Smart Energy Usage Prediction & Anomaly Detection

By: Avanigadda Chhatrapati

Roll No: EE22BTECH11012

Introduction & Motivation

- Rising energy costs and increasing environmental concerns, efficient energy monitoring has become essential for both consumers and utility providers.

Motivation:

- Rising global energy demands require intelligent monitoring system.
- Early detection of abnormal consumption can prevent energy wastage and equipment damage.
- ML approaches can provide some predictions and identify unusual patterns automatically.

Goals

Primary Objective:

- Predict daily consumption using Bayesian MMSE/MAP estimate.
- Classify energy usage patterns(low/normal/high) using SVM.
- Detect anomaly using likelihood ratio test.
- Compare model performance against theoretical bounds(CRLB).

Targets:

- Achieve >80% prediction accuracy.
- Maintain <5% false anomaly detection.

Dataset overview

- Source: https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set
- Date period: December 2006-November 2010 (47 months)
- Frequency of data: Minute level measurement

Data contains:

- Date/time
- Global active/reactive power
- Voltage, Current intensity
- Sub-metering (3 categories)

-Sub-metering 1 – Corresponds to Kitchen

-Sub-metering 2 – Corresponds to Laundry room

-Sub-metering 3 – Corresponds to Electric water heater and AC

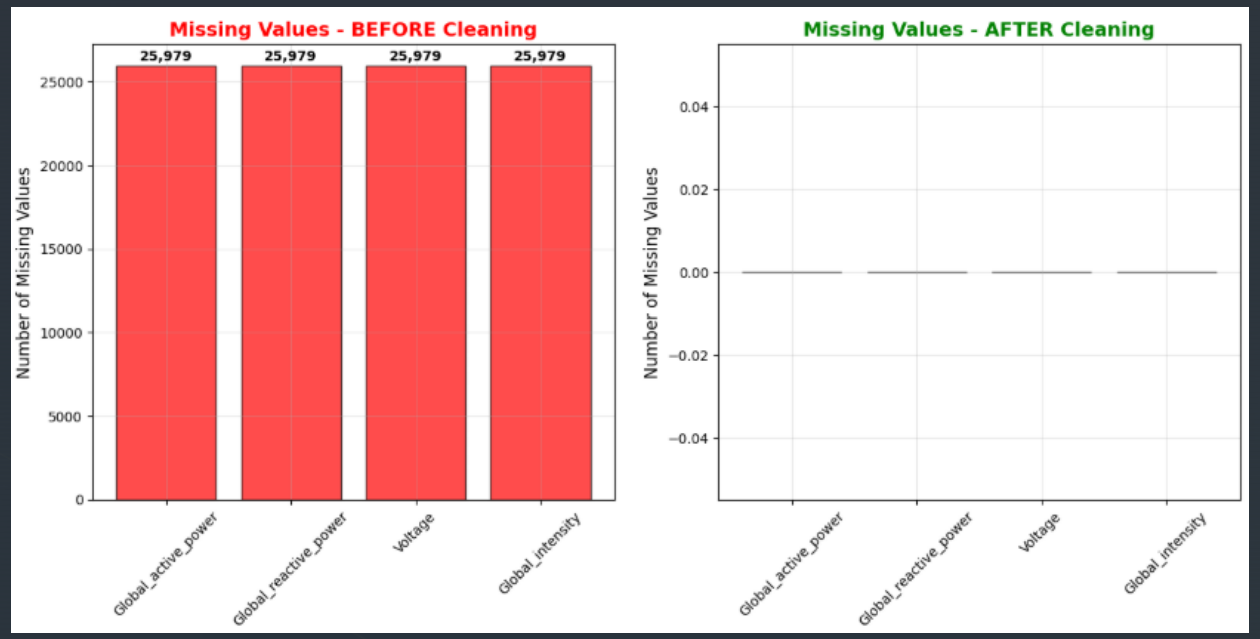
This data is not clean it contains many unknown values given by “?”

Dataset cleaning

- Identify all missing values represented by “?” in the dataset.
- Replace “?” with NaN(Not a number) to mark them as missing entries.
- Fill the missing values using forward fill(copy the previous valid entry).

```
# Fill missing values (forward fill)
data.ffill(inplace=True)
```

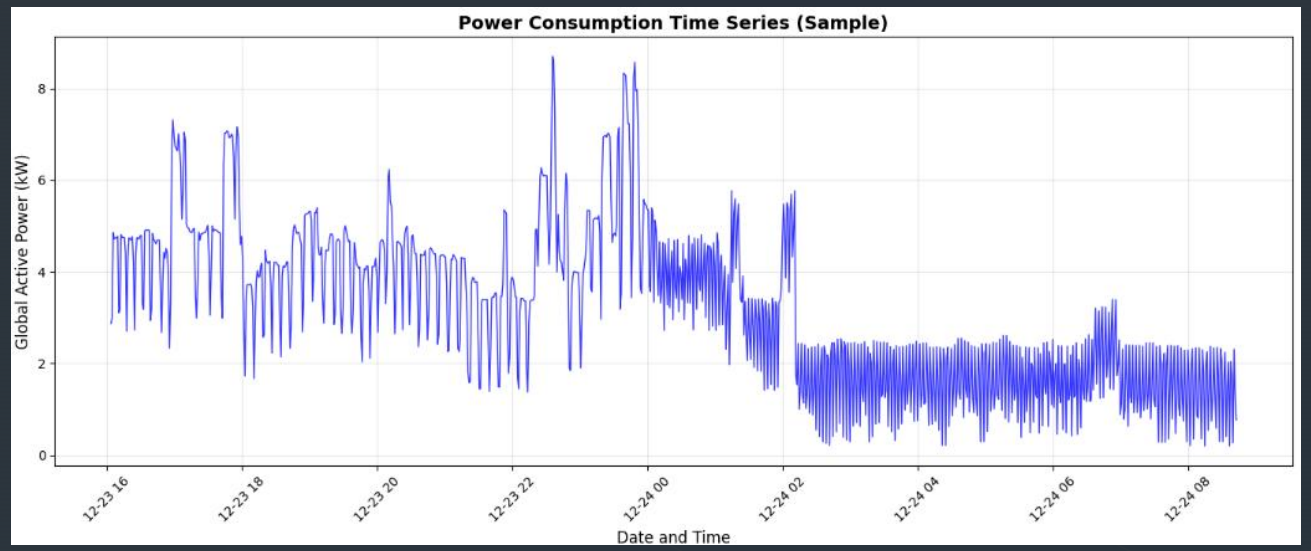
```
🔍 DATASET COMPARISON
================================================
Metric                    Raw Data        Cleaned Data
------------------------------------------------
Total Records             2075259         2075259
Total Columns             9               9
Missing Values            181853          0
```



