**🏙️ PREDICTING THE APARTMENT RENT USING MACHINE LEARNING**

**Prepared by:** Tanvi Desai  
**Date:** October 2025

**1. Business Problem / Discovery Phase**

This project aims to **predict apartment rental prices** based on various property features such as location, size, number of rooms, and amenities. The dataset captures detailed information about rental listings, making it ideal for **regression-based Machine Learning modeling**.

**Objective**

To build a predictive model that accurately estimates the **rental price (target variable)** for apartments listed across different U.S. cities, based on their attributes and listing details.

**Dataset Overview**

The dataset was sourced from **Kaggle**, containing **cleaned and structured rental information** suitable for exploratory data analysis (EDA) and model development.

|  |  |
| --- | --- |
| **Feature Group** | **Description** |
| **Identifiers & Location** | Unique IDs, address, city name, state, latitude, longitude, and listing source. |
| **Property Details** | Apartment category, title, amenities, number of bathrooms, bedrooms, and total square feet. |
| **Pricing Information** | Price (target), displayed price, price type (USD), and any applicable fees. |
| **Additional Attributes** | Photo availability, pet policy, currency, and time of listing creation. |

The dataset ensures that **key numerical columns (price, square feet)** are not empty, supporting robust model training.

**2. Data Cleaning & Preprocessing**

**Handling Null Values**

| **Column** | **Null Count** | **Action Taken** | **Justification** |
| --- | --- | --- | --- |
| **id** | 0 | Dropped | Identifier only, no analytical use |
| **amenities** | 16,044 | Imputed with Mode | Categorical feature |
| **bathrooms** | 63 | Imputed with Median | Outliers present |
| **bedrooms** | 124 | Imputed with Median | Outliers present |
| **pets\_allowed** | 60,424 | Dropped | >50% missing values |
| **price** | 1 | Imputed with Median | Contains outliers |
| **price\_display** | 1 | Same as price | Removed redundancy |
| **address** | 91,549 | Dropped | >50% missing, low value |
| **cityname / state** | ~300 | Filled with Mode | Categorical features |
| **latitude / longitude** | 25 | Dropped | Minimal analytical relevance |

**Duplicate Values**

Found **84 duplicate rows** → removed to prevent bias in training.

**Outlier Detection**

Boxplots were used to visualize outliers for the following numerical columns:

* bathrooms
* bedrooms
* price
* latitude
* longitude

Median imputation ensured robustness against outlier distortion.

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

**Univariate Analysis – Categorical Features**

| **Feature** | **Observation** |
| --- | --- |
| **Category** | Majority listings are housing/rent/apartment. Very few belong to short-term or commercial rentals. |
| **Currency** | All transactions are in USD. |
| **Fee** | ~99% listings have “No Fee”. |
| **Has Photo** | ~56% “Yes”, ~34% “Thumbnail”, remaining “No”. |
| **Pets Allowed** | Primarily “Cats and Dogs” (~97%). |
| **Price Type** | Almost entirely “Monthly”. |
| **State** | 51 U.S. states represented. |
| **Source** | 25 listing platforms (e.g., RentLingo, RentCaf, Listanza, etc.). |

**Univariate Analysis – Numerical Features**

Kernel Density Estimates (KDE) were plotted for:

* bathrooms, bedrooms, price, and square\_feet  
  to understand distribution shape and detect skewness.

**4. Feature Engineering**

To enhance model performance and prevent computational overload (curse of dimensionality):

* Dropped **high-cardinality columns** (id, title, body, time, address, cityname, amenities).
* **Encoding:** Applied **Label Encoding** and **One-Hot Encoding** for categorical features.
* **Standardization:** Scaled numerical features using **StandardScaler** to normalize magnitude differences.

**Rationale:**  
Standardization prevents the model from over-prioritizing larger numerical values and ensures fair weight distribution during model training.

**5. Model Building**

Since the goal is to predict continuous price values, this is a **Regression Problem**.

