Project Report

Introduction:

Pricing short-term accommodation is a complex and dynamic problem influenced by both internal (property-specific) and external (contextual) factors. While hosts typically adjust prices based on intuition or competitor listings, it's unlikely that they adopt the kind of dynamic pricing models that a wide range of enterprises are currently putting in place. Accordingly, it's likely that hosts rarely take into account data such as weather conditions or holidays - all of which can significantly affect demand - when determining the pricing of their property.

Prior research (Casamatta et al., 2022) has shown that pricing behavior on Airbnb varies substantially across seasons and host types, highlighting the importance of incorporating temporal and contextual signals into pricing models. However, these dynamics are rarely applied proactively by hosts as a means by which to boost property revenues.

This project proposes the development of a data-driven smart pricing recommendation system for Airbnb listings to enable such dynamic pricing by hosts. More specifically, we propose a system that will combine internal listing data with external contextual signals to recommend optimal daily prices and generate transparent natural-language explanations for hosts.

Research Question: How can external contextual factors such as weather, holidays, and proximity to points of interest within a city be integrated with Airbnb data to produce accurate, fair, and explainable short-term accommodation pricing recommendations?

Data Collection and Integration:

- To build a robust pricing system, we integrated internal listing data with diverse external datasets to capture the contextual factors that influence demand. We used Airbnb Dataset as our primary foundation. We chose features that held the most important information on properties from our analysis: *location, pricing_details, details, description_items , cancellation_policy, amenities, lat, long, available_dates, property_id, currency,* and *reviews.*
- All external data were collected via public APIs or open data services, processed using PySpark, stored in Parquet format, and later joined with the main dataset.

Weather Data: Historical and forecast data (temperature and weather indicators {snow, clear sky, rainy, etc}) were retrieved via the Open-Meteo API.

Public Holidays: We integrated national holiday schedules using the Nager.Date API, as holidays significantly drive spikes in short-term rental demand.

Points of Interest (POI): Using the OpenStreetMap (Nominatim and Overpass APIs), we identified central landmarks, museums, and transport hubs. We then calculated the count of these POIs within 1km, 5km, and 10km radius of each property using the Haversian distance formula.

- Integration of the data:

- The weather data were stored as a Parquet file and later joined to the main dataset using *location_city* and *date*.

- The resulting city-date holiday table was stored in Parquet format and joined with the main dataset on location_city and date.

- The standardized city center table was saved as a Parquet file and used for geographic alignment.

This external data was integrated into our solution in two main ways:

1. Predict the probability a property will be rented on a certain day, in that day's weather, with a certain price and taking account holiday status.
2. Locations of especial tourist or attraction points, closeness to public transportation, etc. were used as a feature for the XGBoost model

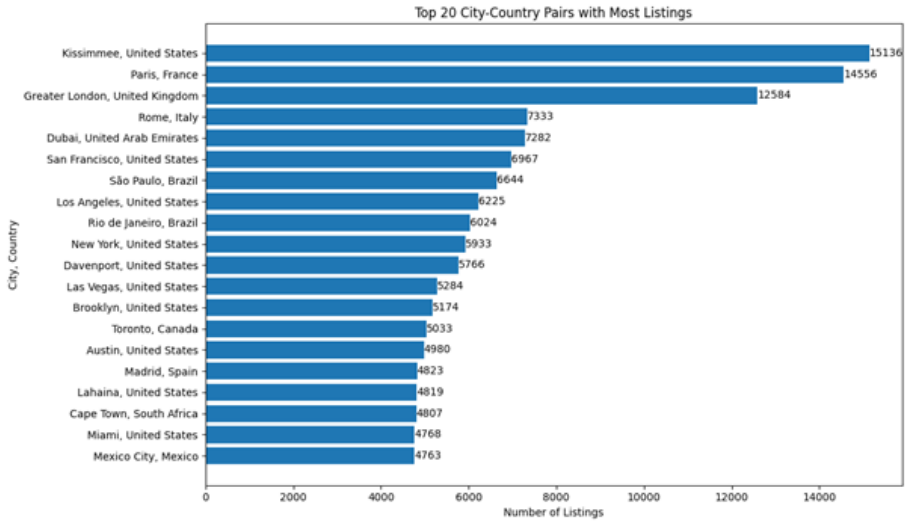
- Definition of an item:

