**Smart water monitoring systems**

**Overview**

In this project, the goal was to predict daily water consumption for individual households. Worked with a dataset that included timestamps, household characteristics, weather data, and water usage information. Our main aim was to build a robust model that could accurately forecast water usage, ultimately helping to support water conservation efforts.

**Data-Preprocessing**

I began by loading the data into Google Colab environment and converting the timestamp information into a proper date format. From the timestamp, I have extracted valuable details like the month, day, and weekday, which can reveal seasonal and daily consumption patterns. And ensured that all numeric fields were correctly formatted and that our categorical fields were consistently treated as text. This step was crucial for ensuring that our subsequent analysis and modeling would run smoothly.

**Feature-Engineering**

To improve the model’s predictive power, I have created additional features that I thought would be useful:

* **Interaction Features:** Multiplied certain numeric columns (like the number of residents and temperature) to capture interactions between variables. This helped the model understand how combinations of features might impact water consumption.
* **Lag Features:** Recognizing that past water usage can be indicative of future consumption, I have introduced a feature that represented the previous period’s water consumption. This allowed the model to take into account the recent trends in usage.
* **Log Transformation:** Some variables, such as water price, had skewed distributions. So I have applied a log transformation to these features to help normalize the data and reduce the impact of extreme values.

**Modeling**

We built our solution using a machine learning pipeline, which helped streamline the entire process—from preprocessing to model training. The pipeline first handled missing values and scaled the numeric data, while categorical data was processed using one-hot encoding. We then used LightGBM as our predictive model because of its efficiency and performance with structured data. We evaluated the model using a holdout validation set, which gave us confidence in the model’s ability to generalize to unseen data.

**Tools & Environment**

The project was developed on Google Colab, taking advantage of its powerful computing resources and collaborative features. The main tools and libraries included:

* **pandas and NumPy** for data handling and manipulation.
* **scikit-learn** for creating pipelines, preprocessing data, and evaluating our model.
* **LightGBM** as a chosen machine learning model.
* **Matplotlib** for generating visualizations that provided insights into the data and model behavior.

**Visualizations and Insights**

Throughout the project, I have created a series of visualizations to better understand the data:

* Plotted histograms and boxplots to inspect the distribution of key features such as the number of residents.
* Bar charts were used to analyze the counts of different apartment types.
* Scatter plots helped us explore relationships between temperature and water consumption.
* A correlation heatmap provided a visual overview of how numeric features related to each other. These visualizations were essential for identifying outliers, checking data quality, and ensuring that our feature engineering steps were on the right track.

**Final Submission**

The final deliverable includes:

* A submission file (submission.csv) containing the predicted water consumption values alongside their corresponding timestamps.
* The complete Notebook that documents the entire process.
* This documentation file, which explains the approach, feature engineering strategies, tools used, and overall workflow.

**Conclusion**

By systematically preprocessing the data, engineering meaningful features, and carefully building a machine learning pipeline, we developed a model that can effectively predict water consumption. This approach not only improves prediction accuracy but also provides valuable insights into the factors influencing water usage.