- From the above plot, 25,979 values are missing in each of the following columns: Global_active_power, Global_reactive_power, Voltage, and Global_intensity.

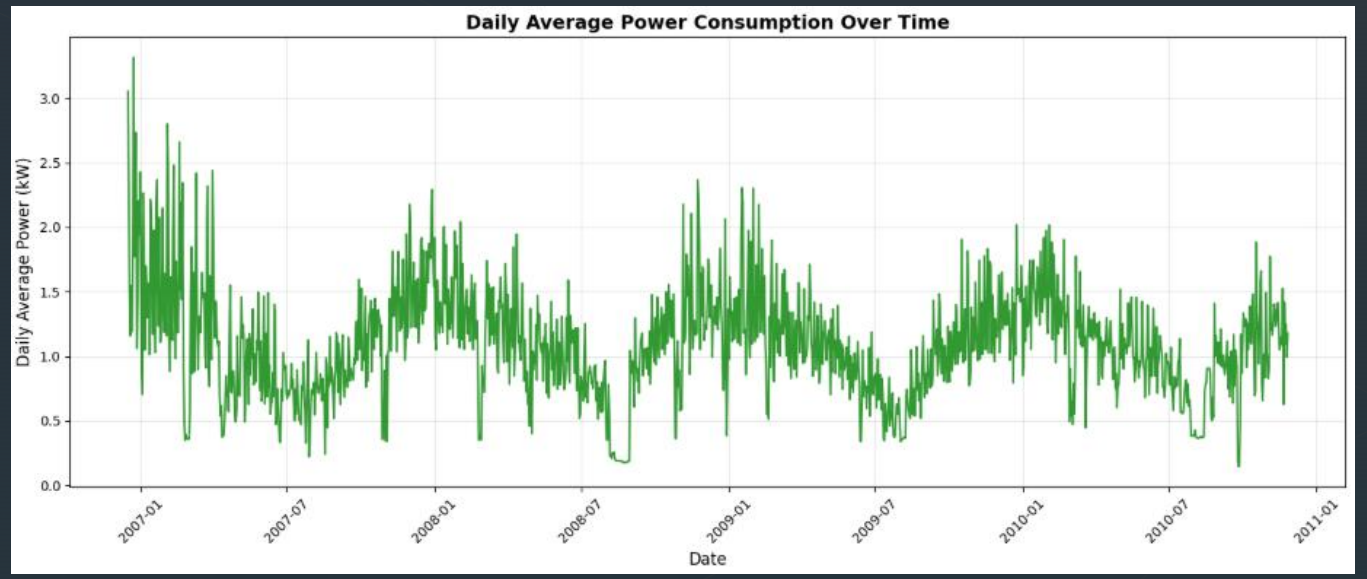
```
📊 MISSING DATA SUMMARY:
   Total missing values before cleaning: 103,916
   Total missing values after cleaning: 0
   Reduction: 100.0%
✅ DATA COMPLETENESS IMPROVEMENT:
   Global_active_power :   98.7% →  100.0% (+  1.3%)
   Global_reactive_power:   98.7% →  100.0% (+  1.3%)
   Voltage             :   98.7% →  100.0% (+  1.3%)
   Global_intensity    :   98.7% →  100.0% (+  1.3%)
```

