Tasks for Data Science Interns

Task 1: EDA and Visualization of a Real-World Dataset

Description:
Perform exploratory data analysis (EDA) on a dataset such as the Titanic Dataset or Airbnb Listings Dataset.

Steps:

1. **Load the Dataset:** Use Pandas to load and explore the dataset.
2. **Data Cleaning:**
    - Handle missing values using imputation techniques or removal.
    - Remove duplicates.
    - Identify and manage outliers using statistical methods or visualizations.
3. **Visualizations:**
    - Create bar charts for categorical variables.
    - Plot histograms for numeric distributions.
    - Generate a correlation heatmap for numeric features.
4. **Summarize Insights:** Document your findings and observations in a clear and concise manner.

Outcome:

- A Jupyter Notebook or Python script containing the EDA process, visualizations, and detailed insights.

Task 2: Text Sentiment Analysis

Description:
Build a sentiment analysis model using a dataset such as IMDB Reviews.

Steps:

1. **Text Preprocessing:**
    - Tokenize text into individual words.
    - Remove stopwords.
    - Perform lemmatization for text normalization.
2. **Feature Engineering:**
    - Convert text data into numerical format using TF-IDF or word embeddings.
3. **Model Training:**

    - Train a classifier such as Logistic Regression or Naive Bayes to predict sentiment.
4. **Model Evaluation:**
    - Evaluate the model's performance using metrics like precision, recall, and F1-score.

Outcome:

- A working Python script that processes input text, predicts sentiment, and provides evaluation metrics.

Task 3: Fraud Detection System

Description:
Develop a fraud detection system using a dataset like the Credit Card Fraud Dataset.

Steps:

1. **Data Preprocessing:**
    - Handle imbalanced data using techniques like SMOTE or undersampling.
2. **Model Training:**
    - Train a Random Forest or Gradient Boosting model to detect fraudulent transactions.
3. **Model Evaluation:**
    - Evaluate the system's precision, recall, and F1-score.
4. **Testing Interface:**
    - Create a simple interface (e.g., a command-line input) to test the fraud detection system.

Outcome:

- A Python script capable of detecting fraudulent transactions with evaluation metrics and a testing interface.

Task 4: Predicting House Prices Using the Boston Housing Dataset

Description:
Build a regression model from scratch to predict house prices using the Boston Housing Dataset.

Steps:

1. **Data Preprocessing:**
    - Normalize numerical features and preprocess categorical variables.
2. **Model Implementation:**
    - Implement Linear Regression, Random Forest, and XGBoost models from scratch (avoid using built-in libraries like `sklearn.linear_model`).
3. **Performance Comparison:**
    - Compare the models using RMSE and R^2 metrics.
4. **Feature Importance:**
    - Visualize feature importance for tree-based models.

Outcome:

- A Python script containing the custom implementation of regression models, performance comparisons, and visualizations.

Submission Requirements

1. **GitHub Repository:**
    - Push your code, datasets, and all related files to a GitHub repository. Share the repository link.
2. **Visuals Submission:**
    - Record a short video or take screenshots of the data visualizations, showing insights and model performance.
3. **Documentation:**
    - Include a `README.md` in your repository, explaining the project steps, how to run the scripts, and your observations.
4. **Submission Deadline:**
    - All tasks must be completed and submitted by **28th May 2025**.
    - You will have to submit all of these tasks collectively on your consoles before deadline.