(0-based) is the channel dimension, which found to be `None`.
```

Solution:

Saving a model (either yourself via model.save() or via checkpoint when save_weights_only = False) saves two things: The trained model weights (for example the best weights found during training) and the model architecture. If the number of channels is not explicitly specified in the Input layer of the model, and is instead defined as a variable, the model architecture will not have the value in the variable stored. Therefore when the model is reloaded, it will complain about not knowing the number of channels. See the code below, in the first line, you need to specify number of channels explicitly:

```
# model architecture:

inputs = keras.Input(shape=(input_size, input_size, 3))
base = base_model(inputs, training=False)
vectors = keras.layers.GlobalAveragePooling2D()(base)
inner = keras.layers.Dense(size_inner, activation='relu')(vectors)
drop = keras.layers.Dropout(droprate)(inner)
outputs = keras.layers.Dense(10)(drop)
model = keras.Model(inputs, outputs)
```

(Memoona Tahira)

How to unzip a folder with an image dataset and suppress output?

Problem:
A dataset for homework is in a zipped folder. If you unzip it within a jupyter notebook by means of ! unzip command, you'll see a huge amount of output messages about unzipping of each image. So you need to suppress this output

Solution:
Execute the next cell:

```
%%capture
! unzip zipped_folder_name.zip -d destination_folder_name
```

Added by Alena Kniazeva

How keras `flow_from_directory` know the names of classes in images?

Problem:
When we run `train_gen.flow_from_directory()` as in video 8.5, it finds images belonging to 10 classes. Does it understand the names of classes from the names of folders? Or, there is already something going on deep behind?

Solution:

- The name of class is the folder name
- If you just create some random folder with the name "xyz", it will also be considered as a class!! The name itself is saying `flow_from_directory`
- a clear explanation below: https://vijayabhaskar96.medium.com/tutorial-image-classification-with-keras-flow-from-directory-and-generators-95f75ebe5720

Added by Bhaskar Sarma

How are numeric class labels determined in flow_from_directroy using binary class mode and what is meant by the single probability predicted by a binary Keras model:

The command to read folders in the dataset in the tensorflow source code is:

```
for subdir in sorted(os.listdir(directory)):
      ...
```

```
Reference:
https://github.com/keras-team/keras/blob/master/keras/preprocessing/image.py, line 563
```

This means folders will be read in alphabetical order. For example, in the case of a folder named *dino*, and another named *dragon*, dino will read first and will have class label 0, whereas dragon will be read in next and will have class label 1.

When a Keras model predicts binary labels, it will only return one value, and this is the probability of class 1 in case of **sigmoid activation function in the last dense layer with 2 neurons**. The probability of class 0 can be found out by:

```
prob(class(0)) = 1- prob(class(1))
```

In case of using `from_logits` to get results, you will get two values for each of the labels.

A prediction of 0.8 is saying the probability that the image has class label 1 (in this case dragon), is 0.8, and conversely we can infer the probability that the image has class label 0 is 0.2.

(Added by Memoona Tahira)

Does the actual values matter after predicting with a neural network or it should be treated as like hood of falling in a class?

It's fine, some small changes are expected

Alexey Grigorev

9. Serverless Deep Learning

How to get started with Week 9?

Executing the command echo ${REMOTE_URI} returns nothing.

Solution description

In the unit **9.6**, Alexey ran the command *echo ${REMOTE_URI}* which turned the URI address in the terminal. There workaround is to set a local variable (REMOTE_URI) and assign your URI address in the terminal and use it to login the registry, for instance, REMOTE_URI=2278222782.dkr.ecr.ap-south-1.amazonaws.com/clothing-tflite-images (fake address). One caveat is that you will lose this variable once the session is terminated.

I also had the same problem on Ubuntu terminal. I executed the following two commands:

```
$ export REMOTE_URI=1111111111.dkr.ecr.us-west-1.amazonaws.com/clothing-tflite-images:clothing-model-xception-v4-001
$ echo $REMOTE_URI
111111111.dkr.ecr.us-west-1.amazonaws.com/clothing-tflite-images:clothing-model-xception-v4-001
```

Note: 1. no curly brackets (e.g. echo ${REMOTE_URI}) needed unlike in video 9.6, 2. Replace `REMOTE_URI` with your `URI`

(Bhaskar Sarma)

Getting a syntax error while trying to get the password from aws-cli

The command `aws ecr get-login --no-include-email` returns an invalid choice error:

The solution is to use the following command instead: `aws ecr get-login-password`
Could simplify the login process with, just replace the <ACCOUNT_NUMBER> and <REGION> with your values:

