National University of Singapore

School of Computing

CS4225/CS5425
Big Data Systems for Data Science

Assignment 2

Outline

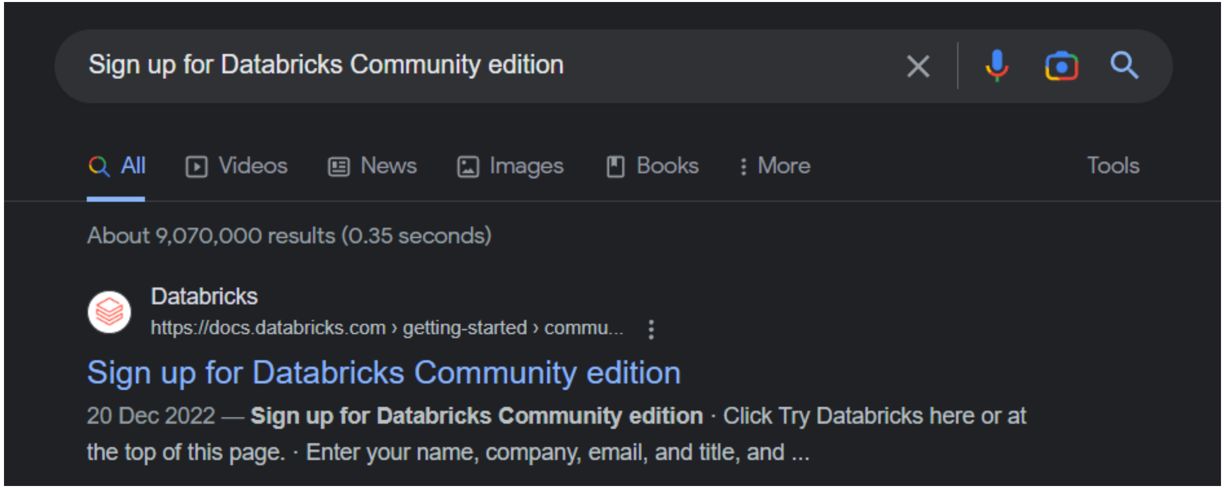
1. Databricks Setup
2. Assignment Description
3. Submission requirements

Using Databricks

- Databricks Community Edition Registration
- Uploading given datasets

Databricks Community Edition Registration

1) Google "Sign up for Databricks Community edition" and click on the first link or click https://docs.databricks.com/getting-started/community-edition.html



Sign up for Databricks Community edition

About 9,070,000 results (0.35 seconds)

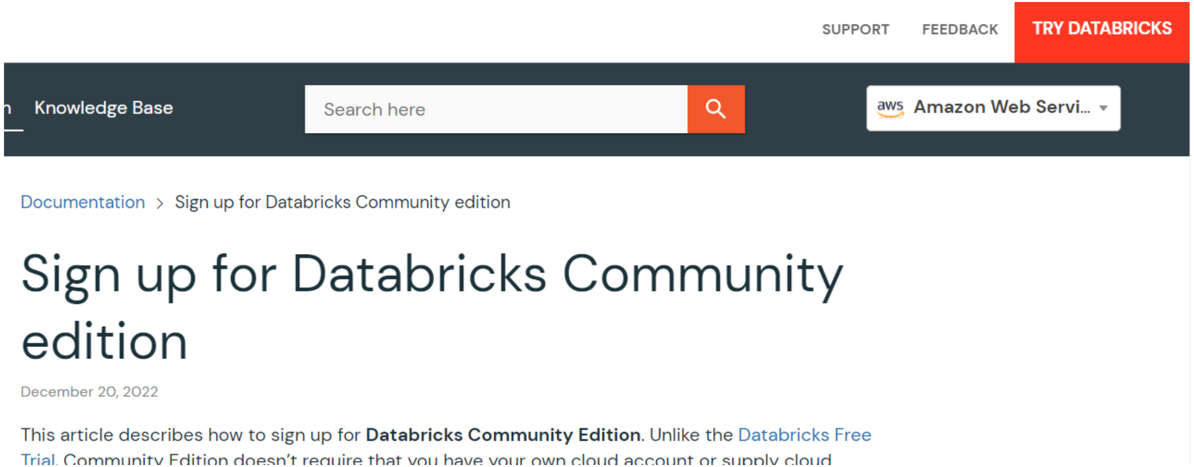
Databricks
https://docs.databricks.com › getting-started › commu...

