House Price Prediction: Business Report

1. Introduction

In the dynamic real estate market, accurately pricing a property is both an art and a science. Homebuyers, sellers, investors, and real estate professionals rely on a multitude of property features—from square footage and the number of bedrooms to the year built and neighborhood amenities—to estimate a house's market value. However, traditional methods based on comparable properties may overlook subtle yet impactful factors, leading to inaccurate pricing decisions.

This project aims to leverage machine learning techniques to build a robust predictive model for estimating house prices. Using historical housing market data, we explore and quantify the most significant variables affecting property values. Through a structured approach involving data cleaning, exploratory analysis, feature engineering, and regression modeling, this project provides actionable insights into the real estate market. Our goal is not only to create a high-performing model but also to enable better, data-driven decision-making for stakeholders.

2. Problem Statement

Homeowners and real estate professionals often face the challenge of determining a fair and realistic price for a property. Setting the price too high may deter potential buyers, while undervaluing a property leads to financial loss. With the availability of structured housing datasets, there is an opportunity to replace intuition-based pricing with evidence-backed valuation models.

The core objective of this project is to identify the key factors that influence house prices and to develop a predictive model capable of estimating property values with high accuracy. We aim to answer the following business-critical questions:

- What are the most influential features that determine the price of a house?

- Can we predict the price of a house based on its features with minimal error?

- How can data visualization and modeling assist stakeholders in understanding market dynamics?

By solving this problem, we aim to streamline property pricing strategies and reduce uncertainty in real estate transactions using a scalable, machine learning-driven approach.

3. Dataset Overview

The dataset contains property records with 23 features per house listing. These include physical attributes (like number of rooms, area), location (latitude, longitude), and quality indicators (condition, construction quality).

Context and Variables:

- **price** *(Target)*: Sale price of the house

- **room_bed**: Number of bedrooms

- **room_bath**: Number of bathrooms

- **living_measure**: Interior area of the house in square feet

- **lot_measure**: Total land area

- **ceil**: Number of floors

- **coast**, **sight**: Proximity to coast and scenic views

- **condition**, **quality**: Evaluations of structural condition and construction quality

- **yr_built**: Year of construction

- **zipcode**: Location-based identifier
- **total_area** *(engineered)*: Sum of living area and basement area

Variables like **sight**, **coast**, and **quality** bring domain-specific context into model development, affecting price based on aesthetic and practical value.

4. Data Cleaning & Preprocessing

Key Steps:

- **Dropped irrelevant columns**: cid, dayhours, long, furnished, yr_renovated
- **Converted variables to numeric** where applicable: ceil, coast, condition, yr_built, total_area
- **Handled missing values**:
    - Numeric: Imputed with mean
    - Categorical: Filled with mode

The cleaned dataset ensured no null values and retained meaningful, numeric-only features for modeling.

5. Exploratory Data Analysis (EDA)

EDA was used to:

- Understand the distribution and skewness of price
- Identify relationships using correlation heatmaps

- Detect outliers, particularly in price, using the **IQR method**

Insights:

- **Living area** and **quality** had strong positive correlation with price
- **Condition** and **sight** showed moderate influence
- Outliers in price were filtered to enhance model reliability

These insights guided feature selection and preprocessing strategy.

6. Feature Engineering & Selection

- Created total_area = living_measure + basement
- Dropped highly collinear or unimportant features based on EDA
- Standardized features for better model convergence (especially for Linear Regression)

7. Model Building

Chosen Models:

1. **Linear Regression**:
    - Acts as a baseline
    - Easy interpretability and low complexity
2. **Decision Tree Regressor**:
    - Captures non-linear relationships

    - Handles feature interaction well

3. **Random Forest Regressor**:
    - Robust ensemble method
    - Reduces overfitting and improves accuracy

These models were selected for their varying complexity and ability to learn both linear and non-linear patterns.

8. Model Evaluation

Each model was evaluated using:

- **R^2 Score**: Variance explained by model
- **RMSE**: Measures average prediction error
- **MAE**: Measures average absolute error

Model	R^2 Score	RMSE
Linear Regression	0.682	125,621
Decision Tree	0.755	102,346

Random Forest	0.837	84,568

Key Insight:

Random Forest showed the best performance with a balance of bias and variance.

9. Results & Comparison

- **Linear Regression** set a baseline
- **Decision Tree** captured more complex patterns
- **Random Forest** was most accurate and robust

The business implication is a reliable, high-performing prediction model that can estimate prices with a good degree of confidence.

10. Conclusion & Future Work

The project demonstrates that:

- Feature selection and preprocessing are key to performance
- Random Forest is effective for house price prediction

Future Enhancements:

- Apply hyperparameter tuning (GridSearchCV)
- Try advanced models (XGBoost, LightGBM)
- Incorporate geospatial analysis

- Deploy model with a Streamlit dashboard for business use
