COMPETENCIES

**4123.1.1** : **Capstone**

The learner integrates and synthesizes competencies from across the degree program, thereby demonstrating the ability to participate in and contribute value to the chosen professional field.

INTRODUCTION

Before starting this task, ensure that Tasks 1 and 2 have a passing score. If Task 1 or 2 has not yet passed, your Task 3 submission will be returned without evaluation.  
  
In this task, you will complete your capstone by writing a report that summarizes your data analytics project, including an overview of the project, a discussion of how the execution of your project differed from your plan, a discussion of your project methodology, and an evaluation of your project results. You will also submit a recorded summary of your project.

REQUIREMENTS

Your submission must represent your original work and understanding of the course material. Most performance assessment submissions are automatically scanned through the WGU similarity checker. Students are strongly encouraged to wait for the similarity report to generate after uploading their work and then review it to ensure Academic Authenticity guidelines are met before submitting the file for evaluation. See [Understanding Similarity Reports](https://cm.wgu.edu/t5/Frequently-Asked-Questions/Understanding-Similarity-Reports/ta-p/252) for more information.    
  
**Grammarly Note:**   
Professional Communication will be automatically assessed through Grammarly for Education in most performance assessments before a student submits work for evaluation. Students are strongly encouraged to review the Grammarly for Education feedback prior to submitting work for evaluation, as the overall submission will not pass without this aspect passing. See [Use Grammarly for Education Effectively](https://cm.wgu.edu/t5/Academic-Coaching-Center/Use-Grammarly-for-Education-Effectively/ta-p/52276) for more information.    
  
**Microsoft Files Note:**  
Write your paper in Microsoft Word (.doc or .docx) unless another Microsoft product, or pdf, is specified in the task directions. Tasks may not be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc.  All supporting documentation, such as screenshots and proof of experience, should be collected in a pdf file and submitted separately from the main file. For more information, please see [Computer System and Technology Requirements.](https://cm.wgu.edu/t5/WGU-Student-Policy-Handbook/Computer-System-and-Technology-Requirements/ta-p/78)    
 *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.*

Write a report summarizing your completed data analytics capstone project by doing the following:

**Project Overview**

A**.  Summarize the following elements of your capstone project:**

**•  the research question or organizational need that your capstone addressed**

Can Lightspeed accurately validate and predict which network assets are missing from its core management systems to improve data quality ahead of a security audit?

**•  the scope of your project**

Simulated network asset data for Observability, Inventory, and IPAM systems; included data generation, preparation, analysis, scenario testing, and reporting. No production data or live system integration.

**•  an overview of your solution, including any tools and methodologies used**

Used Python (with Faker, Pandas, and Scikit-learn) to generate and analyze synthetic asset data. Applied descriptive analysis and a random forest model to assess completeness and identify risk factors. Managed the process with MLflow and documented results in a Jupyter Notebook.

**Project Execution**

**B.  Summarize the execution of your project, including how the execution of the following elements differed from the plan developed in part B of Task 2:**

**•  project plan**

The project was executed as planned, following the CRISP-DM methodology and the original milestone schedule.

**•  project planning methodology**

All major tasks—data generation, preparation, modeling, analysis, and reporting—were completed on schedule. No significant changes to the project plan or methodology were required.

**•  project timeline and milestones**

Minor adjustments were made to the scenario-testing pipeline to improve flexibility, but these did not affect overall timing or outcomes. The project was completed within the original estimated 20-hour timeline and all milestones were met by the planned 7/31 deadline.

**Methodology**

**C.  Discuss your data selection and collection process in detail, including *each* of the following elements:**

**•  how your actual data selection and collection process differed from your plan**

The data was generated using Python scripts and the Faker library to simulate realistic asset records for Observability, Inventory, and IPAM systems. This process matched the initial project plan; no changes were needed. Controlled missingness was introduced via configurable parameters to enable scenario testing.

