Airbnb Price Prediction: Leveraging Sentiment Analysis and NER Enhanced Regression Model

Hanzhi Ding - 20199988, Roffy Shan - 20206876

Abstract—This research presents a comprehensive approach to predicting Airbnb rental prices in New York City by employing a multifaceted data analysis strategy and leveraging advanced machine learning techniques. The project begins with an extensive exploratory data analysis (EDA) of a dataset containing 39,202 Airbnb listings, encompassing a broad array of features ranging from basic property details to nuanced host characteristics and textual descriptions. Through meticulous data preprocessing, we address missing values and engineer new features such as sentiment scores derived from natural language processing (NLP) techniques, specifically sentiment analysis on textual descriptions, and named entity recognition to quantify amenities and location references. Additionally, we utilize one-hot encoding for the presence of popular amenities. To further refine our predictive capabilities, we experiment with several machine learning models, including Lasso Regression, Supported Vector Regression (SVR), and CatBoost Regression. These models are trained and evaluated based on Root Mean Squared Error (RMSE) and R^2 We observe that incorporating NLP-generated features, along with judiciously engineered structured attributes, enhances the predictive power of our models. Our findings indicate that location coordinates (longitude), accommodation capacity, the number of bathrooms, and NLP-created attributes significantly influence rental prices. The best-performing model, CatBoost, achieves an R-squared value of approximately 0.714, indicating a strong relationship between the engineered features and the target variable – the daily rental price. The study offers valuable insights into the complex dynamics regarding Airbnb pricing in New York City's short-term rental market.

Index Terms—Group 21, Short-term rental market analysis, sentiment analysis, named entity recognition, predictive analytics

1 Introduction

Airbnb is a popular C2C platform where property owners could list their properties for short-term rentals. However, pricing the rental property is challenging. Even if Airbnb provides suggested prices for hosts, the prices are straightforwardly calculated based on the average price in the hosts' regions. Therefore, hosts cannot justify if the listings that are used to generate the suggested price share similar features with their own properties or whether the suggested price is inflated or deflated to accommodate Airbnb's business interests. On the other hand, customers are unaware of the suggested prices when booking the properties, leaving them in a vulnerable position in the transaction as they usually lack insights into short-term rental prices at their travel destinations. Without a reference price, customers' demand is believed to be decreased [11], which jeopardizes the efficiency of the C2C platform and thus leads to a loss in the sharing economy.

Motivated by the aforementioned inequalities faced by both hosts and customers, we are interested in developing an objective predictive model using state-of-the-art regression algorithms to output an unbiased suggested price for hosts on Airbnb, taking basic rental criteria (e.g., location, listing types, rooms, amenities, ratings, etc.), property descriptions and host information (text mining of the property, neighborhood, and host description provided by hosts) as the independent attributes. Customer reviews, which may have potential predictive power, are excluded from the modeling process at the current stage due to restraints in computational resources.

Published academic literature on rental price prediction adopted regression algorithms (e.g., linear regression, Lasso regression, and boosting algorithms) and had limited involvement in text mining of property information, which mainly focused on sentiment analysis. In our research, we capture the subtle unstandardized details proposed by the host using Named Entity Recognition (NER) and part-of-speech tagging (POST), enabling more robust prediction in Airbnb pricing in New York City based on the assumption that the cleansed dataset is a representation of the actual market trends. The main contributions of this work are:

- We propose a pipeline to retrieve listings' amenities and geographical information from hosts' descriptions of their properties at a higher granularity compared to the standardized location information on the Airbnb platform.
- We evaluate which regression model performs the best in predicting Airbnb prices with our dataset augmented by sentiment analysis and text mining, and achieve an improvement from the existing literature with an R^2 of 0.7145.
- We evaluate the possible influence of hosts' writing styles on listings' prices to provide guidance on proper writing techniques on the Airbnb platform.

• *Hanzhi Ding and Roffy Shan are with Smith School of Business at Queen's University*
E-mail: hanzhi.ding@queensu.ca, roffy.shan@queensu.ca

2 Related Work

In the corpus of studies examining Airbnb pricing and sentiment analysis, four distinct scholarly works have made substantial contributions. Each paper employs advanced machine learning and natural language processing techniques to unravel the complex dynamics underlying Airbnb rental prices and the influence of user sentiment.