```
export PASSWORD=`aws ecr get-login-password`
docker login -u AWS -p $PASSWORD
<ACCOUNT_NUMBER>.dkr.ecr.<REGION>.amazonaws.com/clothing-tflite-images
```

Added by Martin Uribe

Getting `ERROR [internal] load metadata for public.ecr.aws/lambda/python:3.8`

This error is produced sometimes when building your docker image from the Amazon python base image.

Solution description: The following could solve the problem.

1. Update your docker desktop if you haven't done so.
2. Or restart docker desktop and terminal and then build the image all over again.
3. Or if all else fails, first run the following command: `DOCKER_BUILDKIT=0 docker build .` then build your image.

(optional) Added by Odimegwu David

Problem: 'ls' is not recognized as an internal or external command, operable program or batch file.

When trying to run the command **!ls -lh** in windows jupyter notebook , I was getting an error message that says "'ls' is not recognized as an internal or external command,operable program or batch file.

Solution description :
Instead of **!**ls -lh , you can use this command **!dir ,** and you will get similar output

Asia Saeed

ImportError: generic_type: type "InterpreterWrapper" is already registered!

When I run **import tflite_runtime.interpreter as tflite , I get an error message says "ImportError: generic_type: type "InterpreterWrapper" is already registered!"**

Solution description

This error occurs when you import both tensorflow and tflite_runtime.interpreter **"import tensorflow as tf" and "import tflite_runtime.interpreter as tflite"** in the same notebook. To fix the issue, restart the kernel and import only tflite_runtime.interpreter **" import tflite_runtime.interpreter as tflite".**

Asia Saeed

Windows version might not be up-to-date

Problem description:
In command line try to do `$ docker build -t dino_dragon`
got this

```
Using default tag: latest
[2022-11-24T06:48:47.360149000Z][docker-credential-desktop][W]
Windows version might not be up-to-date: The system cannot
find the file specified.
error during connect: This error may indicate that the docker
daemon is not running.: Post
```

.
Solution description:
You need to make sure that Docker is not stopped by a third-party program.

Andrei Ilin

WARNING: You are using pip version 22.0.4; however, version 22.3.1 is available

When running docker build -t dino-dragon-model it returns the above error

The most common source of this error in this week is because Alex video shows a version of the wheel with python 8, we need to find a wheel with the version that we are working on. In this case python 9. Another common error is to copy the link, this will also produce the same error, we need to download the raw format:

https://github.com/alexeygrigorev/tflite-aws-lambda/raw/main/tflite/tflite_runtime-2.7.0-cp39-cp39-linux_x86_64.whl

Pastor Soto

How to do AWS configure after installing awscli

Problem description:
In video 9.6, after installing aswcli, we should configure it with aws configure . There it asks for Access Key ID, Secret Access Key, Default Region Name and also Default output format. What we should put for Default output format? Leaving it as None is okay?

Solution description:
Yes, in my I case I left everything as the provided defaults (except, obviously, the Access key and the secret access key)

Added by Bhaskar Sarma

Object of type float32 is not JSON serializable

Problem:

While passing local testing of the lambda function without issues, trying to test the same input with a running docker instance results in an error message like

{'errorMessage': 'Unable to marshal response: Object of type float32 is not JSON serializable', 'errorType': 'Runtime.MarshalError', 'requestId': 'f155492c-9af2-4d04-b5a4-639548b7c7ac', 'stackTrace': []}

This happens when a model (in this case the dino vs dragon model) returns individual estimation values as numpy float32 values (arrays). They need to be converted individually to base-Python floats in order to become "serializable".

Solution:

In my particular case, I set up the dino vs dragon model in such a way as to return a label + predicted probability for each class as follows (below is a two-line extract of function predict() in the lambda_function.py):

```
preds = [interpreter.get_tensor(output_index)[0][0], \
          1-interpreter.get_tensor(output_index)[0][0]]
```

In which case the above described solution will look like this:

```
preds = [float(interpreter.get_tensor(output_index)[0][0]), \
        float(1-interpreter.get_tensor(output_index)[0][0])]
```

The rest can be made work by following the chapter 9 (and/or chapter 5!) lecture videos step by step.

Added by Konrad Muehlberg

How do Lambda container images work?

I wanted to understand how lambda container images work in depth and how lambda functions are initialized, for this reason, I found the following documentation

https://docs.aws.amazon.com/lambda/latest/dg/images-create.html
https://docs.aws.amazon.com/lambda/latest/dg/runtimes-api.html

Added by Alejandro aponte

Error building docker image on M1 Mac

Problem:

While trying to build docker image in Section 9.5 with the command:

```
docker build -t clothing-model .
```

It throws a pip install error for the tflite runtime whl

```
#6 0.528 ERROR: tflite_runtime-2.7.0-cp38-cp38-linux_x86_64.whl is not a supported wheel on this platform
```

Solution:

To build the Docker image, use the command:

```
docker build --platform linux/amd64 -t clothing-model .
```

To run the built image, use the command:

```
docker run -it --rm -p 8080:8080 --platform linux/amd64 clothing-model:latest
```

Added by Daniel Egbo

Error invoking API Gateway deploy API locally

Problem: Trying to test API gateway in 9.7 - API Gateway: Exposing the Lambda Function, running: `$ python test.py`

With error message:

```
{'message': 'Missing Authentication Token'}
```

Solution:

Need to get the deployed API URL for the specific path you are invoking. Example:

```
https://<random string>.execute-api.us-east-2.amazonaws.com/test/predict
```

Added by Andrew Katoch

Error: Could not find a version that satisfies the requirement tflite_runtime (from versions:none)

Problem: When trying to install tflite_runtime on Windows OS one gets an error message above. The thing is that tflite_runtime has no versions for Windows

Solution:

Use a virtual machine (with VM VirtualBox, for example) with a Linux system. The other way is to run a code at a virtual machine within cloud service, for example you can use Vertex AI Workbench at GCP (notebooks and terminals are provided there, so all tasks may be performed).

Added by Alena Kniazeva

Running out of space for AWS instance.

Due to experimenting back and forth so much without care for storage, I just ran out of it on my 30-GB AWS instance. It turns out that deleting docker images does not actually free up any space as you might expect. After removing images, you also need to run **docker system prune**

10. Kubernetes and TensorFlow Serving

How to get started with Week 10?

How to install Tensorflow in Ubuntu WSL2

Running a CNN on your CPU can take a long time and once you've run out of free time on some cloud providers, it's time to pay up. Both can be tackled by installing tensorflow with CUDA support on your local machine if you have the right hardware.

I was able to get it working by using the following resources:

- CUDA on WSL :: CUDA Toolkit Documentation (nvidia.com)
- Install TensorFlow with pip
- Start Locally | PyTorch

I included the link to PyTorch so that you can get that one installed and working too while everything is fresh on your mind. Just select your options, and for Computer Platform, I chose CUDA 11.7 and it worked for me.

Added by Martin Uribe

Getting: Allocator ran out of memory errors?

If you are running tensorflow on your own machine and you start getting the following errors:

```
Allocator (GPU_0_bfc) ran out of memory trying to allocate
6.88GiB with freed_by_count=0. The caller indicates that this
is not a failure, but this may mean that there could be
performance gains if more memory were available.
```

Try adding this code in a cell at the beginning of your notebook:

```
config = tf.compat.v1.ConfigProto()
config.gpu_options.allow_growth = True
session = tf.compat.v1.Session(config=config)
```

After doing this most of my issues went away. I say most because there was one instance when I still got the error once more, but only during one epoch. I ran the code again, right after it finished, and I never saw the error again.

Added by Martin Uribe

Problem with recent version of protobuf

In session 10.3, when creating the virtual environment with pipenv and trying to run the script gateway.py, you might get this error:

```
TypeError: Descriptors cannot not be created directly.

If this call came from a _pb2.py file, your generated code is out of date and must
be regenerated with protoc >= 3.19.0.
If you cannot immediately regenerate your protos, some other possible workarounds
are:
 1. Downgrade the protobuf package to 3.20.x or lower.

 2. Set PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION=python (but this will use
pure-Python parsing and will be much slower).


More information:
https://developers.google.com/protocol-buffers/docs/news/2022-05-06#python-updates
```

This will happen if your version of protobuf is one of the newer ones. As a workaround, you can fix the protobuf version to an older one. In my case I got around the issue by creating the environment with:

```
pipenv install --python 3.9.13 requests grpcio==1.42.0 flask gunicorn \
            keras-image-helper tensorflow-protobuf==2.7.0 protobuf==3.19.6
```

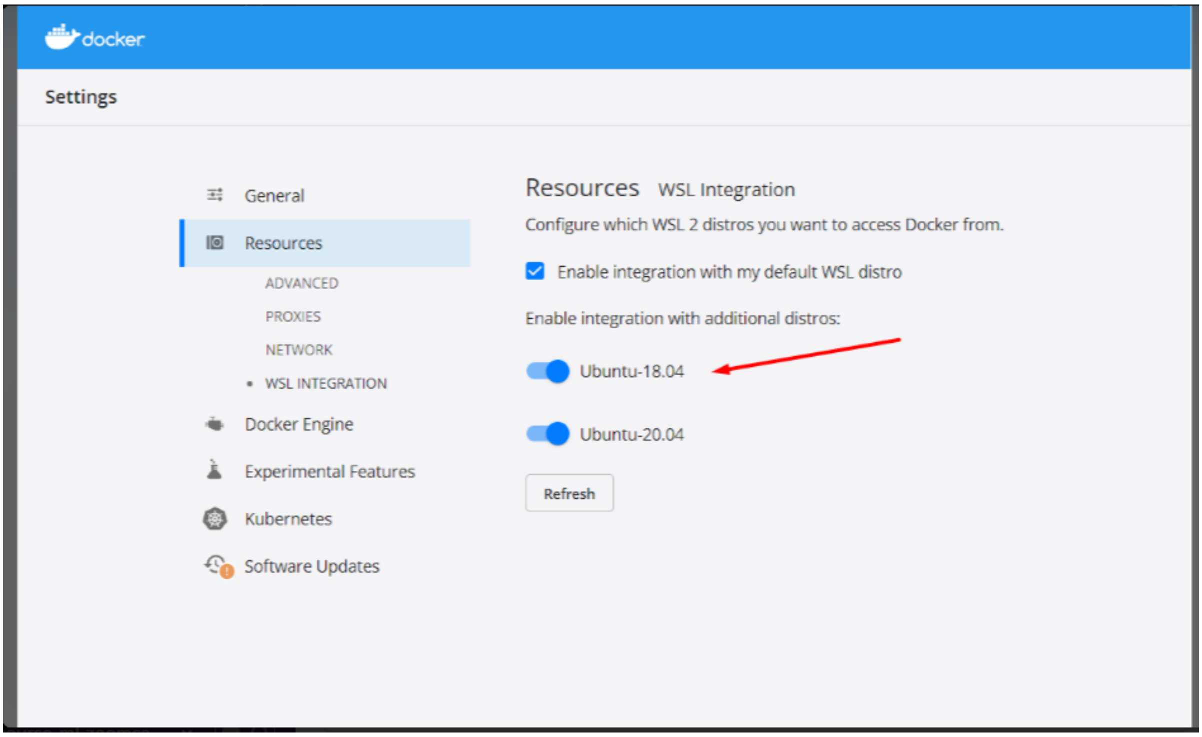
Added by Ángel de Vicente

WSL Cannot Connect To Docker Daemon

Due to the uncertainties associated with machines, sometimes you can get the error message like this when you try to run a docker command:

