Final Project Report

Predicting Household Energy Usage

Hunter Adrian
Brown University
https://github.com/huntera-1/Data1030-Final-Project
December 15, 2024

Introduction

Problem Motivation

Household energy consumption is a significant contributor to global energy demand and carbon emissions. Accurate predictions of energy usage enable efficient resource management, sustainable energy practices, and waste reduction. With the increase in number of smart devices and sensors, vast datasets capturing energy usage and environmental conditions have become available. This project leverages machine learning techniques to predict household energy consumption, offering a path toward more sustainable living.

Dataset Description

The dataset for this project was collected from a household energy monitoring system and contains approximately **20,000 records**. It includes:

- **Target Variable**: Appliance energy consumption (Wh).
- **Features**: Environmental factors such as temperature, humidity, wind speed, visibility, and pressure.
- **Temporal Data**: Timestamp information enabling feature engineering for cyclical and lagged effects.

The dataset required extensive preprocessing, including handling missing values, scaling features, and engineering lagged and rolling statistical features to model temporal dependencies.

Previous Work

Energy consumption prediction has been studied extensively using statistical and ML approaches. Linear models, such as Ridge regression, are valued for their simplicity and interpretability, while ensemble methods, such as Random Forest, have demonstrated robustness in capturing complex relationships. Recently, XGBoost has shown significant promise for its ability to handle non-linear patterns and interactions. This project builds upon these findings, developing a robust pipeline to model household energy usage effectively.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis focuses on uncovering trends and patterns in energy usage and its relationships with environmental factors.

Energy Consumption Trends
Daily energy consumption fluctuates significantly over time, with higher usage during colder months. Peaks align with daily household routines, such as increased evening activity.

Correlation Analysis
A correlation heatmap reveals strong relationships between energy consumption and environmental features:

- **Outdoor Temperature**: Inversely correlated, reflecting heating demands during colder periods.
- **Humidity Levels**: showed a positive relationship in certain areas, which may be linked to systems that control the environment.

