Predicting Airbnb Rental Prices Using Machine Learning

HARIDEEP AEPURI

REGIS UNIVERSITY

DATA SCIENCE PRACTICUM 1

Introduction

Airbnb is one of the most popular rental service provider The main objective of the project it to predict the Airbnb rental prices using machine learning

Problem Definition

- The main problem addressed in this project is the large variation and unpredictability in Airbnb listing prices because of several independent variables.
- Pricing with traditional methods do not process the capacity to analysis these factors and adjust to varying the market conditions
- For this problem a data-driven automated pricing solution is needed which can predict the listings prices
- Regression modeling was necessary.
- Target variable: Price (continuous)

Project Workflow

This workflow describes the entire pipeline that has been followed to create an accurate and optimized machine learning model for predicting Airbnb rental prices.

Overview of Dataset

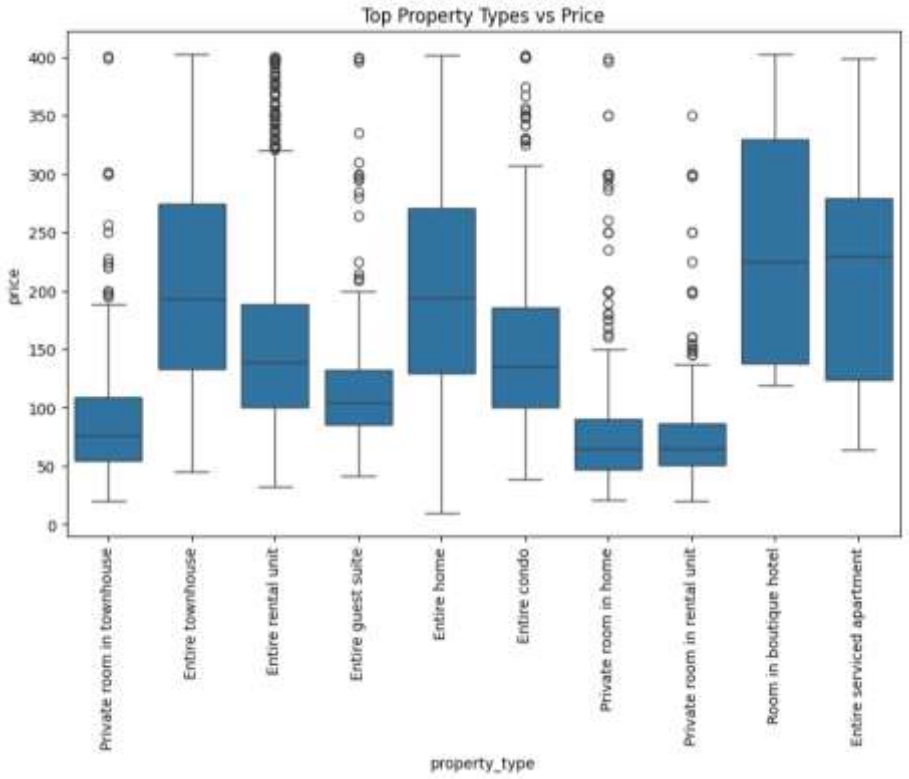
- The dataset I used in this project which is obtained from Kaggle which contains information of Airbnb listing that includes:
- Information of property like room type , number of rooms , locations
- pricing information as listing prices and fees
- However, the dataset contains both numerical and categorial variables

Data Cleaning

- Removed the columns which are not necessary for prediction.
- Price column (cleaned) (no minuses, and no idle symbols).
- Percentages converted into numeric.
- Converted boolean to binary

Outlier & Handling Missing values

- I used IQR method to identify the values which are above normal range.
- Once after removing the outliers price distribution became more consistent
- I used the median imputation which resulting in reducing the missing numerical values.
- Then to improve the data quality the highly missing data was eliminated.



