Report: Optimized Predictive and Inferential Analysis of Swiss Apartment Model

Introduction

This analysis focuses on exploring and modelling the geometries and simulations datasets related to Swiss Apartment Models. The objective is to extract meaningful insights, perform predictive modelling, and apply clustering for segmenting apartments or units. The analysis is designed to handle large datasets efficiently without explicitly merging the geometries and simulations data. By processing them separately, we maintain the integrity of their structure while exploring intra-dataset relationships.

Steps in Analysis

1. Data Optimization

Memory optimization is critical when handling large datasets. A custom optimise_memory function was employed to downcast numerical columns, reducing memory consumption. This step was applied to both geometries and simulations datasets:

- **Geometries**: Contains spatial and structural details like area_id, height, and geometry.
- **Simulations**: Includes results of various metrics like sunlight exposure (sun_*), noise levels (noise_*), and accessibility (connectivity_*).

2. Correlation Analysis

Correlation matrices were computed separately for both datasets to understand relationships among features.

Geometries Correlation

The correlation matrix highlighted weak relationships between:

- area_id and height: $r=0.015r = 0.015$

Simulations Correlation

Correlation analysis revealed significant relationships:

- sun_201803210800_mean and view_ground_mean: $r=0.684r = 0.684$ (moderate correlation)
- Minimal correlation between sunlight and connectivity_entrance_door_distance_mean ($r=0.034r = 0.034$).

3. Regression Analysis

A linear regression model was applied to predict area_id using simulation metrics:

- **Predictor Variables**:
    - sun_201803210800_mean
    - view_ground_mean

- connectivity_entrance_door_distance_mean

- **Results**:
  - The model explained almost none of the variance (R2=0.00R^2 = 0.00), suggesting weak predictive power of these simulation metrics on area_id.

4. Principal Component Analysis (PCA)

To reduce dimensionality and identify the most influential features:

- PCA was applied to simulation features (sun_201803210800_mean, view_ground_mean, and connectivity_entrance_door_distance_mean).
- **Results**:
  - First two principal components explained 89.5%89.5\% of the total variance:
    - **PC1**: 56.2%56.2\%
    - **PC2**: 33.3%33.3\%
- **Visualization**:
  - A scatterplot of the PCA components provided a clear separation of data points in reduced dimensions, simplifying downstream clustering.

5. Cluster Analysis

K-means clustering was performed on simulation data to group apartments/units with similar characteristics:

- **Preprocessing**:
  - Numerical features were standardized using StandardScaler.
  - Missing values were imputed with column means.
- **Clustering**:
  - Clusters were created on a sampled subset of 10,00010,000 data points using MiniBatchKMeans to improve efficiency.
  - **Optimal Number of Clusters**: Determined to be 5 using the silhouette score.
- **Results**:
  - Clusters were visualized in reduced feature space (PCA components), revealing distinct groupings.

Key Insights

1. Correlation Analysis

- view_ground_mean and sun_201803210800_mean are moderately correlated, suggesting a connection between sunlight exposure and visual openness.

- Accessibility features like connectivity_entrance_door_distance_mean showed weak correlations with other metrics, indicating independence from sunlight and views.

2. Regression Analysis

- Simulations metrics failed to predict area_id effectively, suggesting that other unobserved factors influence area_id.

3. PCA Analysis

- PCA successfully reduced dimensionality while preserving a significant proportion of variance, highlighting its value for feature simplification and visualization.

4. Clustering

- Apartments were grouped into 5 clusters based on simulation metrics, enabling segmentation for targeted analysis:
    - Clusters could represent units with similar sunlight exposure, view quality, or accessibility.

Methodological Contributions

This analysis showcases a novel approach to handling large datasets by:

1. **Independent Processing**: Avoids merging datasets, maintaining efficiency and structure integrity.
2. **Dimensionality Reduction**: PCA enables streamlined clustering and visualization.
3. **Memory Optimization**: Reduces computational burden, crucial for large-scale datasets.

Conclusions

- The analysis provides insights into the relationship between apartment geometries and simulation metrics like sunlight and view quality.
- While simulation features failed to predict spatial features like area_id, clustering revealed meaningful groupings that can be leveraged for further exploration or decision-making.
- Future work could involve incorporating domain-specific knowledge to refine feature selection and improve predictive power.

Recommendations

1. **Feature Engineering**:
    - Explore interactions or transformations (e.g., combining sunlight and view metrics) for better predictive modeling.
2. **Enhanced Clustering**:
    - Experiment with other clustering techniques (e.g., DBSCAN) for improved cluster separation.
3. **Integration with Domain Knowledge**:
    - Incorporate insights from urban planning or architecture to contextualize findings.