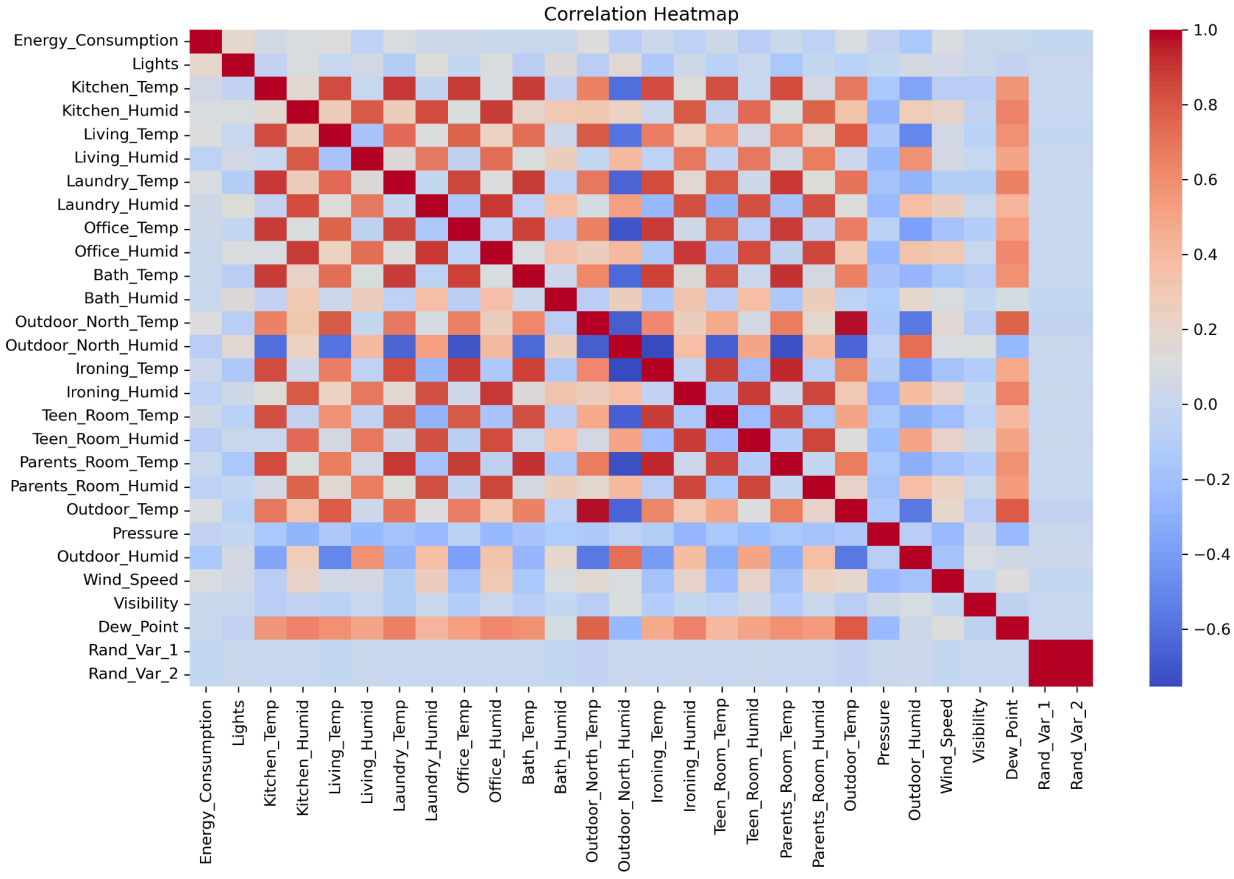


Figure: Correlation heatmap highlighting relationships between features and energy consumption.

Distribution of Energy Consumption

The histogram shows a right-skewed distribution, with most consumption values clustered at lower ranges and occasional spikes reflecting high-demand periods.

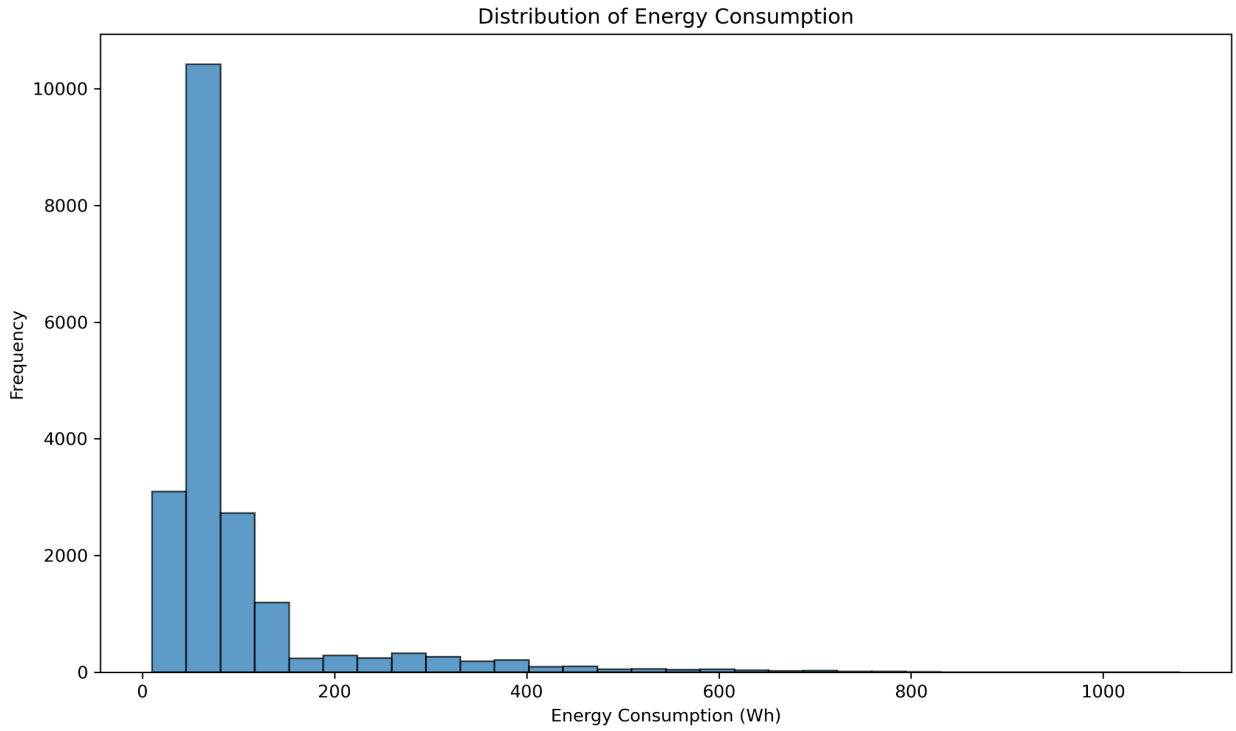


Figure: Histogram of energy consumption showing a right-skewed distribution.

