ML Experiment Management Tools: A Usability & Learnability Study

In this experiment, you will perform a series of tasks with ML experiment management tools. These tools are supposed to assist you in tracking and organizing the assets required and generated when running multiple iterations of ML experiments. The main goal of this experiment is to evaluate usability, learnability, and the level of support offered by these tools. In this experiment, you will be asked to carry out some guided ML tasks while using two tools—Neptune.ai and DVC. This experiment has been designed to guide you through some steps of simple regression tasks to reflect a typical data scientist's workflow, which usually involves multiple runs of experiments before arriving at optimal intended results. For example, you will make changes to the datasets, the learning algorithm, and the learning parameters. These steps will be done using support from the two target tools and an additional manual approach with "no-tool."

After being guided through the ML tasks using the subject tools (i.e., Neptune, DVC, and "No-Tool"), you will be asked traceability-related questions to help evaluate how well you are able to track, query, and retrieve the tracked assets. Finally, you will be asked to answer general questions about your experience with the tools.

Participant's Details

What is your current occupation?

Student

What is your level of education?

● B.Sc.

○ M.Sc.

○ PhD.

○ Other: ____________________

How many years of ML experience do you have?

2 months

Do you have any experience with the Git version control tool?

● Yes

○ No

If your answer is yes, how many years of experience do you have with Git?

4

Do you have any prior experience with any ML experiment tracking tools?

○ Yes

● No

If your answer is yes, name the ML experiment tracking tools.

Write down your email address:

[redacted]

Your Task

We have prepared your tasks based on your assigned group.

Select your group *

○ Group A

● Group B

○ Group C

Group A

The following tasks are designed for participants assigned to group A only. You will be guided through three different phases in the following order:

* Phase 1 - Neptune
* Phase 2 - DVC
* Phase 3 - No-Tool

Phase 1: Neptune

Neptune.ai is a machine learning experiment tool used to track Machine learning assets such as datasets, parameters, metrics, etc. The tool can be used to track, retrieve and query the assets different runs of an experiment by instrumenting code. The assets are mainly tracked as metadata which can be viewed on a web dashboard for post-experiment analysis. The web dashboard allows users to view the experimental runs that have been done, their results, and associated assets.

Tutorial: Getting started with Neptune

You have already received a link to a brief Neptune tutorial that we have prepared for this experiment. We expect that you have familiarized yourself with the key features of Neptune that are relevant to this experiment.

We have also provided the link below in case you would like to review it.

Follow the link for the tutorial: [redacted]

Experiment's Task

Follow the instructions provided in the experiment task document for the Group-A Neptune phase.

Link to the instructions: [redacted]

Link to python script: [redacted]

Questions:

Retrieving & Querying Tracked Data:

You will be asked to answer a number of retrieving and querying related questions.You are expected to use the Neptune tool alone to assist you in answering the questions (Do not consult the task instruction document). Try your best to provide the correct answer.

Which of the 13 runs performed best? (i.e. which has the lowest RMSE score?)

What is the RMSE value for that run?

Which of the algorithms (LR & RFR) performed best (lowest RMSE) in their first run?

○ LR

○ RFR

○ I don't know

What data features were used for the experimental run with the highest r2_score?

Compare Run 4 and Run 1, which one had the highest mean absolute error?

○ Run 4

○ Run 1

○ I don't know

What was the value?

____

Compare “EX-5” ,“EX-7”, "EX-9" and "EX-11" in the dashboard table.

	EX-5	EX-7	EX-9	EX-11
Highest RMSE	○	○	○	○
Highest R2	○	○	○	○
Highest mean absolute error	○	○	○	○

If we want to reproduce the results of previous runs, we need to retrieve the model. Which model was used for “EX-4”?

○ RandomForestRegressor(max_depth=5, min_samples_split=6, n_estimators=50)

○ LinearRegression(normalize=False)

○ RandomForestRegressor(max_depth=10, min_samples_split=12, n_estimators=45)

○ Other: ____

From the dashboard table filter out the to show the tracked experiment runs:

Query Neptune for the model with the worst RMSE (Largest value). Provide Run id, and the normalize parameter.

List all linear regression runs with RMSE value less than 6.5. Provide Run id, and the r2 value for that run.

