

Predicting Airbnb Prices

A Machine Learning Approach for London Listings

-Mohamed Reda Aarsalane
-Mohamed Bilal Abdenouri
-Sunil Kumar Nallani

Project Scope

Objective: Build a machine learning model to accurately predict Airbnb nightly prices (USD) in London using publicly available data from the Inside Airbnb Open Data Initiative.

Dataset: Utilized ~60,000 cleaned listings from listings.csv (originally 90,000+ before cleaning).

Target Variable: price — nightly rate charged by hosts (converted from string to float and log-transformed for modeling).

Feature Sources: Numerical & categorical attributes (e.g., bedrooms, room type, location) Text-based NLP features (from listing descriptions and host info) Sentiment scores from guest reviews Amenities & availability information

Goal: Deliver an interpretable and robust pricing model that generalizes well to unseen listings using advanced regression models like CatBoost, XGBoost, SVR, and MLP.

Overview

Features Used

Our model incorporates diverse data types including tabular, text, spatial, and review information. We analyzed **numerical features** (like number of bedrooms, bathrooms), **categorical features** (property type, room type), and **NLP features** extracted from listing descriptions and reviews to capture the full spectrum of pricing factors. Some of the numerical features are not included in this slide.

Feature	Data Type
neighbourhood_cleansed	Categorical
property_type	Categorical
room_type	Categorical
amenities	Complex (List)

Feature	Data Type
average sentiment score	Numerical
sentiment score count	Numerical
NLP embeddings	Vector

Feature	Data Type
accommodates	int64
bedrooms	float64
beds	float64
calculated_host_listings_count	int64
has_availability	int64
host_acceptance_rate	float64
host_has_profile_pic	int64
host_identity_verified	int64
host_is_superhost	int64
host_listings_count	float64
instant_bookable	int64
latitude	float64
longitude	float64
minimum_nights	int64

Methodology

Our data preparation process involved creating structured datasets from raw listings, removing duplicates, and cleaning price data by removing symbols and converting to numeric values. We engineered specialized features to handle missing reviews, including *'is_unreviewed' flags and 'days_since_last_review'* metrics, with appropriate review score imputations to maintain data integrity.



Review data handling

We developed specialized features to address missing review information in the dataset. This included creating *'is_unreviewed' flags to identify listings without reviews and calculating 'days_since_last_review'* metrics to track review recency. We applied appropriate review score imputations to maintain data integrity for analysis.



Structured dataset creation

We transformed raw listings into organized, structured datasets that could be effectively analyzed. This process included removing duplicate entries to ensure data accuracy and prevent skewed analysis results.



Price data cleaning

We implemented a thorough cleaning process for price data by removing currency symbols and converting string values to numeric format. This standardization allowed for consistent price comparisons and mathematical operations across the dataset.

Feature Engineering :

Feature Engineering :

Date Features: Created `days_since_last_review` & `days_as_host`.

Binary Encoding: Converted 't'/'f' columns to 1/0.

Sentiment Features: Merged sentiment score & count from external file.

Top Amenities: One-hot encoded 50 most frequent amenities.

Categorical Encoding: Grouped rare values as "Other", then one-hot encoded.

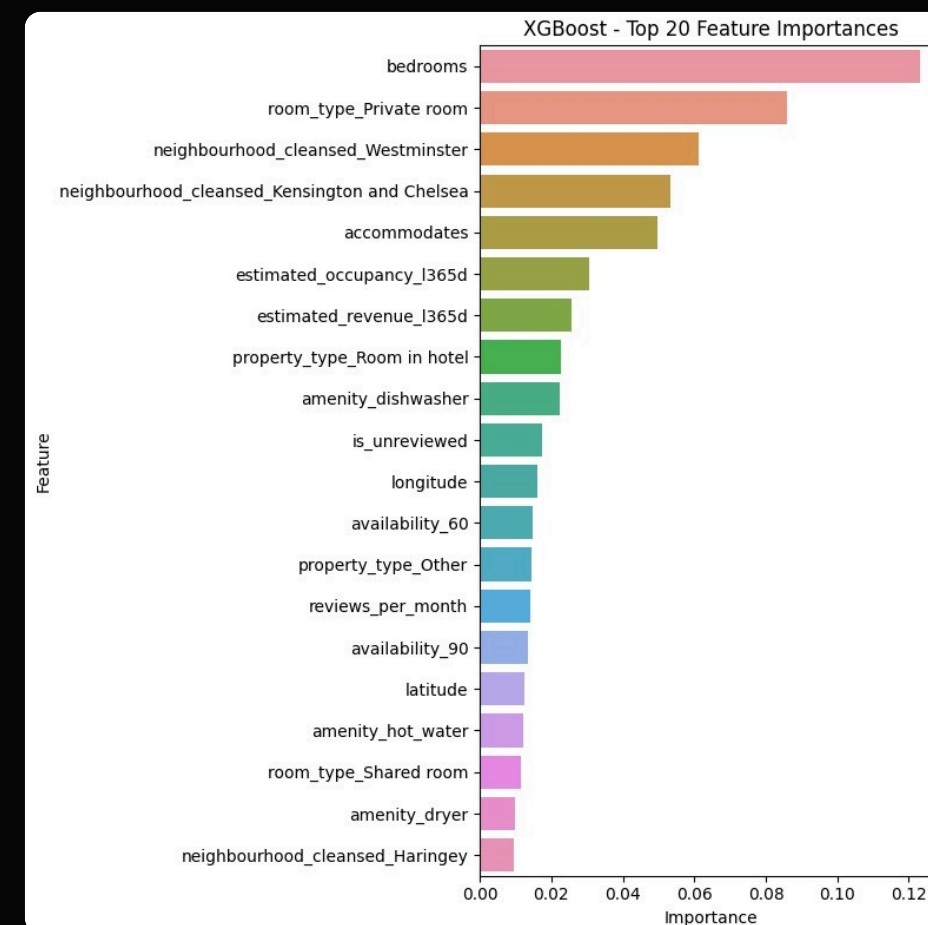
Engineered features from dates, binary flags, amenities, sentiment scores, and categorical columns.

