Water Consumption Analysis and Prediction Using Machine Learning

Introduction

This report summarizes the analysis of water consumption data, focusing on predicting current charges, identifying patterns through clustering, and conducting time series analysis. The objective is to explore the interrelationships among features while employing various statistical and machine learning techniques.

Abstract

Focus:

Integrating IoT, machine learning, and cloud infrastructure for optimized water governance.

Goals:

Optimize water distribution.

Reduce waste and improve accountability.

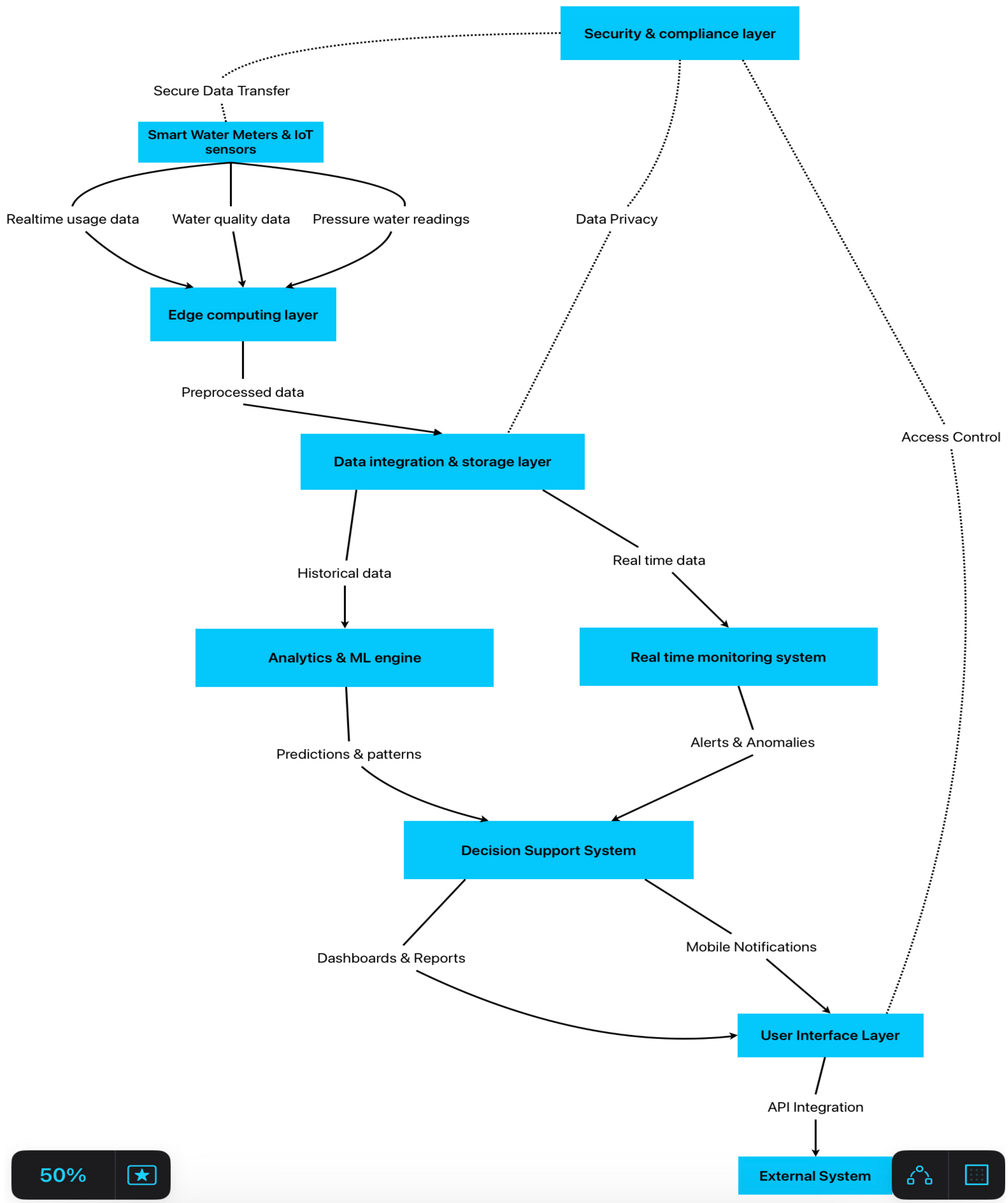
Enable real-time monitoring and predictive analytics