Sign up for Databricks Community edition

20 Dec 2022 — **Sign up for Databricks Community edition** · Click Try Databricks here or at the top of this page. · Enter your name, company, email, and title, and ...

Databricks Community Edition Registration

2) click “TRY DATABRICKS” on the top right corner of the page



Documentation > Sign up for Databricks Community edition

Sign up for Databricks Community edition

December 20, 2022

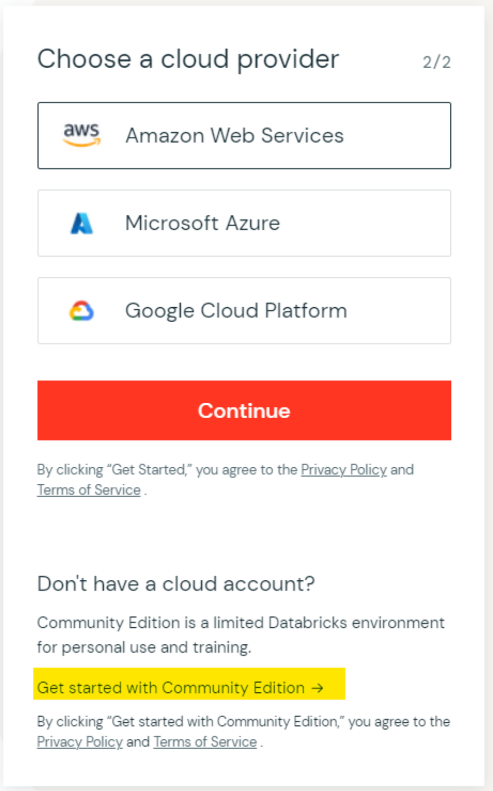
This article describes how to sign up for **Databricks Community Edition**. Unlike the Databricks Free Trial, Community Edition doesn’t require that you have your own cloud account or supply cloud

Databricks Community Edition Registration

3) Fill in the necessary details and click continue to create an account

4) You will reach the page shown and since we are using the Community Edition, please click on the portion highlighted below

5) Validate your email address and login to databricks



Choose a cloud provider 2/2

By clicking "Get Started," you agree to the Privacy Policy and Terms of Service .

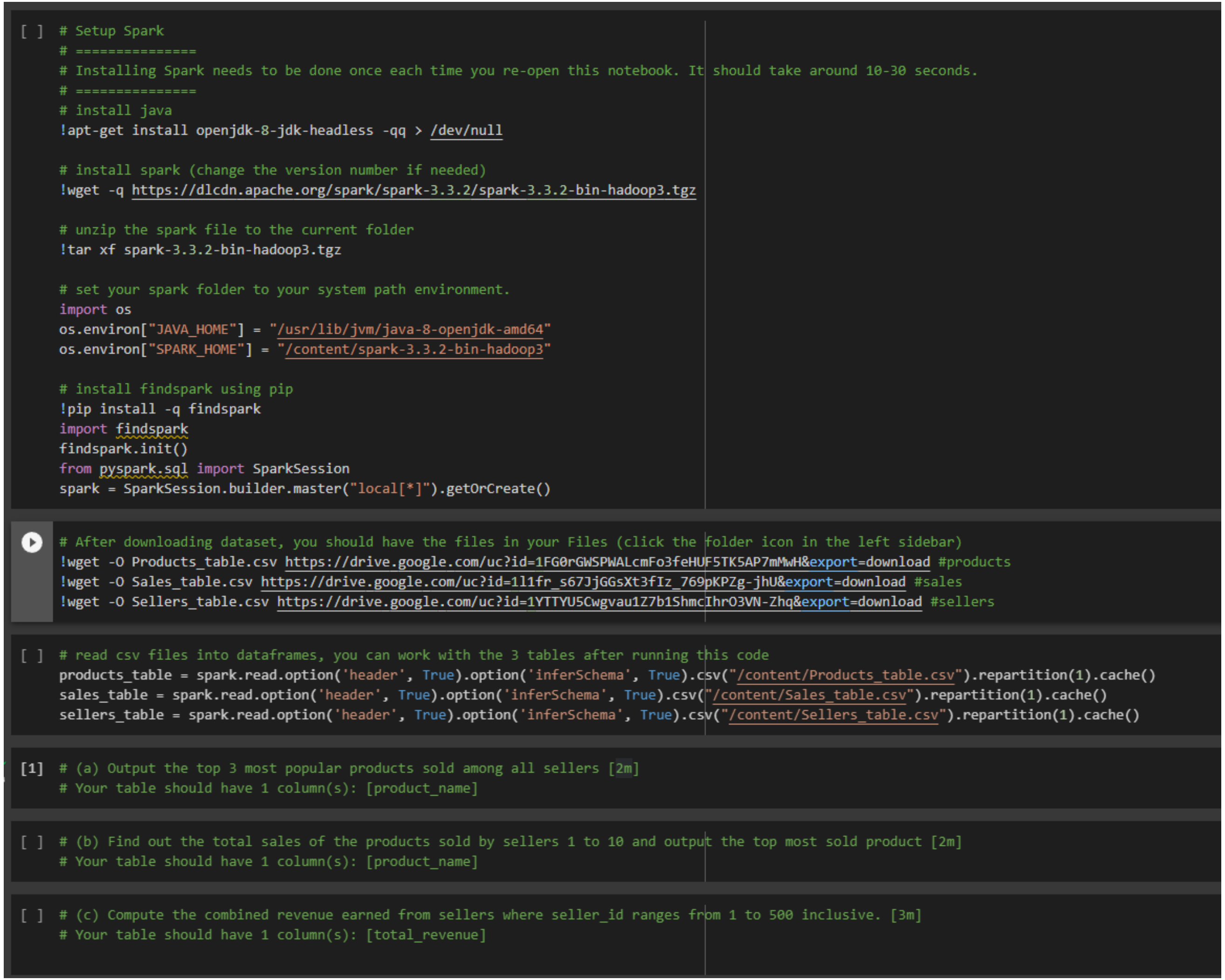
Don't have a cloud account?

