Name: Aisun Zarghami

Project Report: Machine Learning Workflow

Task 1

Objective:

The objective of this project was to create a machine learning model to house price and classify the data which have the parking or not.

Steps Taken:

Dataset Overview

The dataset comprises several attributes including:

- **Area**: Total area of the house (in square meters).
- **Room**: Number of rooms in the house.
- **Parking**: Whether parking is available (boolean).
- **Warehouse**: Presence of a warehouse (boolean).
- **Elevator**: Availability of an elevator (boolean).
- **Address**: The locality or area where the house is located.
- **Price**: The total price of the house.
- **Price(USD)**: The price of the house converted to USD.

1. **Data Preprocessing**
    - **Missing Value Treatment:**
        - Checked for missing values in the dataset. There are 3479 houses that we have information about and 23 of them have missing values.
        - Dropped rows with missing values or filled them with appropriate statistical measures (mean, median, or mode), depending on the context. In this case we preferred to drop those lines that had missing values because we cannot fill them with inappropriate values.

- **Feature Scaling:**
  - Based on the fact that we have numerical and categorical and boolean values, at first we make it clear each column contains what sort of values.
  - StandardScaler ()is used for scaling numerical features.
  - OneHotEncoder() is used for encoding categorical features into a suitable format for machine learning.
- **Train-Test Split:**
  - Split the dataset into training and testing subsets using an 80-20 split to ensure the model could be evaluated on unseen data.

2. **Exploratory Data Analysis (EDA)**

   - Visualized the distribution of features using histograms to identify skewness and outliers.
   - Analyzed feature correlations with a heatmap to understand relationships and possible feature redundancy.
   - Used box plots to detect and visualize outliers within the dataset.

3. **Model Building**

   - Selected several machine learning models to train, including:
     - Random Forest Classifier
     - Support Vector Machine (SVM)
     - Decision Tree Classifier
   - Trained the Random Forest model on the training dataset and performed initial evaluations.

4. **Model Evaluation**

   - Evaluated the models using performance metrics such as accuracy, F1-score, and confusion matrix.
   - Used cross-validation to ensure model stability and assess performance consistency.

5. **Hyperparameter Tuning**

   - Conducted hyperparameter tuning using `GridSearchCV` to optimize the Random Forest model's parameters, contributing to better model performance.

6. **Model Interpretation**

- Analyzed feature importances to interpret which features had the highest impact on model predictions.
- Discussed how the insights from feature importance could guide future business decisions.

Conclusion

Here we tried to create a predictive model for house prices based on various features. While initial results were promising, performance can be improved through hyperparameter tuning and potentially by exploring additional feature engineering or selection strategies. Studies like this shows that by using machine learning and statistical factors, predictions that can be used for real world data are created and by going through of these steps we can get the needed information for a large dataset.