Dataset Overview

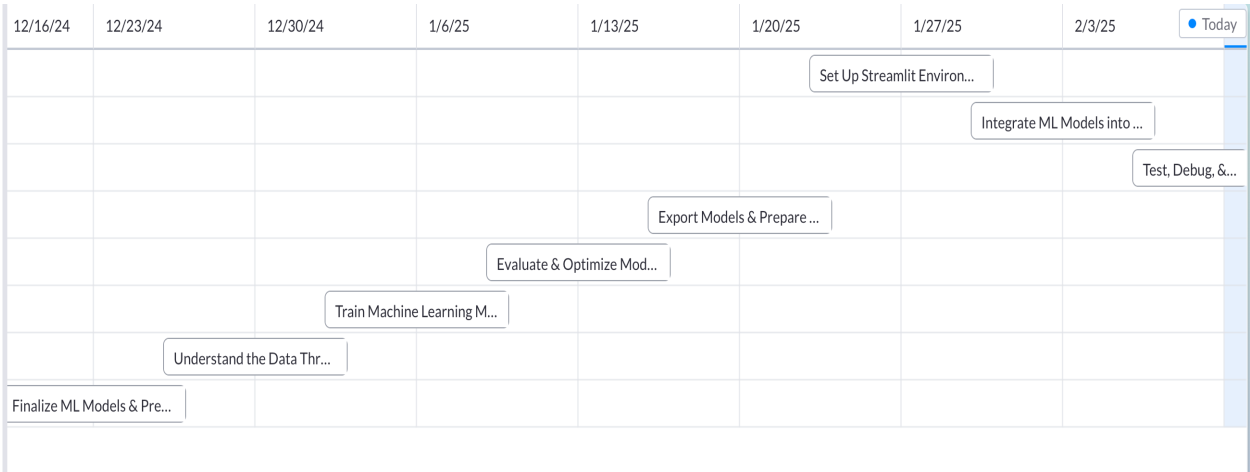
The dataset comprises **50,315** entries with **25 columns**, including features related to water consumption, charges, and account details. Key attributes include:

- **Consumption (HCF)**: The volume of water consumed.
- **Current Charges**: The billing amount associated with consumption.
- **Estimated**: Indicates whether a charge is estimated.
- **Cluster**: Derived from K-Means clustering to segment data.

High Level Architecture



Gantt Chart



Methodology

1. **Data Cleaning**:
   - Addressed missing values and formatted date fields.
   - Introduced new features such as Consumption_per_day to enhance analysis.
2. **Exploratory Data Analysis (EDA)**:
   - Visualized data distributions and relationships.
   - Generated a correlation matrix to understand feature interdependencies.

```
  Development Name   Borough  Account Name                   Location  \
0     BAISLEY PARK    QUEENS  BAISLEY PARK                     BLD 09
1     BAISLEY PARK    QUEENS  BAISLEY PARK                     BLD 09
2     BAISLEY PARK    QUEENS  BAISLEY PARK                     BLD 09
3     BAISLEY PARK    QUEENS  BAISLEY PARK                     BLD 09
4         BAY VIEW  BROOKLYN      BAY VIEW  BLD 25 - Community Center

  Meter AMR       Meter Scope  TDS #  EDP  RC Code      Funding Source  ...  \
0       AMR            BLD 09   91.0  240  Q009100             FEDERAL  ...
1       AMR            BLD 09   91.0  240  Q009100             FEDERAL  ...
2       AMR            BLD 09   91.0  240  Q009100             FEDERAL  ...
3       AMR            BLD 09   91.0  240  Q009100             FEDERAL  ...
4      NONE  Community Center   92.0  670  K209200  MIXED FINANCE/LLC1  ...

  Service End Date # days  Meter Number Estimated Current Charges  \
0       01/26/2020   34.0     K13060723         N          196.35
1       02/24/2020   29.0     K13060723         N          258.35
2       03/23/2020   28.0     K13060723         N          217.02
3       04/23/2020   31.0     K13060723         N          103.34
4       01/26/2020   34.0     E17250205         N           72.34

             Rate Class  Bill Analyzed Consumption (HCF) Water&Sewer Charges  \
0  Basic Water and Sewer           Yes                19              196.35
1  Basic Water and Sewer           Yes                25              258.35
2  Basic Water and Sewer           Yes                21              217.02
3  Basic Water and Sewer           Yes                10              103.34
4  Basic Water and Sewer           Yes                 7               72.34

   Other Charges
0            0.0
1            0.0
2            0.0
3            0.0
4            0.0
```