In the python file, print the R2 value of the very first run. What was the value?

You are expected to retrieve this information via the python script. You can write the code below line 99

Experience with Neptune:

How helpful was the brief Neptune tutorial provided ahead of this experiment?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How do you rate the ease of completing the tasks?

	1	2	3	4	5	
Not Easy	○	○	○	○	○	Very Easy

How helpful is the visual dashboard provided by Neptune.ai when comparing the experimental runs?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How long did it take you to complete the Neptune.ai section?

○ 0 - 30 mins

○ 30 mins - 1 hour

○ 1 hour - 1 hour 30 mins

○ 1 hour 30 mins - 2 hours

○ 2+ hours

Any Additional Comments:

....

Phase 1 Completed

Please proceed to the next phase

Phase 2: DVC

Data Version Control is an experiment management tool that extends Git to make ML models shareable and reproducible. It is designed to handle large files, datasets, machine learning models, metrics as well as code.

Tutorial: Getting started with DVC

You have already received a link to a brief DVC tutorial that we have prepared for this experiment. We expect that you have familiarized yourself with the key features of DVC that are relevant to this experiment.

We have also provided the link below in case you would like to review it.

Follow the link for the tutorial:
[redacted]

Experiment's Task

Follow the instructions provided in the experiment task document for the Group-A DVC phase.

Link to the instructions: [redacted]

Questions (DVC):

Retrieving & Querying data:

You will asked to perform a number of retrieving tasks, and you will use the tool to assist you in performing these tasks. Try your best to provide the correct answer:

Which run performed best, (i.e. has the lowest RMSE score)?

What is the RMSE value for that run?

Which of the algorithms performed best in their first run? LR or RFR i.e. which one had the lowest RMSE.

○ LR

○ RFR

○ I don't know

What data features were used for the run with the lowest r2_score?

....

Between Run 4 and Run 1, which one had the highest mean absolute error?

○ Run 4

○ Run 1

○ I don't know

What was the value?

....

Compare “Run 5”, “Run 7”, "Run 9" and "Run 13":

	Run 5	Run 7	Run 9	Run 13
Highest RMSE	○	○	○	○
Highest R2	○	○	○	○
Highest mean absolute error	○	○	○	○

If we want to reproduce the results of previous runs, we need to retrieve the model. Which model was used for “Run 4”?

○ RandomForestRegressor(max_depth=5, min_samples_split=6, n_estimators=50)

○ LinearRegression(normalize=False)

○ LinearRegression(normalize=True)

○ Other: ______________________________

Using DVC commands, solve the following:

Return the run with the worst RMSE (Largest value)

Provide tag name, and the parameters used.

____

Find the runs that produced models evaluation metric, r2 is greater than 0.32

Provide tag number.

____

Where you used linear regression algorithm and the RMSE was greater than 1.15

Provide tag name, and the normalize parameter used.

What was the R2 value of the very first run?

Hint: Use dvc metrics show, as shown in tutorial

Experience with DVC:

How helpful was the short tutorial in the beginning in completing the tasks?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How do you rate the ease of completing the tasks?

	1	2	3	4	5	
Not Easy	○	○	○	○	○	Very Easy

How helpful were the DVC commands to compare the runs?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How long did it take you to complete the DVC section?

○ 0 - 30 mins

○ 30 mins - 1 hour

○ 1 hour - 1 hour 30 mins

○ 1 hour 30 mins - 2 hours

○ 2+ hours

Any Additional Comments:

Phase 3: No Tool

In this part of the experiment, you will run the experiment with python only. And will you will not have any experiment tools to assist you.

Tutorial: Getting started with No-Tool

You have already received a link to a brief No-Tool tutorial that we have prepared for this experiment. We expect that you have familiarized yourself with running the project with No-Tool.

We have also provided the link below in case you would like to review it.

Follow the link for the tutorial:
[redacted]

Experiment's Task

Follow the instructions provided in the experiment task document for the Group-A Neptune phase.

Link to the instructions: [redacted]

Link to python script: [redacted]

Questions (No-Tool):

Retrieving & Querying data:

You will asked to perform a number of retrieving tasks, and you will use the tool to assist you in performing these tasks. Try your best to provide the correct answer:

Which run performed best, (i.e. has the lowest RMSE score)?

