QBN1 — QBN1 TASK 2: DATA PRODUCTION PIPELINE

DEPLOYMENT — D602
PRFA — QBN1

COMPETENCIES

4162.1.2 : Implements a Data Product Pipeline
The learner implements a data product pipeline to address organizational needs.

INTRODUCTION

As a data analyst, you may be responsible for designing and executing a data pipeline that effectively addresses specific organizational requirements.

In this task, you will write code in Python or R that downloads and filters a selected airport's data. You will then run an MLFlow experiment capturing the features listed in the comments within the poly_regressor file. Next, you will write an MLProject file that links the two scripts. Finally, you will provide an explanation of how you wrote your code.

SCENARIO

You are an analyst at a major airline whose work involves predicting delays from various airports that your airline serves. To capture both current and future needs, a previous analyst developed a polynomial regression to predict the average departure delay in minutes between a given departure airport and all the airports that are served from that departure airport. They tested the model on a single airport but then left the company abruptly, leaving you with the task of finishing their work and deploying the model to other business units within the airline.

You have been provided with the previous analyst's regression model and the Bureau of Transportation Statistics website they used to download the initial data that was used to train the model. You will need to download data for one of the airports listed at the end of this section and test the regression model on that airport. To capture the performance of the model, you will implement MLFlow experiments to record the various parameters with which you test the model and the performance under each set of conditions. You will then combine three steps of model testing into a single ML pipeline such that other analysts can extend your work in other business units.

Your final code and all updates must be stored in a GitLab repository so users in other business units can track the changes you made to the code over time.

Airport list:

- ATL: Atlanta, Georgia
- LAX: Los Angeles, California
- JFK: New York, New York
- MIA: Miami, Florida
- DFW: Dallas, Texas
- ORD: Chicago, Illinois

REQUIREMENTS

Your submission must be your original work. No more than a combined total of 30% of the submission and no more than a 10% match to any one individual source can be directly quoted or closely paraphrased from sources, even if cited correctly. The similarity report that is provided when you submit your task can be used as a guide.

You must use the rubric to direct the creation of your submission because it provides detailed criteria that will be used to evaluate your work. Each requirement below may be evaluated by more than one rubric aspect. The rubric aspect titles may contain hyperlinks to relevant portions of the course.

*Tasks may **not** be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc., unless specified in the task requirements. All other submissions must be file types that are uploaded and submitted as attachments (e.g., .docx, .pdf, .ppt).*

A. Create your subgroup and project in GitLab using the provided web link and the "GitLab How-To" web link by doing the following:

- Clone the project to the IDE.
- Commit with a message and push when you complete each requirement listed in parts B, C, D, and E.

Note: You may commit and push whenever you want to back up your changes, even if a requirement is not yet complete.

- Submit a copy of the GitLab repository URL in the "Comments to Evaluator" section when you submit this assessment.
- Submit a copy of the repository branch history retrieved from your repository, which must include the commit messages and dates.

B. Write a script in either Python or R to import the data you downloaded and format it according to the criteria required by the model script, demonstrating a progression of work on your code. You must run a DVC command to create a metafile for your dataset and submit the metafile to the GitLab repository. Submit *at least* **two** versions of your code to the GitLab repository demonstrating a progression of work on your code.

C. Write a script in either Python or R to filter data to only departures from the chosen airport, then implement *at least* **two** other data cleaning steps. Submit *at least* **two** versions of your code to the GitLab repository demonstrating a progression of work on your code.

D. Using the code template provided in the GitLab repository, implement an MLFlow experiment that captures the features listed in the comments within the poly_regressor file for either Python or R. Submit

at least **two** versions of your code to the GitLab repository demonstrating a progression of work on your code.

E. Using the provided YAML file for either Python or R, write an MLProject file that links the two scripts you wrote in parts B and C with the modified poly_regressor script from part D.

Note: Your MLProject file must be submitted to the GitLab repository.