Airbnb Price Prediction Using Machine Learning and Sentiment Analysis [13] developed a predictive model for Airbnb rental prices in New York City by integrating various machine learning algorithms with sentiment analysis of customer reviews. The authors assembled a dataset consisting of property specifications, owner details, and user feedback. They incorporate sentiment scores using TextBlob as a feature derived from customer reviews, discovering that the Support Vector Regression (SVR) model, when fortified with sentiment analysis, achieved the highest predictive accuracy, as evidenced by an R^2 score of 0.6901.

A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis [1] focused on creating a more sustainable model for Barcelona's Airbnb price prediction by considering a range of factors, including property specifications, owner attributes, and sentiment analysis on customer reviews. Specifically, the study involved a more extensive conversion of categorical features into numerical form; for instance, amenities are formatted into binary variables representing whether a particular amenity was provided. The findings suggested that the inclusion of sentiment analysis improved the predictive power of the model, enabling hosts to better estimate the value of their listings based on the opinions expressed in customer reviews.

Airbnb Price Prediction with Sentiment Classification [8] focused on improving the accuracy of Airbnb price prediction models by integrating sentiment classification with TextBlob. Customer reviews are categorized into positive, neutral, and negative sentiments. The distribution of sentiments showed the association between sentiment and price was unclear: listings with predominantly positive and neutral reviews had a 0.339 higher normalized average price compared to those with negative comments. Overall, the research demonstrated that incorporating sentiment analysis into the price prediction model allowed for a deeper understanding of how users' perceptions and opinions, as captured in reviews, contribute to the dynamic pricing behavior observed in the Airbnb market.

Performance Analysis of Deep Approaches on Airbnb Sentiment Reviews [12] assessed the performance of deep learning algorithms, specifically Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), and Gated Recurrent Units (GRUs), for sentiment analysis of Airbnb reviews. The authors sought to identify and evaluate different aspects of guest reviews with respect to their accuracy in sentiment classification. They discovered that the GRU architecture outperformed RNNs and LSTMs in sentiment classification tasks, emphasizing its suitability for extracting sentiment signals that can guide hosts in refining their services and rental pricing strategies.

In conclusion, these four research endeavors collectively highlight the significant role that sentiment analysis and machine learning play in understanding and predicting Airbnb rental prices. By examining unstructured data of customer reviews and integrating this information with quantitative listing attributes, the studies have expanded the knowledge base on Airbnb price determination, empowering hosts to make more informed decisions and fostering a more sustainable and equitable rental market ecosystem.

3 Dataset

The data analyzed in this project originate from a public dataset of Airbnb listings in New York City, scraped from the Airbnb website [7]. The dataset contains 39,202 records across 75 columns, providing a wealth of information on each listing, including source details, accommodation specifics (e.g., property type, location, and amenities), and extensive information about the hosts, such as their self-introduction, responsiveness, and experience level.

During the preliminary preprocessing stage, several steps were taken to clean, transform, and prepare the data for further feature engineering and modeling:

- Redundant Column Removal

Irrelevant or duplicate columns, such as `listing_url`, `scrape_id`, and `last_scraped`, were removed to streamline the dataset.

- Missing Value Handling

Missing values are addressed, starting with the target variable `price`. We strictly remove rows with missing values due to their indispensability in the prediction task.

We also refer to correlations in missing values as shown in Figure 1, for instance, `host_response_time` and `host_response_rate` columns were found to have high correlation in missing values due to the unstandardized way these metrics are displayed on Airbnb, which leads errors when the dataset was scraped. Consequently, the two columns are properly discretized, and missing instances were replaced with 'Unknown' as they appear to be missing completely at random (MCAR). In terms of missing values in columns related to customer reviews, for instance, `review_score_rating`, are deemed missing not at random (MNAR) as newly listed properties without transactions are impossible to have customer-related information. We derive an assumption that when customers see a listing with no review scores, they will regard the listing's quality as mediocre based on customer perceived value (CPV) theory [5]. Thus, we fill the missing values with the median of each column instead of using predictive imputation methods, which, though they might yield better performance, are not consistent with the assumption.