What is the RMSE value for that run?

Which of the algorithms performed best in their first run? LR or RFR i.e. which one had the lowest RMSE.

○ LR

○ RFR

○ I don't know

What data features were used for the run with the lowest r2_score?

Between Run 4 and Run 1, which one had the highest mean absolute error?

○ Run 4

○ Run 1

○ I don't know

And what was the value?

Compare “Run 5”, “Run 7”, "Run 9" and "Run 13":

	Run 5	Run 7	Run 9	Run 13
Highest RMSE	○	○	○	○
Highest R2	○	○	○	○
Highest mean absolute error	○	○	○	○

If we want to reproduce the results of previous runs, we need to retrieve the model. Which model was used for “Run 4”?

○ RandomForestRegressor(max_depth=5, min_samples_split=6, n_estimators=50)

○ LinearRegression(normalize=False)

○ RandomForestRegressor(max_depth=10, min_samples_split=12, n_estimators=45)

○ Other:

For the following write down the run number along with the value:

Run with the worst RMSE (Largest value)

Provide the run nr and RMSE

Where you used linear regression algorithm and the RMSE was less than 55

Provide Run number, and the r2 value for that run.

What was the R2 value of the very first run?

Experience with 'No-Tool':

How helpful was the short tutorial in the beginning in completing the tasks?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How do you rate the ease of completing the tasks?

	1	2	3	4	5	
Not Easy	○	○	○	○	○	Very Easy

Describe how you manually track, query, and retrieve the experiment runs and their assets.

How helpful was the use of your manual technique in tracking/retrieving the experimental runs and their assets?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How long did it take you to complete the No-Tool section?

○ 0-30 mins

○ 30 mins-1 hour

○ 1 hour-1 hour 30 mins

○ 1 hour 30 mins-2 hours

○ 2+ hours

Any Additional Comments:

Experience with all Tools (Neptune, DVC, and 'No-Tool'

Even though the experiment has guided you through few runs of machine learning experiments, a typical machine learning task usually has many more runs carried out over a longer period of time. Consequently, it becomes harder for users to manage the multiple experiment runs and their assets without any supporting tools.

To answer the following questions, we want you to keep a perspective of a machine learning expert who is probably working on multiple ML tasks with way too many experimental runs.

Describe your experience with each of the tools(Neptune, DVC and No-Tool).

The DVC and Neptune tools provide a significant support for tracking, querying and retrieving generated data from ML experiments over No-tool.

○ Strongly Agree

○ Agree

○ Neutral

○ Disagree

○ Strongly Disagree

Please elaborate on your response above.

....

Which tool do you consider best for tracking data during ML experiments ?

○ Neptune.ai

○ DVC

○ No-tool

Which tool do you consider best for querying and retrieving previously tracked data?

○ Neptune.ai

○ DVC

○ No-tool

Which of Neptune and DVC do you consider least intrusive in completing the tasks?

○ Neptune.ai

○ DVC

Which of Neptune and DVC was the easiest to learn?

○ Neptune.ai

○ DVC

Which of the tools provides the best support for comparing different experiment runs?

○ Neptune.ai

○ DVC

Neptune helps compare different runs using a web dashboard, while DVC uses CLI. Which do you most convenient?

○ Neptune (Web dashboards)

○ DVC (CLI)

Please elaborate.

....

Finally, which tool would you recommend a ML practitioner to use?

○ Neptune.ai

○ DVC

○ No-Tool

Please elaborate.

Group B:

The following tasks are designed for participants assigned to group B only. You will be guided through three different phases in the following order:

* Phase 1 - DVC
* Phase 2 - No-Tool
* Phase 3 - Neptune.ai

Phase 1: DVC

Data Version Control is an experiment management tool that extends Git to make ML models shareable and reproducible. It is designed to handle large files, datasets, machine learning models, metrics as well as code.

Tutorial: Getting started with DVC

You have already received a link to a brief DVC tutorial that we have prepared for this experiment. We expect that you have familiarized yourself with the key features of DVC that are relevant to this experiment.

We have also provided the link below in case you would like to review it.