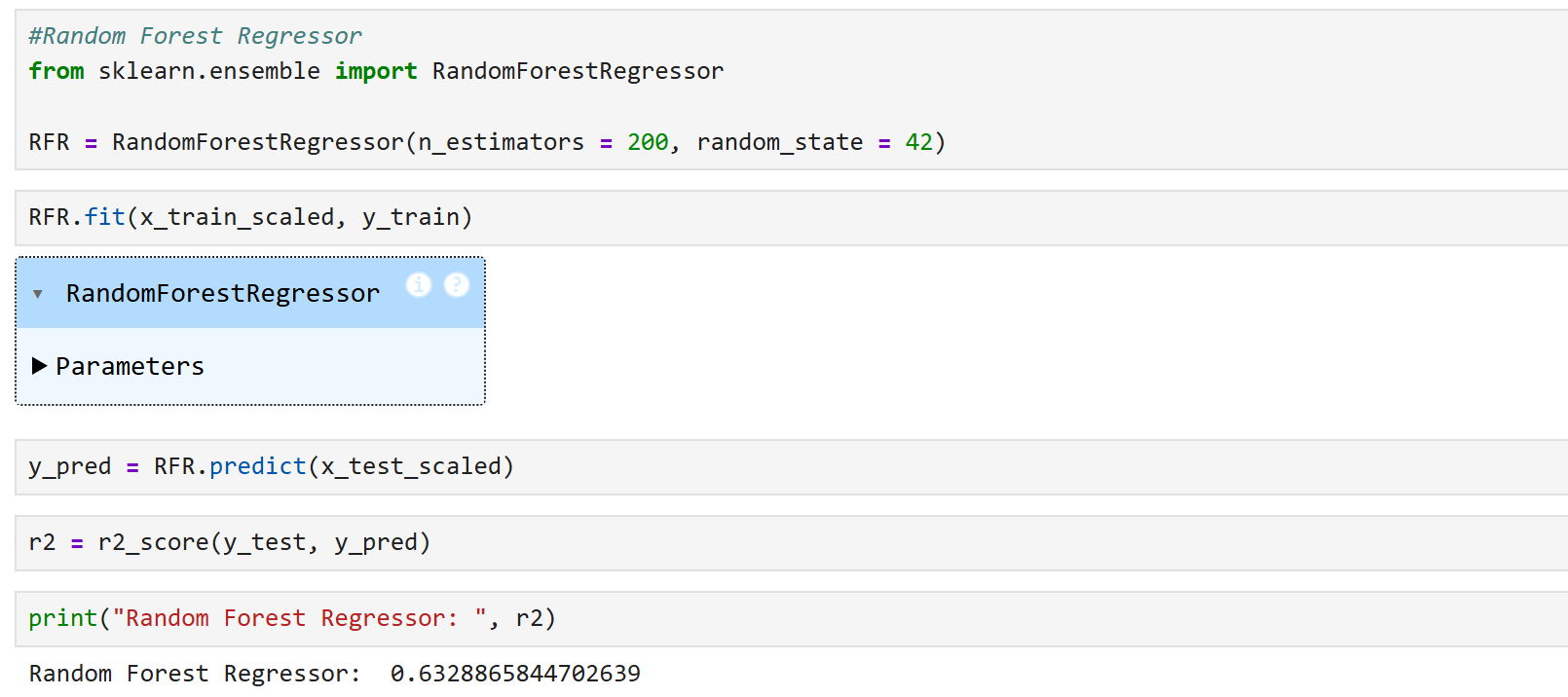
**Data Split**

**Train-Test Split:** 80:20 ratio

**Models Tested**

| **Model** | **Type** | **R² Score** |
| --- | --- | --- |
| Linear Regression | Baseline | 0.43 |
| Decision Tree Regressor | Non-linear | 0.59 |
| **Random Forest Regressor** | Ensemble | **0.633** |
| XGBoost Regressor | Gradient Boosting | 0.61 |