Community Edition is a limited Databricks environment for personal use and training.

By clicking "Get started with Community Edition," you agree to the Privacy Policy and Terms of Service .

Google Colab (Backup)

- https://drive.google.com/file/d/1JCk9gXtAFV5q7PPfX8QH8bIX87hBL9bB/view?usp=sharing



```
# Setup Spark
# ==============
# Installing Spark needs to be done once each time you re-open this notebook. It should take around 10-30 seconds.
# ==============
# install java
!apt-get install openjdk-8-jdk-headless -qq > /dev/null

# install spark (change the version number if needed)
!wget -q https://dlcdn.apache.org/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.tgz

# unzip the spark file to the current folder
!tar xf spark-3.3.2-bin-hadoop3.tgz

# set your spark folder to your system path environment.
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.3.2-bin-hadoop3"

# install findspark using pip
!pip install -q findspark
import findspark
findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[*]").getOrCreate()
```

```
# After downloading dataset, you should have the files in your Files (click the folder icon in the left sidebar)
!wget -O Products_table.csv https://drive.google.com/uc?id=1FG0rGWSPWALcmFo3feHUF5TK5AP7mMwH&export=download #products
!wget -O Sales_table.csv https://drive.google.com/uc?id=1l1fr_s67JjGGsXt3fIz_769pKPZg-jhU&export=download #sales
!wget -O Sellers_table.csv https://drive.google.com/uc?id=1YTTYU5Cwgvau1Z7b1ShmcIhrO3VN-Zhq&export=download #sellers
```

```
# read csv files into dataframes, you can work with the 3 tables after running this code
products_table = spark.read.option('header', True).option('inferSchema', True).csv("/content/Products_table.csv").repartition(1).cache()
sales_table = spark.read.option('header', True).option('inferSchema', True).csv("/content/Sales_table.csv").repartition(1).cache()
sellers_table = spark.read.option('header', True).option('inferSchema', True).csv("/content/Sellers_table.csv").repartition(1).cache()
```

```
# (a) Output the top 3 most popular products sold among all sellers [2m]
# Your table should have 1 column(s): [product_name]
```

