Data Driven Price Estimation for Airbnb Listings in Stockholm

Filippo Muscherà
Blekinge Institute of Technology
19991219-T591
fimu24@student.bth.se

Eneko Isturitz
Blekinge Institute of Technology
20030623-T150
enis24@student.bth.se

Paul Gasnier
Blekinge Institute of Technology
20041108-T110
paga24@student.bth.se

Unai Lalana
Blekinge Institute of Technology
20030802-T310
unla24@student.bth.se

***Abstract*—This project presents a data-driven approach for predicting nightly rental prices for Airbnb listings in Stockholm. Using a dataset of over 5,000 listings, with features like amenities, reviews, host characteristics, and listing descriptions. We trained a ML model to predict prices and based on the predictions some tips are displayed for hosts, including pricing suggestions, optimal listing descriptions, and amenity enhancements. The goal of the resulting tool is to help Airbnb hosts improve their listings to increase both visibility and profitability.**

I. Introduction

The short-term rental market in major cities has grown significantly in the past decade, and platforms like Airbnb allow homeowners to generate income by renting out their properties. However, setting up a successful Airbnb listing, especially in a competitive market like Stockholm, requires more than just offering a space. Hosts must make informed decisions about pricing, minimum stay requirements, amenities, and listing descriptions to attract guests and maximize revenue.

In this project, we used real-world Airbnb data from Stockholm to explore what makes a listing competitive and profitable. By analyzing patterns in pricing, availability, amenities, and listing descriptions, we aim to answer key questions such as:

- What is a fair nightly price for a given apartment?
- How do amenities and location affect the rental value?
- What are the most effective keywords to include in a listing?

The outcome of this project is a data-driven tool that takes the characteristics of an apartment such as address, number of rooms, and available amenities and provides an estimated nightly rental price. In addition, the tool offers recommendations to help a listing stand out in the crowded Stockholm market.

II. Related Work

Pricing strategy and competitiveness on peer-to-peer rental platforms such as Airbnb have been widely studied in various domains, including data science, economics, and urban studies. A growing body of research has explored how different listing features, such as location, amenities, and host behavior, correlate with pricing, occupancy rates, and guest satisfaction.

One major area of focus in the literature is the prediction of prices using machine learning models. For example, studies such as Wang and Nicolau [1] have applied regression models and ensemble learning techniques to predict nightly rental prices based on structured data, including the number of bedrooms, location coordinates and availability. Other works, such as those of Gutiérrez et al. [2], emphasize the role of spatial characteristics and proximity to landmarks or city centers in determining the price.

Another relevant area is the analysis of amenities and the optimization of the list. Research has shown that listings offering certain amenities (e.g., Wi-Fi, kitchen, washer/dryer) tend to receive more bookings and higher ratings. In addition, the presence of unique or luxurious features, such as a balcony or sauna, can allow hosts to charge a premium. Studies have also analyzed the importance of text features, finding that listings with well-written titles and descriptions, particularly those containing attractive keywords, perform better in terms of visibility and booking rates [3].

Additionally, various Airbnb data science competitions and open-source projects have tackled similar objectives, providing public datasets and benchmark models for automated price estimation and ranking listings by attractiveness. Tools developed in these contexts often use linear regression, decision trees, or more advanced models like XGBoost or neural networks, balancing prediction accuracy with interpretability.

Our work builds on these foundations by combining structured Airbnb data with feature analysis and predictive modeling. Unlike some previous studies focused solely on academic insights or highly tuned models, our goal is to develop a functional and practical tool for prospective Airbnb hosts in Stockholm, providing actionable insights for pricing and listing optimization based on real data.

