Data Science Task
ProgressSoft Corporation, Apollo Team

Instructions

The candidate must submit Python script(s) for each task, using valid syntax and appropriate file extensions. All scripts should be executable with minimal setup using only the Python interpreter. If external libraries are required, include them in a requirements.txt file. A README file should also be provided, specifying the Python version and any additional information needed to understand and run the code.

Important:

- Use of LLMs (e.g., ChatGPT, Copilot, Bard, etc.) for solving or generating any part of the assignment is strictly prohibited.

- If any part of this instruction is violated — including failure to meet the script, dependency, or documentation requirements, or evidence of LLM-generated content — the candidate will be disqualified.

Energy Consumption Forecasting and Uncertainty Quantification

Overview

This assignment challenges you to develop highly robust and interpretable time series forecasting models for household energy consumption, incorporating techniques and addressing real-world complexities. Beyond point forecasts, you will be required to quantify forecast uncertainty, which is critical for risk management and operational planning in the energy sector.

Duration: 1 week

Objectives

- **Data Preparation:** Handle large-scale, high-frequency time series data with significant missingness and anomalies, performing sophisticated aggregation.
- **Time Series Analysis:** Identify and model intricate temporal patterns, including multiple seasonalities and change points.
- **Feature Engineering:** Develop features from intrinsic time series properties and integrate diverse external data sources.
- **Ensemble & Probabilistic Modelling:** Build and evaluate a portfolio of forecasting models, including methods capable of providing prediction intervals.
- **Model Evaluation:** Assess model performance using a comprehensive suite of metrics, with a strong emphasis on uncertainty quantification.
- **Actionable Insights & Risk Assessment:** Translate forecasts and their associated uncertainties into concrete business strategies and risk mitigation recommendations.
- **Professional Communication:** Articulate complex methodologies and findings clearly and concisely to both technical and non-technical audiences.

Dataset

You will use the **'Individual household electric power consumption'** dataset from the UCI Machine Learning Repository. This dataset contains highly granular measurements of electric power consumption over several years.

- **Dataset Link:** Individual household electric power consumption Data Set

Note: This dataset is large and contains significant data quality challenges, including missing values and potential recording errors at a very high frequency.

Assignment Tasks

Phase 1: Data Preparation & Time Series Exploration

Data Loading & Cleaning

- Work with a time series dataset containing energy consumption data. Ensure proper datetime handling and assess the raw data quality.
- Identify and address missing values, outliers, and anomalies using well-justified techniques.

Temporal Aggregation & Multi-Frequency Views

- Aggregate the data at various time intervals (e.g., hourly, daily, weekly) to explore different temporal patterns. Choose appropriate aggregation logic per level.

Exploratory Time Series Analysis

- Decompose the time series to uncover trends, seasonality, and residuals.
- Analyze autocorrelation and potential structural shifts.
- Summarize insights on recurring patterns and their forecasting relevance.

Phase 2: Feature Engineering & Forecasting

Feature Design

- Derive time-based and lag features, rolling statistics, and interaction terms.
- Integrate an external data source relevant to energy usage (e.g., weather, holidays).
- Ensure all features respect temporal causality.

Train-Test Strategy

- Use a realistic time-based validation approach. Clearly define forecast horizon(s).

Modelling

- Build at least three diverse models, covering statistical, machine learning, and ensemble approaches.
- Include at least one probabilistic forecasting method with uncertainty estimation.

Evaluation

- Use appropriate metrics to assess forecast accuracy and interval quality.
- Visualize forecasts and analyze residuals.
- Compare models and summarize trade-offs between accuracy, interpretability, and uncertainty.

Submission Guidelines

- Submit your solution as a **Jupyter Notebook (.ipynb)** or a collection of **Python scripts (.py)** organized in a logical project structure, accompanied by a comprehensive **PDF report**.
- The code must be self-contained and fully runnable from start to finish.
- All significant plots and tables should be rendered in the notebook or included in the PDF report.
- Clearly document all assumptions, design choices, and challenges encountered.