F. Provide a written explanation of how you wrote the code and MLProject pipeline, including any challenges you encountered and how you addressed these challenges. Include a screenshot of your MLProject pipeline running successfully.

G. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

H. Demonstrate professional communication in the content and presentation of your submission.

File Restrictions

File name may contain only letters, numbers, spaces, and these symbols: ! - _ . * ' ( )
File size limit: 200 MB
File types allowed: doc, docx, rtf, xls, xlsx, ppt, pptx, odt, pdf, csv, txt, qt, mov, mpg, avi, mp3, wav, mp4, wma, flv, asf, mpeg, wmv, m4v, svg, tif, tiff, jpeg, jpg, gif, png, zip, rar, tar, 7z

RUBRIC

A:GITLAB REPOSITORY

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A GitLab repository is not provided.	The subgroup and project are created in GitLab, but 1 or more of the given actions are not completed, or they are completed incorrectly.	The subgroup and project are created in GitLab correctly, and all of the given actions are completed correctly.

B:IMPORT AND FORMAT SCRIPT

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A script in either Python or R to import and format the downloaded data is not provided.	The provided script fails to import or format the downloaded data according to the criteria required by the model script, or only 1 version of the code is provided.	*At least* 2 versions of the script are provided, with the final version successfully importing and formatting the downloaded data according to the criteria required by the model script.

C:DATA FILTERING SCRIPT

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A script in either Python or R to filter and clean the data is not provided.	The provided script does not filter the data to include only departures from the chosen airport, or the script does not implement *at least* 2 data cleaning steps, or only 1 version of the code is provided.	*At least* 2 versions of the script are provided, with the final version successfully filtering the data to include only departures from the chosen airport and implementing *at least* 2 data-cleaning steps.

D:MLFLOW EXPERIMENT

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
An MLFlow experiment is not provided.	The provided MLFlow experiment does not capture *all* the features listed in the comments within the poly_regressor file, or only 1 version of the code is provided.	*At least* 2 versions of code are provided, with the final version successfully capturing *all* the features listed in the comments of the poly_regressor file.

E:MLPROJECT LINKING FILE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
An MLProject file that links the two scripts written in parts A and B is not provided.	The provided MLProject file does not link the two scripts written in parts A and B or does not incorporate the modified poly_regressor script from part C.	The MLProject file links the two scripts written in parts A and B and incorporates the modified poly_regressor script from part C, and the file is free of errors.

F:EXPLANATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
An explanation of how the code and MLProject pipeline were written is not provided.	The explanation of how the code and MLProject pipeline were written does not address challenges that were encountered, or it does not contain a screenshot of the MLProject pipeline	The explanation of how the code and MLProject pipeline were written includes challenges encountered and a screenshot of the MLProject pipeline running successfully and is free of errors.

running successfully, or it contains errors.

G:SOURCES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not include both in-text citations and a reference list for sources that are quoted, paraphrased, or summarized.	The submission includes in-text citations for sources that are quoted, paraphrased, or summarized and a reference list; however, the citations or reference list is incomplete or inaccurate.	The submission includes in-text citations for sources that are properly quoted, paraphrased, or summarized and a reference list that accurately identifies the author, date, title, and source location as available.

H:PROFESSIONAL COMMUNICATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
Content is unstructured, is disjointed, or contains pervasive errors in mechanics, usage, or grammar. Vocabulary or tone is unprofessional or distracts from the topic.	Content is poorly organized, is difficult to follow, or contains errors in mechanics, usage, or grammar that cause confusion. Terminology is misused or ineffective.	Content reflects attention to detail, is organized, and focuses on the main ideas as prescribed in the task or chosen by the candidate. Terminology is pertinent, is used correctly, and effectively conveys the intended meaning. Mechanics, usage, and grammar promote accurate interpretation and understanding.

WEB LINKS

- WGU GitLab Environment FAQ
- Bureau of Transportation Statistics
- GitLab Environment