III. Methodology

A. Dataset

The dataset utilized for this project consists of Airbnb listing data for Stockholm, Sweden. This raw dataset, comprises 5223 rows and 75 columns, providing a comprehensive overview of various listing attributes. Key features include:

- **Listing Information:** Unique identifiers, URLs, scraping metadata, and textual descriptions such such as `name` and `description`.
- **Host Details:** Information about the host, including `host_id`, `host_since`, `host_response_time`, `host_response_rate`, `host_acceptance_rate`, and `host_is_superhost` status.
- **Property Characteristics:** `latitude`, `longitude` for geographical location, `property_type`, `room_type`, `accommodates`, `bedrooms`, `beds`, `bathrooms_text`, and a list of `amenities`.
- **Pricing and Availability:** The target variable `price`, `minimum_nights`, `maximum_nights`, and various `availability` metrics (e.g., `availability_30`, `availability_365`).
- **Review Scores:** Comprehensive `review_scores` metrics (e.g., `review_scores_rating`, `review_scores_cleanliness`, `review_scores_location`) and `number_of_reviews`.

Initial data exploration revealed the necessity for extensive pre-processing due to missing values, inconsistent data formats, and the presence of categorical variables requiring encoding. The `amenities` column, for instance, was provided as a string representation of a list, which required parsing and transformation into individual boolean features.

B. Data Preprocessing and Feature Engineering

A pre-processing pipeline was established to transform raw data into a suitable format for machine learning. The key steps included:

- **Missing Value Imputation:** Numerical columns such as `host_response_rate`, `host_acceptance_rate`, and various `review_scores` were imputed. For instance, `price` entries of zero were removed, and other missing values were handled appropriately (e.g., `review_scores` were set to 0 for listings with no reviews, and derived features `never_reviewed`, `days_since_last_review`, `days_since_first_review` were created).
- **Data Type Conversion:** The `price` column, initially a string with currency symbols, was converted to a numerical float type. Similarly, `host_response_rate` and `host_acceptance_rate` were converted from percentage strings to numerical values.
- **Feature Extraction from Text:**
  - **Bathrooms:** The `bathrooms_text` column was parsed to extract the numerical quantity of bathrooms, and a boolean indicator `is_shared_bath` was derived.
  - **Amenities:** The `amenities` string was parsed, and individual amenities were extracted to create boolean (binary) features for each unique amenity present in the dataset with the one-hot encoding technique.
  - **Host Duration:** The `host_since` column was used to calculate `host_duration_days`, representing the time the host have been present on the platform.
- **Textual Feature Engineering:**
  - **Description and Name Analysis:** The `name` and `description` fields underwent natural language processing. This involved tokenization, removal of stopwords using NLTK [4], and vectorization using TF-IDF and Count Vectorizers [5].
  - **Sentiment Analysis:** Sentiment scores (`desc_sentiment`) were calculated for the `name` and `description` column after being unified, providing a numerical representation of the sentiment conveyed in the listing text. To do so, we used a model pre-trained for the sentiment analysis of Airbnb reviews (from Assignment 1) [6].
- **Categorical Encoding:** Categorical features such as `neighbourhood_cleansed`, `property_type`, and `room_type` were one-hot encoded to convert them into a numerical format suitable for machine learning models. The use of one-hot encoding in this phase expands significantly the number of features in the dataset, creating binary columns for each unique category.

C. Model Architecture

The predictive model employed for estimating nightly rental prices is a Light Gradient Boosting Machine (LightGBM) Regressor [7]. LightGBM is an ensemble learning method that uses gradient boosting, known for its efficiency, speed, and high performance on tabular data. Its ability to handle a large number of features and its optimized memory usage make it suitable for datasets of this scale.

The model was configured as an `lgb.LGBMRegressor`. The choice of a gradient boosting model is motivated by its proven effectiveness in similar regression tasks, particularly in competitive data science scenarios where complex, non-linear relationships between features and the target variable are common. LightGBM's leaf-wise tree growth algorithm further contributes to its speed and efficiency over other boosting frameworks. Before choosing the final model, other models where also evaluated. We tested Linear Regression and Random Forest Regression models as well, but they showed poorer performance than LighGBM.