Parameters/Features

```
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Development Name      50255 non-null  object 
 1   Borough               50315 non-null  object 
 2   Account Name          50315 non-null  object 
 3   Location              49487 non-null  object 
 4   Meter AMR             49805 non-null  object 
 5   Meter Scope           12782 non-null  object 
 6   TDS #                 50255 non-null  float64
 7   EDP                   50315 non-null  int64  
 8   RC Code               50315 non-null  object 
 9   Funding Source        50239 non-null  object 
 10  AMP #                 50193 non-null  object 
 11  Vendor Name           50315 non-null  object 
 12  UMIS BILL ID          50315 non-null  int64  
 13  Revenue Month         50315 non-null  object 
 14  Service Start Date    50308 non-null  object 
 15  Service End Date      50308 non-null  object 
 16  # days                50308 non-null  float64
 17  Meter Number          50315 non-null  object 
 18  Estimated             50315 non-null  object 
 19  Current Charges       50315 non-null  float64
 20  Rate Class            50279 non-null  object 
 21  Bill Analyzed         50315 non-null  object 
 22  Consumption (HCF)     50315 non-null  int64  
 23  Water&Sewer Charges   50315 non-null  float64
 24  Other Charges         50315 non-null  float64
```

Handle Missing Values

```
missing_counts = data.isnull().sum()
print("Missing values per column:\n", missing_counts[missing_counts > 0])
```

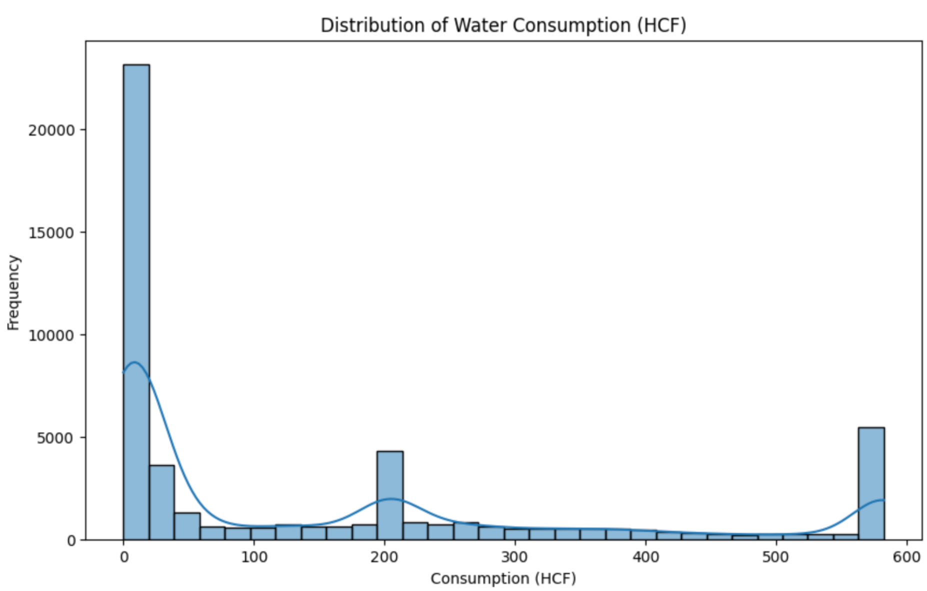
```
Missing values per column:
 Development Name         60
Location                828
Meter AMR               510
Meter Scope           37533
TDS #                    60
Funding Source           76
AMP #                   122
Service Start Date        7
Service End Date          7
# days                    7
Rate Class               36
dtype: int64
```