For other categorical columns, for example, `neighbourhood`, We replaced missing values with a generic category 'Unknown', which allowed us to retain as much information as possible without making strong assumptions.

For other numerical columns, for example, `beds` and `bathrooms`, missing values are filled with 0 as we discover from the Airbnb website that empty values in these columns indicate that the listing is not equipped with such facilities.

- Date Conversion

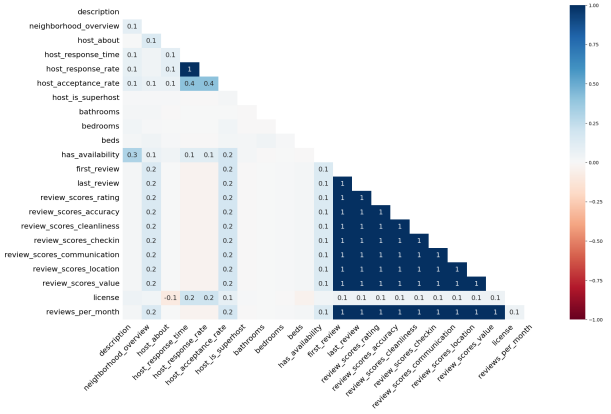

Fig. 1. Missing Value Correlation

`host_since` are converted to a datetime format and used to calculate the experience of hosts in years (`yrs_exp`), which was deemed a potentially important predictor variable.

- Outlier Handling

`price` were cleaned by removing currency symbols and commas before converting them to floats. Referring to the distribution shown in Figure 2, we remove listings with prices below $50 and above $1100 to ensure the model's robustness. The outlier removal is validated by checking the current price range of Airbnb listings in New York City, as the minimum price is $60, and the maximum price is $1091 [6].

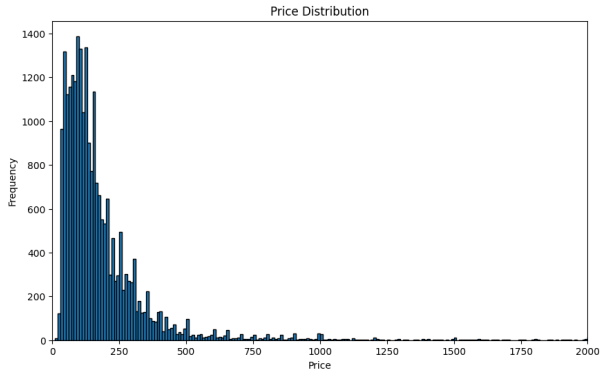

Fig. 2. Price Distribution

- Amenities Extraction

The processing of the amenities feature in the dataset starts with transforming the string representation of a list into an actual list using Python's `eval()` function. Subsequently, the list is flattened to count the occurrences of each amenity using the `Counter` class from the `collections` module. This count data is then organized into a Pandas DataFrame for better visualization and analysis. Next, to reduce dimensionality and focus on the most prevalent amenities, the DataFrame is filtered to include only those amenities that occur above a specified threshold (600 times among all observations). After filtering, one-hot encoding is applied to create binary features for each filtered amenity, assigning a value of 1 when the amenity is present in the property listing and 0 if absent. Lastly, the newly generated dummy variables are converted to the category data type for optimized memory usage and computation. Overall, this sequential approach converts the raw, unstructured amenities data into a structured format that can be effectively utilized in statistical models and machine learning algorithms to analyze and predict various aspects of property listings.

Column Name	Description
description	Detailed description of the property written by host
neighborhood_overview	Summary of the neighborhood written by host
host_about	Host's self-introduction
host_response_time	Host's typical response time
host_response_rate_bins	Discretized host's response ratio
host_acceptance_rate_bins	Discretized rate at which host accepts reservations
host_is_superhost	Indicator if host is a Superhost
host_total_listings_count	Total listings managed by host across all cities
neighbourhood_group_cleansed	Neighborhood group classification
latitude & longitude	Geographic coordinates
accommodates	Maximum number of guests allowed
bedrooms	Number of bedrooms
property_type	Type of property
room_type	Type of room offered (Entire home/apt, Private room, Shared room)
review_scores_ (7 features)	Average rating from guests, including cleanliness, communication, etc.
reviews_per_month	Average monthly review count
amenity_dummies (118 features)	Dummy features documenting the presence of popular amenities

TABLE 1
A Section of Independent Features

4 METHODOLOGY

4.1 RQ1: How to distill information from property descriptions and host information?

Motivation: Property descriptions and host information contain valuable information regarding some qualitative metrics of the listings (e.g., decoration quality, living comfort, host's characteristics, etc.) that might be correlated with the price, the extraction of which may provide valuable insights for hosts when building their profiles on Airbnb.

