Final Technical Report: Airbnb Price Prediction Intelligence

Student Name: Mina Ezach Naeem Faltos Student No: 34388

Course: Machine Learning Algorithms (MAAI)

A. Introduction and Problem Definition (Phase 1)

The fast growth of the sharing economy has introduced nightly pricing as one of the primary concerns of Airbnb hosts. Exceeding the necessary price causes low occupancy whereas underpricing means loss of revenue. The project aims to determine the fair market value of the short-term rental listing in Barcelona using objective characteristics. The process transations from classical statistics to artificial intelligence to derive meaningful insights and advance the explainable AI based on sound reasoning, giving the ability to use the model to make predictions for other locations as well.

B. Methodology

1 Dataset Description and Preparation (Phase 2)

The data source was obtained from inside airbnb (through kaggle) for Barcelona listing.

- **Structure and Scope**: The dataset were initially 19,833 observations and 25 variables.
- **Variable Classification**: Ten core variables were selected, which are either numerical (latitude, longitude, accommodates, bathrooms, review_scores_rating) or categorical (neighbourhood, room_type).
- **Cleaning and Engineering**: Price strings (e.g., "$130.00") were turned into floats and the absences of any numerical values were filled in with the median to make sure its resistant to outliers.
- **Filtering**: In order to make the model more stable, information on listings that were higher than 500 were filtered out.

2 Model Selection and Training (Phase 3)

Three regression algorithms were compared to meet the demand of multiple techniques:

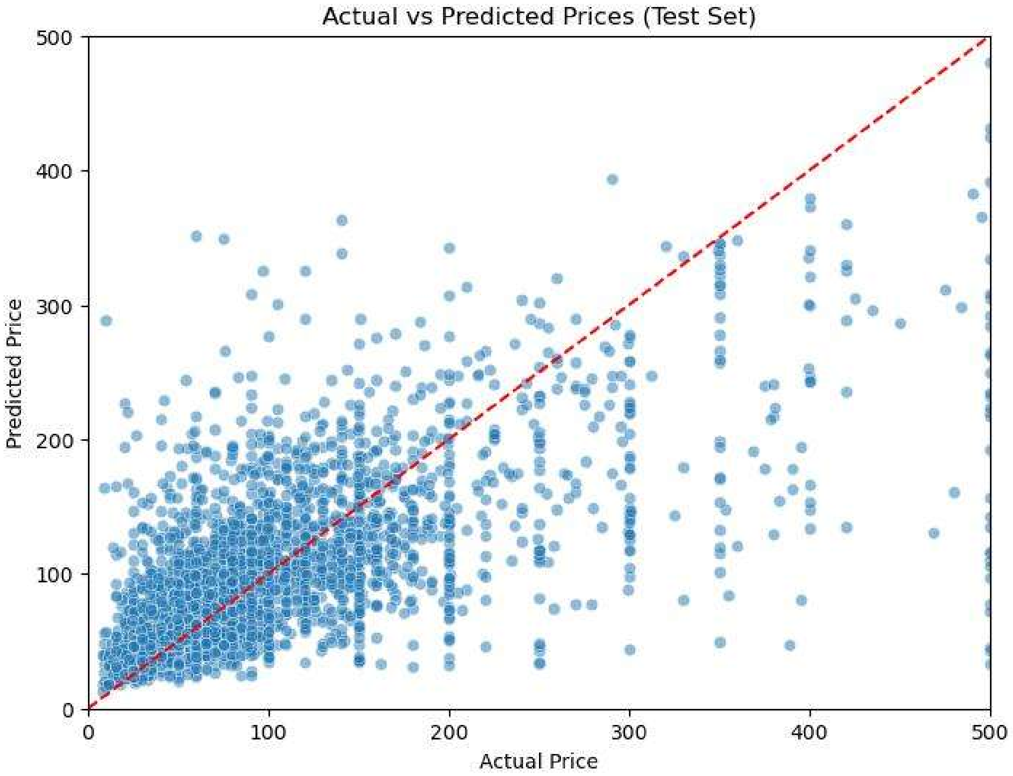
- **Linear Regression**: Was used as a base of simple linear associations.
- **Decision Tree**: A non-linear model that reflects complicated rules.
- **Random Forest**: An ensemble technique that is selected in order to maximize accuracy and reduce overfitting.
- **Hyperparameter Tuning**: The Hyperparameter Optimalization was achieved through GridSearchCV, which optimized the Random Forest with max_depth: 20 and n_estimators: 100.