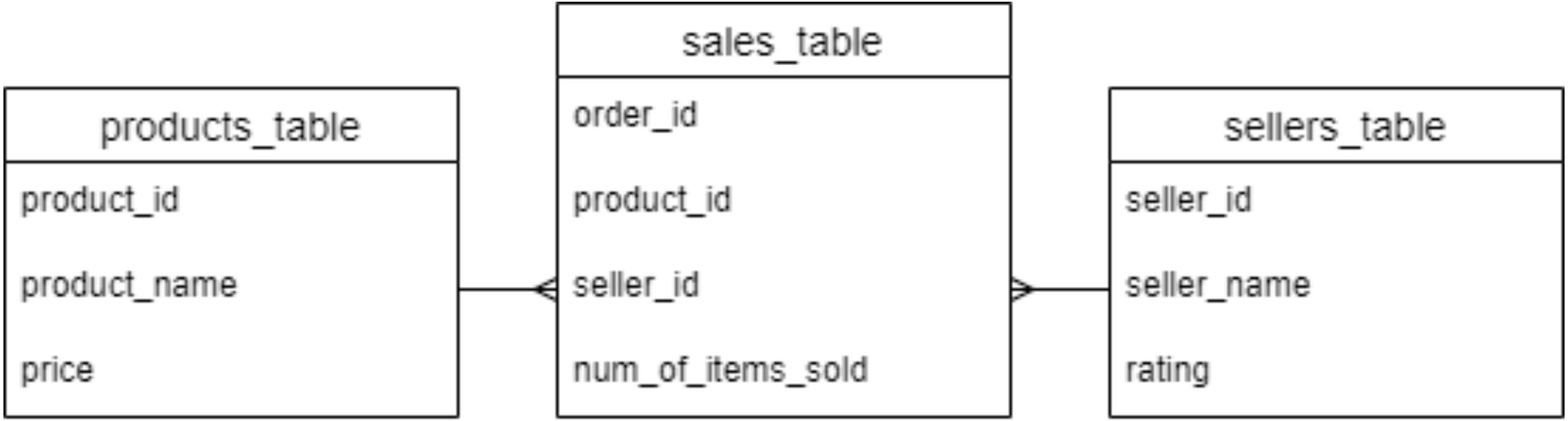
```
# (b) Find out the total sales of the products sold by sellers 1 to 10 and output the top most sold product [2m]
# Your table should have 1 column(s): [product_name]
```

```
# (c) Compute the combined revenue earned from sellers where seller_id ranges from 1 to 500 inclusive. [3m]
# Your table should have 1 column(s): [total_revenue]
```

Assignment 2

- Task 1: Spark SQL (15 Marks)
- Task 2: Spark ML (10 Marks)

Task1: Dataset



products_table

- product_id
- product_name
- price

sales_table

- order_id
- product_id
- seller_id
- num_of_items_sold

sellers_table

- seller_id
- seller_name
- rating

Task1: Questions

a) Output the top 3 most popular products sold among all sellers [2m]

   a. Your table should have 1 column(s): [product_name]

b) Find out the total sales of the products sold by sellers 1 to 10 and output the top most sold product [2m]

   a. Your table should have 1 column(s): [product_name]

c) Compute the combined revenue earned from sellers where seller_id ranges from 1 to 500 inclusive. [3m]

   a. Your table should have 1 column(s): [total_revenue]

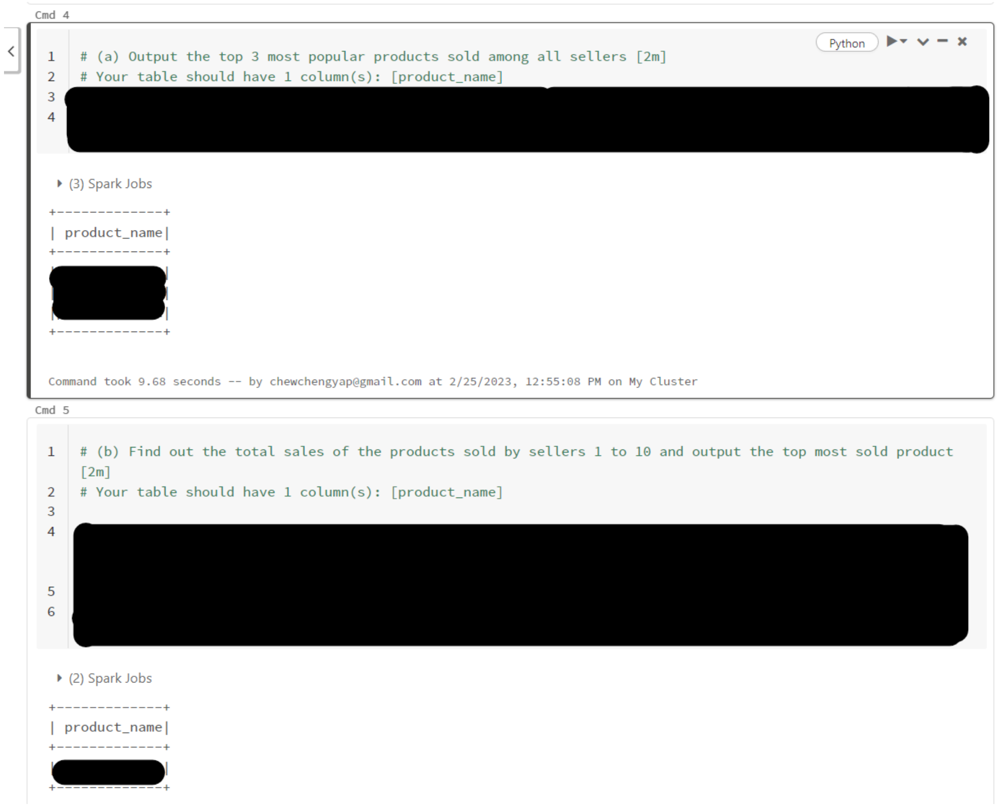
d) Among sellers with rating >= 4 who have achieved a combined number of products sold >= 3000, find out the top 10 most expensive product(product_name) sold by any of the sellers. (If there are multiple products at the same price, please sort them in ascending order of product_id) [8m]