- Evaluating the quality of a listing relies heavily on the content of the description, which includes details about the property, amenities, location, and overall appeal. Such information adds a subjective layer of feedback that reflects actual guest experiences and satisfaction.
- Understanding the sentiment expressed in both descriptions and the host's self-introduction can help gauge the emotional appeal of a listing. Positive or negative sentiment could indicate strong preferences among customers and thereby impact demand.
- Distilling keywords, phrases, and entities can possibly inform recommendation systems to suggest similar or relevant listings to users based on their search history or interests.
- Analyzing descriptions and reviews can help segment the market according to various property types, neighborhood qualities, and guest profiles, enabling targeted marketing strategies.

Proposed Methodology:

1. Preprocess the text data by tokenizing, removing stop words, and applying word embeddings (e.g., pre-trained tokenizer).

2. Utilize NLP techniques to extract relevant features (e.g., sentiment scores, geographical information, writing

styles, etc.) from property descriptions and the host's self-introduction.

- NER: Use NLP libraries such as NLTK to identify named entities like locations, facilities, or amenities mentioned in the descriptions and reviews. This step helps quantify the importance of certain features and geographical areas [3].
- Sentiment Analysis: Apply sentiment analysis algorithms to calculate positivity, negativity, and neutrality scores for the entire description and the host's self-introduction [2].
- POST: Use part of speech tagging algorithms to extract the usage of specific adjectives, verbs, nouns, and foreign words, which illustrate the host's specific writing styles.

4.2 RQ2: Do the proposed independent attributes (basic rental criteria, property descriptions, host information, and customer ratings) objectively reflect the actual value of the rental (price per night)?

Motivation: Evaluating the alignment between predicted and actual market prices is essential for understanding the accuracy of our predictive model and its potential to address asymmetric information in the market.

The motivation for examining whether the suggested independent attributes (including fundamental rental criteria, property descriptions, host details, and customer ratings) accurately represent the true value of a rental (price per night) extends beyond mere curiosity; it holds critical implications for creating a robust predictive model that can help tackle issues around market inefficiency and inequality. By thoroughly evaluating the alignment between the estimated prices derived from these attributes and the actual market prices, we aim to validate the model's effectiveness in capturing the underlying value drivers and dynamic pricing mechanisms of short-term rental properties.

An accurate predictive model enables stakeholders, such as property owners, to make informed decisions about pricing strategies, ultimately contributing to a more transparent and balanced marketplace. Moreover, by addressing potential discrepancies between predicted and actual prices, we can identify potential factors that could contribute to further improvements in the model's performance. This could highlight disparities in the treatment of various properties or hosts based on factors unrelated to the quality or value provided to guests. For instance, a model that successfully accounts for the influence of host responsiveness, property location, and customer reviews on price could reveal cases where listings with comparable features are undervalued or overpriced due to market asymmetries or lack of information transparency.

Ultimately, refining our model to better align predicted and observed prices empowers us to build a more equitable system where properties are priced according to their true market worth, reducing market inequalities and enhancing customer satisfaction. It also allows for interventions that can correct misaligned valuations, benefiting both consumers and providers in the short-term rental economy. Through rigorous evaluation and continuous improvement of our model, we can foster a more rational and fair market environment while providing actionable insights for strategic decision-making in the sector.

Proposed Methodology:

1. We assume that the current price per night is the equilibrium price, which is an unbiased reflection of the market value of each rental.

2. We conduct extensive Exploratory Data Analysis (EDA) to analyze the distribution of critical independent attributes and the price per night according to domain knowledge. For example, neighborhood, property types, customer review scores, and host responsiveness.