After replacing missing and NaN values

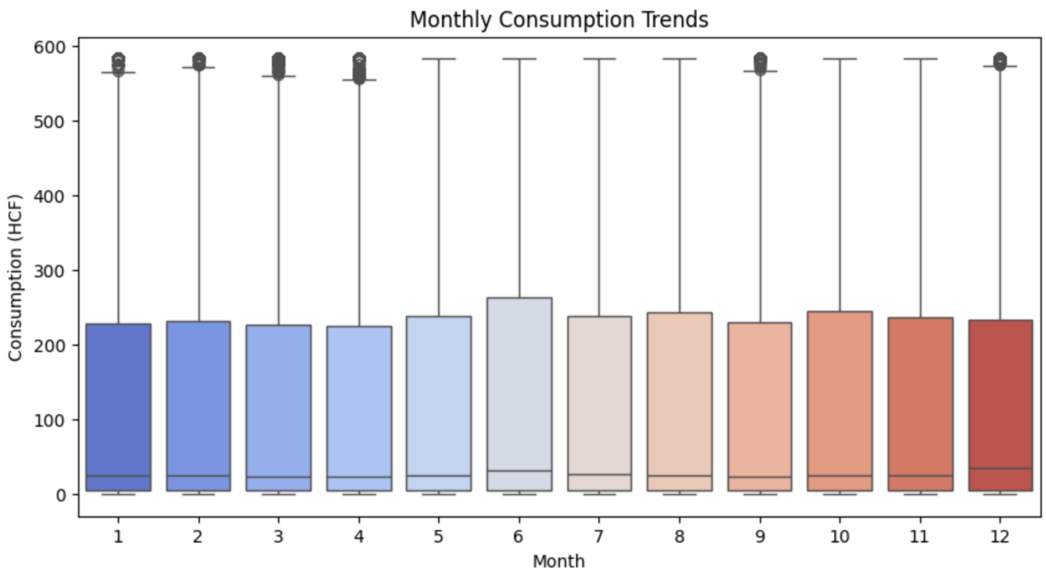
```
# Verify that there are no more missing values
print("Missing values after imputation:\n", data.isnull().sum().sum())
```

```
Missing values after imputation:
 0
```