**•  how you handled any obstacles you encountered while collecting your data**

* **MLflow Parameter Bug:** While parameterizing the pipeline to accept alternative data generation configs, we discovered an MLflow bug where custom config values were not always propagated through multi-step pipeline runs. To resolve this for the capstone, we executed pipeline steps individually when scenario-specific configs were required, ensuring results matched the intended parameters.
* **Feature Overlap and Variable Reuse:** During notebook development, we encountered variable overlap between default and alternative scenarios, which risked inaccurate reporting. This was addressed by storing scenario results in separate dictionaries, maintaining clarity between runs.
* **Model Output Verification:** At one stage, model evaluation metrics appeared unexpectedly high (perfect accuracy), revealing an issue with the dataset splitting and label assignment logic. We resolved this by auditing the data preparation steps, realigning features and labels, and verifying all splits matched the intended design.

**•  how you handled any unplanned data governance issues**

There were no unplanned data governance issues, since only synthetic data was used. All data, scripts, and reports were clearly labeled as synthetic to avoid any confusion.

1. **Discuss the advantages and limitations of the dataset you used.**

* **Advantages:** The dataset is fully reproducible, free of privacy or security risks, and can be easily tailored to different scenarios.
* **Limitations:** As synthetic data, it may not capture all nuances or unexpected complexities of real production systems.

**D.  Explain your data extraction and data preparation processes, including the tools and techniques you used, and why these processes were appropriate for your data.**

All asset records were generated and labeled programmatically using Python scripts and the Faker library, then stored as CSV files for maximum portability. Data preparation included creating binary flags for asset presence in Inventory and IPAM, assigning labels based on configurable missingness probabilities, and ensuring consistent formats for key fields (e.g., device role, region). Pandas was used for all data manipulation and cleaning steps, supporting efficient filtering, joining, and feature engineering.

These processes were appropriate because they ensured full control, reproducibility, and transparency throughout the pipeline. Automated, script-based extraction and preparation minimized errors and allowed for easy scenario testing, while the use of widely adopted open-source tools (Python, Pandas) provided robustness and scalability.

**E.  Report on your data analysis process by doing the following:**

**1.  Describe the methods you used to analyze the data.**

* **Descriptive analysis:** Calculated presence rates and completeness metrics for assets across systems.
* **Predictive modeling:** Trained a random forest classifier to predict missing asset records based on device attributes and scenario parameters.
* **Scenario analysis:** Re-ran the pipeline with alternative configuration files to measure how risk assumptions impact completeness.

1. **Discuss the advantages and limitations of the tools and techniques you used to analyze the data.**

**Advantages:**

* **Python/Pandas**: Fast, flexible, widely used for data analysis and preparation.
* **Scikit-learn**: Provides robust, interpretable machine learning models and evaluation metrics.
* **MLflow**: Orchestrates pipeline steps and ensures reproducibility.
* **Jupyter Notebook**: Enables clear documentation and visualization.

**Limitations:**

* Results are based on synthetic, not real, production data—some edge cases or system-specific issues may not be captured.
* Pipeline execution required a manual workaround for parameter passing due to an MLflow bug.

1. **Provide a step-by-step explanation of how you applied the analytical methods in part E1 to the data, including how you verified that the data satisfied the assumptions or requirements for any analytical methods you used.**

* **Generated synthetic asset data** using Python and Faker with scenario-driven missingness.
* **Loaded and prepared data** in Pandas, assigning presence flags and ensuring clean, formatted features.
* **Performed descriptive analysis** to compute presence percentages and verify they matched expected scenario settings.
* **Split data** into features and labels for modeling; checked for class balance and consistent label assignment.
* **Trained random forest model** on the prepared data; used scikit-learn’s built-in functions for splitting, training, and validation.
* **Evaluated model** using accuracy, precision, recall, F1-score, and feature importance plots.
* **Verified assumptions** by ensuring all feature columns were properly encoded and that class imbalance was acceptable for model reliability.