3. We build various regression models, including CatBoost, Lasso, and SVR using the identified independent attributes as input features. Split the dataset into training, validation, and testing subsets to avoid overfitting and ensure the model's generalizability. Perform hyperparameter tuning to find the optimal model configuration [4] [10].

4. We evaluate the model's performance using RMSE and R^2. A lower RMSE, along with a higher R^2, would indicate a better fit and suggest that the chosen model is capturing a significant portion of the variance in rental prices.

4.3 RQ3: How do the proposed independent attributes impact the price of rentals, especially the attributes created by NLP techniques? What are the most important aspects of them?

Motivation: In terms of value perception, we want to find out if detailed and appealing property descriptions actually attract more interest, leading to increased demand and justifying a higher asking price, and whether listings that emphasize unique features, prime locations, or desirable amenities tend to command better prices. On the other hand, we want to quantify the role that social trust and credibility play during renting; in other words, we want to determine whether reviews from previous guests can act as social proof, establishing credibility and trustworthiness.

We believe the research of the question will mitigate risks in that potential customers use previous customer reviews to assess the likelihood of a satisfactory stay, reducing perceived risk. Rentals with consistently high ratings may be priced higher due to the reduced uncertainty associated with the booking.

The common belief that well-written descriptions and favorable reviews differentiate listings from competitors, which can translate into a pricing advantage for unique, well-maintained, or highly reviewed properties, motivates us to delve into the question.

Lastly, we postulate that the feature importance obtained is crucial for both hosts looking to optimize their listings and customers aiming to understand market dynamics. Property descriptions can offer explicit information on factors like accommodation capacity, nearby attractions, and unique selling points.

Proposed Methodology:

Create plots to visually illustrate the relationship between the rental price and prominent independent attributes and provide insights for hosts on building Airbnb profiles.

5 Experiments and Results

5.1 Experiment Setup

After experimenting with different combinations of the original attributes and the NLP-created attributes mentioned above with CatBoost, we recognize that all the NLP-created attributes have feature importance. Therefore, we decide to train, tune, and evaluate the regression models on two datasets. One contains only the original attributes with `amenity_dummies`, excluding unstructured `description`, `host_about`, and `neighborhood_overview`; one contains the original attributes and augmented by the NLP-created attributes. Both datasets have 21,583 observations, and the former one contains 162 independent attributes, the latter one contains 180 independent attributes. The datasets are split into training (70%), validation (10%), and testing (20%) sets to evaluate the performance of various regression models, including Lasso, SVR, and CatBoost. Hyperparameter tuning was performed to optimize these models, aiming to strike a balance between model complexity and predictive power. The ultimate success of the models was measured using RMSE and R^2.

5.2 Results and analysis

5.2.1 RQ1: How to distill information from property descriptions and host information?

Our text mining consists of three techniques, sentiment analysis, NER, and POST, which pivot around `amenities`, `description`, `host_about`, and `neighborhood_overview`.

- `amenities`: NER is conducted on this attribute to further extract insights after we retrieve and store the most popular amenities in one-hot attributes in an effort to prevent loss of information. Utilize wordnet from NLTK, we conclude the synonyms of the most popular amenities which is used as a vocabulary list to recognize the less popular amenities that are excluded from one-hot attributes (`amenity_dummies`). We store the number of these amenities in `recognized_amenities_count` for each listing. Also, `num_amenities` documents the total number of amenities provided by each listing. In summary, for each listing, `amenity_dummies` label the most popular amenities, `recognized_amenities_count` counts the presence of less popular amenities, and `num_amenities` counts the total number of amenities.

- `description`: Sentiment analysis, NER, and POST are conducted on this attribute, as it is rich in both objective information (e.g., geographical information of the listing) and subjective information (e.g., the host's writing style). We derive the negativity, neutrality, and positivity scores of `description` as numerical attributes, extract `named_place_count` to store the frequency of mentioning geographical information of the listing, and document the occurrence of nouns, verbs, adjectives, and foreign words as an additional illustration of the host's writing styles.

- `host_about`: Sentiment analysis and NER are conducted, focusing on extracting the writing style of the host due to the lack of property-related information in this attribute. Similar to `description`, the negativity, neutrality, and positivity scores are derived, and the usage of adjective words is documented.