- For price prediction (XGBoost), an "item" is defined as a unique *property_id*. The dataset includes static features like *amenities, location scores, average sentiment*. The Label was *property_price_per_night*

- For rental probability (Logisitc Regression), an "item" is defined as *a property_id* paired with a specific *date*. This captures how a single property's attractiveness changes based on shifting variables like the weather or whether that specific day is a holiday, the price per night, number of reviews, etc. The **size** of the enriched data is $5220\ rows\ X\ 7\ colomns$

Data Analysis:

- **City Scope Selection**: We used *Location* to extract *City* and *Country*. We then focused our analysis on the top 20 cities with the highest density of listings to ensure statistical significance and data richness.



Date Scope Selection: because of the use of external weather data, we could not integrate every single day of the year. We analyzed the most significant days in the year (in terms of availability) by calculating the weekly mean of available properties. This became our threshold from which we chose only dates that exceeded it (in blue, appendix).

Outlier Detection: We performed a quantile analysis on the *price_per_night* field, identifying and removing extreme outliers (below the 2% and above the 98.55% quantiles) to prevent the models from being skewed by data entry errors or ultra-luxury listings *(in appendix).*

Price Normalization: Nightly prices were extracted from the *pricing_details* field. Those with missing values were not taken into account. We than made sure all values were converted into a single currency (USD) to allow for cross-city comparisons.

- **Feature selection and engineering:**

We extracted from *description_items* key predictors such as the number of bedrooms, bathrooms, whether a property is private or shared (in terms of amenities use) etc. From *cancellation_policy* we created a *cancellation grade* to score properties, where a full refund gave a score of 2 and no refund a score of 0. A partial refund earned the score of 1.

Amenity Scoring: We calculated an *Amenity_Score* by weighting features based on their rarity (e.g., WiFi is common and weighted lower, while a pool is weighted higher).

Sentiment Analysis: We performed sentiment analysis on the *reviews* column, assigning each property a "majority vote" sentiment (Positive, Neutral, or Negative) to quantify guest satisfaction.

Methodology:

An XGBoost regression model was employed. While XGBoost handles the base price for a new property, a Logistic Regression model predicts the probability of a property being rented on a specific day.

Our implementation evolved through two distinct stages:

Initial Optimization Approach: We first attempted a full optimization strategy where a range of prices was tested for every property-day combination. The goal was to select the price that maximized expected revenue, defined as $Price\ X\ Rental\ Probability$.

Observations: Results indicated that the model's sensitivity to price was relatively moderate. Consequently, the optimization frequently converged on the same price across different days, failing to produce significant dynamic variance

To overcome this, we adopted a more robust, **demand-based** ranking approach that directly leverages the logistic model's output.

We calculated the rental probability for each property on every given day using its original baseline price. Days were then ranked based on these predicted probabilities to identify peak and off-peak periods. For days identified with exceptionally high/low/ demand, the price is increased/decreased by 10%

This refined approach ensures clear price differentiation between high and low-demand periods while utilizing the logistic model's predictive power without the need for computationally heavy optimization.

Evaluation and Results:

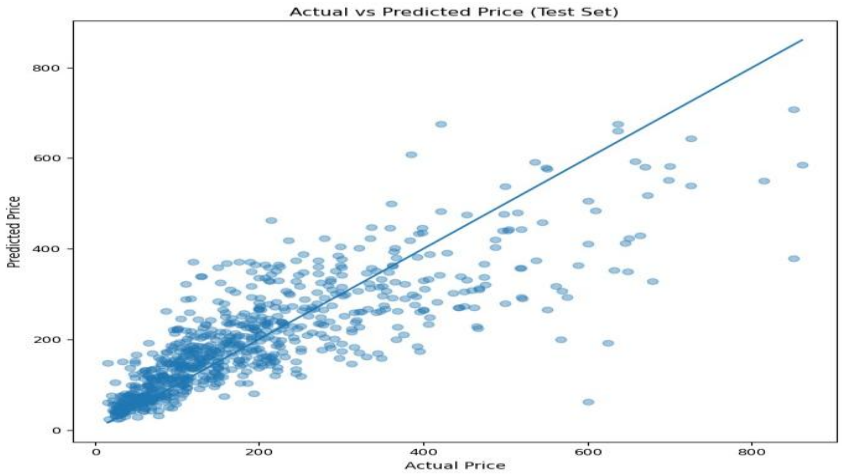
The XGBoost model achieved an RMSE of approximately 88$, an MAE of about 61$, and an R^2 of 0.68, indicating that it explains roughly 68% of the variance in nightly prices. Compared to a city-level median price baseline, the XGBoost model substantially outperformed the baseline across all metrics (RMSE ≈ 143, $R^2 \approx 0.15$), demonstrating the added value of incorporating property-level and location-based features.