   a. Your table should have 1 column(s): [product_name]

   b. To get the full mark, your query should not run for more than 1 min

Task 1

If you follow the instructions of the questions, there should only be one unique solution for all the questions



Cmd 4

```
# (a) Output the top 3 most popular products sold among all sellers [2m]
# Your table should have 1 column(s): [product_name]
[redacted]
```

▸ (3) Spark Jobs

```
+-------------+
| product_name|
+-------------+
[redacted]
+-------------+
```

Command took 9.68 seconds -- by chewchengyap@gmail.com at 2/25/2023, 12:55:08 PM on My Cluster

Cmd 5

```
# (b) Find out the total sales of the products sold by sellers 1 to 10 and output the top most sold product [2m]
# Your table should have 1 column(s): [product_name]

[redacted]
```

▸ (2) Spark Jobs

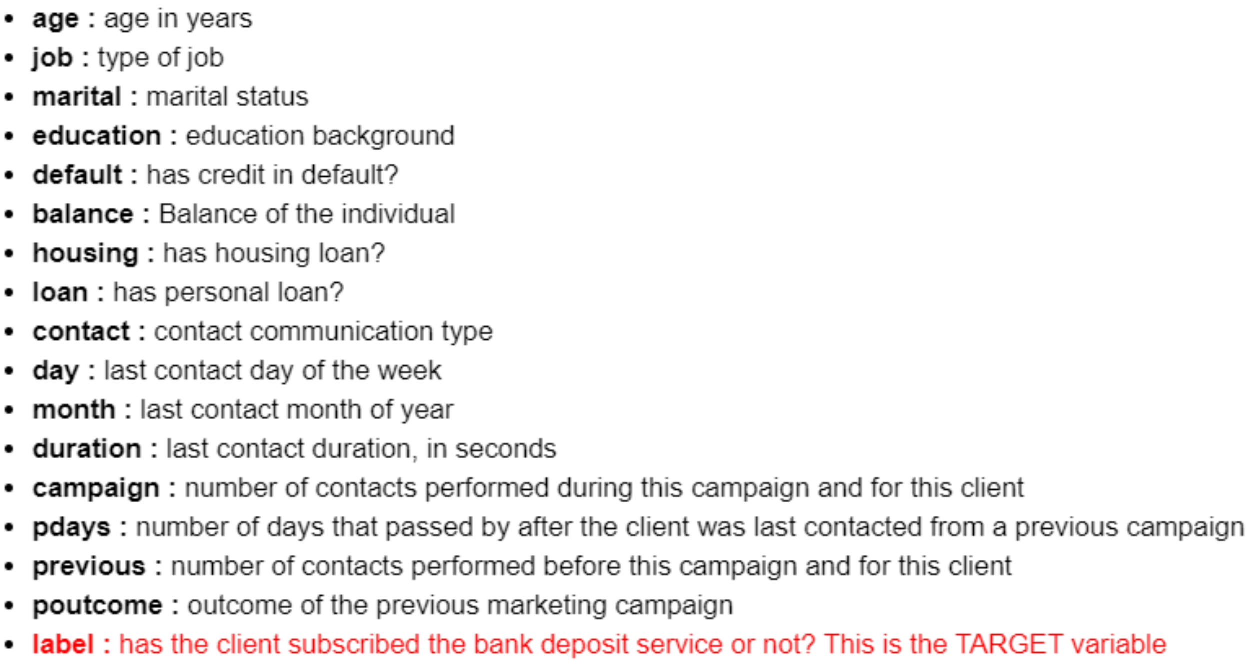
```
+-------------+
| product_name|
+-------------+
[redacted]
+-------------+
```