D. Experimental Setup

The experimental setup involved a standard machine learning pipeline for regression tasks:

- **Data Splitting:** The preprocessed dataset was partitioned into training and testing sets. A typical split ratio (80% for training, 20% for testing) was applied

to evaluate the model's generalization performance on unseen data.

- **Hyperparameter Tuning:** To optimize the LightGBM model's performance, `RandomizedSearchCV` was utilized. This technique efficiently explores a defined range of hyperparameter values, sampling random combinations to find an optimal set. The objective function for optimization was the Mean Squared Error (MSE), which measures the average squared difference between the estimated values and the actual values. This process helps to mitigate overfitting and to improve predictive accuracy.
- **Evaluation Metric:** Mean Squared Error (MSE) was selected as the primary metric for model evaluation during training and tuning. Lower MSE values indicate better model performance, implying that predictions are closer to the true values.

The entire model training and tuning process was encapsulated within a Jupyter Notebook (`FinalAssignment.ipynb`), allowing for iterative development and analysis. The best performing model, identified through `RandomizedSearchCV`, was then saved using `joblib` for later deployment in the graphical user interface (GUI).

IV. Results

The trained LightGBM model provides a robust estimation of nightly rental prices for Airbnb listings in Stockholm. The specific performance metrics for the best model are detailed within the Jupyter Notebook. Feature importance analysis, derived from the correlation analysis, highlights the most influential factors in price prediction. Fig.1 illustrates the Pearson Correlation Coefficient of various numerical features with respect to price. Derived from this correlation analysis, a feature importance analysis highlights the most influential factors in price prediction. The tool presents these analytical results to the user, guiding them on how to make their listing more competitive.

Beyond price prediction, the project delivers actionable recommendations for hosts:

- **Predicted Price:** The core output is a precise nightly rental price estimation in Swedish Krona (SEK), convertible to other common currencies via an external API for user convenience.
- **Minimum Nights Recommendation:** Analysis of successful listings suggested optimal `minimum_nights` to be between 1 and 3, balancing flexibility with booking stability. Furthermore, analyzing top listings, we noticed a slight decrease in listing with a minimum number of nights equal to one. Top-performing listings, were identified as those in the top 25% by review scores, or in the top 25% by price that simultaneously exhibited bottom 25% availability over 365 days. This second condition was chosen assuming that a listing with a high price and a low availability has a high number of bookings, and is thus successful.

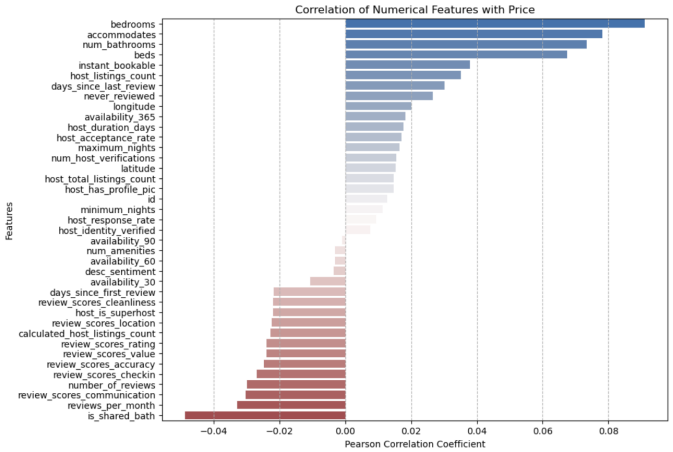

Fig. 1. Pearson Correlation Coefficient of features with respect to price. We can see from the figure that features like `bedrooms`, `accommodates`, `num_bathrooms` and `beds` are the ones with the highest positive correlation. Similarly, `is_shared_bath` is the feature with the highest negative correlation. Features that appear to have a low influence on price are the ones with a correlation close to 0, like `num_amenities`, `availability_90` and `host_identity_verified`.

