DevelopersHub Corporation

Data Science & AI/ML Engineering – Advanced Internship Tasks

Due Date: 31st Aug 2025

Overview

As part of your **Data Science & AI/ML Engineering Internship** at **DevelopersHub Corporation**, you are required to complete **at least 3 out of the 5 advanced tasks** listed below within the given timeline.

These tasks are designed to provide you with practical exposure to **real-world data science problems**, including **classification, clustering, time series forecasting, explainable AI (XAI), and business intelligence**.

You are encouraged to complete all tasks for deeper learning and a stronger portfolio. You'll work with popular datasets and industry-standard tools and libraries such as `pandas`, `numpy`, `scikit-learn`, `xgboost`, `shap`, `matplotlib`, `seaborn`, `Prophet`, and `Streamlit`.

Advanced Task Set

Task 1: Term Deposit Subscription Prediction (Bank Marketing)

Objective:
Predict whether a bank customer will subscribe to a term deposit as a result of a marketing campaign.

Dataset:
Bank Marketing Dataset (UCI Machine Learning Repository)

Instructions:

- Load and explore the dataset
- Encode all categorical features properly
- Train classification models (e.g., Logistic Regression, Random Forest)

- Evaluate the models using Confusion Matrix, F1-Score, and ROC Curve
- Use SHAP or LIME to explain at least 5 model predictions

Skills Gained:

- Classification modeling
- Feature encoding
- Model interpretability (Explainable AI - XAI)
- Customer behavior analysis

Task 2: Customer Segmentation Using Unsupervised Learning

Objective:
Cluster customers based on spending habits and propose marketing strategies tailored to each segment.

Dataset:
Mall Customers Dataset

Instructions:

- Conduct Exploratory Data Analysis (EDA)
- Apply K-Means Clustering to segment customers
- Use PCA or t-SNE to visualize the clusters
- Suggest relevant marketing strategies for each identified segment

Skills Gained:

- Unsupervised learning (K-Means)
- Dimensionality reduction (PCA, t-SNE)
- Customer segmentation

- Strategy development based on data insights

Task 3: Energy Consumption Time Series Forecasting

Objective:
Forecast short-term household energy usage using historical time-based patterns.

Dataset:
Household Power Consumption Dataset

Instructions:

- Parse and resample the time series data
- Engineer time-based features (e.g., hour of day, weekday/weekend)
- Compare performance of ARIMA, Prophet, and XGBoost models
- Plot actual vs. forecasted energy usage for visualization

Skills Gained:

- Time series forecasting
- Feature engineering
- Model comparison and evaluation (MAE, RMSE)
- Temporal data visualization

Task 4: Loan Default Risk with Business Cost Optimization

Objective:
Predict the likelihood of a loan default and optimize the decision threshold based on cost-benefit analysis.

Dataset:
Home Credit Default Risk Dataset

Instructions:

- Clean and preprocess the dataset
- Train binary classification models (e.g., Logistic Regression, CatBoost)
- Define business cost values for false positives and false negatives
- Adjust the model threshold to minimize total business cost

Skills Gained:

- Binary classification modeling
- Cost-based evaluation metrics
- Risk modeling and scoring
- Feature importance analysis

Task 5: Interactive Business Dashboard in Streamlit

Objective:
Develop an interactive dashboard for analyzing sales, profit, and segment-wise performance.

Dataset:
Global Superstore Dataset

Instructions:

- Clean and prepare the dataset
- Build a Streamlit dashboard with filters (Region, Category, Sub-Category)
- Display key performance indicators (KPIs) using charts:
    - Total Sales
    - Profit

- Top 5 Customers by Sales

Skills Gained:

- Business Intelligence (BI) dashboarding
- Data storytelling
- User interactivity with Streamlit
- Visual KPI analysis

Submission Requirements (Per Task)

Each completed task must be **uploaded to your GitHub repository** and shared via **Google Classroom**.

Checklist for Each Task:

1. **Jupyter Notebook**
    - Problem Statement and Objective
    - Dataset description and loading
    - Data cleaning and preprocessing
    - Exploratory Data Analysis (EDA)
    - Model building and evaluation
    - Visualizations (charts, plots, graphs)
    - Final conclusion with insights
2. **Code Quality**
    - Well-structured, readable, and commented code

3. **GitHub Repository**
    - Clear and descriptive repository name
    - README.md file including:
        - Task objective
        - Your approach
        - Results and findings
4. **Submission on Google Classroom**
    - Share the link to your GitHub repository after completing each task

---

Important Notes

- **Complete at least 3 out of the 5 tasks** by 31**st Aug 2025**
- You are encouraged to complete all 5 tasks
- Reach out to mentors if you need help or feedback
- Showcase your work on LinkedIn or GitHub for better visibility

---