Feature Engineering

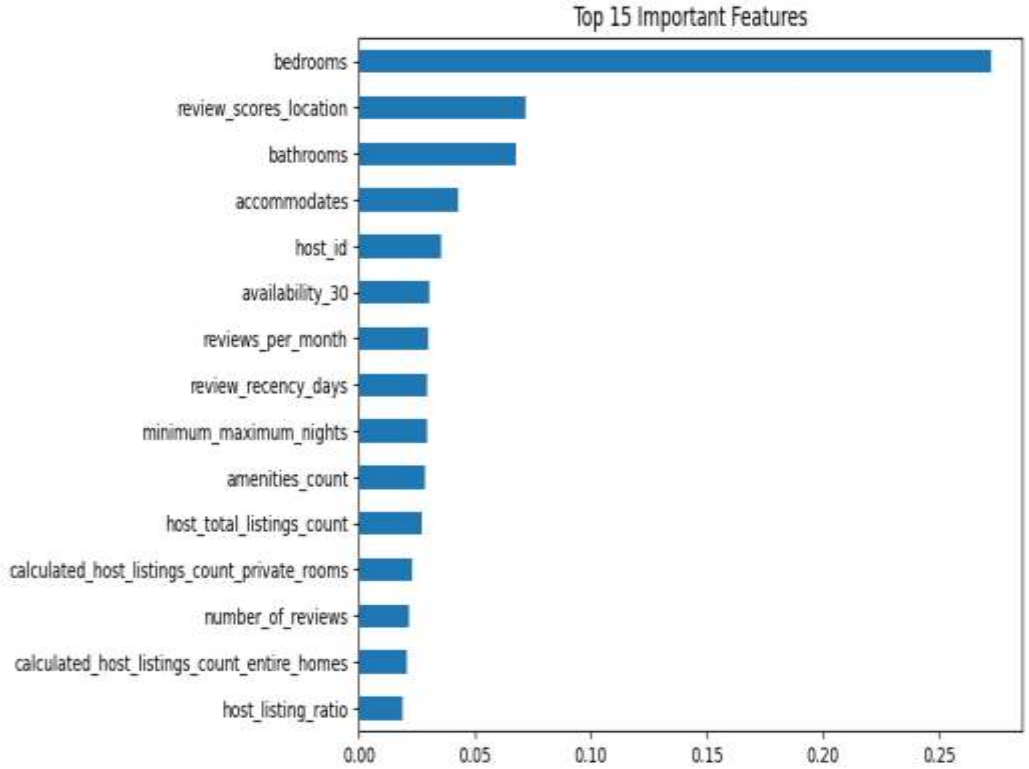
- In feature engineering the raw features like room type, number of bedrooms , amenities and location all the variables were converted into meaningful numerical representations so that it can be understood by machine learning algorithm
- Created host_experience_years to measure the how long host have been active
- Generated host_verification_count
- Feature engineering is a very critical step as it directly affects the accuracy, reliability and performance of the machine learning model.
- Proper feature engineering enables the model to learn meaningful patterns to make better predictions of Airbnb prices and decision-making.

Feature Scaling

- Feature scaling is the procedure of converting the range of the independent variables to bring them into the same order range so that every numerical feature can have the same contribution to the machine learning model.
- In airbnb price predictory. features such as price, availability which may contain thousands and bedrooms and bathrooms which may contain between 1-10.
- Applied StandardScaler
- Normalized in terms of means and standard deviation.
- needed in SVR and linear models.
- Enhances convergence & stability.

