Predicting Airbnb Listing Prices with MLflow and AWS S3

Project Overview

This project predicts optimal nightly prices for Airbnb listings on StayWise, a global vacation rental platform. Accurate price prediction helps hosts set competitive rates and improves booking efficiency. The project uses:

AWS S3: Data storage
MLflow: Experiment tracking and model registry
Machine Learning Models: Regression models for price prediction

Objectives:
Retrieve Airbnb listings from AWS S3
Clean and preprocess the data
Develop and compare multiple regression models
Track experiments and register the best model in MLflow.

Dataset Description

The dataset contains information about Airbnb listings, including price, location, amenities, reviews, and host details. The dataset requires preprocessing to handle missing values, categorical variables, and outliers.

Data Preprocessing:

Handled missing values in 'name' and 'host_name'.
Converted categorical variables using encoding.
Created new features like 'amenities_count'.
Removed outliers using the IQR method.
Split the dataset into training and testing sets.

MLflow Experiment Tracking:

All models were tracked using MLflow, logging metrics, parameters, and artifacts. The best model (Random Forest) was registered in the MLflow Model Registry. Attach MLflow UI screenshots here.

Repository Structure

```
airbnb-price-prediction/
── notebook.py
── requirements.txt
── .gitignore
── README.pdf
```

Setup Instructions:

1. Clone the repository

```
git clone https://github.com/yourusername/airbnb-price-prediction.git
cd airbnb-price-prediction
```

2. Install dependencies

```
pip install -r requirements.txt
```

3. Run Jupyter notebooks

```
jupyter notebook
```

Workflow Diagram:

1. Retrieve Data from S3
2. Data Cleaning & Preprocessing
3. Feature Engineering & Encoding
4. Train Regression Models
5. MLflow Experiment Tracking
6. Compare Models & Register Best
7. Deploy/Use Model

Data Preprocessing Steps

- Handle missing values (name, host_name, etc.)
- Encode categorical variables (neighbourhood, room_type)
- Handle outliers in price
- Feature engineering: reviews_per_month, amenities_count, etc.
- Normalize and scale numeric features

Model Development & MLflow

Tested regression models:

1. Linear Regression
2. Random Forest Regressor
3. Gradient Boosting Regressor

Metrics tracked using MLflow:

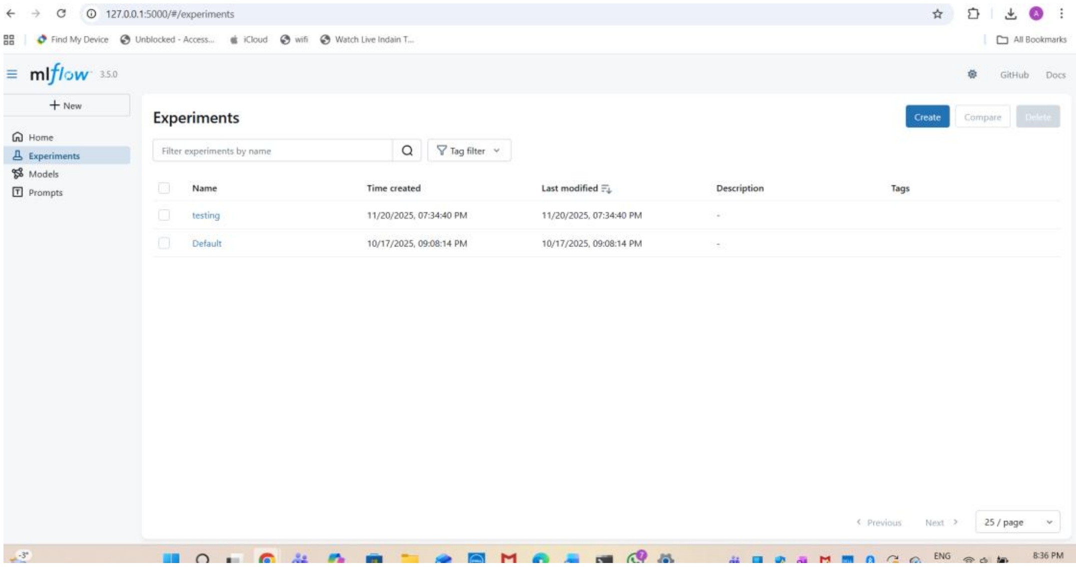
- RMSE
- R^2 Score

Model Performance

Model	RMSE	R^2
Linear Regression	45.56	0.54
Random Forest Regressor	43.88	0.57
Gradient Boosting	44.55	0.56