- **Keyword Suggestions:** Important keywords for listing titles and descriptions were identified. These include words with high frequency in top listings (e.g., 'close', 'central', 'kitchen') and those with high differentiating ratios (e.g., 'Sköndal', 'tennis', 'Södermalm'). (See Figure 2 for examples of word cloud visualizations).
- **Amenity Recommendations:** Based on feature importance and commonality in top-performing listings, specific amenities are suggested to enhance a listing's appeal and potentially increase its value. These include features like 'Dishwasher', 'Blender', 'outdoor dining area', and 'Bathtub', as we can see in fig.3

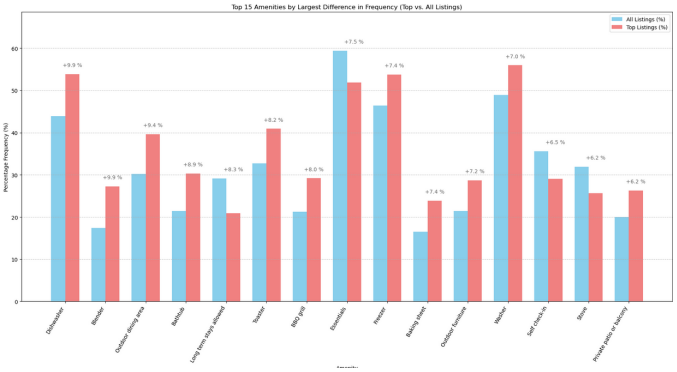

Fig. 3. Feature Importance (Top 20). This graph shows the features that have a greater difference, in percentage, between the "top" listings and all the listings. This is used to highlight those amenities that possibly contribute to making a listing a top one. In fact, we can see how the features that appear the most in the listings are all accessories or amenities that contribute to a sense of "non-home home". Finally, we notice how top listings tend to specify less the "essentials," which they probably take for granted, in order to convey more of a sense of coziness, where there is not a direct focus on essentials, but rather on amenities.

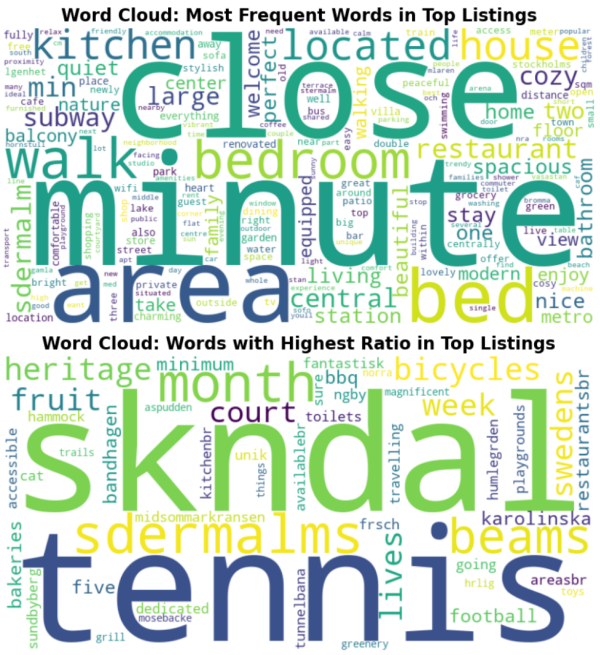

Fig. 2. Example Word Clouds: (Left) Top 20 Most Common Words in Descriptions, (Right) Top 20 Words with Highest Ratio in Top 10% Listings. 'skndal' and 'sdermalms' stands for the neighboroods of Sköndal and Södermalm, highlighting the influence that being close to these areas has on the fact that a listing is considered a top listing.

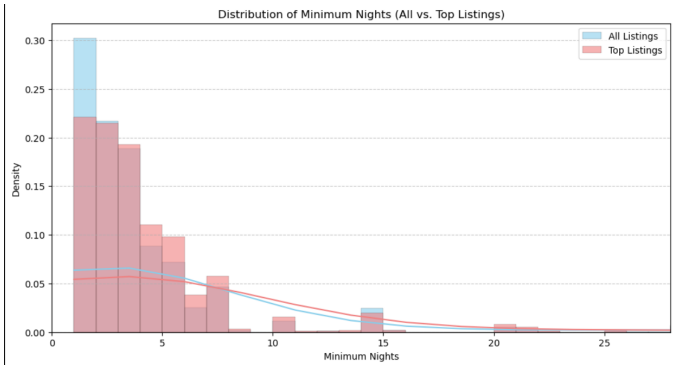

Fig. 4. Distribution of Minimum Nights for All Listings and Top Listings.