- `neighborhood_overview`: Sentiment analysis and POST are not conducted as we assume the information is subjective and the presence of different writing styles is not obvious. Similar to `description`, the negativity, neutrality, and positivity scores are derived, and we extract `neighbor_place_count` to store the frequency of mentioning geographical information of the listing's neighborhood.

ANOVA Test	neighbourhood _cleansed (smaller neighborhood)	neighbourhood _group_cleansed (larger neighborhood)
F-statistic	1.372	20.774
P-value	2×10^{-4}	4.11×10^{-17}

TABLE 2
ANOVA Test: Neighborhood and Price

5.2.2 RQ2: Do the proposed independent attributes (basic rental criteria, property descriptions, host information, and customer ratings) objectively reflect the actual value of the rental (price per night)?

After conducting an extensive Exploratory Data Analysis (EDA) on the Airbnb dataset, the following insights were gained:

Neighborhood Distribution and Price (Table 2): We conduct ANOVA tests to evaluate if the means of the rental price is different across various neighborhoods and discover that the listing's neighborhood does not have a significant impact on the price. This could potentially be due to the fact that hosts might set the price based on attributes other than the neighborhood. For instance, based on our discovery on the Airbnb website, in some expensive neighborhoods, there might be disturbingly cheap listings for basements, sofa beds in the living room, etc. In other words, Airbnb pricing is much more delicate than property prices in the real estate market, which requires sophisticated modeling.

Property Types and Pricing: Different property types, such as entire homes, private rooms, and shared rooms, displayed distinct price distributions. Entire homes generally commanded the highest prices, while shared rooms were priced lower.

Customer Review Scores and Host Responsiveness: Listings with higher average review scores and quicker host response times were associated with higher rental prices, indicating a premium on positive guest experiences and efficient communication.

Upon building and evaluating regression models with Lasso, SVR, and CatBoost, the following model performances were noted (Table 3):

Model name	Original Dataset		NLP Augmented Dataset	
	RMSE	R^2 Score	RMSE	R^2 Score
Lasso	123.355	0.232	123.305	0.162
SVR	145.163	-0.064	145.163	-0.064
CatBoost	78.027	0.693	75.175	0.715

TABLE 3
Model Performance

- CatBoost Model Performance: The model achieved an RMSE of 78.027 and an R^2 value of 0.693 on the original dataset. Comparatively, an RMSE of 75.175 and an R^2 value of 0.715 is achieved on the NLP augmented dataset, signifying that the NLP-created attributes captured an additional portion of the variance in rental prices.
- Lasso Model Performance: The model's performance is lower than CatBoost, possibly due to the prominent non-linearity in the dataset with a considerable amount of categorical attributes. However, as the RMSE value decreased from 123.355 on the original dataset to 123.305 on the NLP augmented dataset, it also confirms the potential predictive power of the NLP-created attributes. Additionally, the decrease in R^2 value suggests Lasso's inefficiency in handling datasets with high dimensionality where all attributes have certain predictive power.
- SVR Performance: The model's performance is the lowest among all models, and no improvements are observed when NLP-created attributes are added. The result is potentially due to the model's rigorous regularization setting, which excludes attributes with relatively small predictive power.
- Overall, the results suggest that regression modeling successfully identified and quantified the impact of various independent attributes on rental prices. In particular, the text mining efforts help the regression models capture more variance in the New York City Airbnb market.

5.2.3 RQ3: How do the proposed independent attributes impact the price of rentals, especially the attributes created by NLP techniques? What are the most important aspects of them?

By utilizing SHAP, we are able to identify the most influential attributes affecting the rental price (Figure 4), showcasing their direct contribution to either increasing or decreasing the estimated cost. For instance, attributes like location coordinates (longitude), accommodation capacity, and number of bathrooms consistently ranked high in the positive impact on the rental price [9].

All of the NLP-created attributes contribute positively to the predicted price at different magnitudes. It is notable that negativity in writing styles is of greater importance than positivity. The counterintuitive fact could be explained by the customer's perception of social trust and credibility, which implies that the hosts should be objective and critical about describing their properties and neighborhoods. Correspondingly, the number of nouns used in property descriptions also positively contributes to the predicted price at a relatively large magnitude, suggesting hosts should exhaustively list the characteristics of their properties, such as the number of amenities and the named locations surrounding the neighborhood.