Follow the link for the tutorial:
[redacted]

Experiment's Task

Follow the instructions provided in the experiment task document for the Group-A DVC phase.

Link to the instructions: [redacted]

Questions (DVC):

Retrieving & Querying data:

You will asked to perform a number of retrieving tasks, and you will use the tool to assist you in performing these tasks. Try your best to provide the correct answer:

Which run performed best, (i.e. has the lowest RMSE score)?

Runs: 6 and 7

What is the RMSE value for that run?

58.84

Which of the algorithms performed best in their first run? LR or RFR i.e. which one had the lowest RMSE.

○ LR

● RFR

○ I don't know

What data features were used for the run with the highest r2_score?

's1','sex'

Between Run 4 and Run 1, which one had the highest mean absolute error?

● Run 4

○ Run 1

○ I don't know

What was the value?

Compare “Run 5”, “Run 7”, "Run 9" and "Run 13":

	Run 5	Run 7	Run 9	Run 13
Highest RMSE	○	●	○	○
Highest R2	●	○	○	○
Highest mean absolute error	○	●	○	○

If we want to reproduce the results of previous runs, we need to retrieve the model. Which model was used for “Run 4”?

○ RandomForestRegressor(max_depth=5, min_samples_split=6, n_estimators=50)

● LinearRegression(normalize=False)

○ RandomForestRegressor(max_depth=10, min_samples_split=12, n_estimators=45)

○ Other:

Using DVC commands, solve the following:

Return the run with the worst RMSE (Largest value)

Provide tag name, and the parameters used.

run 3

Where you used linear regression algorithm and the RMSE was less than 60

Provide tag name, and the parameters used.

run 6, True, run 7, True

What was the R2 value of the very first run?

Hint: Use dvc metrics show, as shown in tutorial

0.07234

Experience with DVC:

How helpful was the short tutorial in the beginning in completing the tasks?

	1	2	3	4	5	
Not Helpful	○	○	○	○	●	Very Helpful

How do you rate the ease of completing the tasks?

	1	2	3	4	5	
Not Easy	○	○	●	○	○	Very Easy

How helpful were the DVC commands to compare the runs?

	1	2	3	4	5	
Not Helpful	○	○	○	○	●	Very Helpful

How long did it take you to complete the DVC section?

- ○ 0 - 30 mins
- ○ 30 mins - 1 hour
- ○ 1 hour - 1 hour 30 mins
- ○ 1 hour 30 mins - 2 hours
- ● 2+ hours

Any Additional Comments:

I prefer using CLI to track the assets, easy and short commands

Phase 2: No Tool

In this part of the experiment, you will run the experiment with python only. And will you will not have any experiment tools to assist you.

Tutorial: Getting started with No-Tool

You have already received a link to a brief No-Tool tutorial that we have prepared for this experiment. We expect that you have familiarized yourself with running the project with No-Tool.

We have also provided the link below in case you would like to review it.

Follow the link for the tutorial:
[redacted]

Experiment's Task

Follow the instructions provided in the experiment task document for the Group-A Neptune phase.

Link to the instructions: [redacted]

Link to python script: [redacted]

Questions (No-Tool):

Retrieving & Querying data:

You will asked to perform a number of retrieving tasks, and you will use the tool to assist you in performing these tasks. Try your best to provide the correct answer:

Which run performed best, (i.e. has the lowest RMSE score)?

run 5

What is the RMSE value for that run?

Not sure

Which of the algorithms performed best in their first run? LR or RFR i.e. which one had the lowest RMSE.

○ LR

○ RFR

● I don't know

What data features were used for the run with the lowest r2_score?

not sure

Between Run 4 and Run 1, which one had the highest mean absolute error?

○ Run 4

○ Run 1

● I don't know

And what was the value?

not sure

Compare “Run 5”, “Run 7”, "Run 9" and "Run 13":

	Run 5	Run 7	Run 9	Run 13
Highest RMSE	●	○	○	○
Highest R2	○	○	●	○
Highest mean absolute error	●	○	○	○

If we want to reproduce the results of previous runs, we need to retrieve the model. Which model was used for “Run 4”?