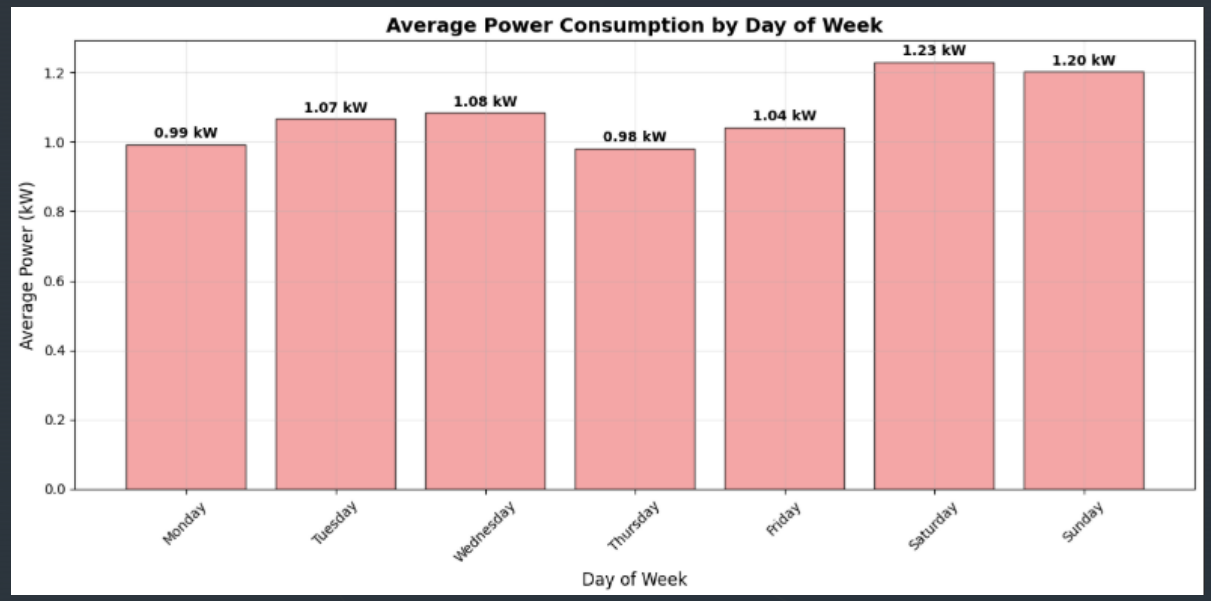
Data overview



Power consumption for a span of time

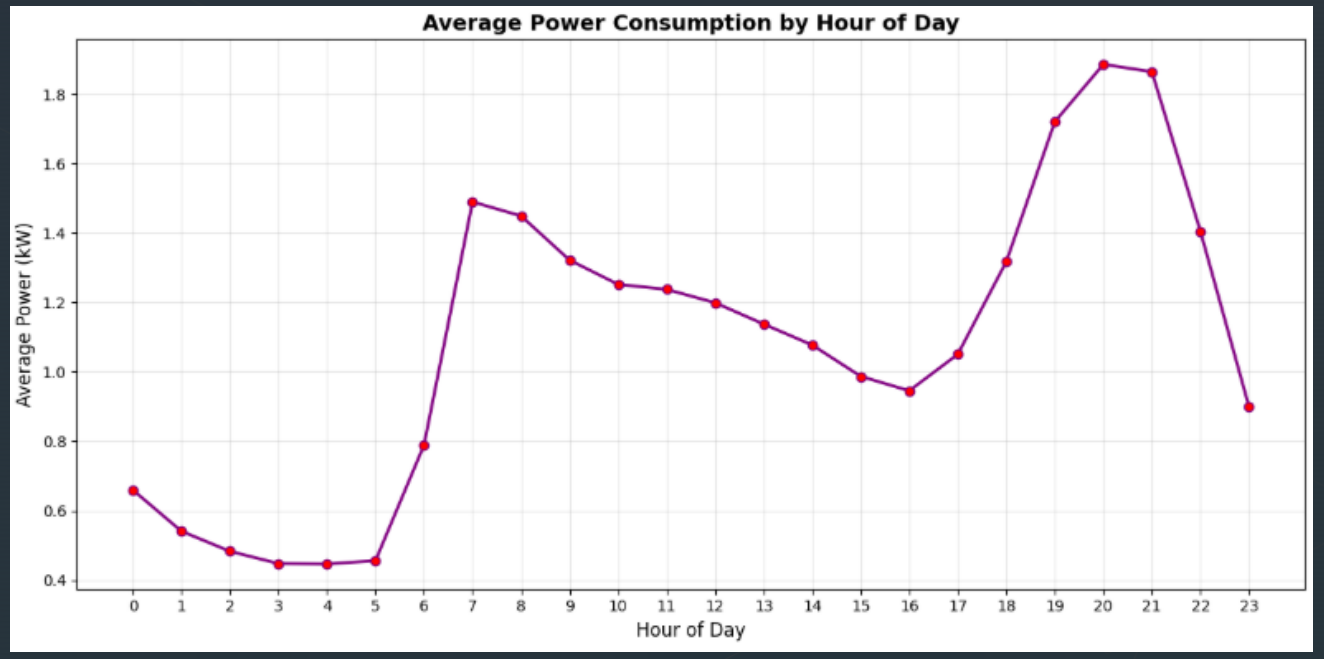
Average power consumption on daily basis