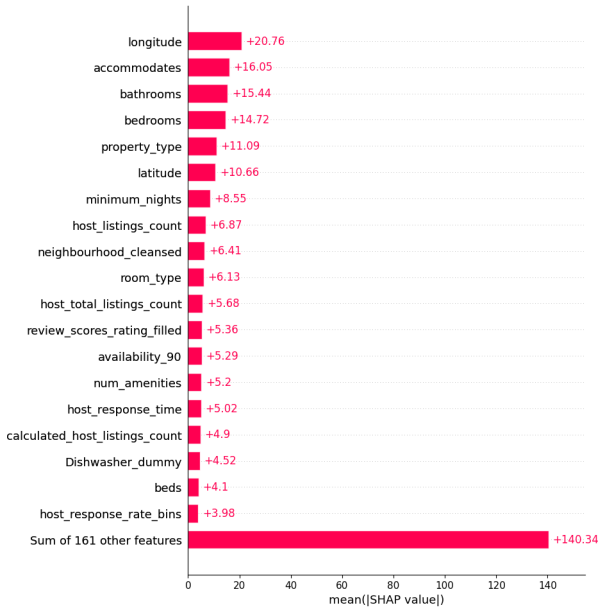

Fig. 3. CatBoost SHAP Values

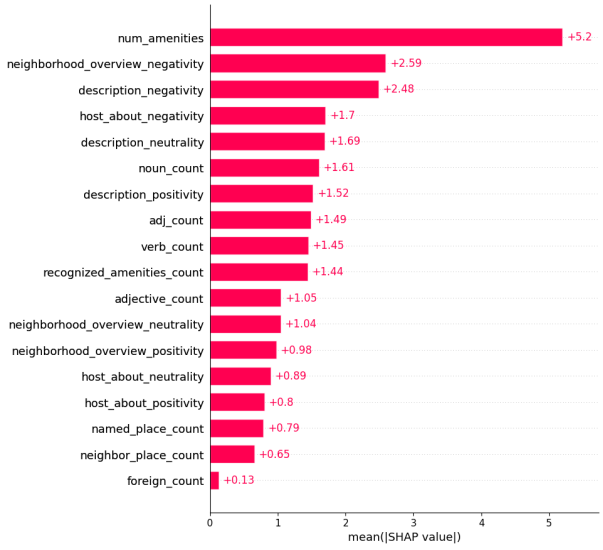

Fig. 4. Catboost SHAP Values – NLP-created Attributes

In conclusion, employing SHAP in answering RQ3 added a layer of transparency to the model predictions and facilitated the interpretation of how different independent attributes interact and combine to determine the rental price of a listing. This not only strengthened the credibility of the quantitative findings but also provided actionable insights for stakeholders, such as hosts and Airbnb management, to make strategic decisions based on the factors driving rental prices in the marketplace. Overall, the experiments provide clear evidence that NLP-created attributes from property descriptions and hosts' self-introduction significantly contribute to the rental price, reinforcing the importance of

crafting compelling narratives and maintaining a positive online reputation.

6 Group Member Contributions

At the initial stages of the project, Hanzhi Ding and Roffy Shan collaborated closely to conduct a thorough exploratory data analysis (EDA) on the New York City Airbnb dataset. Both members were actively involved in understanding the features present in the dataset, assessing their quality, and deciding on the preprocessing steps required to prepare the data for modeling.

Hanzhi Ding and Roffy Shan equally contributed to tasks such as data cleaning, handling missing values, and feature engineering, which included converting categorical variables into usable formats, deriving new features, and possibly dealing with skewed data distributions. They also participated in the joint decision-making processes concerning which features to include and exclude based on their potential impact on the final predictive models.

As the project progressed into more specialized areas, the responsibilities became more distinct. Hanzhi Ding took charge of building and optimizing regression models to predict Airbnb listing prices. This involved selecting appropriate algorithms (such as CatBoostRegressor and RandomForestRegressor), setting up the experimental design, conducting cross-validation, and fine-tuning hyperparameters to maximize model performance.

