**Title: Cloud-Based Machine Learning Project on AWS**

**Objective:** Throughout this course you have completed hands-on labs, where you used the features of different AWS services in guided activities. In this capstone project, you are challenged to build a solution that uses many AWS services that are familiar to you without being given step-by-step guidance, as you solve a problem or tackle an opportunity using the machine learning lifecycle.

**Prerequisites:**

* Knowledge of AWS cloud computing foundational concepts
* Knowledge of AWS data engineering concepts
* Knowledge of machine learning concepts
* Familiarity with programming languages (e.g., Python or Java).
* AWS account with appropriate permissions (AWS Academy Learner Lab)
* Ability to apply AWS Well Architected Framework Guidelines

**Task Description:**

**PHASE 1 Planning (15%)**

1. **Project Scope and Business Goal:** Clearly define the scope and objectives of the project.
   * **Project Scope:** What specific problem, task or opportunity will the machine learning component address? Ensure that it aligns with the educational program goals of AWS Academy Cloud Foundations and Data Engineering, and our course.
   * **Domain:** Clearly define the domain or industry you are working in. Understand its key characteristics, challenges, and opportunities. This includes understanding the specific problem or task you are addressing within that domain. Include an identification of the stakeholders in the domain who will benefit from your project. This could be healthcare providers, financial institutions, or any relevant parties.
   * **Literature Review:** Conduct a literature review of research papers, case studies, articles, and books related to the chosen domain. The purpose is to stay informed about the latest advancements and best practices in the field. Sources should be current and at least 5 sources should be documented and summarized.
   * **Data Source(s):** Identify potential data sources within the domain (see Canvas for sources). This could include databases, APIs, public datasets, or data collected specifically for the project. Understand and document the availability, nature and quality of the data available.  
       
     Your dataset will need to be approved and must meet the following requirements:
     1. Publicly available, not “overused” (see suggestions for Datasets)
     2. Sufficient size
     3. Relevant to project scope
     4. Requires some merging, cleansing, preparation for machine learning
   * **Domain-specific Challenges:** Highlight any unique challenges or considerations in the domain, including regulations, ethical considerations and bias. For example, in healthcare, privacy and data security are of utmost importance. In Finance, handling imbalanced datasets or detecting fraud might be key challenges.
   * **KPI’s:** Determine what KPIs (Key Performance Indicators) Seand measures of success are important in the chosen domain. Is it accuracy, precision, recall, or some other metric? This will set the stage for the metrics that will be used to optimize models.

Phase 1 Deliverable 1:

Domain and Dataset Submitted to Professor for Approval

Phase 1 Deliverable 2:

Project Repository on GitHub (Read Me with Table of Contents)  
Deliverable 1 Document accessible in Github

Groups will move forward once they receive approval on Project Scope and data

**PHASE 2 Machine Learning Problem Framing, Data Pre-Processing (15%)**

**Set Up AWS Environment** - Account Setup: Team Access to AWS Learner Lab and Github, create IAM Roles

**S3 Data Storage:**

Create an S3 bucket to store the dataset(s)  
Upload data to the S3 Bucket

**Data Exploration for Insight and Pre-processing**

Use Amazon Athena to query the transformed data.  
 Use SQL queries for meaningful insights from the dataset for data exploration  
Create visualizations using Amazon QuickSight.

**AWS Glue ETL Job:**

Create an ETL job using AWS Glue to transform the dataset(s) in S3  
 Perform basic data transformations (e.g., filtering, aggregation, type conversions)

**AWS Pipeline/Solution Chart**

Phase 2 Deliverable:

Project Repository on GitHub (Updated Table of Contents)  
Deliverable 2 Document accessible in Github (include screenshots)

**PHASE 3 Modeling, Evaluation, Tuning (20%)**

**Choose Machine Learning Frameworks:**

Decide on the machine learning apps, frameworks and libraries that your group will use, such as Sagemaker, AutoML, TensorFlow, PyTorch, scikit-learn, etc..

**Develop and Train Models:**

Perform additional data preprocessing, model development, and training. Explore different algorithms and techniques to solve the problem.

**Evaluation and Validation:**

Evaluate and validate their machine learning models. Use appropriate metrics such as ROC curves, MSE, precision, recall, F1-score, and accuracy to measure model performance.

Hyperparameter Tuning:

If appropriate, fine-tune hyperparameters to optimize model performance. This can involve using AWS SageMaker's hyperparameter tuning capability.

Phase 3 Deliverable:

Project Repository on GitHub (Updated Table of Contents)  
Deliverable 3 Document accessible in Github (include screenshots)

**PHASE 4 Optimization, Scalability and Deployment (10%)**

**AWS Cost Analysis and Optimization:**

* + Analyze the cost of running your AWS resources.
  + Suggest cost optimization strategies (e.g., using AWS Cost Explorer and AWS Trusted Advisor).

**Scaling the Pipeline (optional):**

Investigate options for automating and scaling the pipeline, such as AWS Lambda   
for triggering ETL jobs.

**Deployment**

Phase 4 Deliverable:

Project Repository on GitHub (Updated Table of Contents)  
Deliverable 4 Document accessible in Github (include screenshots)

**FINAL PROJECT DELIVERABLES (40%)**

**Completed Github Project Repository with relevant code, documents for each phase, supporting materials to include SQL queries, visualizations, etc.**

**Presentation (video capture)**

**Assessment Criteria:**

Students will be assessed based on the following criteria:

* Project Scope and Choice of Data
* Completeness and accuracy of the data analysis.
* Creativity in data visualization and report design.
* Effective use of AWS services.
* Optimization and cost analysis (if applicable)
* Presentation Skills