○ RandomForestRegressor(max_depth=5, min_samples_split=6, n_estimators=50)

● LinearRegression(normalize=False)

○ RandomForestRegressor(max_depth=10, min_samples_split=12, n_estimators=45)

○ Other:

For the following write down the run number along with the value:

Run with the worst RMSE (Largest value)

Provide the run nr and RMSE

IDK

Where you used linear regression algorithm and the RMSE was less than 6.5

Provide Run number, and the r2 value for that run.

IDK

What was the R2 value of the very first run?

IDK

Experience with No-Tool:

How helpful was the short tutorial in the beginning in completing the tasks?

	1	2	3	4	5	
Not Helpful	○	●	○	○	○	Very Helpful

How do you rate the ease of completing the tasks?

	1	2	3	4	5	
Not Easy	●	○	○	○	○	Very Easy

Describe how you manually track, query, and retrieve the experiment runs and their assets.

I did not track the data in any way, which made it difficult to complete the tasks

How helpful was the use of your manual technique in tracking/retrieving the experimental runs and their assets?

	1	2	3	4	5	
Not Helpful	●	○	○	○	○	Very Helpful

How long did it take you to complete the No-Tool section?

○ 0-30 mins

● 30 mins - 1 hour

○ 1 hour - 1 hour 30 mins

○ 1 hour 30 mins - 2 hours

○ 2+ hours

Any Additional Comments:

Phase 3: Neptune

Neptune.ai is a machine learning experiment tool used to track Machine learning assets such as datasets, parameters etc. The tool can be used to track, retrieve and query the assets from each run. The assets are tracked as metadata which can be viewed in the GUI. It provides the user with a GUI where you're able to view the runs that have been done and view the results of each run.

Tutorial: Getting started with Neptune

You have already received a link to a brief Neptune tutorial that we have prepared for this experiment. We expect that you have familiarized yourself with the key features of Neptune that are relevant to this experiment.

We have also provided the link below in case you would like to review it.

Follow the link for the tutorial: [redacted]

Experiment's Task

Follow the instructions provided in the experiment task document for the Group-C Neptune phase.

Link to the instructions: [redacted]

Follow the link for the python: [redacted]

Questions (Neptune):

Retrieving & Querying Tracked Data:

You will be asked to answer a number of retrieving and querying related questions.You are expected to use the Neptune tool alone to assist you in answering the questions (Do not consult the task instruction document). Try your best to provide the correct answer.

Which of the 13 runs performed best? (i.e. which has the lowest RMSE score?)

Run 7

What is the RMSE value for that run?

0.512

Which of the algorithms (LR & RFR) performed best (lowest RMSE) in their first run?

○ LR

● RFR

○ I don't know

What data features were used for the experimental run with the highest r2_score?

Empty

Compare Run 4 and Run 1, which one had the highest mean absolute error?

● Run 4

○ Run 1

○ I don't know

What was the value?

0.602

Compare “EX-5” ,“EX-7”, "EX-9" and "EX-13" in the dashboard table.

	EX-5	EX-7	EX-9	EX-11
Highest RMSE	●	○	○	○
Highest R2	○	●	○	○
Highest mean absolute error	●	○	○	○

If we want to reproduce the results of previous runs, we need to retrieve the model. Which model was used for “EX-4”?

○ RandomForestRegressor(max_depth=5, min_samples_split=6, n_estimators=50)

● LinearRegression(normalize=False)

○ RandomForestRegressor(max_depth=10, min_samples_split=12, n_estimators=45)

○ Other: ________________

From the dashboard table filter out the to show the tracked experiment runs:

Query Neptune for the model with the worst RMSE (Largest value). Provide Run id, and the parameters used.

EX-5, normalize=False

List all linear regression runs with RMSE value greater than 0.9. Provide Run id, and the r2 value for that run.

EX-5: 0.3, EX-11: 0.3

In the python file, print the R2 value of the very first run. What was the value?

You are expected to retrieve this information via the python script.

0.50123

Experience with Neptune:

How helpful was the brief Neptune tutorial provided ahead of this experiment?

	1	2	3	4	5	
Not Helpful	○	○	○	○	●	Very Helpful

How do you rate the ease of completing the tasks?

	1	2	3	4	5	
Not Easy	○	○	○	○	●	Very Easy

How helpful is the visual dashboard provided by Neptune.ai when comparing the experimental runs?