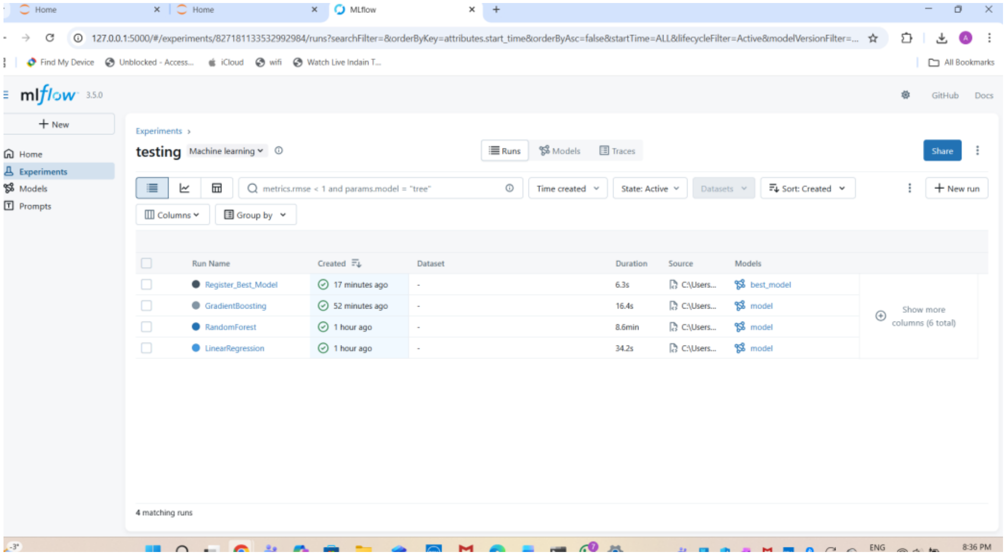
ML FLOW:

Experiment Tracking UI



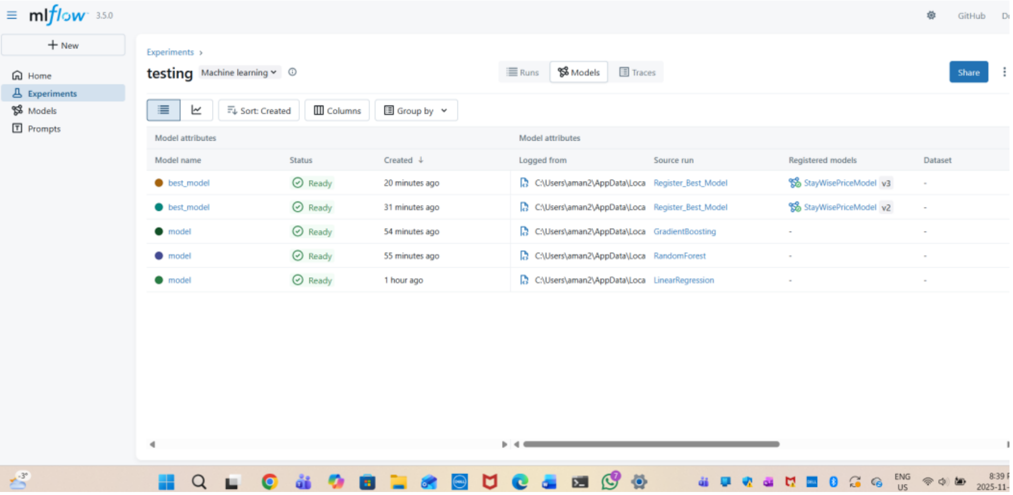
	Name	Time created	Last modified	Description	Tags
□	testing	11/20/2025, 07:34:40 PM	11/20/2025, 07:34:40 PM	-	
□	Default	10/17/2025, 09:08:14 PM	10/17/2025, 09:08:14 PM	-	

Metrics Comparison:



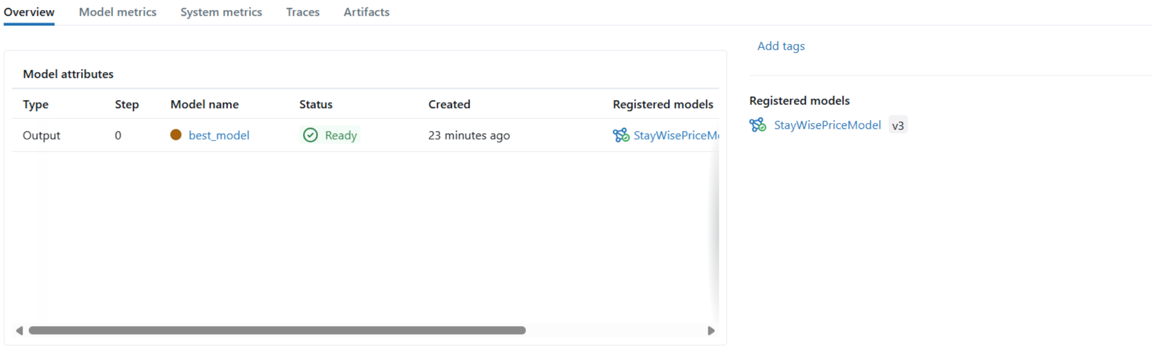
	Run Name	Created	Dataset	Duration	Source	Models
□	● Register_Best_Model	17 minutes ago	-	6.3s	C:\Users...	best_model
□	● GradientBoosting	52 minutes ago	-	16.4s	C:\Users...	model
□	● RandomForest	1 hour ago	-	8.6min	C:\Users...	model
□	● LinearRegression	1 hour ago	-	34.2s	C:\Users...	model

4 matching runs



testing Machine learning

Model attributes			Model attributes			
Model name	Status	Created ↓	Logged from	Source run	Registered models	Dataset
● best_model	✓ Ready	20 minutes ago	C:\Users\aman2\AppData\Loca	Register_Best_Model	StayWisePriceModel v3	-
● best_model	✓ Ready	31 minutes ago	C:\Users\aman2\AppData\Loca	Register_Best_Model	StayWisePriceModel v2	-
● model	✓ Ready	54 minutes ago	C:\Users\aman2\AppData\Loca	GradientBoosting	-	-
● model	✓ Ready	55 minutes ago	C:\Users\aman2\AppData\Loca	RandomForest	-	-
● model	✓ Ready	1 hour ago	C:\Users\aman2\AppData\Loca	LinearRegression	-	-



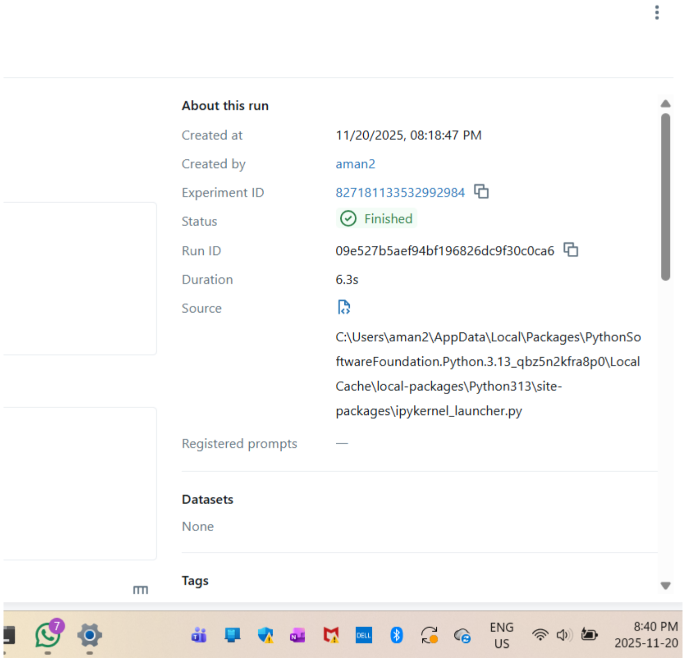
Model attributes

Type	Step	Model name	Status	Created	Registered models
Output	0	● best_model	✓ Ready	23 minutes ago	StayWisePriceM

Registered models

StayWisePriceModel v3

Model Registres:



About this run

Created at	11/20/2025, 08:18:47 PM
Created by	aman2
Experiment ID	827181133532992984
Status	✓ Finished
Run ID	09e527b5aef94bf196826dc9f30c0ca6
Duration	6.3s
Source	C:\Users\aman2\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\ipykernel_launcher.py
Registered prompts	—

Datasets

None

Tags

Dependencies:

pandas
numpy
scikit-learn
mlflow
boto3
matplotlib
seaborn

Notes

Ensure AWS credentials have read access to S3.
Large datasets and MLflow artifacts are excluded via .gitignore.
Follow notebook execution order for smooth workflow.

Key insights and observations:

Features like location, number of amenities, and reviews significantly influence price.
Random Forest captures complex relationships better than Linear Regression or Gradient Boosting in this dataset.
Outlier removal improved model performance slightly.
MLflow provides a convenient way to compare model runs and manage model versions.