3/7/2025

Assignment 1

Supervised Learning – Paris price prediction

By ML Mavericks

INTRODUCTION

This dataset provides a comprehensive overview of housing properties in Paris, containing 10,000 entries with 17 attributes. It includes key property details such as size in square meters, number of rooms, and additional features like yards, pools, basements, attics, and garages. The dataset also records whether the property is newly built, has storm protection, and includes guest rooms or storage spaces.

Each property is associated with a unique city code and a city part range, which may indicate different areas or districts within Paris. Other important attributes include the number of previous owners, the year the property was built, and the number of floors. The target variable in the dataset is the price of the property, expressed as a floating-point value.

This dataset can be useful for real estate analysis, pricing predictions, and market trend assessments in Paris. It provides structured data that can be leveraged for machine learning models to predict house prices based on various influencing factors. The presence of categorical and numerical variables makes it suitable for exploratory data analysis and regression modeling.

Problem Statement

The primary problem this dataset aims to address is predicting house prices in Paris based on various property characteristics. Accurate real estate price prediction is essential for buyers, sellers, and investors to make informed decisions. Property values are influenced by multiple factors, including size, number of rooms, amenities, location, and historical ownership, making it crucial to develop reliable models for price estimation.

Why Is This Problem Worth Solving?

1. Better Decision-Making – Buyers can determine fair property prices, while sellers can set competitive prices based on data-driven insights.
2. Investment Strategy – Real estate investors can analyze trends and make profitable investments.
3. Market Transparency – Predictive models can help reduce uncertainty and speculation in the housing market.
4. Urban Planning & Policy Making – Government agencies can use housing price data to plan infrastructure and housing policies.

Source and Type of Data

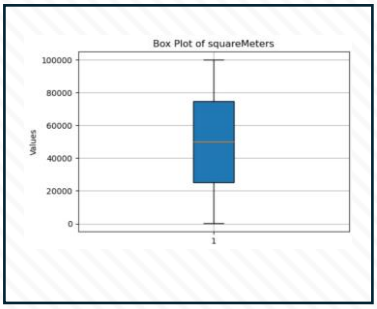
- Source: The dataset appears to be a structured dataset specifically prepared for housing price analysis in Paris.

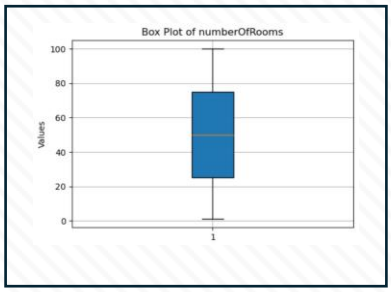
- Type of Data:
    - **Numerical Data:** Square meters, number of rooms, number of previous owners, number of floors, basement size, attic size, garage size.
    - **Categorical/Binary Data:** Presence of a yard, pool, storm protection, storage room, guest room, and whether the property is newly built.
    - **Target Variable:** Property price (continuous numerical value).

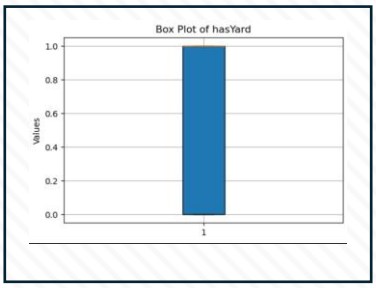
Methodology: Supervised Learning — Regression

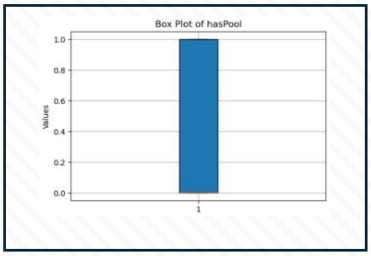
What exploratory analysis, data engineering, or data wrangling did you need to do?