```
”Cannot connect to the Docker daemon at
unix:///var/run/docker.sock. Is the docker daemon
running?”
```

Solution: The solution is simple. The Docker Desktop might no longer be connecting to the WSL Linux distro. What you need to do is go to your Docker Desktop setting and then click on resources. Under resources, click on WSL Integration. You will get a tab like the image below:



Resources WSL Integration

Configure which WSL 2 distros you want to access Docker from.

☑ Enable integration with my default WSL distro

Enable integration with additional distros:

● Ubuntu-18.04

● Ubuntu-20.04

Just enable additional distros. That’s all. Even if the additional distro is the same as the default WSL distro.

Odimegwu David

HPA instance doesn’t run properly

In case the HPA instance does not run correctly even after installing the latest version of Metrics Server from the components.yaml manifest with:

```
>>kubectl apply -f
https://github.com/kubernetes-sigs/metrics-server/releases/latest/download/components.yaml
```

And the targets still appear as <unknown>

Run >>kubectl edit deploy -n kube-system metrics-server

And search for this line:

```
args:
 - --kubelet-preferred-address-types=InternalIP,ExternalIP,Hostname
```

Add this line in the middle: - --kubelet-insecure-tls

So that it stays like this:

```
args:
 - --kubelet-insecure-tls
 - --kubelet-preferred-address-types=InternalIP,ExternalIP,Hostname
```

Save and run again >>kubectl get hpa

Added by Marilina Orihuela

HPA instance doesn't run properly (easier solution)

In case the HPA instance does not run correctly even after installing the latest version of Metrics Server from the components.yaml manifest with:

```
>>kubectl apply -f
https://github.com/kubernetes-sigs/metrics-server/releases/latest/download/components.yaml
```

And the targets still appear as <unknown>

Run the following command:

```
kubectl apply -f
https://raw.githubusercontent.com/Peco602/ml-zoomcamp/main/10-kubernetes/kube-config/metrics-server-deployment.yaml
```

Which uses a metrics server deployment file already embedding the - --kubelet-insecure-tls option.

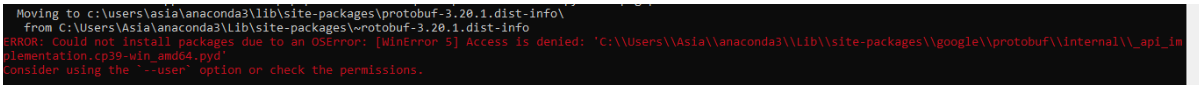
Added by Giovanni Pecoraro

Could not install packages due to an OSError: [WinError 5] Access is denied

When I run **pip install grpcio==1.42.0 tensorflow-serving-api==2.7.0** to install the libraries in windows machine, I was getting the below error :

ERROR: Could not install packages due to an OSError: [WinError 5] Access is denied: 'C:\\Users\\Asia\\anaconda3\\Lib\\site-packages\\google\\protobuf\\internal\\_api_implementation.cp39-win_amd64.pyd'

Consider using the `--user` option or check the permissions.



```
  Moving to c:\users\asia\anaconda3\lib\site-packages\protobuf-3.20.1.dist-info\
   from C:\Users\Asia\anaconda3\Lib\site-packages\~rotobuf-3.20.1.dist-info
ERROR: Could not install packages due to an OSError: [WinError 5] Access is denied: 'C:\\Users\\Asia\\anaconda3\\Lib\\site-packages\\google\\protobuf\\internal\\_api_im
plementation.cp39-win_amd64.pyd'
Consider using the `--user` option or check the permissions.
```

Solution description :

I was able to install the libraries using below command:

pip **--user** install grpcio==1.42.0 tensorflow-serving-api==2.7.0

Asia Saeed

TypeError: Descriptors cannot not be created directly.

Problem description

I was getting the below error message when I run gateway.py after modifying the code & creating virtual environment in video 10.3 :

```
 File "C:\Users\Asia\Data_Science_Code\Zoompcamp\Kubernetes\gat.py", line 9, in <module>
    from tensorflow_serving.apis import predict_pb2
  File "C:\Users\Asia\.virtualenvs\Kubernetes-Ge6Ts1D5\lib\site-packages\tensorflow_serving\apis\predict_pb2.py", line 14, in <module>
    from tensorflow.core.framework import tensor_pb2 as tensorflow_dot_core_dot_framework_dot_tensor__pb2
  File "C:\Users\Asia\.virtualenvs\Kubernetes-Ge6Ts1D5\lib\site-packages\tensorflow\core\framework\tensor_pb2.py", line 14, in <module>
    from tensorflow.core.framework import resource_handle_pb2 as tensorflow_dot_core_dot_framework_dot_resource__handle__pb2
  File "C:\Users\Asia\.virtualenvs\Kubernetes-Ge6Ts1D5\lib\site-packages\tensorflow\core\framework\resource_handle_pb2.py", line 14, in <module>
    from tensorflow.core.framework import tensor_shape_pb2 as tensorflow_dot_core_dot_framework_dot_tensor__shape__pb2
  File "C:\Users\Asia\.virtualenvs\Kubernetes-Ge6Ts1D5\lib\site-packages\tensorflow\core\framework\tensor_shape_pb2.py", line 36, in <module>
    _descriptor.FieldDescriptor(
  File "C:\Users\Asia\.virtualenvs\Kubernetes-Ge6Ts1D5\lib\site-packages\google\protobuf\descriptor.py", line 560, in __new__
    _message.Message._CheckCalledFromGeneratedFile()
TypeError: Descriptors cannot not be created directly.
If this call came from a _pb2.py file, your generated code is out of date and must be regenerated with protoc >= 3.19.0.
If you cannot immediately regenerate your protos, some other possible workarounds are:
 1. Downgrade the protobuf package to 3.20.x or lower.
 2. Set PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION=python (but this will use pure-Python parsing and will be much slower).
```

Solution description:
Issue has been resolved by downgrading protobuf to version 3.20.1.