Methodology

Splitting Strategy

The dataset was divided into three subsets:

- **Training Set** (60%): Used to train the models.
- **Validation Set** (20%): Used for tuning hyperparameters and selecting the best model.
- **Test Set** (20%): Held back for the final evaluation.

The split was done chronologically to preserve the time-based structure of the data. This ensures the approach mimics real-world forecasting and avoids data leakage.

Data Preprocessing

1. **Feature Engineering**:
    - **Lagged Features**: Added lag_1, lag_3, and lag_6 to capture dependencies over time.

- **Rolling Statistics**: Smoothed short-term variations with rolling means (rolling_mean_3, rolling_mean_6) and standard deviations (rolling_std_3, rolling_std_6).
- **Cyclical Encoding**: Represented hour and day_of_week as sine and cosine values to account for patterns that repeat over time.

2. **Scaling**:
    - Standardized numerical features using StandardScaler to ensure all features had consistent scales.

Models and Hyperparameter Tuning
Three models were trained and optimized using RandomizedSearchCV:

1. **Ridge Regression**:
    - Tuned alpha (regularization strength) between 0.1 and 300.
2. **Random Forest Regressor**:
    - Adjusted n_estimators (50, 100, 200) and max_depth (10, 20, 50, None).
3. **XGBoost Regressor**:
    - Tuned learning_rate (0.01, 0.05, 0.1), max_depth (3, 6, 9), and n_estimators (50, 100, 200).

Evaluation Metric
The main metric was Root Mean Squared Error (RMSE). RMSE penalizes larger errors more, making it suitable for energy forecasting where large deviations can have significant consequences.

Cross-Validation
Time Series Cross-Validation (TimeSeriesSplit) with three splits was used during hyperparameter tuning. This method ensures that the validation data always comes after the training data in the timeline, preserving temporal order.

Ethical Considerations

- A chronological split ensured that no future data influenced past predictions, reducing the risk of overfitting.
- Cross-validation and regularization techniques were applied to build robust and reliable models.

4. Results

Model Performance
The models were assessed using Root Mean Squared Error (RMSE) and R^2 on the test set. The results are summarized below:

Model	Validation RMSE	Test RMSE	Test R^2
Baseline (Mean)	—	90.4786	—
Ridge Regression	43.1295	41.2233	0.7924
Random Forest	38.1613	40.0705	0.8039
XGBoost	38.0525	40.9152	0.7955

The baseline model, which predicts the mean energy consumption for all instances, had an RMSE of **90.4786 Wh**. All three machine learning models significantly outperformed the baseline, highlighting the value of advanced modeling and feature engineering.