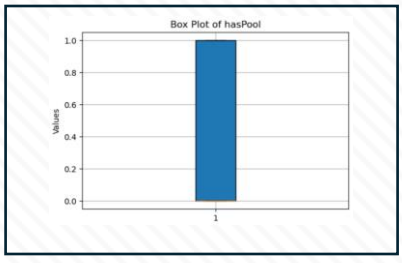
We did not perform any data wrangling, as the dataset we received was already cleaned. The data was used as is without any modifications. However, we checked for outliers using box plots and did not find any, so we proceeded with the same dataset without any additional processing.

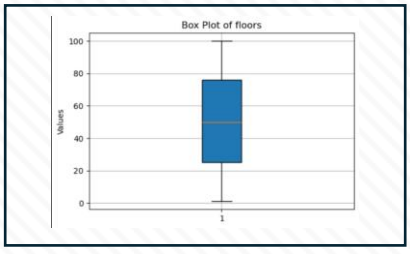






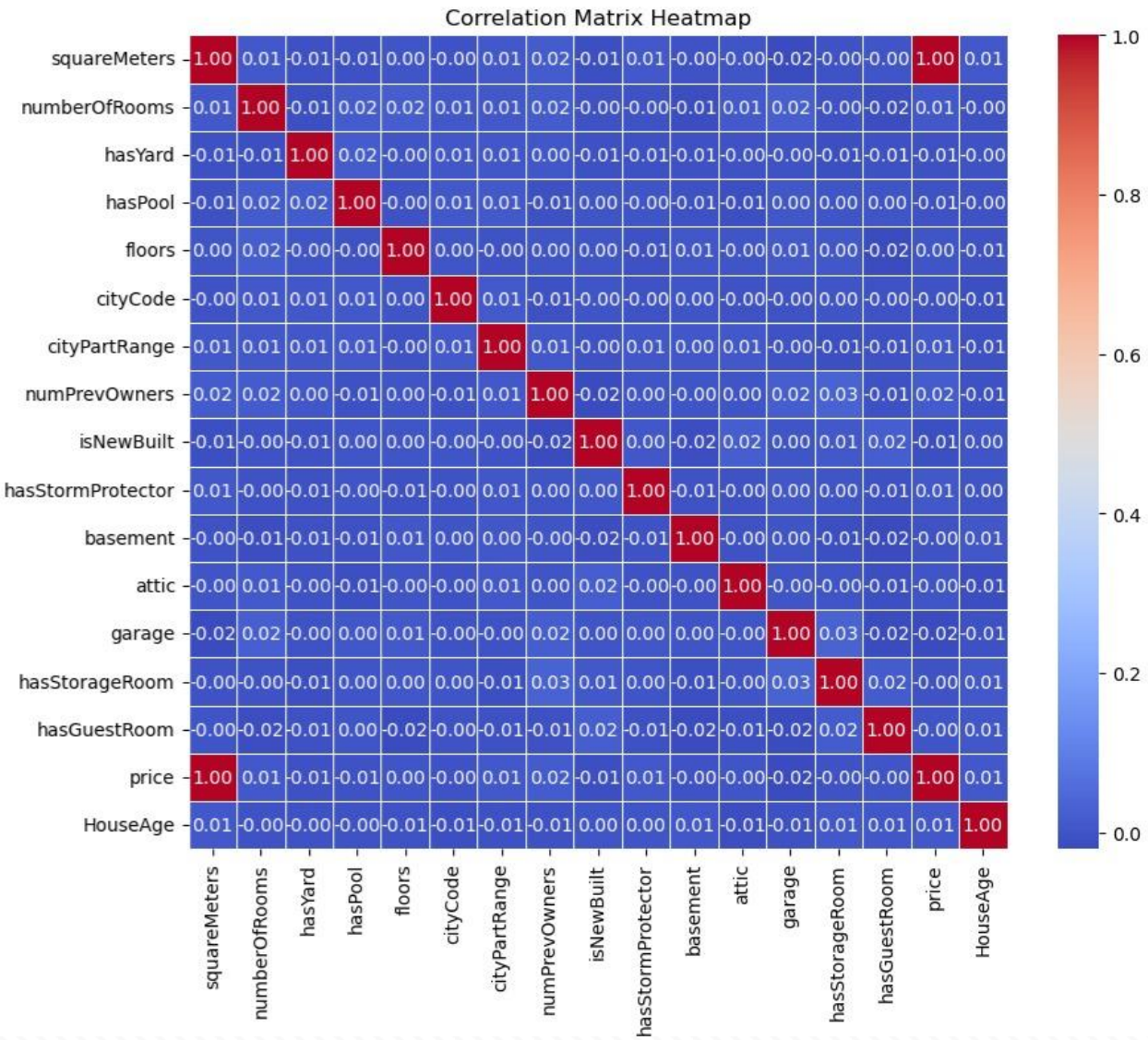