	1	2	3	4	5	
Not Helpful	○	○	○	○	●	Very Helpful

How long did it take you to complete the Neptune.ai section?

○ 0 - 30 mins

○ 30 mins - 1 hour

○ 1 hour - 1 hour 30 mins

● 1 hour 30 mins - 2 hours

○ 2+ hours

Any Additional Comments:

Experience with all Tools (DVC, No-Tool, and Neptune)

Even though the experiment has guided you through few runs of machine learning experiments, a typical machine learning task usually have many more runs carried out over a longer period of time. Consequently, it becomes harder for users to manage the multiple experiment runs and their assets without any supporting tools.

To answer the following questions, we want you to keep a perspective of a machine learning expert who is probably working on multiple ML tasks with way too many experimental runs.

Describe your experience with each of the tools(Neptune, DVC, and No-Tool).

Dvc was my favorite because it's run through the CLI, Neptune dashboard was good but the python file was too messy, no-tool was difficult to track

The DVC and Neptune tools provide a significant support for tracking, querying and retrieving generated data from ML experiments over No-tool.

● Strongly Agree

○ Agree

○ Neutral

○ Disagree

○ Strongly Disagree

Please elaborate on your response above.

When using the no-tool, I found it hard to complete the tasks, whereas neptune and dvc helped me look for the answers

Which tool do you consider best for tracking data during ML experiments ?

○ Neptune.ai

● DVC

○ No-tool

Which tool do you consider best for querying and retrieving previously tracked data?

● Neptune.ai

○ DVC

○ No-tool

Which of Neptune and DVC do you consider least intrusive in completing the tasks?

○ Neptune.ai

● DVC

Which of Neptune and DVC was the easiest to learn?

○ Neptune.ai

● DVC

Which of the tools provides the best support for comparing different experiment runs?

● Neptune.ai

○ DVC

Neptune helps compare different runs using a web dashboard, while DVC uses CLI. Which do you most convenient?

○ Neptune (Web dashboards)

● DVC (CLI)

Please elaborate.

I use git a lot, so using dvc was not that different

Finally, which tool would you recommend a ML practitioner to use?

○ Neptune.ai

● DVC

○ No-Tool

Please elaborate.

Because it would take long to learn it, especially if you have a git background

Group C:

The following tasks are designed for participants assigned to group C only. You will be guided through three different phases in the following order:

* Phase 1 - No-Tool
* Phase 2 - Neptune.ai
* Phase 3 - DVC

Phase 1: No Tool

In this part of the experiment, you will run the experiment with python only. And will you will not have any experiment tools to assist you.

Tutorial: Getting started with No-Tool

You have already received a link to a brief No-Tool tutorial that we have prepared for this experiment. We expect that you have familiarized yourself with running the project with No-Tool.

We have also provided the link below in case you would like to review it.

Follow the link for the tutorial:
[redacted]

Experiment's Task

Follow the instructions provided in the experiment task document for the Group-A Neptune phase.

Link to the instructions: [redacted]

Link to python script: [redacted]

Questions (No-Tool):

Retrieving & Querying data:

You will asked to perform a number of retrieving tasks, and you will use the tool to assist you in performing these tasks. Try your best to provide the correct answer:

Which run performed best, (i.e. has the lowest RMSE score)?

What is the RMSE value for that run?

Which of the algorithms performed best in their first run? LR or RFR i.e. which one had the lowest RMSE.

○ LR

○ RFR

○ I don't know

What data features were used for the run with the lowest r2_score?

Between Run 4 and Run 1, which one had the highest mean absolute error?

○ Run 4

○ Run 1

○ I don't know

And what was the value?

....

Compare “Run 5”, “Run 7”, "Run 9" and "Run 13":

	Run 5	Run 7	Run 9	Run 13
Highest RMSE	○	○	○	○
Highest R2	○	○	○	○
Highest mean absolute error	○	○	○	○

If we want to reproduce the results of previous runs, we need to retrieve the model. Which model was used for “Run 4”?

○ RandomForestRegressor(max_depth=5, min_samples_split=6, n_estimators=50)

○ LinearRegression()

○ RandomForestRegressor(max_depth=10, min_samples_split=12, n_estimators=45)

○ Other: ....