The logistic regression model demonstrates stable and well-calibrated performance, with an AUC-ROC of approximately 0.62 (appendix, calibration curve) and consistent results across validation and test sets, indicating reliable generalization. While its discriminative power is moderate, the model excels in producing meaningful probability estimates rather than sharp binary decisions. At a lower classification threshold, the model achieves high recall, successfully identifying most rented days, though at the cost of lower precision and a higher false-positive rate. This behavior is expected given the overlapping nature of rental and non-rental patterns in the data. Importantly, calibration metrics and calibration curves show strong

alignment between predicted probabilities and observed rental rates, validating the probabilistic interpretation of the outputs. As a result, the model is well suited for the project's goal of demand estimation and revenue optimization, where accurate probabilities are more valuable than strict classification accuracy.

The Actual vs. Predicted Price plot, shown for a 5% sample of the test data, reveals strong alignment in low- to mid-price ranges, with increased dispersion for high-priced listings, reflecting the inherent difficulty of modeling extreme prices.

Model performance:



Limitations and Reflection:

Label Noise: Lacking transaction logs, we used "availability" as a proxy for demand (unavailable = rented), potentially misclassifying host-blocked dates as bookings.

Restricted Scope: To maintain computational feasibility, the analysis was limited to the top 20 data-rich cities and a specific six-month window, which may affect generalizability to smaller markets.

Resource Constraints: Real-time integration was simulated with historical data due to API rate caps and budget limitations.

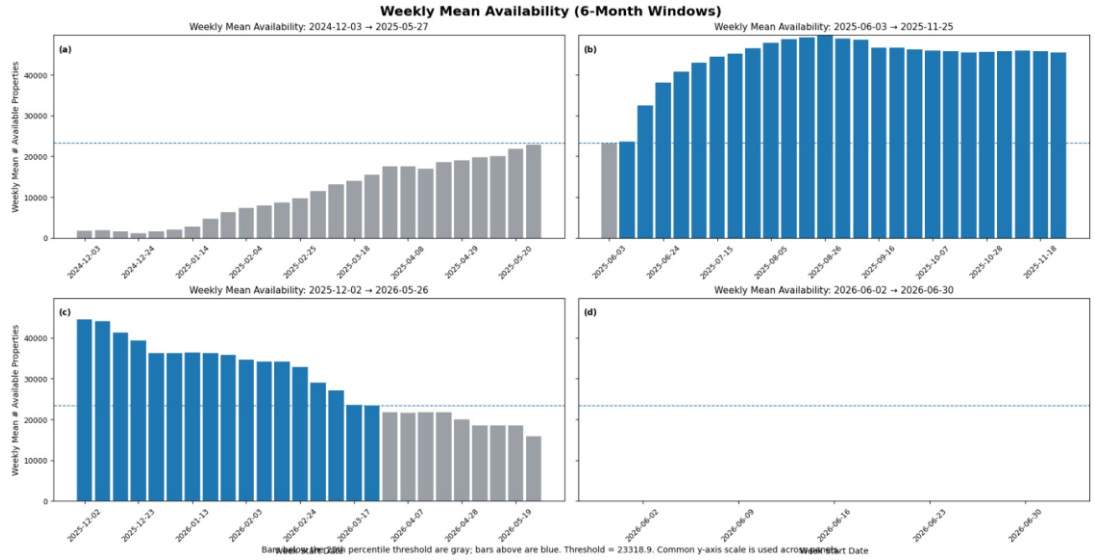
Conclusion:

The project turned out to be excessively challenging. With many limitations and algorithmic and mathematical challenges, we had a real empirical trial-and-error journey. Yet, despite these limitations, the project confirms that integrating external contextual signals, particularly weather and holidays, provides a significant advantage over static pricing. The transition from a full optimization approach to a demand-ranked (dynamic) pricing strategy highlights the importance of balancing theoretical complexity with practical, interpretable results in a real-world data science application.

