PES
UNIVERSITY

Machine Learning Assignment

PROJECT REPORT

TEAM ID : 17

Airbnb Rental Price prediction

Name	SRN
Cherukuri Venkata Kartik	PES2UG23CS148
Gavini Prithvi Ananya	PES2UG23CS198
Fairly Sorathiya	PES2UG23CS189

Problem Statement

The project aims to predict the rental price of Airbnb listings in Melbourne based on various features such as location, property type, number of bedrooms, amenities, and reviews. Accurately predicting rental prices helps hosts optimize their listings and helps potential guests understand pricing trends, improving transparency and decision-making.

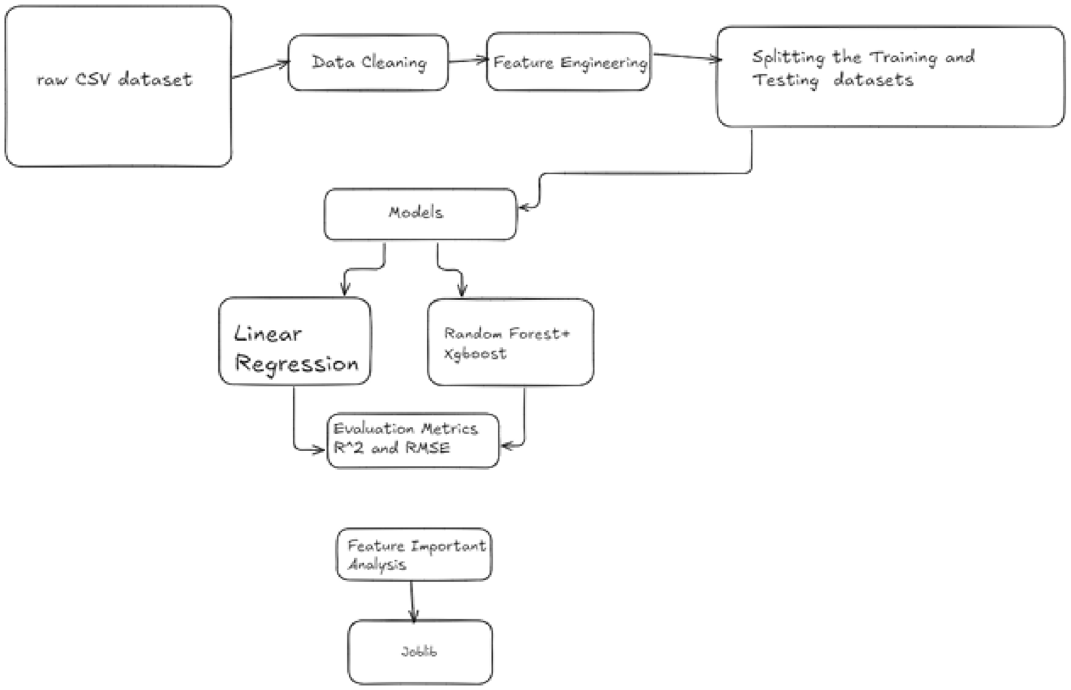
Objective / Aim

Develop a machine learning model that can accurately estimate the price of Airbnb listings in Melbourne.
Identify key features that most influence rental prices.
Evaluate model performance using standard regression metrics.

Dataset Details

- **Source:** http://insideairbnb.com/get-the-data/
- Find Melbourne and choose listings.csv
- **Size:** ~20,000 samples, 16–20 features
- **Key Features:**
- property_type – Type of property (Apartment, House, etc.)
- room_type – Entire home, private room, etc.
- bedrooms, bathrooms, beds – Property capacity
- amenities – List of amenities
- reviews_count, review_scores_rating – User reviews
- neighbourhood – Location information
- latitude, longitude – Geographical location
- **Target Variable:** price (in AUD)

Architecture Diagram


raw CSV dataset
Data Cleaning
Feature Engineering
Splitting the Training and Testing datasets
Models
Linear Regression
Random Forest+ Xgboost
Evaluation Metrics R^2 and RMSE
Feature Important Analysis
Joblib

Methodology

The project follows a standard machine learning workflow, with the key steps outlined below:

Data Preparation: Load the pre-cleaned and feature-engineered data.

Target Scaling: Check the target variable (price) for standardization (mean, std). If scaled, create and save a StandardScaler object to allow for inverse-transforming predictions back to original currency units.

Data Imputation: Handle any remaining missing numeric values in the features by applying median imputation.