```
  pipenv install protobuf==3.20.1
```

Asia Saeed

Connection refused

I ran into an issue where **kubectl** wasn't working.
I kept getting the following error:

```
kubectl get service
The connection to the server localhost:8080 was refused - did you specify the right host or port?
```

I searched online for a resolution, but everyone kept talking about creating an environment variable and creating some **admin.config** file in my home directory. All hogwash.

The solution to my problem was to just start over.

```
kind delete cluster
rm -rf ~/.kube
kind create cluster
```

Now when I try the same command again:

```
kubectl get service
```

```
NAME         TYPE        CLUSTER-IP   EXTERNAL-IP   PORT(S)   AGE
kubernetes   ClusterIP   10.96.0.1    <none>        443/TCP   53s
```

Added by Martin Uribe

Running out of storage after building many docker images

Problem description

Due to experimenting back and forth so much without care for storage, I just ran out of it on my 30-GB AWS instance.

My first reflex was to remove some zoomcamp directories, but of course those are mostly code so it didn't help much.

Solution description

> docker images

revealed that I had over 20 GBs worth of superseded / duplicate models lying around, so I proceeded to > docker rmi
a bunch of those — but to no avail!

It turns out that deleting docker images does not actually free up any space as you might expect. After removing images, you also need to run

> docker system prune

See also:
https://stackoverflow.com/questions/36799718/why-removing-docker-containers-and-images-does-not-free-up-storage-space-on-wind

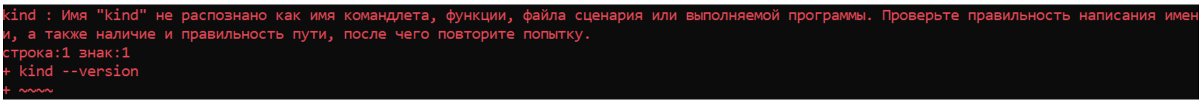
Added by Konrad Mühlberg

In HW10 Q6 what does it mean "correct value for CPU and memory"? Aren't they arbitrary?

Yes, the question does require for you to specify values for CPU and memory in the yaml file, however the question that it is use in the form only refers to the port which do have a define correct value for this specific homework.

Pastor Soto

Problem with kind



```
kind : Имя "kind" не распознано как имя командлета, функции, файла сценария или выполняемой программы. Проверьте правильность написания имен
и, а также наличие и правильность пути, после чего повторите попытку.
строка:1 знак:1
+ kind --version
+ ~~~~
```

Solution: run from the folder in which you put kind.exe and use cmd. I have a problem with PowerShell and this command.

Kind cannot load docker image

Problem: Failing to load docker-image to cluster (when you'ved named a cluster)

```
kind load docker-image zoomcamp-10-model:xception-v4-001
ERROR: no nodes found for cluster "kind"
```

Solution: Specify cluster name with `-n`

```
kind -n clothing-model load docker-image
zoomcamp-10-model:xception-v4-001
```

Andrew Katoch

'kind' is not recognized as an internal or external command, operable program or batch file. (In Windows)

Problem: I download kind from the next command:

```
curl.exe -Lo kind-windows-amd64.exe
https://kind.sigs.k8s.io/dl/v0.17.0/kind-windows-amd64
```

```
When I try
   -  kind --version
```

I get: 'kind' is not recognized as an internal or external command, operable program or batch file

Solution: The default name of executable is kind-windows-amd64.exe, so that you have to rename this file to kind.exe. Put this file in specific folder, and add it to PATH

Alejandro Aponte

Kubernetes-dashboard

Deploy and Access the Kubernetes Dashboard

Luke

General issues with eksctl

Make sure you are on AWS CLI v2 (check with `aws --version`)
https://docs.aws.amazon.com/cli/latest/userguide/cliv2-migration-instructions.html

Regarding the ports which can be used or assign?

TypeError: __init__() got an unexpected keyword argument 'unbound_message' while importing Flask

Problem Description:
In video 10.3, when I was testing a flask service, I got the above error. I ran `docker run ..` in one terminal. When in second terminal I run `python gateway.py`, I get the above error.

Solution: This error has something to do with versions of Flask and Werkzeug. I got the same error, if I just import flask with `from flask import Flask`.
By running `pip freeze > requirements.txt`, I found that their versions are Flask==2.2.2 and Werkzeug==2.2.2. This error appears while using an old version of werkzeug (2.2.2) with new version of flask (2.2.2). I solved it by pinning version of Flask into an older version with `pipenv install Flask==2.1.3`.