NLP Embeddings: Used MiniLM to convert text fields into 384 numeric features.

Converted text fields :

`name`, `description`, `bathrooms_text`, `neighborhood_overview`, `host_about`

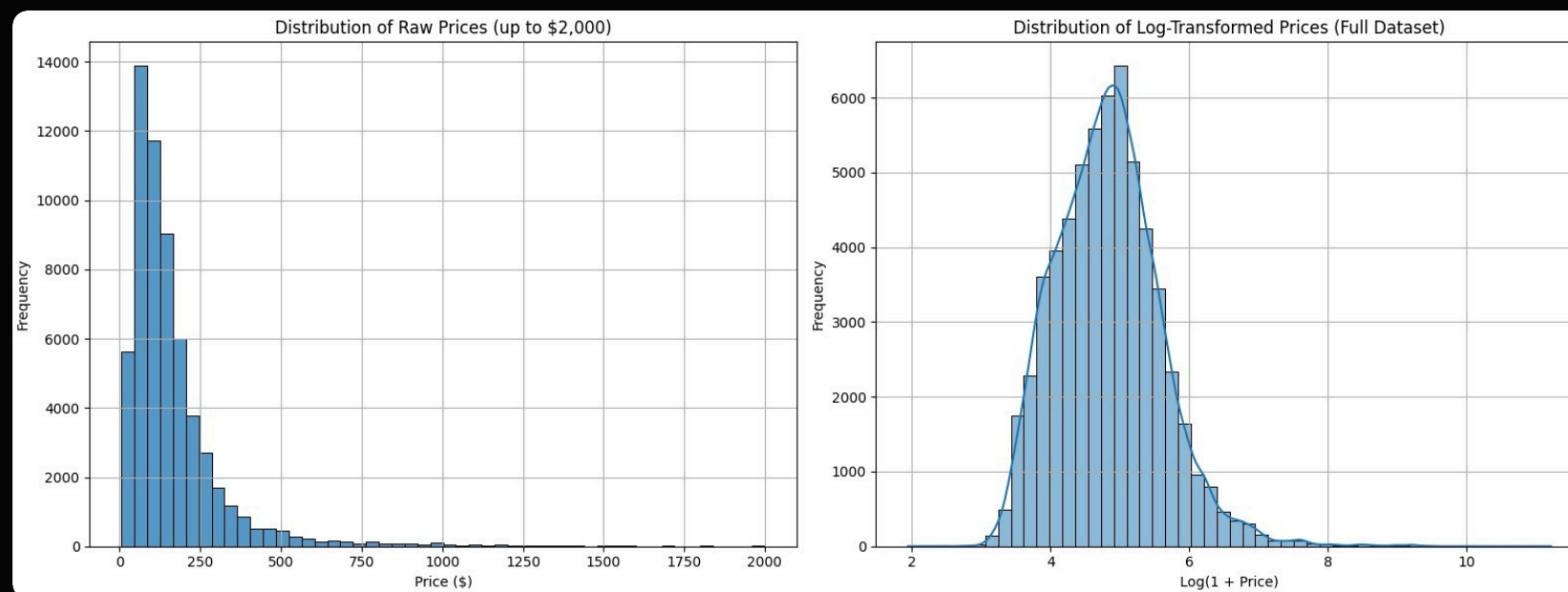
into 384-dimensional sentence embeddings using the all-MiniLM-L6-v2 model.



Price Distribution (Raw vs. Log-Transformed)

Left Plot: Raw price values are heavily right-skewed, with most listings priced under \$250 and a long tail of expensive outliers.

Right Plot: After applying $\log(1 + \text{price})$, the distribution becomes much more symmetric and bell-shaped, helping models better learn patterns.



Model Arsenal: Four Powerful Prediction Algorithms Deployed

CatBoost

Categorical handling

CatBoost efficiently processes categorical and text data without extensive preprocessing. Each model underwent rigorous evaluation using MAE, RMSE, and R^2 metrics to ensure we maintained high performance standards across all algorithms.