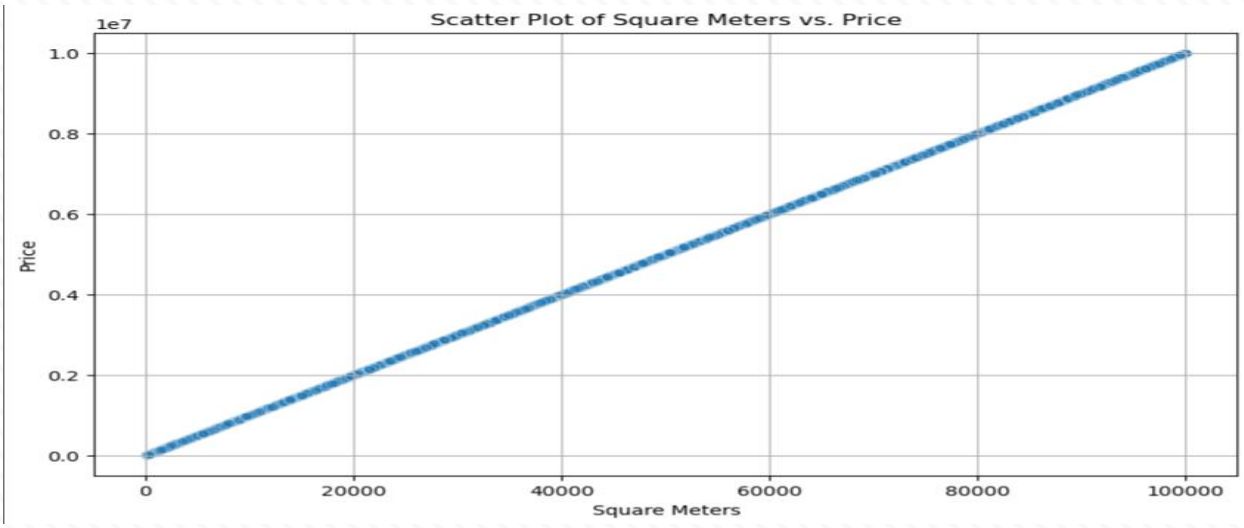




How did you prepare the data for modeling?

We used MLOS software for data modeling and Jupyter Notebook for generating box plots and correlation matrices. For the correlation matrix, we utilized a heatmap, as it is the most suitable visualization for illustrating relationships between variables. The lighter and more intense the color, the stronger the correlation with the target variable.