**Results**

**F.  Evaluate the success of your data analytics project by doing the following:**

**1.  Evaluate the output of your data analytics solution or model, including calculations or metrics for accuracy.**

The random forest model successfully predicted missing asset records with strong performance. In the baseline scenario, the model achieved an accuracy of approximately 88%, with precision and recall above 80% for the primary classes. Feature importance analysis highlighted device role and region as key drivers of missingness, aligning with scenario expectations.

**2.  Evaluate the practical significance of your data analytics solution, including specific examples.**

The analytics solution provided clear, actionable insights for the organization. For example, when scenario parameters were changed to increase missingness risk in distribution devices, the model correctly identified these assets as high risk and quantified the resulting drop in completeness below the critical 75% threshold. This enabled targeted recommendations for audit and remediation, improving operational readiness for security compliance.

1. **Evaluate the overall success and effectiveness of the project.**

The project met all objectives, including validating asset record completeness, identifying high-risk factors, and supporting scenario-based decision-making. All deliverables—reports, visuals, model artifacts, and reproducible code—were produced as planned. The solution is robust, transparent, and ready to inform future data quality initiatives within the organization.

**G.  Summarize the key takeaways of your analysis by doing the following:**

**1.  Summarize the conclusions drawn from your analysis.**

The analysis confirmed that, under baseline conditions, asset presence rates exceed the 75% threshold required for security audit readiness. Scenario analysis showed that changes in risk assumptions can quickly lower completeness rates, highlighting the importance of ongoing monitoring. Predictive modeling identified which asset characteristics (e.g., device role, region) most contribute to missing records.

**2.  Explain why your chosen tools and graphical representations for visually communicating findings support effective storytelling.**

Using Python, Pandas, and Jupyter Notebooks enabled clear documentation and dynamic, interactive reporting. Visuals such as bar charts, summary tables, and feature importance plots made it easy to communicate complex findings to both technical and non-technical stakeholders, supporting effective storytelling and actionable insights.

**3.  Recommend two courses of action based on your findings.**

* 1. **Prioritize targeted data remediation:** Focus on high-risk asset categories and regions identified by the model to quickly improve data completeness before the audit.
  2. **Implement continuous data quality monitoring:** Regularly run this analytics pipeline as asset and system conditions change, ensuring ongoing compliance and audit readiness.

**H.  Provide a link to a Panopto recording in which you present a summary of your capstone project and findings from the analysis for an audience of data analytics peers. Your summary should include the following elements:**

•  a summary of your research question or organizational need

•  a demonstration of the functionality of any code you used for your data analytics solution

•  an outline of the findings and implications of your analysis

*Note: Please submit a single recording that covers all elements. The suggested length of your video should be approximately 15–20 minutes.*

*Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the "WGU" option. If prompted, log in using your WGU student portal credentials, which should forward you to Panopto’s website.*

*Note: To submit your recording, upload it to the Panopto drop box titled "BSDA Capstone – BHN1 | D502." Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the "Links" option. Upload the remaining task requirements using the "Attachments" option.*

**Appendices**

**I.  Provide evidence of project completion by submitting *any* of the following elements that are relevant to your project:**

**•  the code you used to support your project**

All code is available to review at this [Project Github](https://github.com/Baelfur/lightspeed)

**•  a copy of or link to *all* data you used for the data analytics solution**

All data is available to review at this [Github Link](https://github.com/Baelfur/lightspeed/tree/main/data)

**•  web sources that were used to acquire the data or segments of third-party code to support the application**

* Facebook/Meta Robotron paper: <https://research.facebook.com/publications/robotron-top-down-network-management-at-facebook-scale/>
* CRISP-DM methodology: <https://www.scirp.org/reference/referencespapers?referenceid=1592779>
* Udacity ML DevOps Nanodegree: <https://www.udacity.com/course/machine-learning-dev-ops-engineer-nanodegree--nd0821>

**•  other relevant project deliverables**

* Jupyter Notebook (analysis.ipynb) documenting the entire workflow, with live code, outputs, and narrative.
* HTML/Markdown reports generated from the notebook for submission or review.

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

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