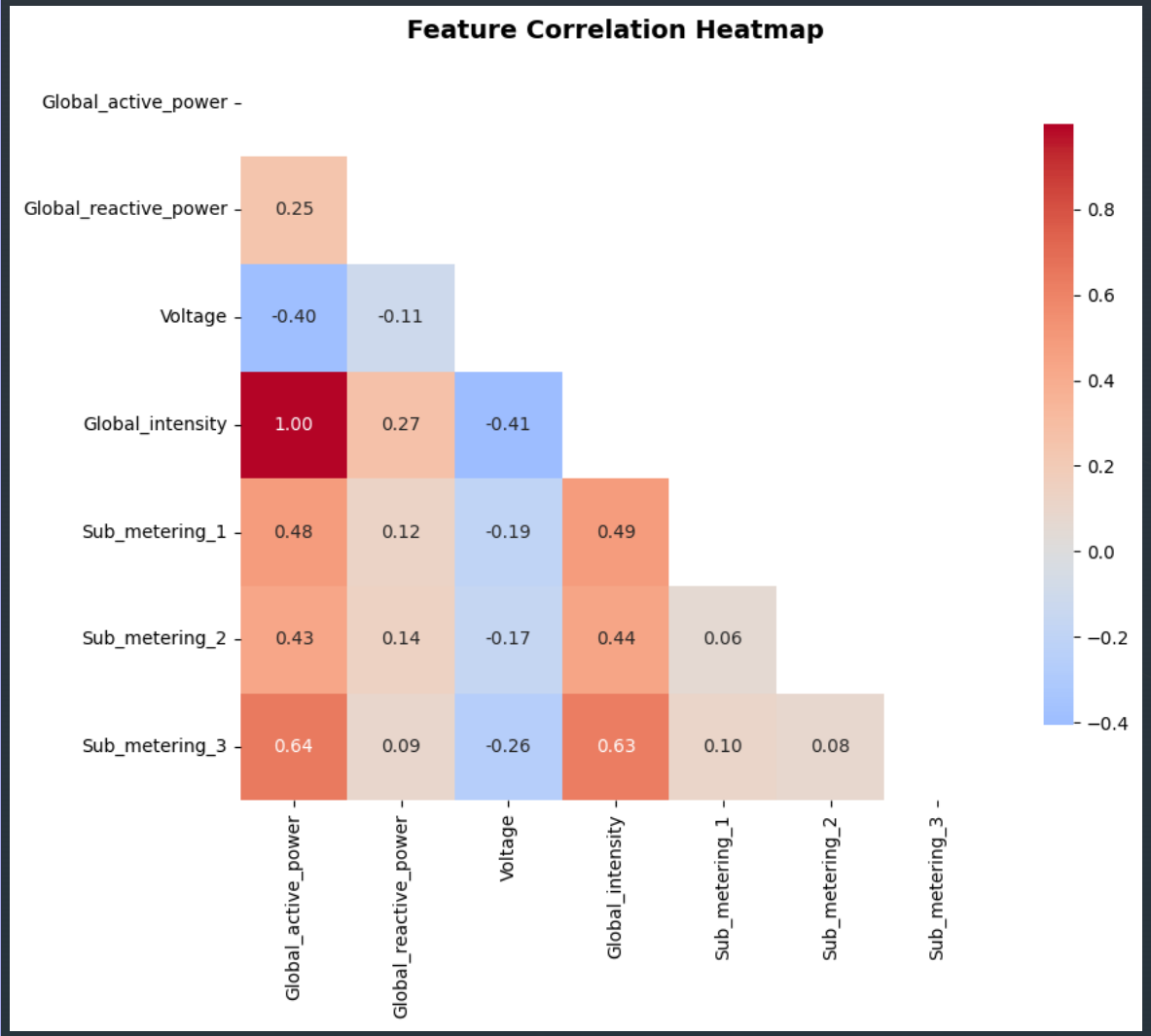


Average power consumption per day of week(weekends has more value)

Average power consumption per hour of a day(night and early morning has more value)



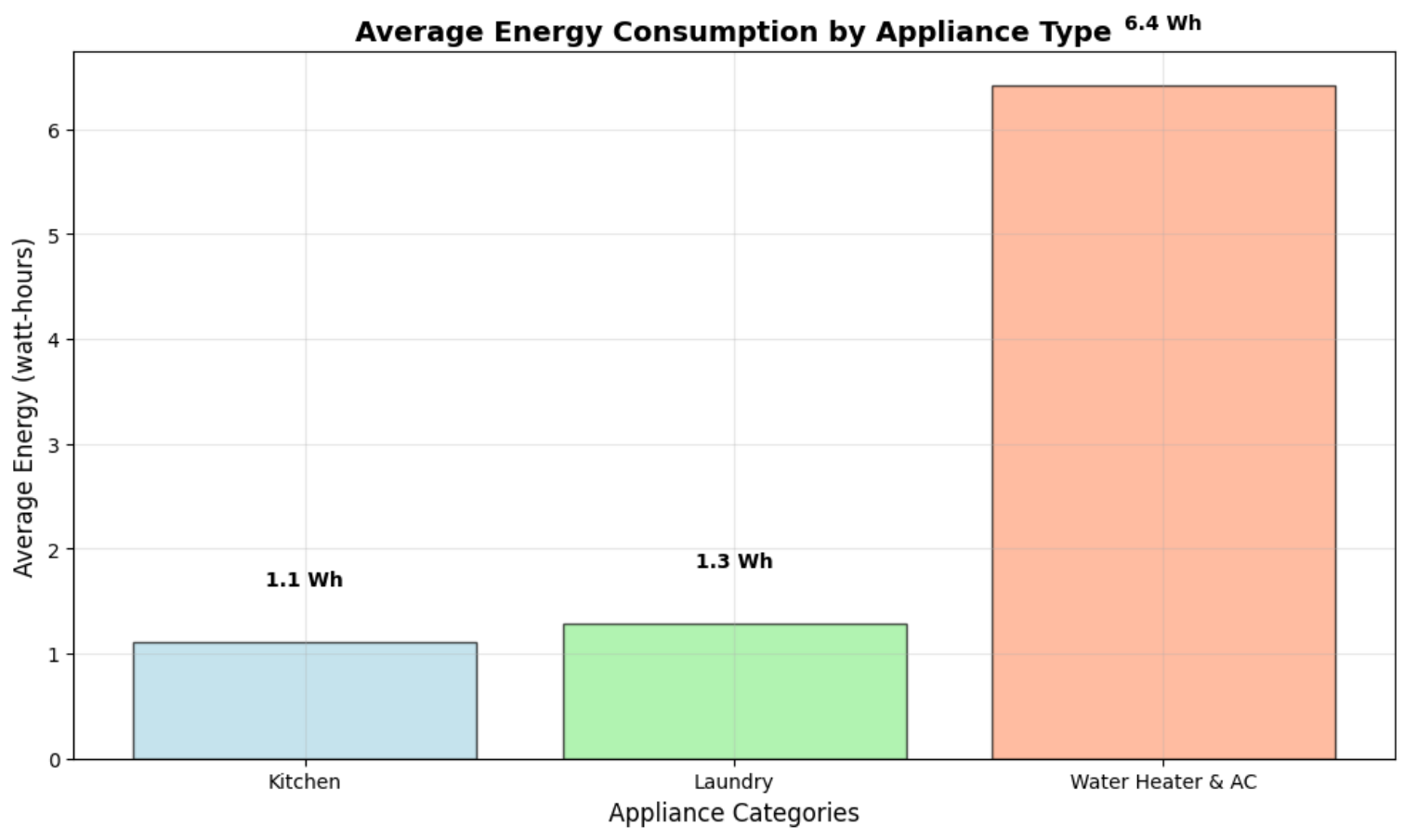
Correlation Heatmap of Energy Consumption Features:



- **Correlation Heatmap of Energy Consumption Features:**
- **Red Colors**: Positive correlation (features increase together)
- **Blue Colors**: Negative correlation (features move in opposite directions)

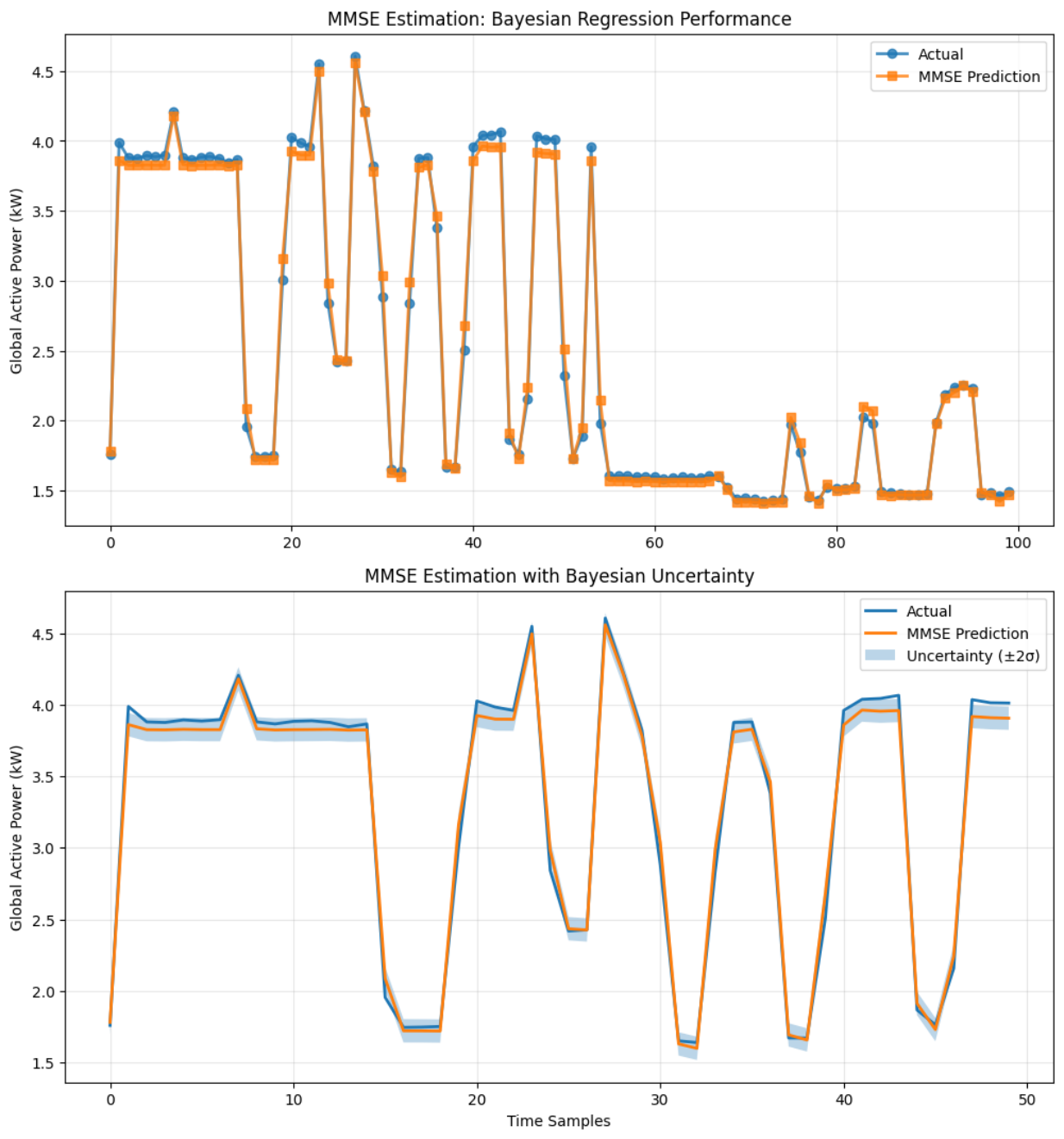
- Global_active_power-Global_intensity(r=1)
- Very strong relationship – current intensity directly relates to power consumption.
- Global_active_power-Sub_metering_3(r=0.64)
- Water heater and AC contribute significantly to total power and intensity.

Plot showing energy consumption by appliances



- Here most energy is consumed by water heater and AC.
- The least energy is consumed by Kitchen.