Task 2: Spark ML

- Build ML model to predict whether the customer will subscribe bank deposit service or not.
- Train the model using training set and evaluate the model performance (e.g. accuracy) using testing set.

Task 2: Dataset

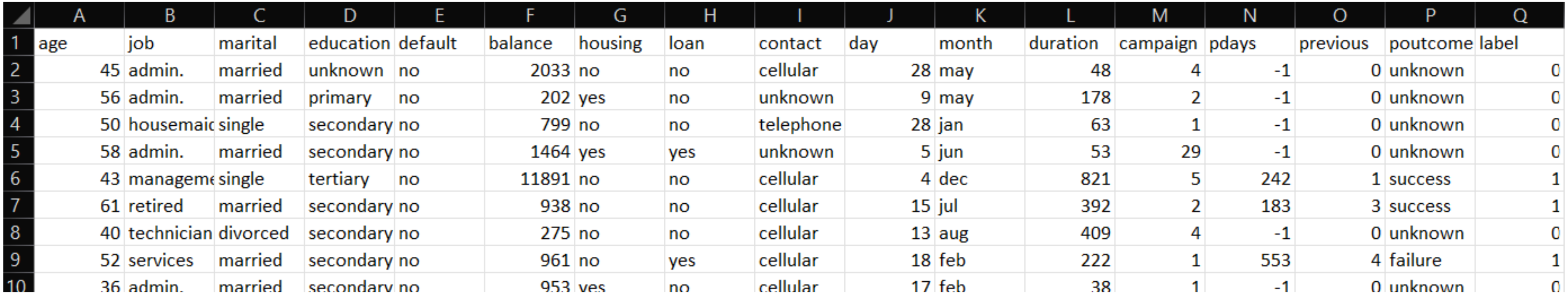
bank marketing campaign data



- **age** : age in years
- **job** : type of job
- **marital** : marital status
- **education** : education background
- **default** : has credit in default?
- **balance** : Balance of the individual
- **housing** : has housing loan?
- **loan** : has personal loan?
- **contact** : contact communication type
- **day** : last contact day of the week
- **month** : last contact month of year
- **duration** : last contact duration, in seconds
- **campaign** : number of contacts performed during this campaign and for this client
- **pdays** : number of days that passed by after the client was last contacted from a previous campaign
- **previous** : number of contacts performed before this campaign and for this client
- **poutcome** : outcome of the previous marketing campaign
- **label** : has the client subscribed the bank deposit service or not? This is the TARGET variable