Train/Test Split: Split the data into an 80% training set and a 20% testing set to ensure unbiased model evaluation.

Model Training: Train multiple regression models, starting with Linear Regression, Ridge, Lasso, and moving to more complex ensemble methods like Random Forest and Gradient Boosting.

Model Evaluation: Evaluate the trained models on the test set using defined metrics.

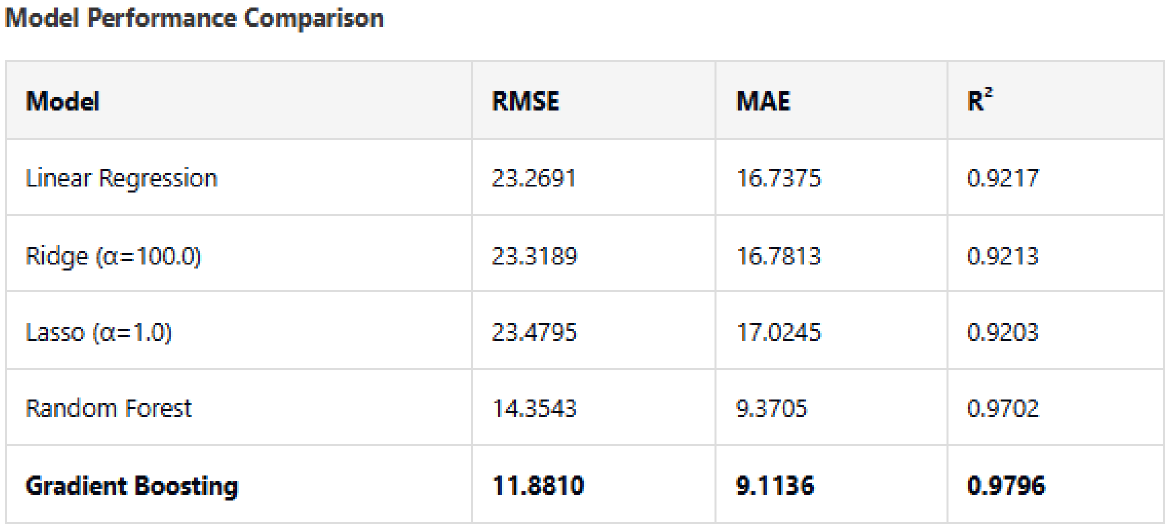
Persistence: Save the final, best-performing model to a file using the joblib library.

Results & Evaluation

Evaluation Metrics Used

The model performance is quantified using three key metrics:

Root Mean Squared Error (RMSE): Measures the average magnitude of the errors, providing an estimate of the typical error size in the predicted price.
Mean Absolute Error (MAE): Represents the average absolute difference between predicted and actual values.
R-squared (R^2): Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. A score closer to 1.0 indicates a better fit.



Model Performance Comparison

Model	RMSE	MAE	R^2
Linear Regression	23.2691	16.7375	0.9217
Ridge (α=100.0)	23.3189	16.7813	0.9213
Lasso (α=1.0)	23.4795	17.0245	0.9203
Random Forest	14.3543	9.3705	0.9702
Gradient Boosting	**11.8810**	**9.1136**	**0.9796**

Conclusion

The project successfully implemented a comprehensive Airbnb price prediction pipeline, including data preprocessing, feature engineering, and training of multiple regression models.

Linear models (Linear Regression, Ridge, and Lasso) achieved respectable R^2 scores around 0.92, showing that linear relationships do exist in the data. However, these models were limited in capturing the complex interactions between features.

The ensemble-based methods demonstrated superior performance. Random Forest achieved an R^2 score of 0.9702 with RMSE of 14.35, while Gradient Boosting emerged as the best-performing model with an R^2 score of 0.9796 and RMSE of 11.88. These models effectively captured non-linear dependencies and feature interactions, resulting in highly accurate price predictions.

The Gradient Boosting model's ability to explain nearly 98% of the price variance with minimal error validates its suitability for real-world deployment. The model can provide hosts with accurate pricing recommendations and help guests understand the factors driving rental costs in Melbourne.

The pipeline's modular design allows for easy experimentation with additional models, hyperparameter tuning, and feature enhancements. Future work could explore deep learning approaches, incorporate additional external data sources (such as local events or seasonal trends), and develop a user-friendly interface for real-time price predictions.

The project concludes with the selection and saving of the Gradient Boosting model for future deployment and integration into a production-level Airbnb price prediction system.