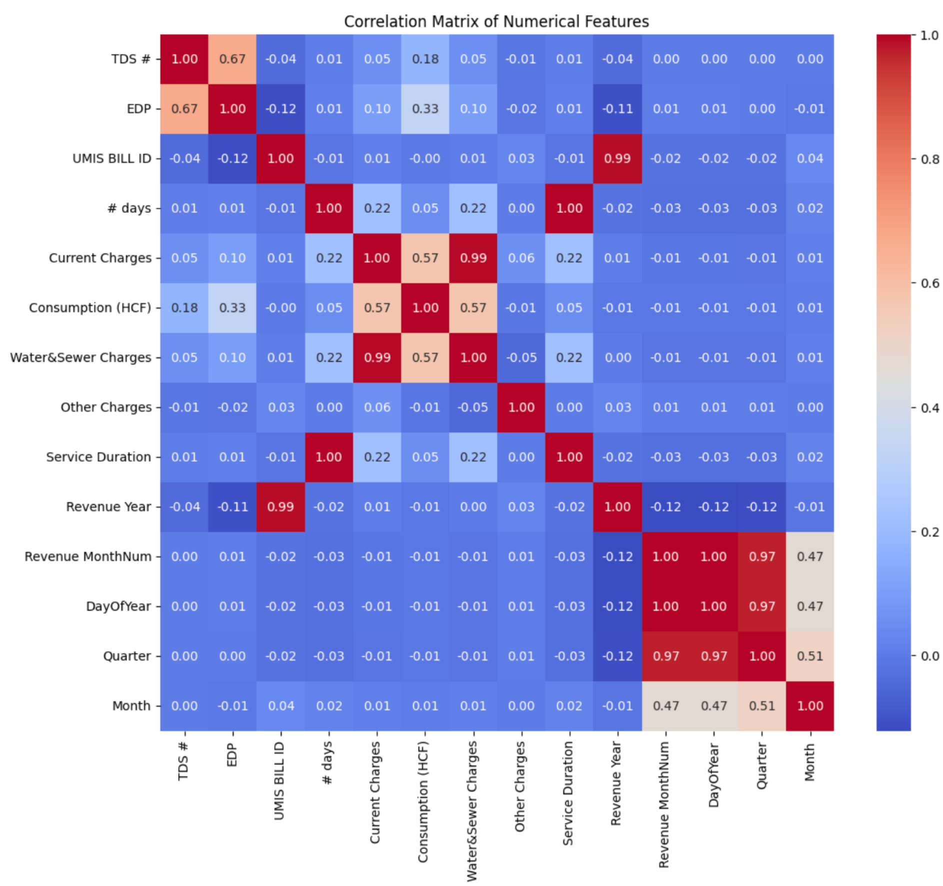
Visualizations



Histogram of Water Consumption



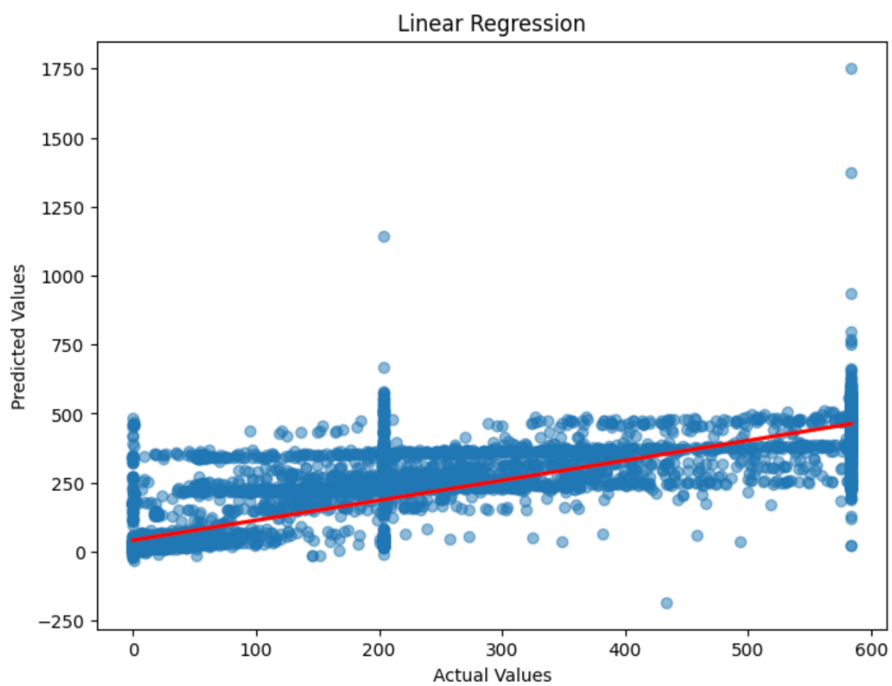
Monthly Consumption Trends



Correlation Matrix of Numerical features

3. **Modeling Approaches**:
    - Linear Regression
    - Random Forest
    - Gradient Boosting
    - Support Vector Machine

Linear Regression



Model Overview:
Predicting Current Charges using Consumption (HCF).
Fitted LinearRegression() model.

Evaluation Metrics:

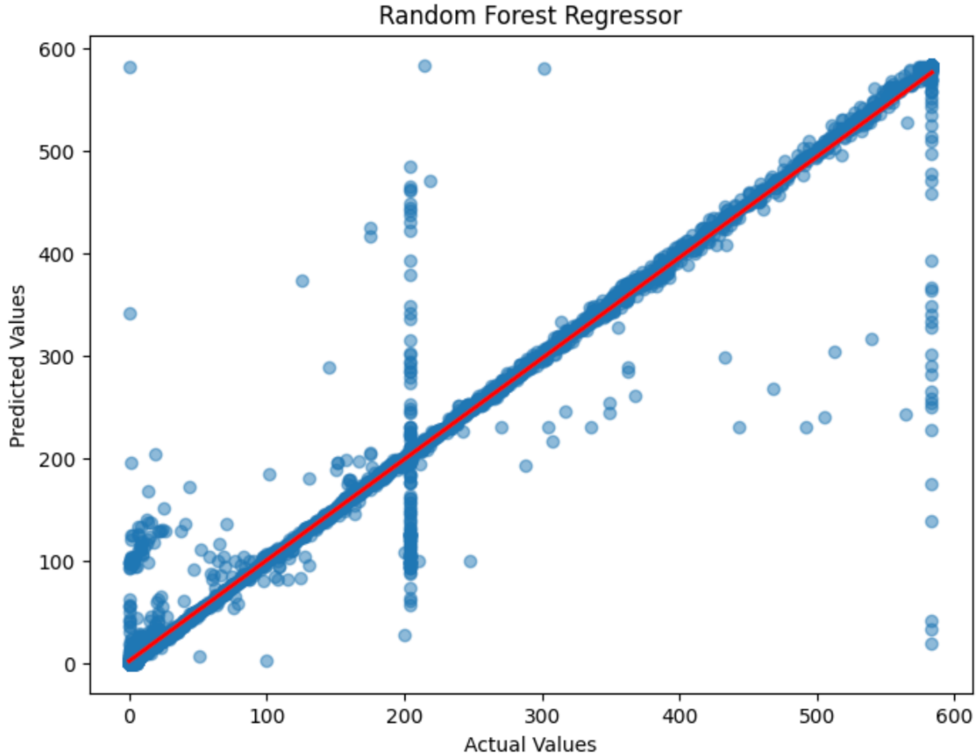
R-squared: 0.735

Mean Squared Error (MSE): 10322.455

Mean Absolute Error (MAE): 64.767

Insights:
Linear Regression struggles to capture data complexities, showing the lowest R^2 and highest errors. It may be useful as a baseline model but lacks the predictive power of other approaches.

Random Forest



Model Overview:
Predicting Current Charges using Consumption (HCF).
Fitted RandomForestRegressor() model.

Evaluation Metrics:

- R-squared: 0.982
- Mean Squared Error (MSE): 708.090
- Mean Absolute Error (MAE): 5.069

Insights:
The Random Forest model provides strong predictive accuracy with a high R^2 score, indicating a good fit. Minimal error suggests robust generalization.

Gradient Boosting

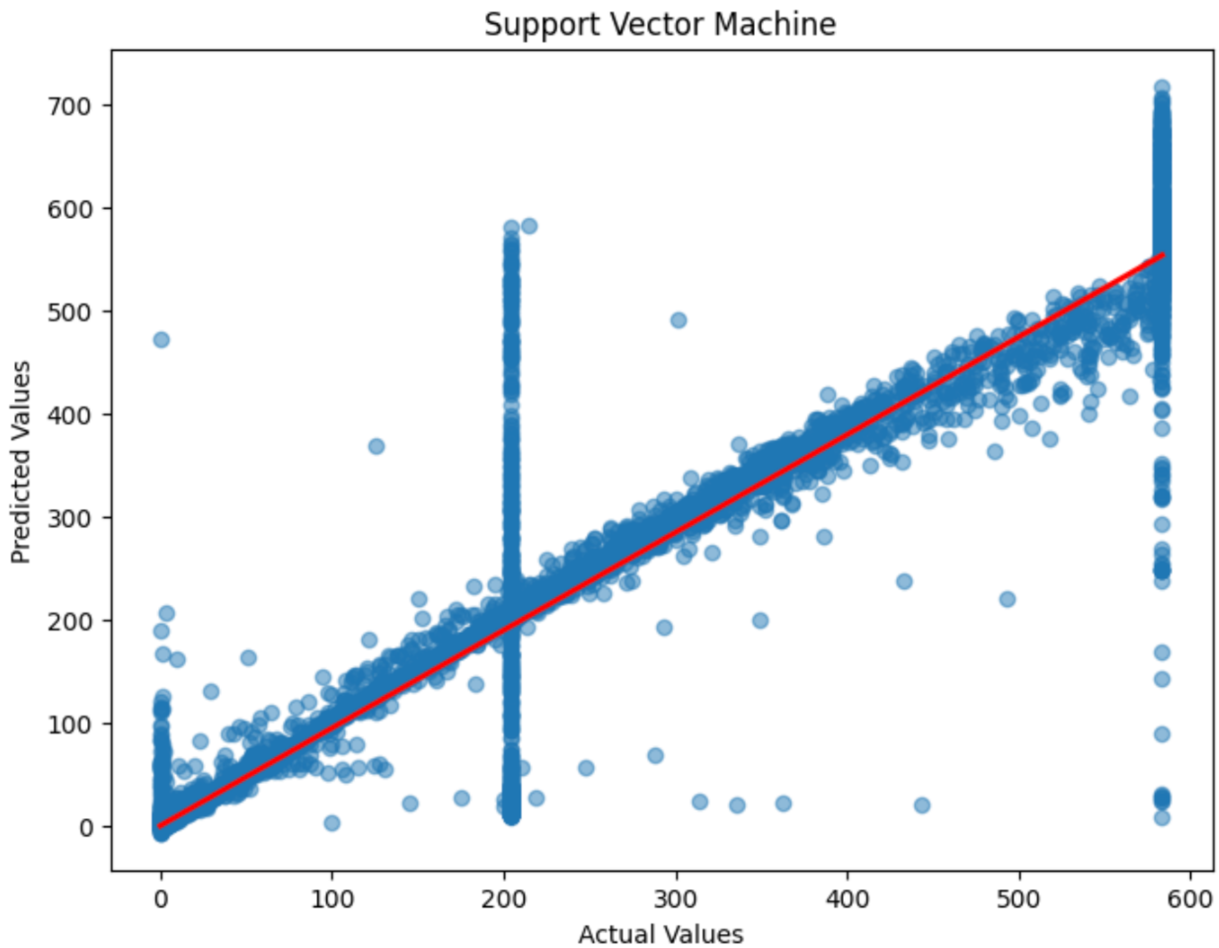
Model Overview:
Predicting Current Charges using Consumption (HCF).
Fitted GradientBoostingRegressor() model.