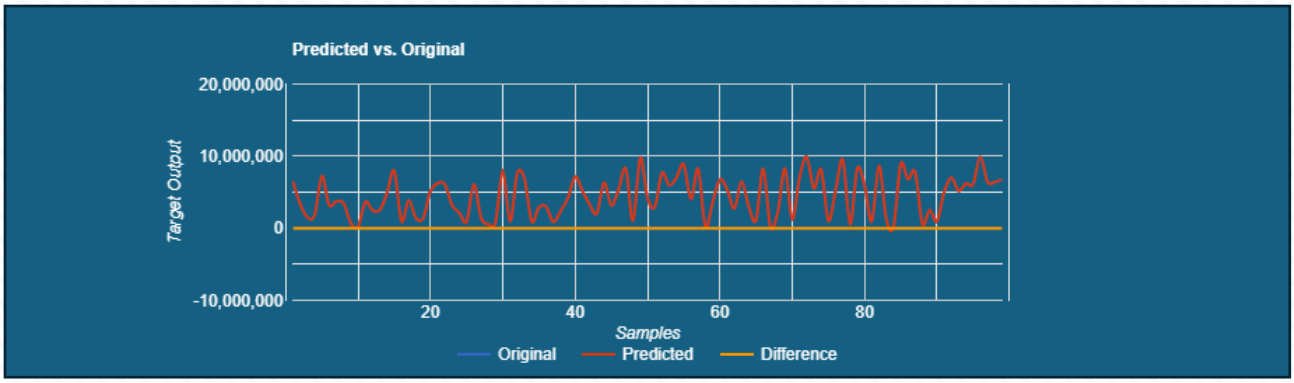


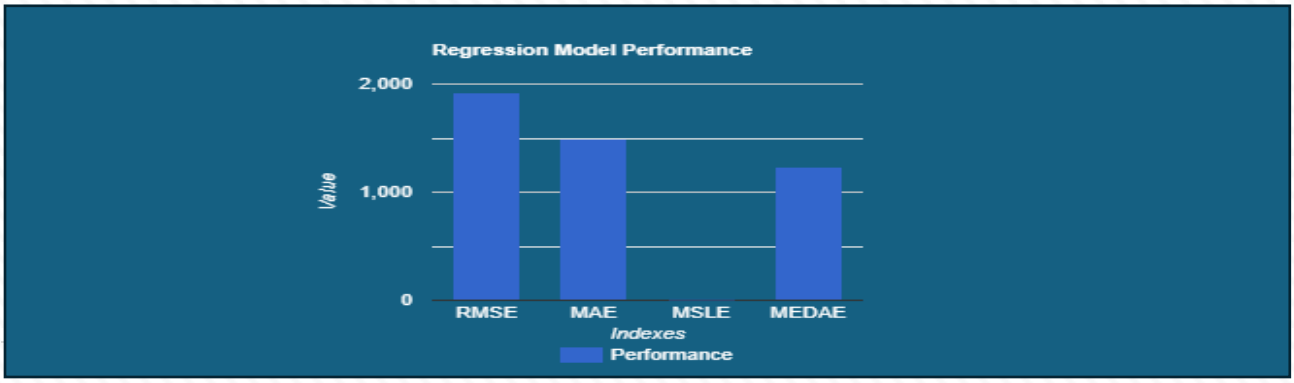


What was your modeling process? Specifically, which algorithms and parameters did you use and why?

We have tested four different regression algorithms on the dataset:

1. **Linear Regression** (Rank 1, Error: 1497.56)
2. **Random Forest Regressor** (Rank 2, Error: 3081.43)
3. **K-Neighbors Regressor** (Rank 3, Error: 20847.67)
4. **XGBoost Regressor** (Rank 4, Error: 2,506,004.72)





Result

Performance Evaluation:

To evaluate the performance of the models, we used **error metrics** such as **Mean Absolute Error (MAE)** or **Root Mean Squared Error (RMSE)**. These metrics quantify the difference between the predicted and actual property prices, providing a clear measure of model accuracy.

- **Linear Regression** performed the best, with the lowest error (1497.56), indicating it was the most accurate model for this dataset.
- **Random Forest Regressor** came in second, with a slightly higher error (3081.43), but still performed reasonably well.
- **K-Neighbors Regressor** had a significantly higher error (20847.67), suggesting it was less suitable for this dataset.
- **XGBoost Regressor** performed the worst, with an extremely high error (2,506,004.72), likely due to overfitting or improper hyperparameter tuning.

Conclusions

Improvements: In the future, I would focus on hyperparameter tuning, feature engineering, and testing advanced models like Gradient Boosting or Neural Networks to enhance accuracy. Cross-validation and error analysis could further refine the models.

Real-Life Use: This solution can help buyers, sellers, and investors make data-driven decisions, set competitive prices, and identify profitable real estate opportunities in Paris.

Client Value: The solution provides accurate price predictions, improving market transparency and supporting informed decision-making for real estate transactions and investments.

Learning: This project highlighted the importance of exploratory analysis, model selection, and the impact of feature relationships on predictive performance.