For the following write down the run number along with the value:

Run with the worst RMSE (Largest value)

Provide the run nr and RMSE

Where you used linear regression algorithm and the RMSE was less than 0.7

Provide Run number, and the r2 value for that run.

What was the R2 value of the very first run?

Experience with No-Tool:

How helpful was the short tutorial in the beginning in completing the tasks?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How do you rate the ease of completing the tasks?

	1	2	3	4	5	
Not Easy	○	○	○	○	○	Very Easy

Describe how you manually track, query, and retrieve the experiment runs and their assets.

How helpful was your use of a manual technique in tracking/retrieving the experimental runs and their assets?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How long did it take you to complete the No-Tool section?

○ 0-30 mins

○ 30 mins-1 hour

○ 1 hour-1 hour 30 mins

○ 1 hour 30mins-2 hours

○ 2+ hour

Any Additional Comments:

Phase 2: DVC

Data Version Control is an experiment management tool that extends Git to make ML models shareable and reproducible. It is designed to handle large files, datasets, machine learning models, metrics as well as code.

Tutorial: Getting started with DVC

You have already received a link to a brief DVC tutorial that we have prepared for this experiment. We expect that you have familiarized yourself with the key features of DVC that are relevant to this experiment.

We have also provided the link below in case you would like to review it.

Follow the link for the tutorial:
[redacted]

Experiment's Task

Follow the instructions provided in the experiment task document for the Group-C DVC phase.

Link to the instructions: [redacted]

Questions (DVC):

Retrieving & Querying data:

You will asked to perform a number of retrieving tasks, and you will use the tool to assist you in performing these tasks. Try your best to provide the correct answer:

Which run performed best, (i.e. has the lowest RMSE score)?

....

What is the RMSE value for that run?

....

Which of the algorithms performed best in their first run? LR or RFR i.e. which one had the lowest RMSE.

○ LR

○ RFR

○ I don't know

What data features were used for the run with the highest r2_score?

Between Run 4 and Run 1, which one had the highest mean absolute error?

○ Run 4

○ Run 1

○ I don't know

What was the value?

Compare “Run 5”, “Run 7”, "Run 9" and "Run 13":

	Run 5	Run 7	Run 9	Run 13
Highest RMSE	○	○	○	○
Highest R2	○	○	○	○
Highest mean absolute error	○	○	○	○

If we want to reproduce the results of previous runs, we need to retrieve the model. Which model was used for “Run 4”?

○ RandomForestRegressor(max_depth=5, min_samples_split=6, n_estimators=50)

○ LinearRegression()

○ RandomForestRegressor(max_depth=10, min_samples_split=12, n_estimators=45)

○ Other: ____________________

Using DVC commands, solve the following:

Return the run with the worst RMSE (Largest value)

Provide tag name, and the parameters used.

____

Where you used linear regression algorithm and the RMSE was less than 5

Provide tag name, and the parameters used.

____

What was the R2 value of the very first run?

Hint: Use dvc metrics show, as shown in tutorial

Experience with DVC:

How helpful was the short tutorial in the beginning in completing the tasks?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How do you rate the ease of completing the tasks?

	1	2	3	4	5	
Not Easy	○	○	○	○	○	Very Easy

How helpful were the DVC commands to compare the runs?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How long did it take you to complete the DVC section?

○ 0 - 30 mins

○ 30 mins - 1 hour

○ 1 hour - 1 hour 30 mins

○ 1 hour 30 mins - 2 hours

○ 2+ hours

Any Additional Comments:

Phase 3: Neptune

Neptune.ai is a machine learning experiment tool used to track Machine learning assets such as datasets, parameters etc. The tool can be used to track, retrieve and query the assets from each run. The assets are tracked as metadata which can be viewed in the GUI. It provides the user with a GUI where you're able to view the runs that have been done and view the results of each run.

Tutorial: Getting started with Neptune

You have already received a link to a brief Neptune tutorial that we have prepared for this experiment. We expect that you have familiarized yourself with the key features of Neptune that are relevant to this experiment.

We have also provided the link below in case you would like to review it.