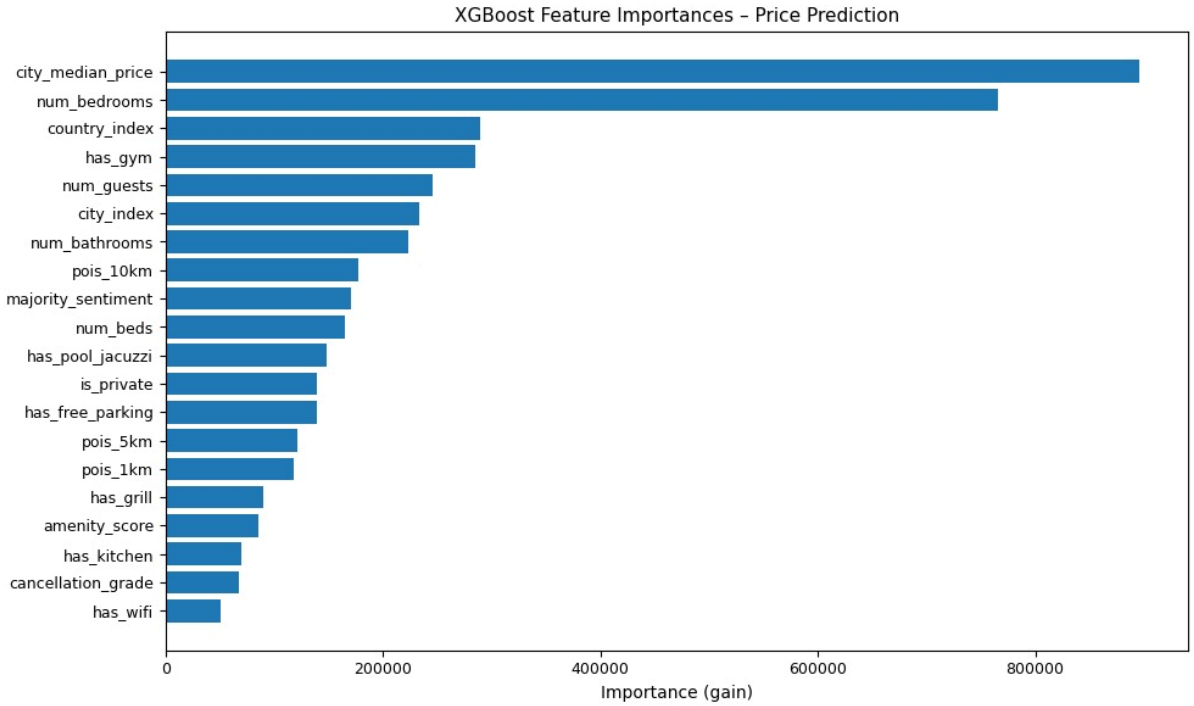
Appendix:

- **Date Scope Selection**

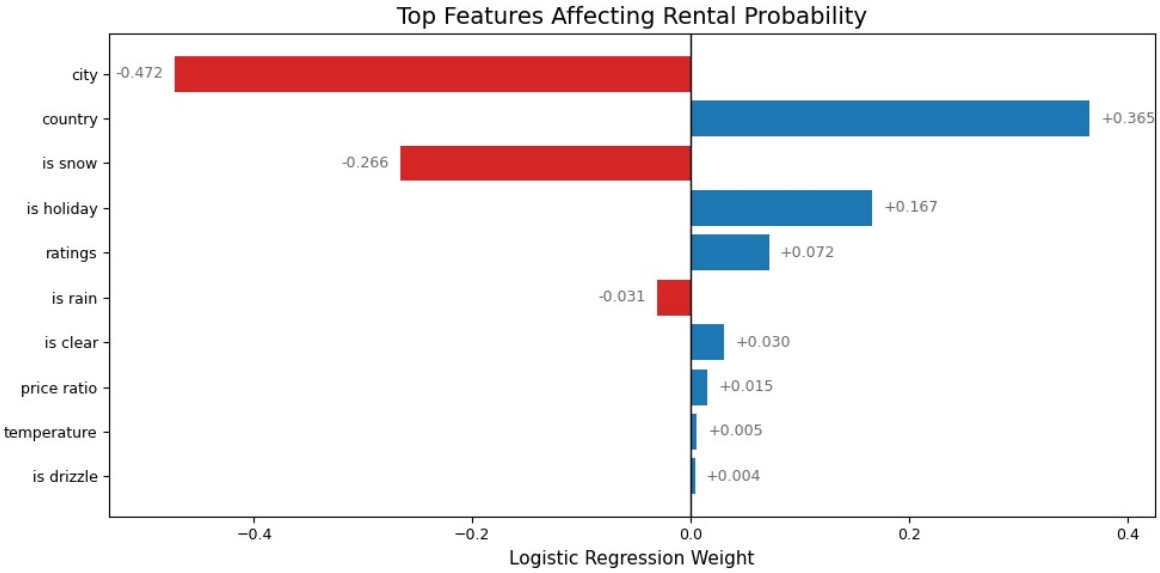
We analyzed the most significant days in the year (in terms of availability) by calculating the weekly mean of available properties. This became our threshold from which we chose only dates that exceeded it (in blue).



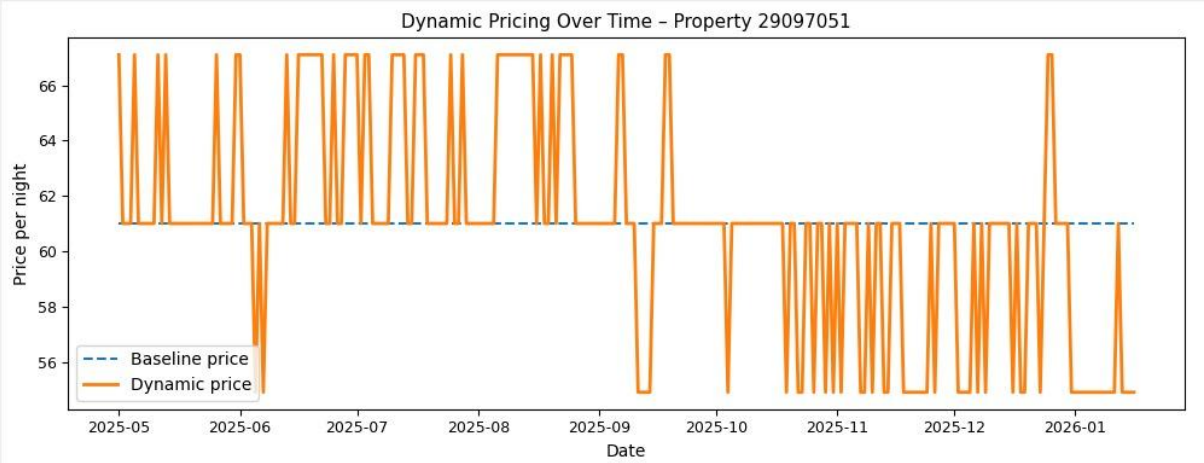
- **Feature Importance XGBoost**: Full ranking of variables influencing the XGBoost price prediction model.



- **Feature Importance Logistic Regression**: Full ranking of variables influencing the Logistic Regression Rental Probability.