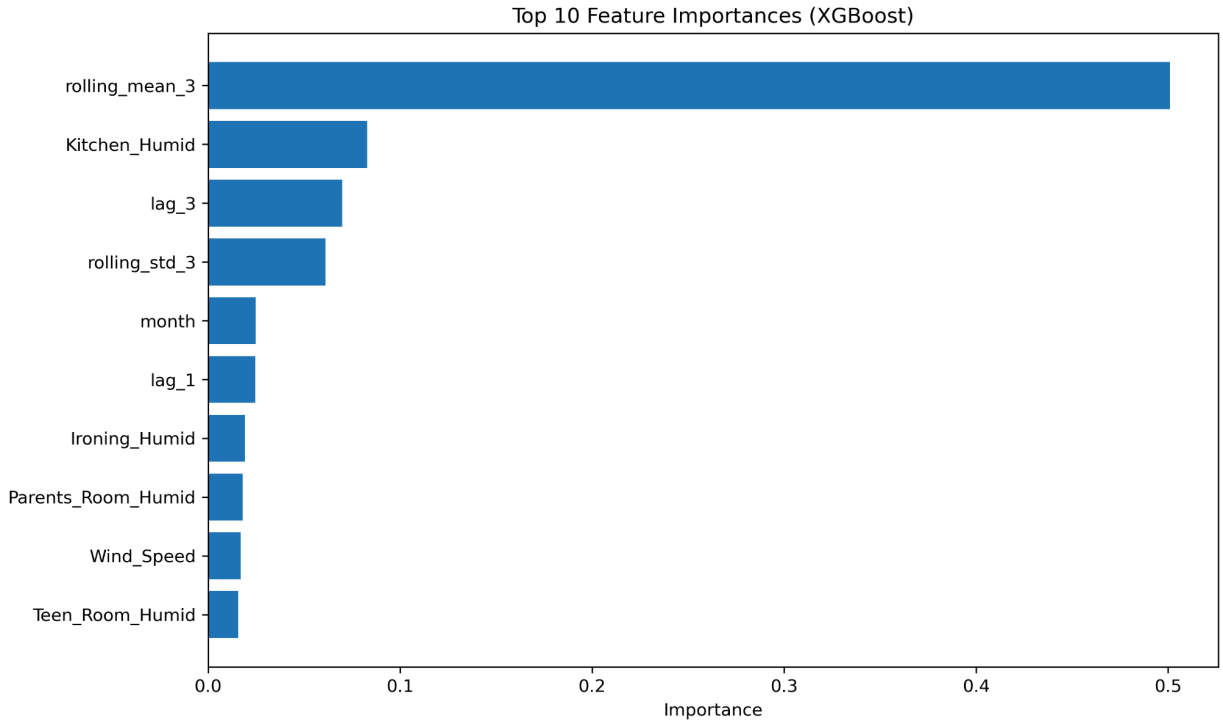
Best Performing Model

The **Random Forest Regressor** emerged as the best performer with the lowest test RMSE of **40.0705 Wh** and the highest R^2 of **0.8039**. This indicates its ability to capture the dataset's non-linear patterns and temporal dependencies.

The top 10 important features identified by the Random Forest model are displayed in Figure 1. Key features include:

- **Outdoor Temperature**: Influences energy use by affecting heating and cooling demand.
- **Lagged Features (Lag 1, Lag 3)**: Reflects the dependency of current energy usage on recent trends.

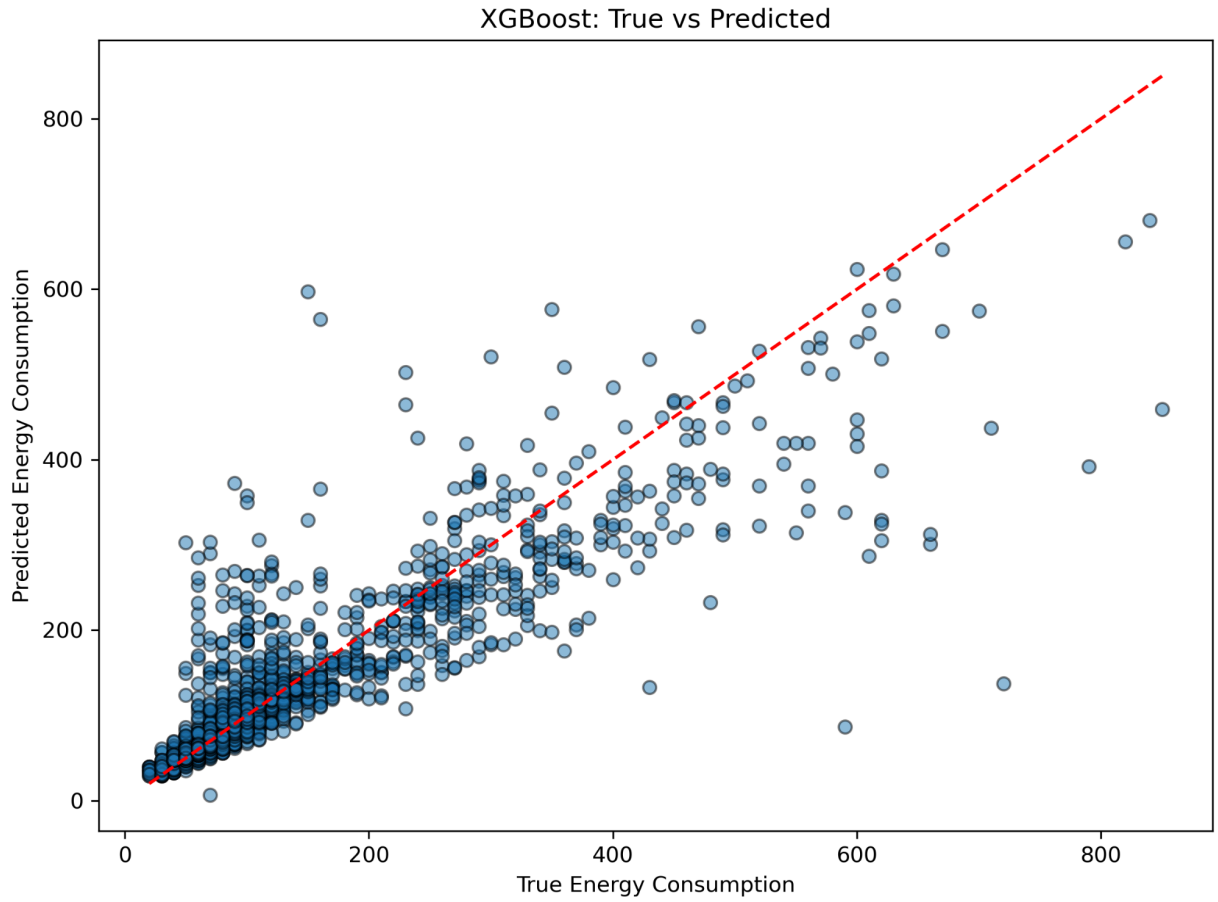
- **Rolling Means**: Smoother features that reveal broader patterns in the data.