Evaluation Metrics:

- R-squared: 0.983
- Mean Squared Error (MSE): 653.342
- Mean Absolute Error (MAE): 6.523

Insights:
Gradient Boosting achieves the best performance among tested models, slightly outperforming Random Forest. It captures complex patterns well but may require tuning to reduce slight MAE increase.

Support Vector Machine



Model Overview:
Predicting Current Charges using Consumption (HCF).
Fitted SVR() model.

Evaluation Metrics:

- R-squared: 0.905
- Mean Squared Error (MSE): 3687.852
- Mean Absolute Error (MAE): 24.480

Insights:
SVM exhibits lower predictive power compared to ensemble models, with a notable increase in error. It may benefit from kernel tuning and hyperparameter optimization.

R-squared, MSE, and MAE for Models

```
Random Forest Results:
R-squared: 0.982
Mean Squared Error: 708.090
Mean Absolute Error: 5.069

Gradient Boosting Results:
R-squared: 0.983
Mean Squared Error: 653.342
Mean Absolute Error: 6.523

Support Vector Machine Results:
R-squared: 0.905
Mean Squared Error: 3687.852
Mean Absolute Error: 24.480

Linear Regression Results:
R-squared: 0.735
Mean Squared Error: 10322.455
Mean Absolute Error: 64.767
```

Best Model

```
The best performing model is: Gradient Boosting
With MSE: 653.34 and R-squared: 0.98
```

Recommendations

1. **Model Enhancement**: Consider using more advanced regression techniques (e.g., Ridge, Lasso, and ensemble methods) to improve predictions, especially for high-charge scenarios.
2. **Feature Engineering**: Introduce additional features, such as demographic information or weather data, to enrich the dataset and improve model performance.
3. **Imbalance Handling**: Address class imbalance in the target variable through techniques like SMOTE to enhance decision tree performance.
4. **Further Analysis**: Conduct additional clustering analyses or segmentation strategies to gain insights into different consumer behaviors and preferences.

Conclusion

Summary:

- Successfully built predictive models including **Linear Regression, Random Forest, Gradient Boosting, and Support Vector Machines (SVM)**.
- Applied **ensemble learning** for improved accuracy and **evaluated model performance** using multiple metrics.

Impact:

- Provides insights into **consumption patterns**, **charge prediction**, and potential **cost optimizations**.
- Helps in **identifying anomalies** and **forecasting future usage trends**.
- Can support **smart city initiatives** for **efficient water resource management and governance**.