V. Analysis and Discussion

The analysis of correlation between features and price, highlights the significant influence of granular features on Airbnb pricing in Stockholm. Location, specifically latitude and longitude, consistently ranked among the most important features, confirming the strong impact of geographical positioning on rental value. This is further refined by the `neighbourhood_cleansed` feature, which captures localized market dynamics. The impact of location on prices was also visualized with geo-spatial graphs (fig.5.)

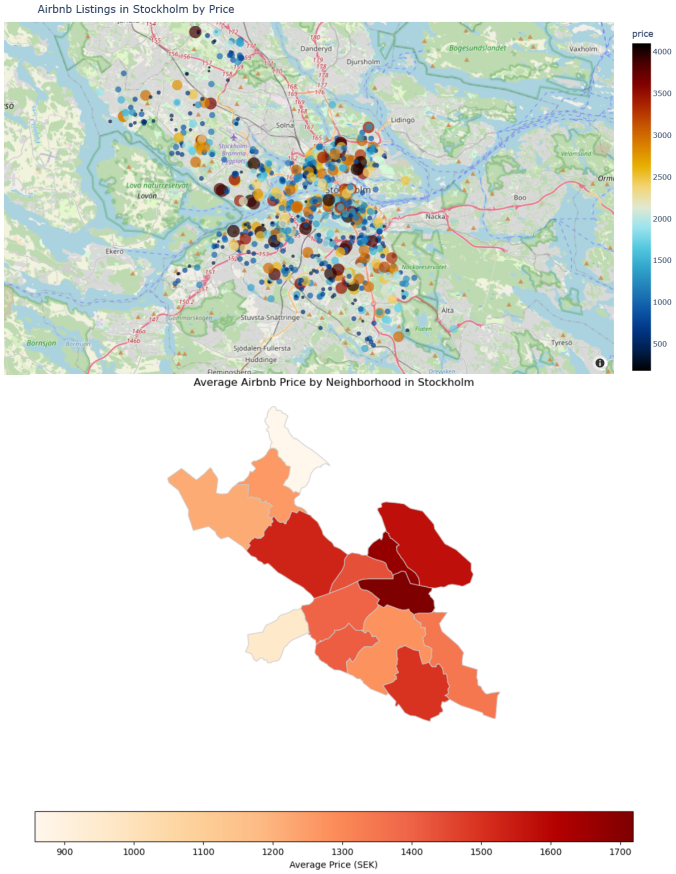

Fig. 5. Geo-spatial visualization of listing prices in Stockholm

The number of rooms (`accommodates`, `bedrooms`, `beds`) and bathrooms (`num_bathrooms`) are also primary drivers of price, intuitively reflecting the capacity and comfort level of the accommodation. The importance of amenities demonstrates that guests are willing to pay a premium for convenience and specific features, highlighting the value of amenities like dishwasher, outside dining area, and leisure facilities (e.g., `BBQ_grill`, `hot_tub`). The analysis of `amenities` not only served as input to the model but also informed direct recommendations for hosts, suggesting features that can differentiate a listing.

The textual analysis provides valuable insights into effective communication strategies. Keywords with high commonality and discriminative presence in high-performing listings offer practical insights for title and description development. This indicates that while general descriptive terms are important, specific keywords related to local attractions or unique aspects can significantly enhance a listing's appeal. Furthermore, from a brief analysis of the word clouds (Fig.2), we can deduce that in a listing it is also important to put focus on what is **close** in the **area** of the apartment, especially in a **walking distance** of a few **minutes**. Sentiment analysis of descriptions, although a less impactful feature in the model, provides a qualitative measure of the listings' presentation.