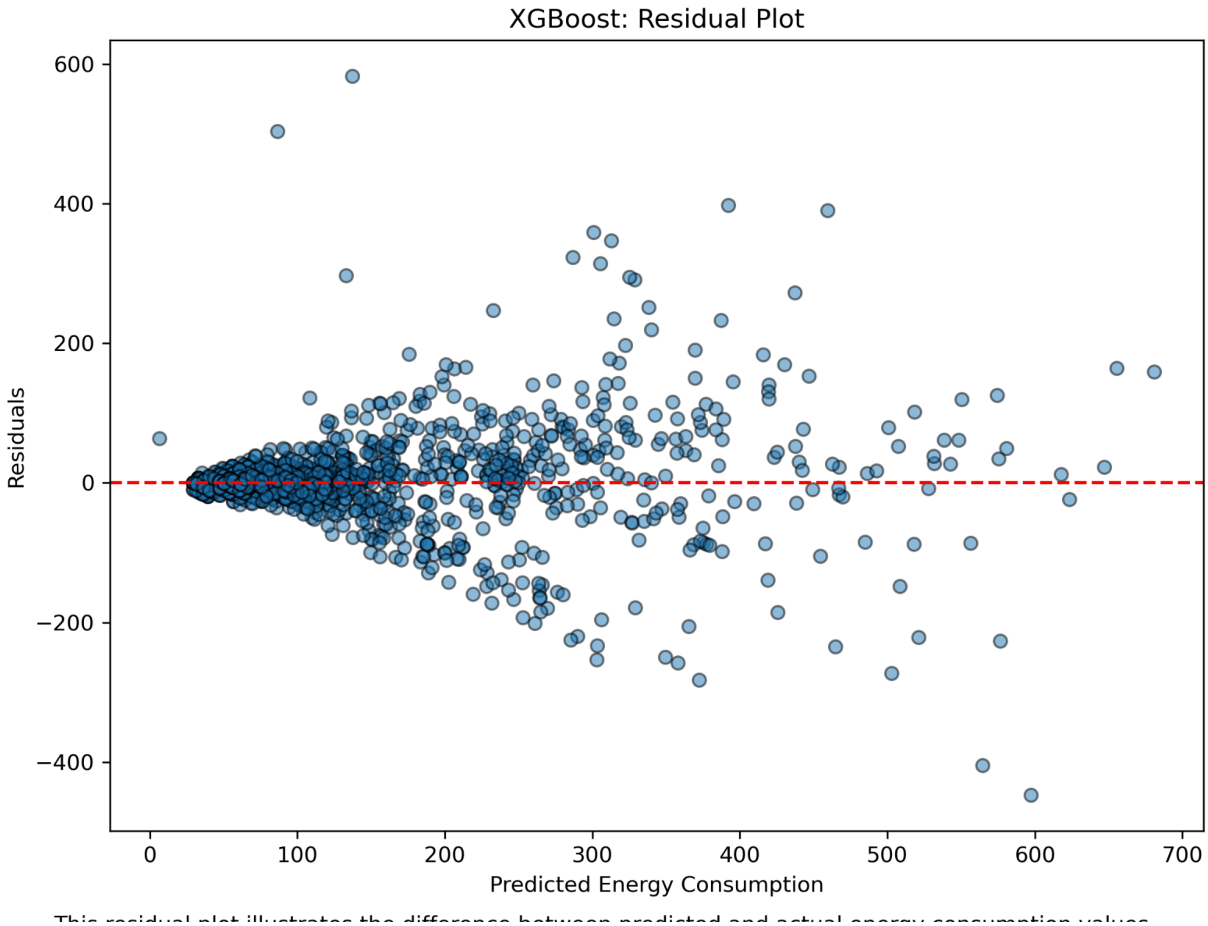
Model Interpretation

The XGBoost model also performed well, with a test RMSE of **40.9152 Wh** and R^2 of **0.7955**. Figure 2 compares the true and predicted energy consumption values for XGBoost, showing

good alignment, which demonstrates the model's ability to generalize effectively to unseen data.



Residual analysis (Figure 3) indicates minimal bias in XGBoost predictions, with errors centered around zero. A slight over-prediction is noticeable for lower energy usage values.



This residual plot illustrates the difference between predicted and actual energy consumption values

Key Observations

- Tree-based models, such as Random Forest and XGBoost, outperformed Ridge Regression, underscoring the importance of non-linear modeling for energy forecasting.
- Feature engineering, particularly lagged and rolling statistical features, played a key role in enhancing model performance.
- The alignment of predictions with actual energy consumption trends validates the modeling approach and highlights its effectiveness.

Outlook

Limitations

While the models demonstrated strong predictive performance, several limitations were identified:

- **Single Household Data**: The dataset is limited to one household, which may restrict the applicability of the models to other households or regions with different energy usage patterns.
- **Reliance on Lagged and Rolling Features**: The models depend heavily on these features, which may not fully capture long-term dependencies or seasonal variations in energy usage.
- **Feature Collinearity**: Analysis of feature importance revealed potential collinearity among environmental variables, which could obscure the true impact of individual predictors on energy consumption.

Improvements

To address these limitations and enhance model performance:

- **Expand Dataset**: Collect data from multiple households across diverse regions to improve the generalizability of the models.
- **Add Features**: Incorporate additional variables, such as electricity prices and broader weather metrics (e.g., wind speed, solar radiation), to better capture factors influencing energy usage.
- **Advanced Temporal Modeling**: Explore more sophisticated models, such as LSTM or Transformer-based architectures, to capture longer-term dependencies and seasonal patterns.

Future Work

Future research can focus on several areas to refine and expand this study:

- **Interpretability**: Apply interpretability frameworks like SHAP or LIME to provide deeper insights into model predictions and the influence of specific features.
- **Real-Time Applications**: Develop a real-time energy monitoring and prediction system based on the trained models to assist households in optimizing energy consumption. This practical application could enhance the utility of the research.
- **Out-of-Sample Testing**: Evaluate model performance on unseen data, including different time periods or households, to assess robustness and further validate generalizability.

References

1. Candanedo, L., Feldheim, V., & Deramaix, D. (2017). *Appliances Energy Prediction Dataset.* UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction
2. Tableau. (n.d.). *What is Time Series Analysis?* Retrieved from https://www.tableau.com/analytics/what-is-time-series-analysis
3. Dexma. (n.d.). *Forecasting Energy Consumption Using Machine Learning and AI.* Retrieved from https://www.dexma.com/blog-en/forecasting-energy-consumption-using-machine-learning-and-ai/

4. Pokharel, S., & Ghimire, S. (n.d.). *Data-driven ML Models for Accurate Prediction of Energy Consumption in Buildings.*Semantic Scholar. Retrieved from https://www.semanticscholar.org/paper/84f4e487a4210ceb7daf88bbc4aa3a96d6ac861a