**Documentation and Analysis of the Enhanced Machine Learning Workflow**

This document provides an in-depth explanation of the steps, rationale, and analysis of the machine learning workflow implemented for the given dataset. Each step is discussed with its purpose and how it contributes to achieving a more accurate predictive model.

**1. Data Cleaning**

**Objective:**

To ensure the dataset is consistent and devoid of missing or erroneous values that could hinder the model's performance.

**Methods:**

* **Handling Missing Values**:
  + **Numerical Features**: Missing values were filled with the median to maintain stability against outliers.
  + **Categorical Features**: Missing values were replaced with "unknown" to retain the feature while avoiding data loss.
* **Outlier Removal**:
  + Outliers in price\_doc were filtered out to focus the model on the most relevant price ranges, preventing skewed predictions.

**Impact:**

* Improved data quality ensured the model trained on a dataset that accurately represented the problem, reducing noise from inconsistencies.

**2. Feature Engineering**

**Objective:**

To derive meaningful features from the raw dataset, capturing complex relationships and trends for improved model accuracy.

**Methods:**

**A. Ratios and Interaction Features:**

* **full\_sq\_ratio\_life\_sq**: Captures the proportionality between full area and living area, which may indicate layout efficiency or luxury.
* **green\_zone\_ratio & indust\_zone\_ratio**: Represent environmental conditions in the region, correlating with property desirability.

**B. Age-Based Features:**

* **dependency\_ratio**: Measures the ratio of non-working (young and elderly) to working populations, indicating economic support structures.
* **elderly\_ratio**: Highlights the proportion of elderly, potentially affecting demand for certain property types.

**C. Accessibility Features:**

* **transport\_accessibility**: Averages distances to metro and railroad stations, emphasizing connectivity.
* **healthcare\_density**: Number of healthcare centers relative to population, influencing location attractiveness.

**D. Density-Based Features:**

* **population\_density**: Highlights the concentration of people per unit area, correlating with urbanization levels.

**E. Categorical Encoding:**

* One-hot encoding transformed categorical variables into numeric format, enabling their use in the model.

**Impact:**

* These engineered features captured nuanced aspects of the data that the raw variables could not express, enabling the model to identify trends and patterns more effectively.

**3. Data Transformation**

**Objective:**

To prepare the dataset for machine learning models by handling missing values, scaling, and dimensionality reduction.

**Methods:**

**A. Imputation:**

* SimpleImputer filled missing values with the median, ensuring no feature was discarded due to missing data.

**B. Scaling:**

* StandardScaler normalized numerical features to have a mean of 0 and a standard deviation of 1, improving model convergence and performance.

**C. Dimensionality Reduction:**

* **UMAP (Uniform Manifold Approximation and Projection)**:
  + Reduced high-dimensional data to 20 components while preserving global and local structure.
  + Helped in mitigating the curse of dimensionality and improving model training speed and generalization.

**Impact:**

* These transformations ensured that the dataset was clean, normalized, and compact, allowing models to learn efficiently without overfitting or being influenced by noise.

**4. Model Selection and Hyperparameter Optimization**

**Objective:**

To select and fine-tune machine learning models that best predict property prices.

**Methods:**

**A. Model Choice:**

1. **Random Forest Regressor**:
   * Chosen for its robustness and ability to capture non-linear relationships.
2. **Extra Trees Regressor**:
   * Selected for its efficiency and strength in handling high-dimensional datasets.

**B. Hyperparameter Optimization:**

* **Optuna**:
  + An automated optimization framework was used to explore the hyperparameter space efficiently.
  + Key hyperparameters like the number of estimators, max depth, and feature splits were tuned.

**C. Cross-Validation:**

* Ensured models were tested on diverse subsets of data, providing a reliable measure of performance.

**Impact:**

* Fine-tuning hyperparameters and employing robust models enhanced predictive accuracy while avoiding overfitting.

**5. Evaluation and Comparison**

**Objective:**

To assess the model's predictive accuracy and select the best-performing one.

**Methods:**

* **Root Mean Squared Error (RMSE)**:
  + Used to evaluate model performance on validation data.
* **Model Comparison**:
  + Random Forest and Extra Trees models were compared based on RMSE, with the better-performing model chosen for final predictions.

**Results:**

* Both models performed well, but one consistently outperformed the other based on RMSE, ensuring the best model was used for final predictions.

**6. Predictions and Submission**

**Objective:**

To generate predictions on the test set and save them for submission.

**Methods:**

* The test data was preprocessed similarly to the training data, ensuring consistency.
* The best model generated predictions, which were saved in the required format.

**Impact:**

* The systematic approach ensured accurate predictions, contributing to higher rankings in evaluations or competitions.

**Analysis of Results and Improvements**

1. **Enhanced Feature Representation**:
   * Engineering ratios, interactions, and temporal trends allowed the model to capture relationships not evident in raw data.
2. **Handling Missing Data**:
   * Imputation prevented data loss while maintaining feature integrity.
3. **Dimensionality Reduction**:
   * UMAP reduced noise and computational complexity, enabling models to focus on the most relevant aspects of the data.
4. **Hyperparameter Tuning**:
   * Optuna systematically explored parameter combinations, improving model performance.

**Overall Impact:**

The combination of thoughtful data engineering, advanced preprocessing, and robust model tuning led to significantly improved predictive accuracy. This approach ensures a scalable and generalizable solution for similar datasets in real-world applications.