MLP

Deep learning capabilities

The Multi-layer Perceptron provides basic deep learning functionality, serving as a foundational neural network approach in our prediction arsenal. This model contributes to our comprehensive performance assessment framework.

SVR

Non-linear relationships

Support Vector Regression excels at capturing complex non-linear relationships within the data. Its ability to map inputs to a higher-dimensional feature space makes it particularly valuable for complex prediction tasks.

XGBoost

Tree analysis

Our implementation of XGBoost leverages gradient-boosted decision trees to provide powerful predictive capabilities. This algorithm sequentially builds trees that correct errors from previous models, enhancing overall prediction accuracy.

Performance Metrics

CatBoost emerged as our top-performing model, particularly after incorporating all text features through NLP techniques.

Accuracy

Final model achieved an impressive MAE of 17.14.

Reliability

Demonstrated R² of 0.8882 in predictions.

Capability

Exceptional generalization across London market.

Integration

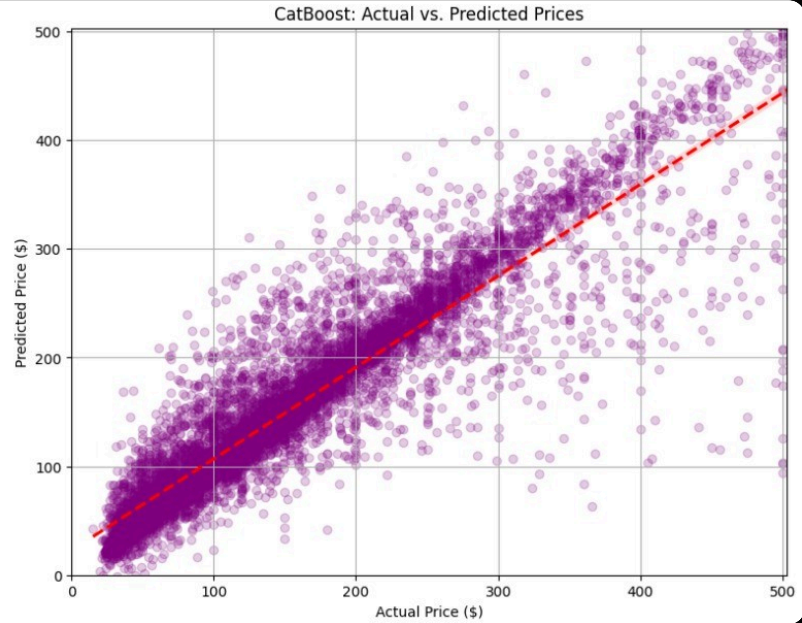
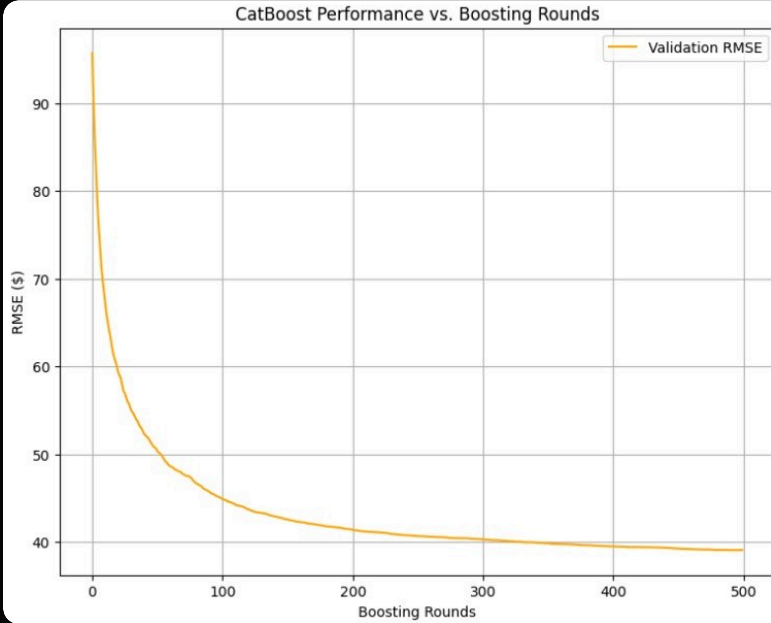
Successfully incorporated all text features through NLP.

Model	MSE	R ² Score	MAE
CatBoost	1489.73	0.8882	17.14
Linear Regression	4009.33	0.6146	42.15
Random Forest	2379.51	0.7713	27.09
MLP Regressor	2529.22	0.7569	26.48
SVM	4742.08	0.5442	40.78
XGBoost	1683.80	0.8382	22.68
LightGBM	1723.32	0.8344	23.02
ElasticNet	4540.56	0.5636	44.21
GradientBoosting	2763.76	0.7343	33.70


Findings


Key Insights


Our project revealed that thorough data preprocessing and feature engineering were critical success factors. Text data from listing descriptions significantly improved prediction accuracy, while CatBoost proved most robust among all tested models. Future enhancements could incorporate image data analysis and more granular location clustering to further refine price predictions.



 Data preprocessing critical

 CatBoost most robust

 Text data improved accuracy

 Future image analysis