On the other hand, Roffy Shan focused on Natural Language Processing (NLP) and text-mining aspects of the project. This included tasks like tokenizing and analyzing the textual data in the descriptions and neighborhood overviews to extract meaningful insights, using tools including sentiment analysis, NER, and POST, which are incorporated into the final regression models to enhance their predictive capabilities.

7 Replication Package

Link to GitHub repository for datasets and code

8 Conclusion and Future Work

The project focused on a dataset of Airbnb listings in New York City, which contains 39,202 records with 75 attributes detailing various aspects of each listing, such as property characteristics, host information, and customer review scores. The EDA involved handling missing values and creating new features.

Numerical features were analyzed for skewness and kurtosis to guide the choice of models. Tree-based models were favored due to the skewed and heavy-tailed distributions of some features. Textual data from property descriptions and amenities were processed to extract meaningful information, with amenities being one-hot encoded into binary variables.

The team trained a CatBoost regression model using a combination of numerical, categorical, and engineered features. During model training, the best iteration was recorded, and the model was shrunk to contain only those iterations up to the best-performing point. The best model achieves an RMSE value of 75.175 and an R^2 value of 0.715 on the testing set. The unexplained variance could result from excluded attributes, for example, customer reviews and photos of the listings.

8.1 Limitations

1. Nature of the Dataset: The dataset used in this research is created by scrapers, which might have irrecoverable loss of information and errors.

2. Assumptions: The decision to treat listings with no review scores as average is based on customer perceived value theory, which is not a data-driven assumption. Future work could explore alternative imputation methods grounded in machine learning or further investigate the relationship between missing review scores and listing characteristics.

3. Feature Selection: Although numerous features were included in the model, a more systematic feature selection process could be implemented to mitigate computation costs and increase scalability on larger datasets.

4. Overfitting: The discrepancy between the R^2 values on the training set (0.984) and testing set (0.715) suggests the CatBoost regression model is overfitting on the training data. However, if we address the overfitting with appropriate hyperparameters, the R^2 values on the testing set will decrease to 0.682. Therefore, we assume the current hyperparameters ensure the maximum generalizability.

8.2 Improvements for Future Design

1. Advanced Imputation Techniques: Implement advanced imputation techniques for missing values in the dataset, which may include KNN imputation, mean/mode imputation based on similar listings, or deep learning-based imputation methods.

2. Feature Engineering: Incorporate text mining on customer reviews and image captioning on photos of the listings.

3. Robustness Checks: Verify the model's robustness to outliers and missing data patterns by performing sensitivity analyses and comparing results across multiple imputation methods.

4. Qualitative Insights: Combine quantitative analysis with qualitative insights gathered from interviews or surveys with hosts and guests to enrich the interpretation of the data and improve the relevance of the model to real-world scenarios.

References

[1] Z. H. Alharbi. A sustainable price prediction model for airbnb listings using machine learning and sentiment analysis. volume 15, 2023.

[2] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, Nov. 2020. Association for Computational Linguistics.

[3] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009.

[4] A. V. Dorogush, V. Ershov, and A. Gulin. Catboost: gradient boosting with categorical features support. https://catboost.ai, 2018. Accessed: ¡insert date here¿.

[5] M. Holbrook. *Consumer Value: A Framework for Analysis and Research (Routledge Interpretive Market Research)*. Routledge, Abingdon, United Kingdom, 1st edition, 1999.

[6] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[7] Inside Airbnb. Airbnb public dataset. http://insideairbnb.com/get-the-data.html, 2024. Accessed: 15 Mar 2024.

[8] P. Liu. Airbnb price prediction with sentiment classification. Master's projects, San José State University, 2021.

[9] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[10] 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 E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] D. S. Putler. Incorporating reference price effects into a theory of consumer choice. In *Marketing Science*, pages 287–309. INFORMS, 1992.

[12] M. R. Raza, W. Hussain, and A. Varol. Performance analysis of deep approaches on airbnb sentiment reviews. In *2022 10th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–5, 2022.

[13] P. Rezazadeh Kalehbasti, L. Nikolenko, and H. Rezaei. Airbnb price prediction using machine learning and sentiment analysis. In *Machine Learning and Knowledge Extraction*, page 173–184. Springer International Publishing, 2021.