The recommendations on `minimum_nights` are derived from the observed patterns (fig.4) in successful listings, balancing host preferences with guest booking behavior. The

tool's ability to offer these data-driven recommendations, beyond just a price prediction, addresses a key problem for new hosts in a competitive market. The tool will recommend a value between 1 to 3 nights, specifying that top listing tends to prefer 2 or 3 nights, probably optimizing also the number of cleanings necessary for the apartment, while still being flexible for the guests.

Limitations and Validity Threats: While the model demonstrates strong predictive capabilities, several limitations should be acknowledged.

- **Data Freshness:** Airbnb data is dynamic. The model's accuracy relies on the recency of the training data. Market conditions, new developments, and changes in demand can influence optimal pricing over time. Continuous retraining with updated data would be necessary to maintain optimal performance.
- **Geocoding Accuracy:** The reliance on `Nominatim` for geocoding, while generally robust, can sometimes provide imprecise coordinates for very ambiguous addresses or when the address is outside of Stockholm's typical geocoding scope. This could lead to incorrect neighborhood assignments and impact location-based feature accuracy.
- **Amenity Categorization:** The extraction of amenities relies on parsing free-form text. While robust regex patterns were applied, nuances or new amenity descriptions could be missed, potentially under representing certain features.
- **Unseen Feature Combinations:** The model's performance on entirely novel combinations of features (e.g., a property type or amenity not present in the training data) might be less reliable, as it can only generalize from patterns observed during training. Additionally, given the extraction process of amenities, the user can only choose amenities from a vast but not infinite list of amenities. So a listing with a very unique amenity, that is not present on the list, might not benefit from its presence during price prediction
- **External Factors:** The model does not account for external, unquantifiable factors such as local events, host personality, specific photography quality (beyond what's implied by high ratings), or real-time market fluctuations that are not captured in the dataset.

These limitations highlight areas for future work, including implementing continuous learning pipelines and incorporating more sophisticated text or image analysis techniques.

VI. Conclusion

This project successfully developed a data-driven tool to assist prospective Airbnb hosts in Stockholm with pricing and listing optimization. By leveraging real-world Airbnb data, a LightGBM regression model was trained to predict nightly rental prices based on a comprehensive set of property characteristics, host attributes, and location data. The model demonstrated strong predictive power, with key features such as location, number of rooms, and specific amenities being significant drivers of price.

Beyond numerical prediction, the tool provides actionable recommendations for hosts, including optimal minimum stay durations, important keywords for listing titles and descriptions, and high-impact amenities to consider. These recommendations are derived from statistical analyses of high-performing listings, offering a competitive edge in the Stockholm market.

While limitations regarding data freshness and external factors exist, the developed tool provides a good foundation for informed decision-making in the short-term rental market. Future enhancements could involve incorporating real-time data feeds and more advanced deep learning for textual and image analysis.

References

[1] D. Wang and J. L. Nicolau, "Price determinants in the sharing economy: How airbnb hosts adjust their prices based on demand levels," *International Journal of Hospitality Management*, vol. 67, pp. 120–129, 2017.

[2] J. Gutiérrez, J. C. García-Palomares, G. Romanillos, and M. H. Salas-Olmedo, "Why do some airbnb listings become more popular than others? analyzing a listing's attractiveness using interpretable machine learning," *Journal of Travel Research*, vol. 56, no. 5, pp. 612–625, 2017.

[3] Q. Ke, Y. Yang, and X. Wang, "Attention-based neural network for learning to rank in e-commerce search," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1983–1986, ACM, 2018.

[4] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] U. Lalana, F. Muscherà, E. Isturitz, and P. Gasnier, "Airbnb-Data-Science: Sentiment Analysis." https://github.com/UnaiLalana/Airbnb-Data-Science/tree/master/Sentiment%20Analysis, 2025.

[7] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154, 2017.