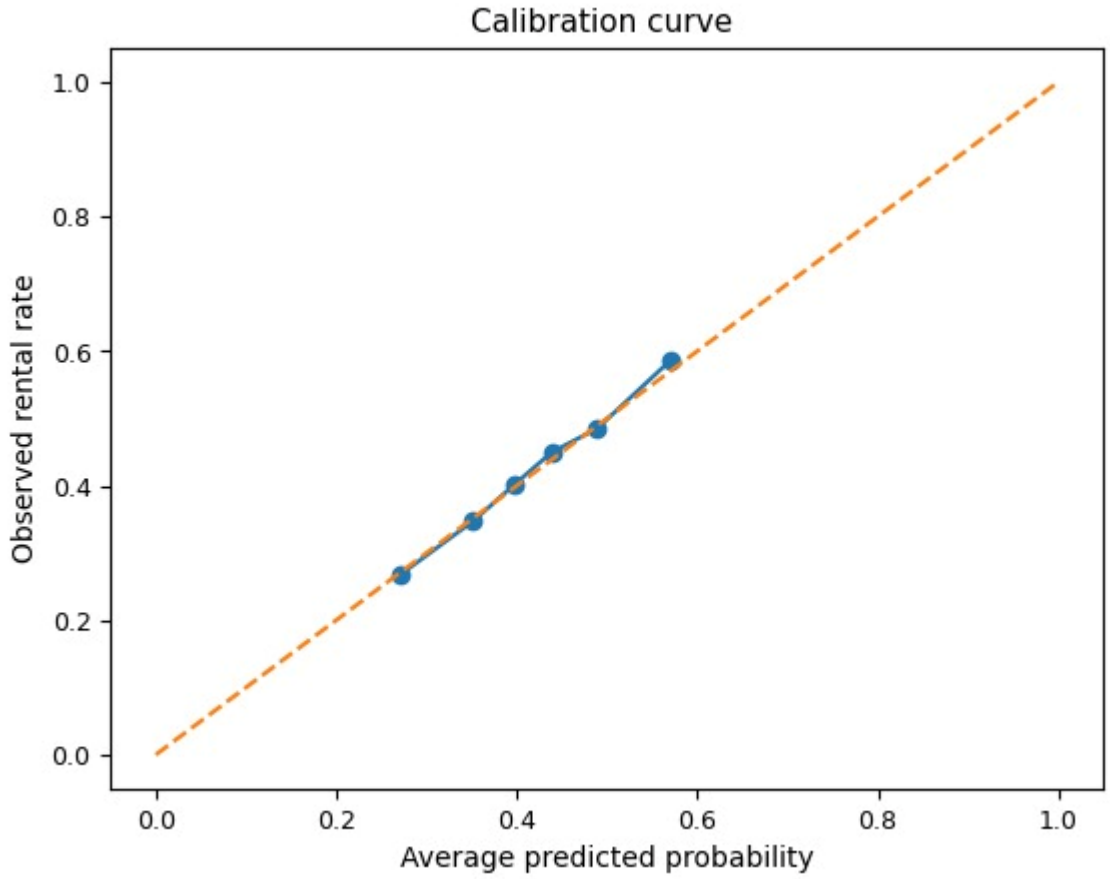


- **Dynamic Pricing Over Time (Property 29097051)**: This chart demonstrates the real-world application of the demand-based pricing strategy. The orange line reflects how the price fluctuates by **±10%** relative to the baseline (blue dashed line) in response to contextual demand signals such as weather changes and public holidays.



The calibration curve measures the reliability of the model's probability outputs by comparing predicted probabilities to actual observed rental rates.

- **High Reliability**: The blue line closely follows the dashed diagonal identity line, indicating that the model is well-calibrated.
- **Statistical Accuracy**: For example, when the model predicts a **49.29%** chance of a rental, the actual observed rental rate in that bin is also **49.29%**, ensuring the demand signals are trustworthy for automated pricing decisions



External data links:

- Open-Meteo API for historical weather and forecasts https://open-meteo.com/
- Nager.Date API for global public holidays. https://date.nager.at/
- OpenStreetMap (Nominatim and Overpass) for geospatial point-of-interest data. https://nominatim.openstreetmap.org/search, https://overpass-api.de/api/interpreter

GitHub Repository: github link

References

Casamatta, G., Giannoni, S., Brunstein, D., & Jouve, J. (2022). Host type and pricing on Airbnb: Seasonality and perceived market power. *Tourism Management*, 88, 104433.

Elmachtoub, A. N., & Grigas, P. (2022). Smart "predict, then optimize". *Management Science*, *68*(1), 9-26.