C. Results (Phase 4)

1 Final Performance (Test Set)

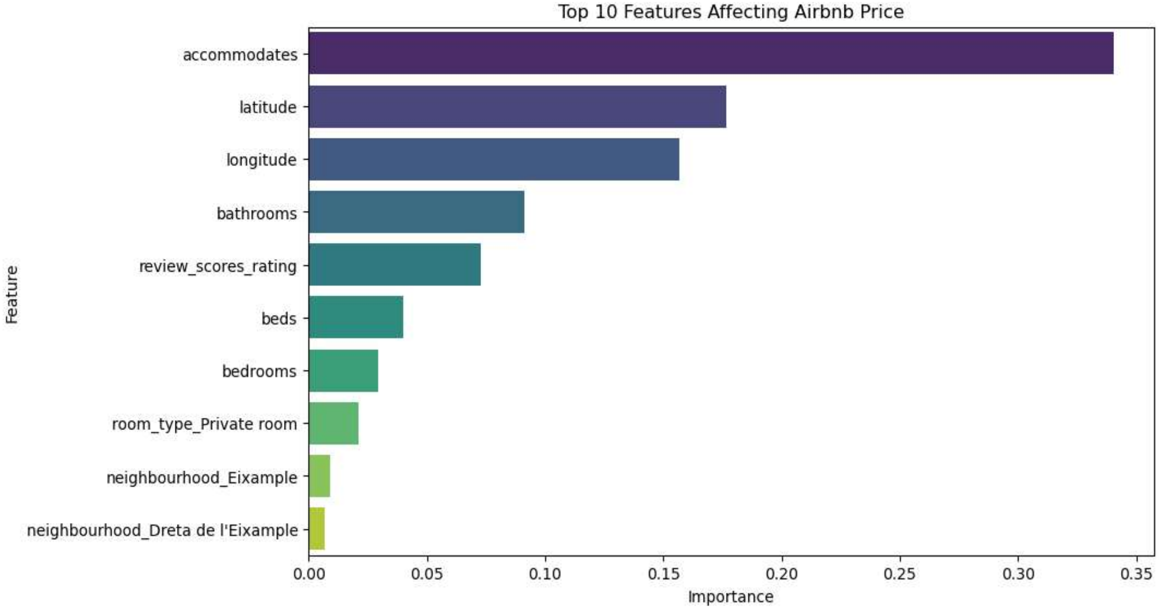
The tuned model was tested on a held-out test set (20% of data):

- **RMSE**: €58.13.
- **MAE**: €34.71.
- Score: 0.4817, meaning that the model is able to explain about 48 percent of price variability.
- **Validation**: The model experimental predictive prices were inspected manually against known listings; these were always within a realistic market range, which validated the practical validity of the model.



2 Visualizations and Feature Importance

Accommodates and geographic location (latitude and longitude) were discovered to be the strongest predictors. This enables the system to rationalize suggestions to the hosts in terms of capacity and central proximity.



D. Lifecycle and Deployment (Phase 5 and 6)

1 Scalability and Persistence

The champion Random Forest model (v1.1, which is the one chosen at Phase 5) was stored as airbnb_price_model.pkl and deployed unchanged in Phase 6. FastAPI was used to start a live RESTful API, which serves as batch predictions in real-time. Using FastAPI with Uvicorn allow parallel processing of requests, which meets the scaling requirement since many users can use the model at the same time.

2 Industrial Sanitizer and Monitoring

- **Sanitization**: A custom "Industrial Sanitizer" engine handles raw inputs, mapping synonyms (Like: 'lat' to "latitude") and repairing the currency strings to ensure the JSON compliance.

- **Monitoring Plan**: An alert for retaining is triggered by a performance threshold of **€70.00 RMSE** due to potential market or data drift.

- **Retraining Strategy**: The model follows a quarterly (3-month) schedule in terms of retraining, to reflect variations in Barcelona.

E. Discussion and Ethical Considerations

1 Methodological Limitations

The main weakness is that the model is focused on the standard market segment (under €500) which is why it could not fit the luxury listing. Also, the strength of objective features is accompanied by subjective aspects as the quality of decor that cannot be captured.

2 Ethical Risks

Market Estimates are used to denote predictions to reduce the risks of gentrification rather than objective truths. Quarterly fairness audits are planned in order to make sure that review scores do not indicate some implicit bias against minority hosts.

F. Conclusions

The project successfully put into practice an end-to-end machine learning pipeline, fulfilling all the technical, methodological, and ethical requirements. The last system offers an effective, professional, and explainable foundation for predictive pricing in Barcelona.

G. Code Availability

The full source code, Jupyter Notebooks (.ipynb), and the project history are available at github through the:

- **Git Repository**: https://github.com/MinaEzach/Airbnb-Price-Prediction-MAAI

H. References

- Inside Airbnb. (2025). *Barcelona Airbnb Listings Dataset*. Kaggle.
- Zervas, G., Proserpio, D., and Byers, J. (2017). *The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry*.