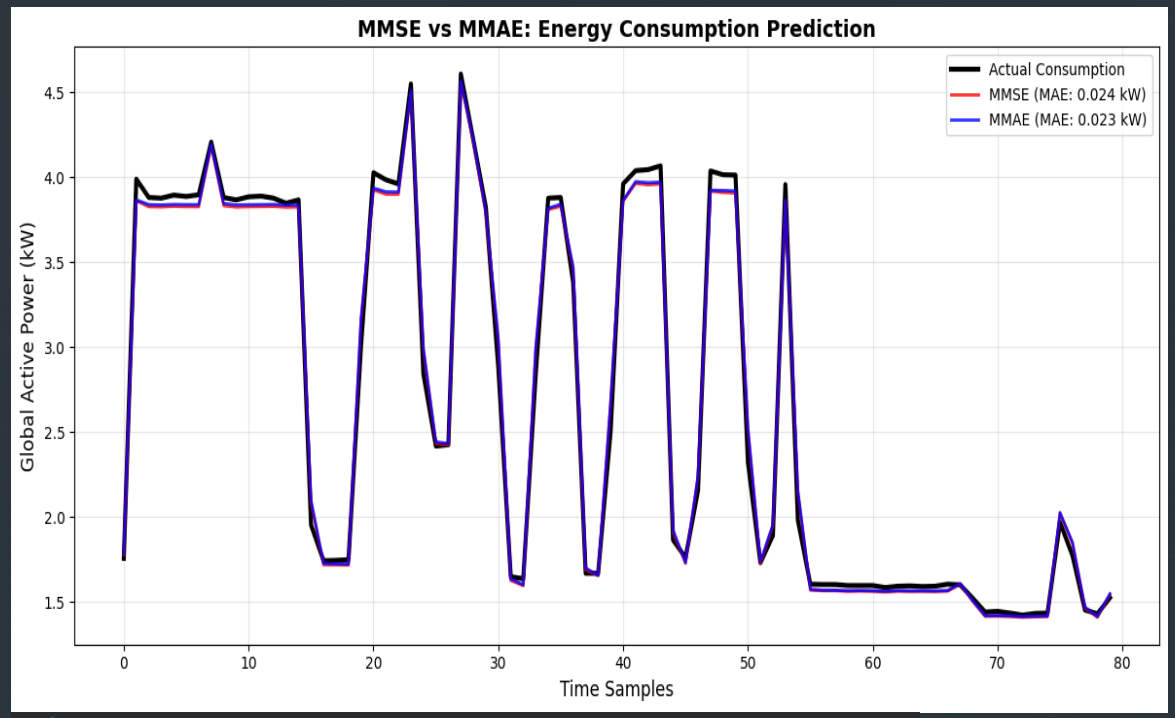
MMSE implementation results



- Used Bayesian approach for probabilistic predictions
- Provides uncertainty quantification with confidence intervals
- **Results:**
  - MSE: 0.0015
  - MAE: 0.0243
  - R^2 Score: 0.9984 (99.84% variance explained)

```
🎯 IMPLEMENTING MMSE ESTIMATION (Bayesian Regression)
Training Bayesian model...
📊 MMSE ESTIMATION RESULTS:
   Mean Squared Error (MSE): 0.0015
   Mean Absolute Error (MAE): 0.0243
   R² Score: 0.9984
```

MMAE implementation result




```
🎯 IMPLEMENTING MMAE ESTIMATION (Quantile Regression)
Training MMAE (Quantile Regression) model...
📊 MMAE ESTIMATION RESULTS:
   Mean Absolute Error (MAE): 0.0233 kW
   Mean Squared Error (MSE): 0.0015
   R² Score: 0.9984

📈 COMPARISON WITH MMSE:
   MMSE MAE: 0.0243 kW
   MMAE MAE: 0.0233 kW
   Difference: +3.90%
```

- **Goal**: Find conditional median - robust to outliers
- Used Quantile Regression (quantile=0.5) as it is handles extreme consumption values in a better way.
- **Results:**
  - Mean Absolute Error (MAE): 0.0233 kW
  - Mean Squared Error (MSE): 0.0015

  Difference: +3.90% (comparison with MMSE)

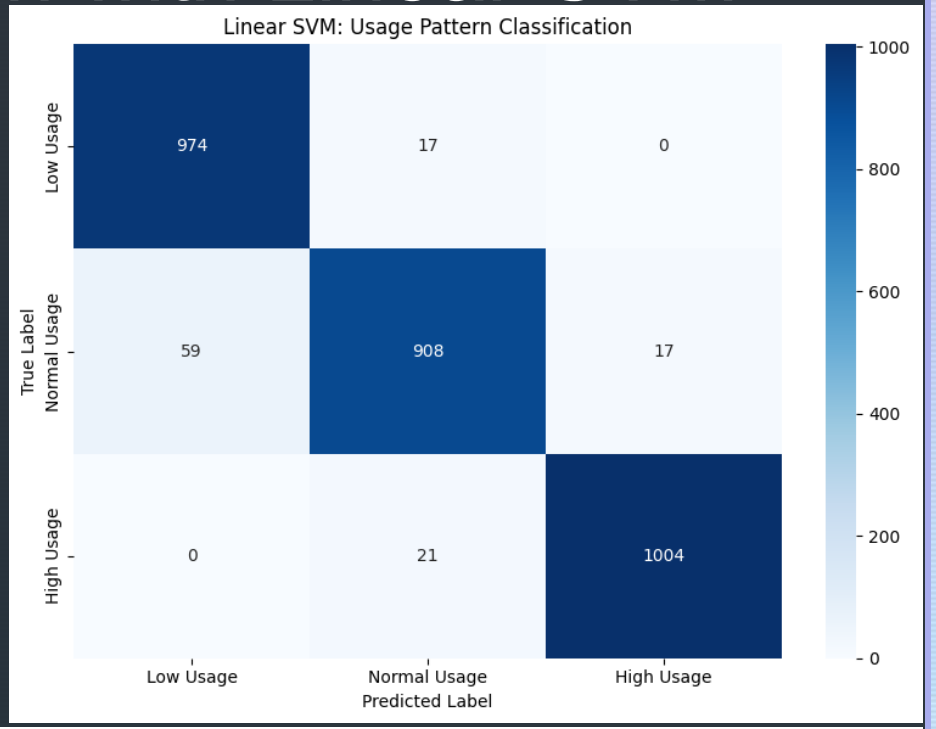
Usage pattern classification with Linear SVM