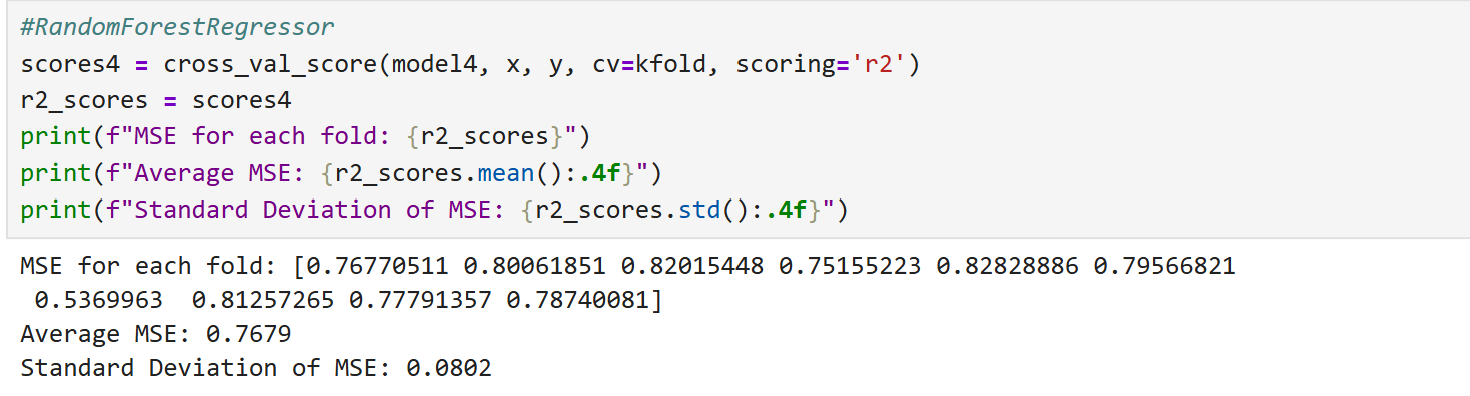
**Best Performing Model:**  
**Random Forest Regressor** with **R² = 0.633**



**Cross Validation**

Applied **K-Fold Cross Validation** to ensure model stability.

Random Forest consistently achieved the highest R² across folds, confirming robustness.



**6. Model Evaluation**

**R² Score:** 0.633

**RMSE (Root Mean Squared Error):** Moderate (indicating acceptable variance in predictions)

**Feature Importance:** Top features influencing price prediction:

* 1. square\_feet
  2. bedrooms
  3. bathrooms
  4. category
  5. state

These results indicate the model’s capability to generalize well while capturing the key drivers of apartment pricing.

**7. Key Insights**

| **Insight** | **Interpretation** |
| --- | --- |
| Apartment size directly impacts price | Larger square footage increases rental price proportionally. |
| Bedroom and bathroom count | More rooms → higher price, especially in urban states. |
| Amenities not strongly predictive | Limited variation within dataset. |
| Pet policies & photos | Minor impact; listings with photos may attract slightly higher rents. |

**8. Conclusion**

This project demonstrates how **Machine Learning regression techniques** can effectively predict apartment rental prices using property-level data.

Key takeaways:

Data preprocessing (imputation, outlier handling, encoding) critically affects model accuracy.

Ensemble models like **Random Forest** provide reliable, high-performing results for heterogeneous datasets.

The final model achieved an **R² of 0.633**, suggesting it explains ~63% of the price variance.

This workflow can be extended for **price optimization**, **market segmentation**, and **real estate valuation** use cases.

**9. Tools & Technologies**

| **Category** | **Tools / Libraries** |
| --- | --- |
| Data Preprocessing | Pandas, NumPy |
| Visualization | Matplotlib, Seaborn |
| Modeling | Scikit-learn (RandomForest, DecisionTree, Linear Regression, KNN, LASSO, Ridge, ElasticNet ) |
| Validation | Cross Validation (K-Fold) |
| IDE | Jupyter Notebook |