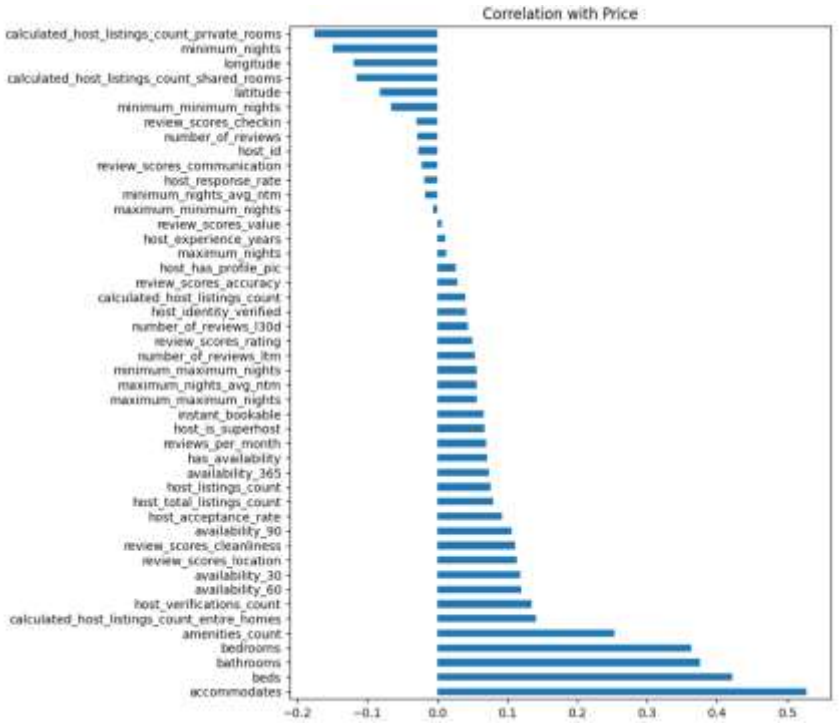
Exploratory Data Analysis

- The EDA show the top variables which shows Bedrooms, bathrooms and accommodates which shows a strong impact on price
- There are some variables like Availability and number of revies which are less impact on price
- Some features shows minor impact towards price like host related features compare to property features.



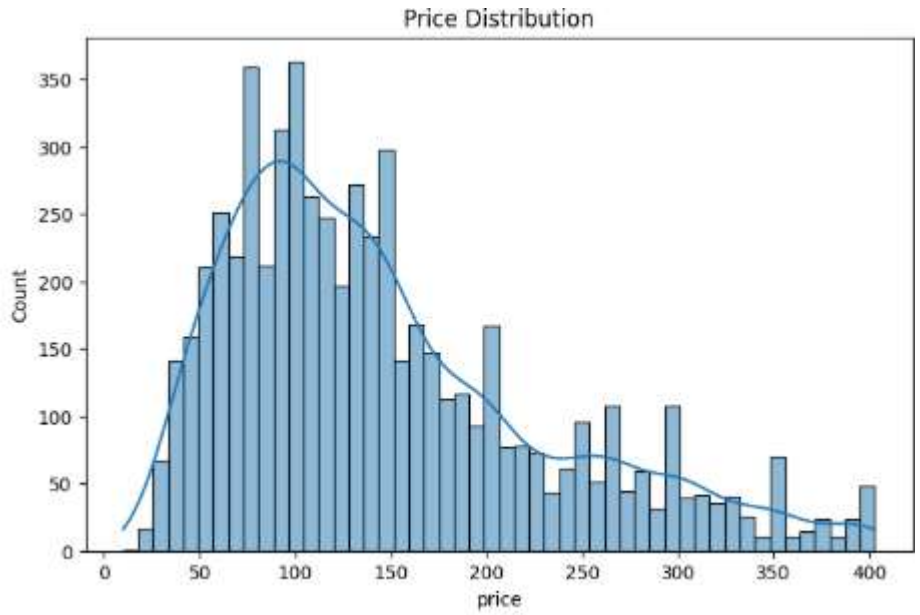
Correlation with price

- This plot shows the correlations with price which shows a strong positive and negative correlations
- Positive correlations like beds, bedrooms, bathrooms and accommodates also increase in amenities listing prices
- The features like review related shows negative correlations