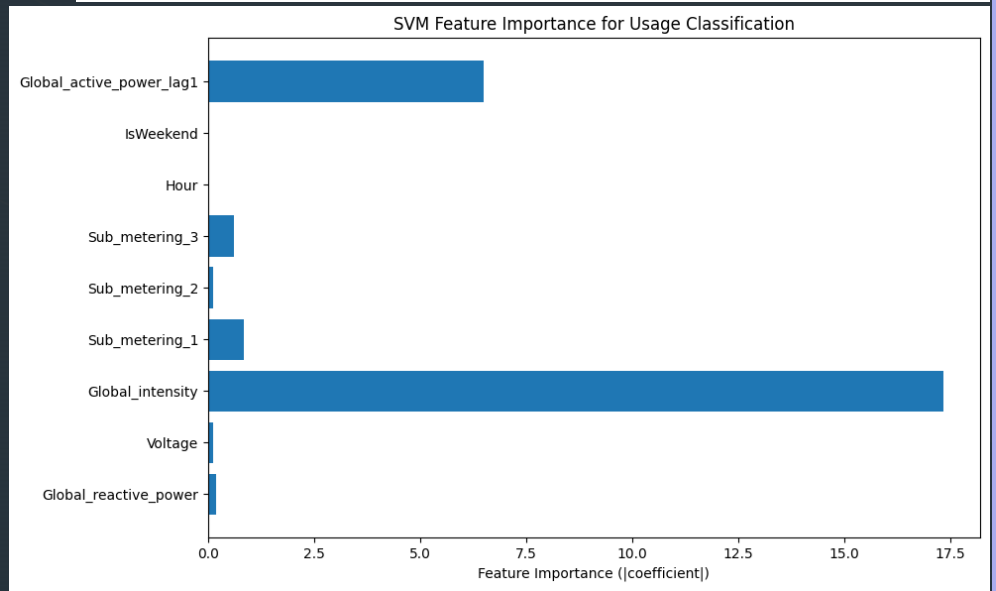
- **Method: Linear Support Vector Machine**
- **Goal**: Classify energy usage into Low/Normal/High patterns
- **Approach**: Linear kernel for fast training on large dataset
- **Classes**: 0=Low, 1=Normal, 2=High usage
- Accuracy of 96.2%

```
🎯 IMPLEMENTING SVM FOR USAGE PATTERN CLASSIFICATION (FAST)
Class distribution: {2: 703322, 1: 691550, 0: 680386}
Training Linear SVM classifier...
Training completed in 0.27 seconds
📊 SVM CLASSIFICATION RESULTS:
   Accuracy: 0.962
   Training time: 0.27 seconds
              precision    recall  f1-score   support

   Low Usage       0.94      0.98      0.96       991
Normal Usage       0.96      0.92      0.94       984
  High Usage       0.98      0.98      0.98      1025

    accuracy                           0.96      3000
   macro avg       0.96      0.96      0.96      3000
weighted avg       0.96      0.96      0.96      3000
```