Added by Bhaskar Sarma

Command `aws ecr get-login --no-include-email` returns "aws: error: argument operation: Invalid choice…"

As per AWS documentation:

https://docs.aws.amazon.com/AmazonECR/latest/userguide/docker-push-ecr-image.html

You need to do: (change the fields in red)

```
aws ecr get-login-password --region region | docker login --username
AWS --password-stdin aws_account_id.dkr.ecr.region.amazonaws.com
```

Added by Humberto Rodriguez

Error downloading tensorflow/serving:2.7.0 on Apple M1 Mac

While trying to run the docker code on M1:

```
docker run --platform linux/amd64 -it --rm \
 -p 8500:8500 \
 -v $(pwd)/clothing-model:/models/clothing-model/1 \
 -e MODEL_NAME="clothing-model" \
 tensorflow/serving:2.7.0
```

It outputs the error:

```
Error:
Status: Downloaded newer image for tensorflow/serving:2.7.0
[libprotobuf FATAL
external/com_google_protobuf/src/google/protobuf/generated_message_reflection
.cc:2345] CHECK failed: file != nullptr:
terminate called after throwing an instance of
'google::protobuf::FatalException'
 what():  CHECK failed: file != nullptr:
qemu: uncaught target signal 6 (Aborted) - core dumped
/usr/bin/tf_serving_entrypoint.sh: line 3:     8 Aborted
tensorflow_model_server --port=8500 --rest_api_port=8501
--model_name=${MODEL_NAME} --model_base_path=${MODEL_BASE_PATH}/${MODEL_NAME}
"$@"
```

Solution

```
docker pull emacski/tensorflow-serving:latest



docker run -it --rm \
 -p 8500:8500 \
 -v $(pwd)/clothing-model:/models/clothing-model/1 \
 -e MODEL_NAME="clothing-model" \
 emacski/tensorflow-serving:latest-linux_arm64
```

See more here: https://github.com/emacski/tensorflow-serving-arm

Added by Daniel Egbo

11. KServe

Errors with istio during installation

Problem description:
Running this:

```
curl -s "https://raw.githubusercontent.com/kserve/kserve/release-0.9/hack/quick_install.sh" | bash
```

Fails with errors because of istio failing to update resources, and you are on kubectl > 1.25.0.
Check kubectl version with `kubectl version`

Solution description
Edit the file “quick_install.bash” by downloading it with curl without running bash. Edit the versions of Istio and Knative as per the matrix on the KServe website.
Run the bash script now.

Added by Andrew Katoch

What If I submitted only two projects and failed to submit the third?

If you submitted two projects, you will get the certificate for the course. According to the course coordinator, Alexey Grigorev, only two projects are needed to get the course certificate.

(optional) David Odimegwu

Problem title

Problem description
Solution description

(optional) Added by Name

Projects (Midterm and Capstone)

What are the project deadlines?

You can see them here (it’s taken from the 2022 cohort page)

Mid-Term Project (Crucial Links)

- Midterm Project Sample: https://github.com/alexeygrigorev/mlbookcamp-code/tree/master/course-zoomcamp/cohorts/2021/07-midterm-project
- MidTerm Project Deliverables: https://github.com/alexeygrigorev/mlbookcamp-code/tree/master/course-zoomcamp/projects
- Submit MidTerm Project: https://docs.google.com/forms/d/e/1FAIpQLSfgmOk0QrmHu5t0H6Ri1Wy_FDVS8I_nr5lY3sufkgk18I6S5A/viewform
- Datasets:
  - https://www.kaggle.com/datasets and https://www.kaggle.com/competitions
  - https://archive.ics.uci.edu/ml/index.php
  - https://data.europa.eu/en
  - https://www.openml.org/search?type=data
  - https://newzealand.ai/public-data-sets
  - https://datasetsearch.research.google.com
- What to do and Deliverables

  - Think of a problem that's interesting for you and find a dataset for that
  - Describe this problem and explain how a model could be used
  - Prepare the data and doing EDA, analyze important features
  - Train multiple models, tune their performance and select the best model
  - Export the notebook into a script
  - Put your model into a web service and deploy it locally with Docker
  - Bonus points for deploying the service to the cloud

~~~ Added by Nukta Bhatia ~~~

Computing the hash for project review

See the answer here.

Learning in public links for the projects

For the learning in public for this midterm project it seems that has a total value of 14!, Does this mean that we need make 14 posts?, Or the regular seven posts for each module and each one with a value of 2?, Or just one with a total value of 14?

14 posts, one for each day

Why do I need to provide a train.py file when I already have the notebook.ipynb file?

Answer: The train.py file will be used by your peers to review your midterm project. It is for them to cross-check that your training process works on someone else's system. It should also be included in the environment in conda or with pipenv.

Odimegwu David
~~~

Is a train.py file necessary when you have a train.ipynb file in your midterm project directory?

Ans: **train.py** has to be a python file. This is because running a python script for training a model is much simpler than running a notebook and that's how training jobs usually look like in real life.

Is there a way to serve up a form for users to enter data for the model to crunch on?

Yes, you can create a mobile app or interface that manages these forms and validations. But you should also perform validations on backend.