Task 2: Dataset

- bank_train.csv
- bank_test.csv

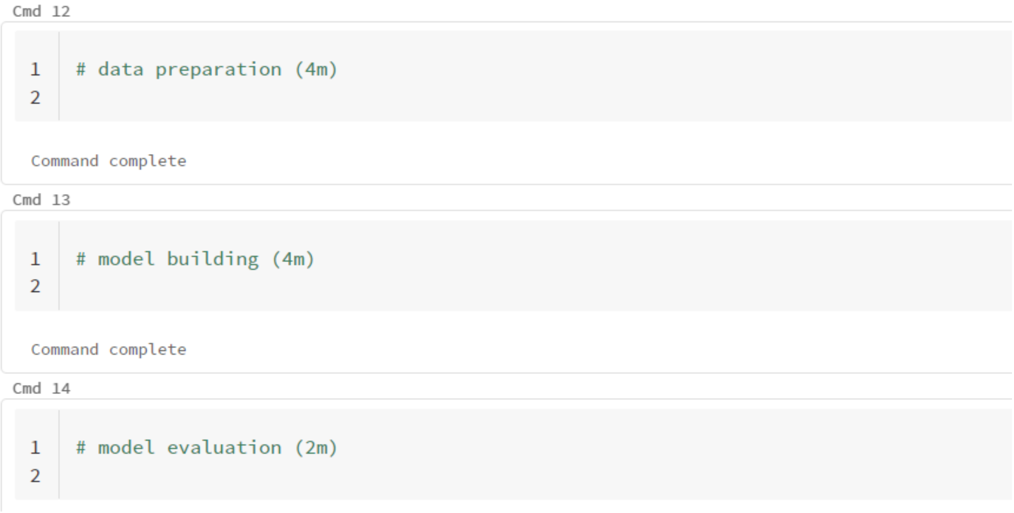


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	label
2	45	admin.	married	unknown	no	2033	no	no	cellular	28	may	48	4	-1	0	unknown	0
3	56	admin.	married	primary	no	202	yes	no	unknown	9	may	178	2	-1	0	unknown	0
4	50	housemaic	single	secondary	no	799	no	no	telephone	28	jan	63	1	-1	0	unknown	0
5	58	admin.	married	secondary	no	1464	yes	yes	unknown	5	jun	53	29	-1	0	unknown	0
6	43	manageme	single	tertiary	no	11891	no	no	cellular	4	dec	821	5	242	1	success	1
7	61	retired	married	secondary	no	938	no	no	cellular	15	jul	392	2	183	3	success	1
8	40	technician	divorced	secondary	no	275	no	no	cellular	13	aug	409	4	-1	0	unknown	0
9	52	services	married	secondary	no	961	no	yes	cellular	18	feb	222	1	553	4	failure	1
10	36	admin.	married	secondary	no	953	yes	no	cellular	17	feb	38	1	-1	0	unknown	0

Task 2: Question

Build ML model to predict whether the customer will subscribe bank deposit service or not. Train the model using training set and evaluate the model performance (e.g. accuracy) using testing set.

- You can explore different methods to pre-process the data and select proper features
- You can utilize different machine learning models and tune model hyperparameters
- Present the final testing accuracy.



Cmd 12

```
# data preparation (4m)

```

Command complete

Cmd 13

```
# model building (4m)

```

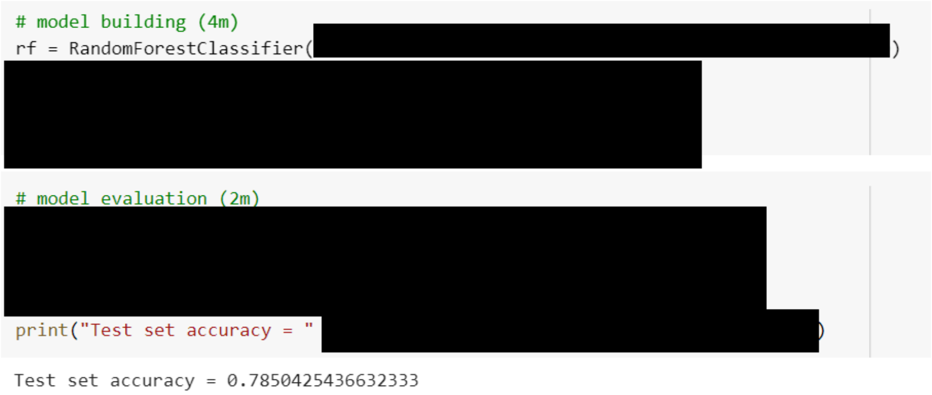
Command complete

Cmd 14

```
# model evaluation (2m)

```

Task 2



```
# model building (4m)
rf = RandomForestClassifier([redacted])
[redacted]
```

```
# model evaluation (2m)
[redacted]
print("Test set accuracy = " [redacted])
```

```
Test set accuracy = 0.7850425436632333
```

Focus is on building a proper ml pipeline so that you can adjust the different parts of pipeline to achieve the best performance.

A baseline model using Random Forest Classifier achieved a test accuracy of around 78.5%. You should at least create a Random Forest model and achieve the similar or higher test accuracy. On top of that, you can also utilize other machine learning models and achieve similar or better test accuracy.

Marking

- Total: 25% of total grade.
  - Task 1: 15m (Code + explanation from video)
  - Task 2: 10m (Code + explanation from video)

Submission Requirements

- Deadline: Apr 3, 2023, Mon 11:59pm

- Submit a zip file which includes:
  - a .ipynb notebook file
  - a recorded video file

- Video Requirements:
  - a video no longer than 5 minutes to show us you understand your codes
  - make sure the video captures your face throughout the full duration

Questions regarding assignment

- Post your questions in your Group at Canvas
- Contact your tutor in charge if you have questions

The End