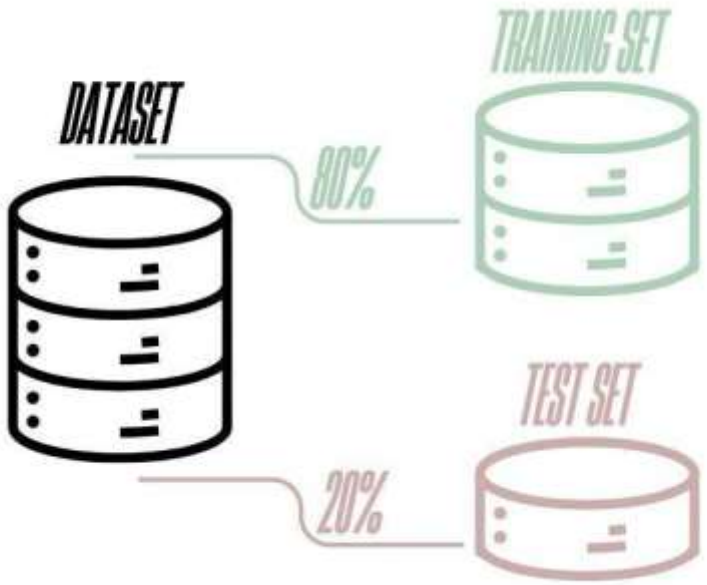
Price Distribution Insights

- The price distribution which is Right skewed distribution.
- Many of listings are falling between 50-150 which show moderate
- Luxury listings are fewer
- After removing outliers, the distribution becomes more balanced



Model Development Strategy

- 80% training, 20% testing split
- Determined random state of reproducibility.
- Different comparison regression models.
- Best model chosen on basis of metrics.



Regression Models Implemented

- Linear Regression (baseline)
- Ridge & Lasso (regularization)
- Decision Tree
- Random Forest
- Gradient Boosting
- Support Vector Regressor

Evaluation Metrics

- MAE – average absolute error
- RMSE – penalizes large errors
- R^2 – variance explained
- Lower error & higher R^2 = better model

Baseline Results

- Gradient Boosting performed best
  - $R^2 \approx 0.55$
  - MAE ≈ 39
  - RMSE ≈ 55

Model Comparison

- The ensemble models performed better in comparison with the linear models.
- Overfitting was exhibited by Decision Tree.
- SVR underperformed
- Gradient Boosting proximity to Random Forest.

	Model	MAE	RMSE	R2 Score
5	Gradient Boosting	39.536109	55.363046	0.549025
4	Random Forest	39.794697	55.489192	0.546967
1	Ridge Regression	46.516554	62.912160	0.417653
2	Lasso Regression	46.516039	62.913763	0.417623
0	Linear Regression	46.515893	62.913819	0.417622
3	Decision Tree	56.062171	81.162218	0.030784
6	SVR	61.051260	83.558430	-0.027290

Cross Validation

- The main benefit of cross validation is that you can get a better estimate of the model's performance, because it evaluates the performance of the model using many combinations of training and testing data.
- This helps in reducing sampling bias and also ensures that the model does not fit too well to a particular data split.
- As aiding, it is responsible for the improvement of stability and robustness of evaluation metrics such as R2, MAE, RMSE.

Hyperparameter Tuning

- Hyperparameter tuning which is important process in the creation of efficient and accurate machine learning model. Hyperparameters are parameters in configuration that control the process of learning a model, and they need to be specified prior to training a model.
- In this project, I used GridSearchCV function which discovers many combinations of parameters with the help of cross validation techniques and selects the best performing combination with the help of predefined evaluation metric.

Final Conclusion

This project finally developed a machine learning model which will predict the rental prices of Airbnb by using the characteristics like room type, amenities and host information. After following a systematic data science process such as data preprocessing, feature engineering and testing the multiple regression models, gradient boosting model produced best performance which archived an R2 of 0.55 variation. The results show that machine learning will identify the patterns and supports the data driven decisions which helps the Airbnb Hosts to set the listing prices.

References

1. Benzing : Airbnb related image
   https://media.zenfs.com/en/benzinga_79/755380940e89d953e22a98d0da665420
2. SlideTeam : Artificial intelligence, machine learning concept image
   https://www.slideteam.net/media/catalog/product/cache/1280x720/a/r/artificial_intelligence_machine_learning_deep_learning_ppt_power_Slide22.jpg
3. Medium : machine learning illustration.
   https://miro.medium.com/v2/resize:fill:320:214/1*eUi0ZHwA6d8xSMGkiF6dpQ.jpeg
4. Encord. : machine learning pipeling

https://images.prismic.io/encord/d70c6d15-2c35-4a19-955c-91da01a30572_image1.png

Thank you