You can also check Streamlit: https://github.com/DataTalksClub/project-of-the-week/blob/main/2022-08-14-frontend.md

Alejandro Aponte

How to get feature importance for XGboost model

Using model.feature_importances_ can gives you an error:

```
AttributeError: 'Booster' object has no attribute
'feature_importances_'
```

Answer: if you train the model like this: model = xgb.train you should use get_score() instead

Ekaterina Kutovaia

[Errno 12] Cannot allocate memory in AWS Elastic Container Service

In the Elastic Container Service task log, error “[Errno 12] Cannot allocate memory” showed up.

Just increase the RAM and CPU in your task definition.

Humberto Rodriguez

Pickle error: can’t get attribute XXX on module __main__

When running a docker container with waitress serving the app.py for making predictions, pickle will throw an error that can't get attribute <name_of_class> on module __main__.

This does not happen when Flask is used directly, i.e. not through waitress.

The problem is that the model uses a custom column transformer class, and when the model was saved, it was saved from the __main__ module (e.g. python train.py). Pickle will reference the class in the global namespace (top-level code): __main__.<custom_class>.

When using waitress, waitress will load the predict_app module and this will call pickle.load, that will try to find __main__.<custom_class> that does not exist.

Solution:
Put the class into a separate module and import it in both the script that saves the model (e.g. train.py) and the script that loads the model (e.g. predict.py)

Note: If Flask is used (no waitress) in predict.py, and predict.py has the definition of the class, When it is run: python predict.py, it will work because the class is in the same namespace as the one used when the model was saved (__main__).

Detailed info:
https://stackoverflow.com/questions/27732354/unable-to-load-files-using-pickle-and-multiple-modules

Marcos MJD

How to handle outliers in a dataset?

There are different techniques, but the most common used are the next:

- Dataset transformation (for example, log transformation)
- Clipping high values
- Dropping these observations

Alena Kniazeva

Failed loading Bento from directory /home/bentoml/bento: Failed to import module "service": No module named 'sklearn'

I was getting the below error message when I was trying to create docker image using bentoml
[bentoml-cli] `serve` failed: Failed loading Bento from directory /home/bentoml/bento: Failed to import module "service": No module named 'sklearn'

Solution description
The cause was because , in bentofile.yaml, I wrote sklearn instead of scikit-learn. Issue was fixed after I modified the packages list as below.

```
packages: # Additional pip packages required by the service
  - xgboost
  - scikit-learn
  - pydantic
```

Asia Saeed

BentoML not working with –production flag at any stage: e.g. with bentoml serve and while running the bentoml container

You might see a long error message with something about sparse matrices, and in the swagger UI, you get a code 500 error with “” (empty string) as output.

Potential reason: Setting DictVectorizer or OHE to sparse while training, and then storing this in a pipeline or custom object in the benotml model saving stage in `train.py`. This means that when the custom object is called in `service.py`, it will convert each input to a different sized sparse matrix, and this can't be batched due to inconsistent length. In this case, bentoml model signatures should have `batchable` set to `False` for production during saving the bentoml mode in `train.py`.

(Memoona Tahira)

Reproducibility

Problem description:

Do we have to run everything?

You are encouraged, if you can, to run them. As this provides another opportunity to learn from others.

Not everyone will be able to run all the files, in particular the neural networks.

Solution description:

Alternatively, can you see that everything you need to reproduce is there: the dataset is there, the instructions are there, are there any obvious errors and so on.

Related slack conversation here.

(Gregory Morris)

Permissions to push docker to Google Container Registry

When you try to push the docker image to Google Container Registry and get this message **"unauthorized: You don't have the needed permissions to perform this operation, and you may have invalid credentials."**, type this below on console, but first install https://cloud.google.com/sdk/docs/install, this is to be able to use gcloud in console:

```
gcloud auth configure-docker
```

(Jesus Acuña)

Tflite_runtime unable to install

I am getting this error message when I tried to install tflite in a pipenv environment

Error: An error occurred while installing tflite_runtime!
Error text:
ERROR: Could not find a version that satisfies the requirement tflite_runtime (from versions: none)
ERROR: No matching distribution found for tflite_runtime

This version of tflite do not run on python 3.10, the way we can make it work is by install python 3.9, after that it would install the tflite_runtime without problem.

Pastor Soto

Error when running ImageDataGenerator.flow_from_dataframe

Error: ImageDataGenerator name 'scipy' is not defined.
Check that scipy is installed in your environment.
Restart jupyter kernel and try again.

Marcos MJD

How to pass BentoML content / docker container to Amazon Lambda

Tim from BentoML has prepared a dedicated video tutorial wrt this use case here:
https://www.youtube.com/watch?v=7gI1UH31xb4&list=PL3MmuxUbc_hIhxl5Ji8t4O6lPAOpHaCLR&index=97

Konrad Muehlberg

Error UnidentifiedImageError: cannot identify image file

In deploying model part, I wanted to test my model locally on a test-image data and I had this silly error after the following command:

```
url = 'https://github.com/bhasarma/kitchenware-classification-project/blob/main/test-image.jpg'
X = preprocessor.from_url(url)
```

I got the error:

```
UnidentifiedImageError: cannot identify image file <_io.BytesIO object at 0x7f797010a590>
```

Solution:

Add `?raw=true` after `.jpg` in url. E.g. as below

```
url =
‘https://github.com/bhasarma/kitchenware-classification-projec
t/blob/main/test-image.jpg?raw=true’
```

Bhaskar Sarma

[pipenv.exceptions.ResolutionFailure]: Warning: Your dependencies could not be resolved. You likely have a mismatch in your sub-dependencies

Problem: If you run pipenv install and get this message. Maybe manually change Pipfile and Pipfile.lock.
Solution: Run: ` pipenv lock` for fix this problem and dependency files

Alejandro Aponte

Miscellaneous

Other course-related questions that don’t fall into any of the categories above

Getting day of the year from day and month column

Problem description: I have one column `day_of_the_month` . It has values 1, 2, 20, 25 etc. and `int` . I have a second column `month_of_the_year`. It has values jan, feb, ..dec. and are `string`. I want to convert these two columns into one column `day_of_the_year` and I want them to be `int`. 2 and jan should give me 2, i.e. 2nd day of the year, 1 and feb should give me 32, i.e. 32 nd day of the year. What is the simplest pandas-way to do it?

Solution description:

1. convert `dtype` in `day_of_the_month` column from `int` to `str` with `df['day_of_the_month'] = df['day_of_the_month'].map(str)`
2. convert `month_of_the_year` column in `jan, feb ...,dec` into `1,2, ..,12` string using `map()`
3. convert day and month into a `datetime` object with:

```
df['date_formatted'] = pd.to_datetime(
     dict(
          year='2055',
          month=df['month'],
          day=df['day']
     )
)
```

4. get day of year with:
   `df['day_of_year']=df['date_formatted'].dt.dayofyear`

(Bhaskar Sarma)

Chart for classes and predictions

How to visualize the predictions per classes after training a neural net

Solution description

```
classes, predictions = zip(*dict(zip(classes,
predictions)).items())
plt.figure(figsize=(12, 3))
plt.bar(classes, predictions)
```

Luke

Convert dictionary values to Dataframe table

You can convert the prediction output values to a datafarme using
df = pd.DataFrame.from_dict(dict, orient='index' , columns=["Prediction"])

Edidiong Esu

Kitchenware Classification Competition Dataset Generator

The image dataset for the competition was in a different layout from what we used in the dino vs dragon lesson. Since that's what was covered, some folks were more comfortable with that setup, so I wrote a script that would generate it for them

It can be found here: kitchenware-dataset-generator | Kaggle

Martin Uribe

CUDA toolkit and cuDNN Install for Tensorflow

Install Nvidia drivers: https://www.nvidia.com/download/index.aspx.
Windows:

- Install Anaconda prompt https://www.anaconda.com/
- Two options:
    - Install package 'tensorflow-gpu' in Anaconda
    - Install the Tensorflow way https://www.tensorflow.org/install/pip#windows-native

WSL/Linux:

- WSL: Use the Windows Nvida drivers, do not touch that.
- Two options:
    - Install the Tensorflow way https://www.tensorflow.org/install/pip#linux_1
        - Make sure to follow step 4 to install CUDA **by environment**
        - Also run:
            - `echo 'export XLA_FLAGS=--xla_gpu_cuda_data_dir=$CONDA_PREFIX/lib/> $CONDA_PREFIX/etc/conda/activate.d/env_vars.sh`
    - Install CUDA toolkit 11.x.x https://developer.nvidia.com/cuda-toolkit-archive
    - Install https://developer.nvidia.com/rdp/cudnn-download

Now you should be able to do training/inference with GPU in Tensorflow

(Learning in public links Links to social media posts where you share your progress with others (LinkedIn, Twitter, etc). Use #mlzoomcamp tag. The scores for this part will be capped at 7 point. Please make sure the posts are valid URLs starting with "https://" Does it mean that I should provide my linkedin link? or it means that I should write a post that I have completed my first assignement? (ANS (by ezehcp7482@gmail.com): Yes, provide the linkedIN link to where you posted.

ezehcp7482@gmail.com:
PROBLEM: Since I had to put up a link to a public repository, I had to use Kaggle and uploading the dataset therein was a bit difficult; but I had to 'google' my way out.
ANS: See this link for a guide (https://www.kaggle.com/code/dansbecker/finding-your-files-in-kaggle-kernels/notebook)

About getting the wrong result when multiplying matrices

When multiplying matrices, the order of multiplication is important.
For example:

A (m x n) * B (n x p) = C (m x p)
B (n x p) * A (m x n) = D (n x n)

C and D are matrices of different sizes and usually have different values. Therefore the order is important in matrix multiplication and changing the order changes the result.

Baran Akın

None of the videos have how to install the environment in Mac, does someone have instructions for Mac with M1 chip?

Refer to
https://github.com/DataTalksClub/machine-learning-zoomcamp/blob/master/01-intro/06-environment.md

(added by Rileen Sinha)

I may end up submitting the assignment late. Would it be evaluated?

Depends on whether the form will still be open. If you're lucky and it's open, you can submit your homework and it will be evaluated. if closed - it's too late.

(Added by Rileen Sinha, based on answer by Alexey on Slack)

Does the github repository need to be public?

Yes. Whoever corrects the homework will only be able to access the link if the repository is public.

(added by Tano Bugelli)

How to install Conda environment in my local machine?

Which ide is recommended for machine learning?