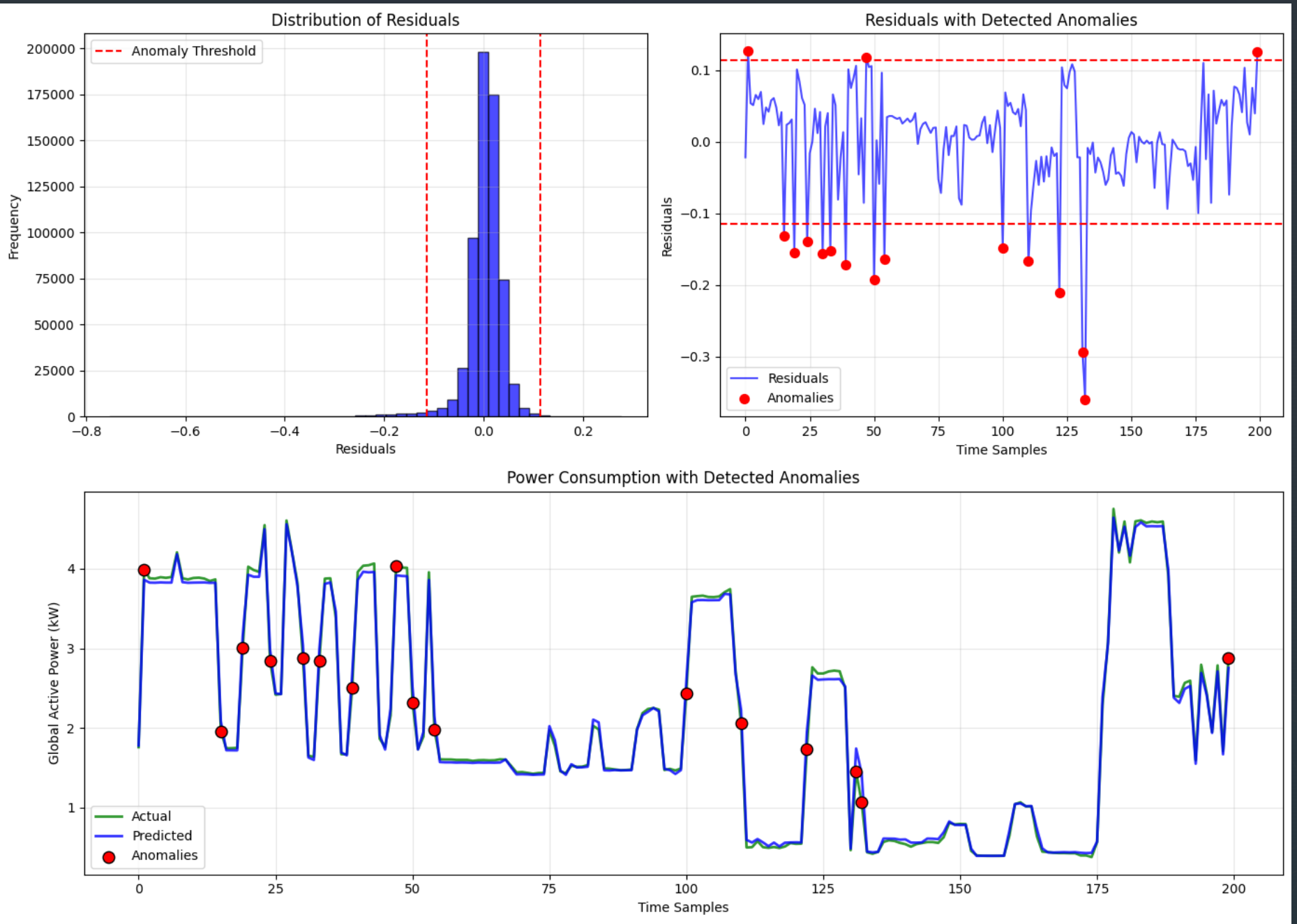


Anomaly detection

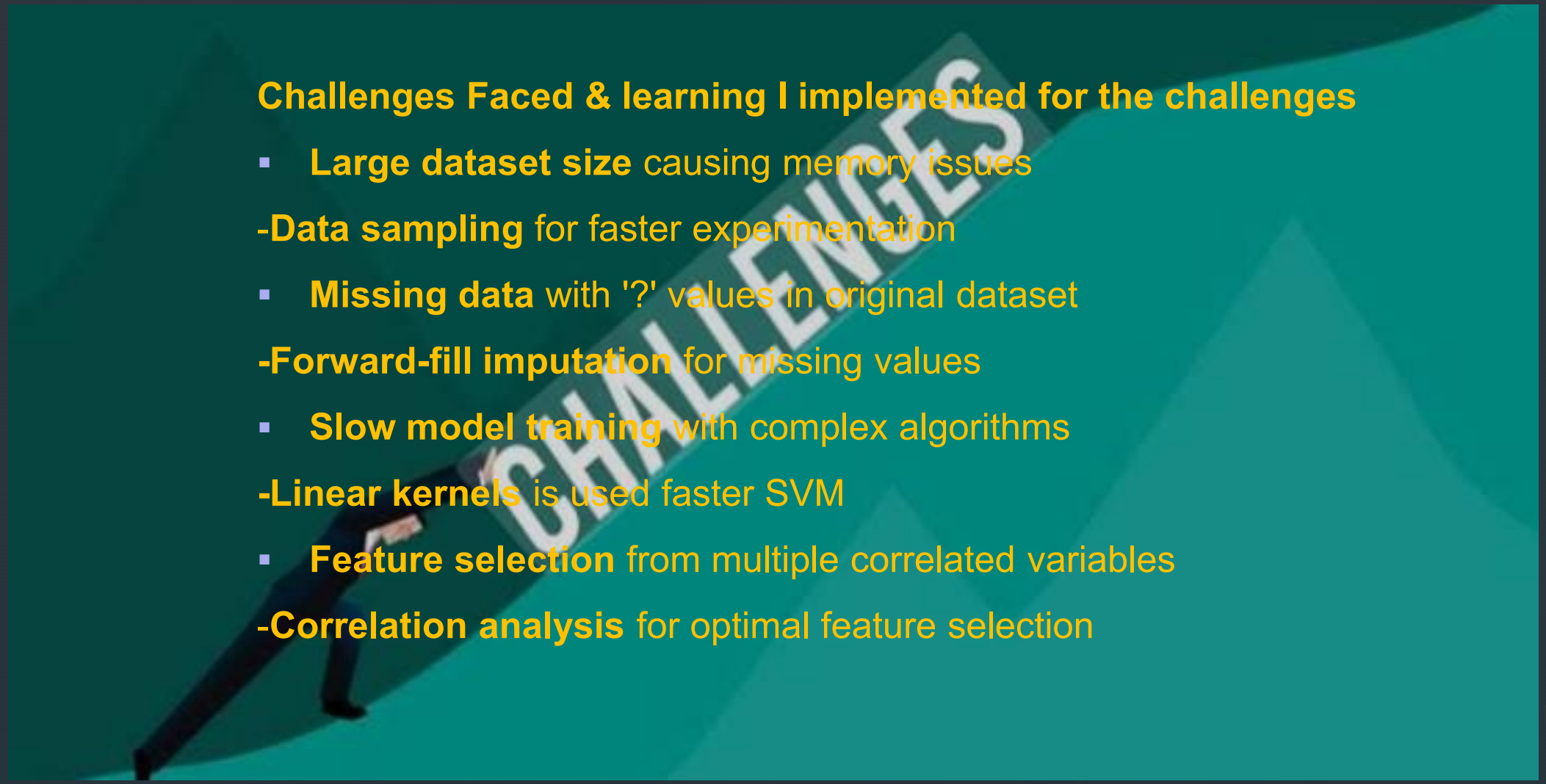
- Analyse prediction errors from MMSE errors.
- Flag points beyond 3 standard deviations.
- Unusually large prediction errors indicate anomalies.

```
🎯 IMPLEMENTING STATISTICAL ANOMALY DETECTION
📊 ANOMALY DETECTION RESULTS:
   Total samples: 622578
   Anomalies detected: 11286
   Anomaly rate: 1.81%
   Residual statistics - Mean: 0.0039, Std: 0.0380
```

Intermediary results



Learning & Challenges



Challenges Faced & learning I implemented for the challenges

- **Large dataset size** causing memory issues

-Data sampling for faster experimentation

- **Missing data** with '?' values in original dataset

-Forward-fill imputation for missing values

- **Slow model training** with complex algorithms

-Linear kernels is used faster SVM

- **Feature selection** from multiple correlated variables

-Correlation analysis for optimal feature selection

Planned work ahead…

Immediate next steps:

- Implement Bayesian Decision Theory with cost matrices
- Calculate CRLB for theoretical performance bounds