Follow the link for the tutorial: [redacted]

Experiment's Task

Follow the instructions provided in the experiment task document for the Group-C Neptune phase.

Link to the instructions:[redacted]

Follow the link for the python: [redacted]

Questions (Neptune):

Retrieving & Querying Tracked Data:

You will be asked to answer a number of retrieving and querying related questions.You are expected to use the Neptune tool alone to assist you in answering the questions (Do not consult the task instruction document). Try your best to provide the correct answer.

Which of the 13 runs performed best? (i.e. which has the lowest RMSE score?)

What is the RMSE value for that run?

Which of the algorithms (LR & RFR) performed best (lowest RMSE) in their first run?

○ LR

○ RFR

○ I don't know

What data features were used for the experimental run with the highest r2_score?

Compare Run 4 and Run 1, which one had the highest mean absolute error?

○ Run 4

○ Run 1

○ I don't know

What was the value?

Compare “EX-5” ,“EX-7”, "EX-9" and "EX-13" in the dashboard table.

	EX-5	EX-7	EX-9	EX-11
Highest RMSE	○	○	○	○
Highest R2	○	○	○	○
Highest mean absolute error	○	○	○	○

If we want to reproduce the results of previous runs, we need to retrieve the model. Which model was used for “EX-4”?

○ RandomForestRegressor(max_depth=5, min_samples_split=6, n_estimators=50)

○ LinearRegression()

○ RandomForestRegressor(max_depth=10, min_samples_split=12, n_estimators=45)

○ Other:

From the dashboard table filter out the to show the tracked experiment runs:

Query Neptune for the model with the worst RMSE (Largest value). Provide Run id, and the parameters used.

List all linear regression runs with RMSE value less than 54. Provide Run id, and the r2 value for that run.

In the python file, print the R2 value of the very first run. What was the value?

You are expected to retrieve this information via the python script.

Experience with Neptune:

How helpful was the brief Neptune tutorial provided ahead of this experiment?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How do you rate the ease of completing the tasks?

	1	2	3	4	5	
Not Easy	○	○	○	○	○	Very Easy

How helpful is the visual dashboard provided by Neptune.ai when comparing the experimental runs?

	1	2	3	4	5	
Not Helpful	○	○	○	○	○	Very Helpful

How long did it take you to complete the neptune.ai section?

○ 0 - 30 mins

○ 30 mins - 1 hour

○ 1 hour - 1 hour 30 mins

○ 1 hour 30 mins - 2 hours

○ 2+ hours

Any Additional Comments:

....

Experience with all Tools (No-Tool, DVC, Neptune)

Even though the experiment has guided you through few runs of machine learning experiments, a typical machine learning task usually have many more runs carried out over a longer period of time. Consequently, it becomes harder for users to manage the multiple experiment runs and their assets without any supporting tools.

To answer the following questions, we want you to keep a perspective of a machine learning expert who is probably working on multiple ML tasks with way too many experimental runs.

Describe your experience with each of the tools(Neptune, DVC and No-Tool).

....

The DVC and Neptune tools provide a significant support for tracking, querying and retrieving generated data from ML experiments over No-tool.

○ Strongly Agree

○ Agree

○ Neutral

○ Disagree

○ Strongly Disagree

Please elaborate on your response above.

Which tool do you consider best for tracking data during ML experiments ?

○ Neptune.ai

○ DVC

○ No-tool

Which tool do you consider best for querying and retrieving previously tracked data?

○ Neptune.ai

○ DVC

○ No-tool

Which of Neptune and DVC do you consider least intrusive in completing the tasks?

○ Neptune.ai

○ DVC

Which of Neptune and DVC was the easiest to learn?

○ Neptune.ai

○ DVC

Which of the tools provides the best support for comparing different experiment runs?

○ Neptune.ai

○ DVC

Neptune helps compare different runs using a web dashboard, while DVC uses CLI. Which do you most convenient?

○ Neptune (Web dashboards)

○ DVC (CLI)

Please elaborate.

Finally, which tool would you recommend a ML practitioner to use?

○ Neptune.ai

○ DVC

○ No-Tool

Please elaborate.

Conclusion

Thank you for participating.

This content is neither created nor endorsed by Google.

Google Forms