Cloud-Based Airbnb Recommendation System

Xueyuan Li
Data Science
Vanderbilt University
Nashville, USA
xueyuan.li@vanderbilt.edu

Oscar Yu
Data Science
Vanderbilt University
Nashville, USA
zihao.yu@vanderbilt.edu

Xishan Deng
Data Science
Vanderbilt University
Nashville, USA
xishan.deng@vanderbilt.edu

***Abstract*—This paper presents the development and evaluation of a hotel recommendation system for Airbnb, utilizing collaborative and content-based filtering techniques to enhance personalized guest experiences. Leveraging the scalable infrastructure of Google Cloud and PySpark's distributed data processing capabilities, the system addresses the competitive demands of the hospitality industry. We first implement collaborative filtering methods such as Alternating Least Squares (ALS) and Singular Value Decomposition (SVD), achieving significant improvements in prediction accuracy with an RMSE as low as 0.89. Subsequently, content-based filtering is employed, utilizing item features and user preferences to predict hotel ratings, achieving an RMSE of 0.3741 and high precision, recall, and F1-score metrics. This dual approach not only improves recommendation precision but also aids in optimizing offerings, pricing, and marketing strategies, thereby driving growth and profitability in the hospitality sector. The results demonstrate the system's capability to deliver highly personalized accommodation recommendations, thereby enhancing user experience and fostering customer loyalty.**

***Index Terms*—PySpark, ALS, Recommendation System, Collaborative filtering, Content-based filtering**

I. Problem Statement

The hospitality industry is intensely competitive, making personalized guest experiences crucial for success. Our hotel recommendation system for Airbnb addresses this need by leveraging machine learning on Google Cloud's scalable infrastructure and PySpark's distributed data processing capabilities.

The system employs collaborative filtering to analyze user behavior patterns and content-based filtering to match listing attributes with user preferences. This dual approach enables highly tailored accommodation recommendations that significantly enhance user experience.

Providing such personalization fosters customer loyalty, increases bookings and revenue. Moreover, mining large data volumes uncovers valuable consumer insights for optimizing offerings, pricing, and marketing strategies - driving growth and profitability.

Implementing this advanced, cloud-based recommendation system positions our business as an innovator in the hospitality sector, attracting customers seeking exceptional experiences and top industry talent.

By solving a critical industry pain point through technological innovation, data-driven insights, and unparalleled personalization, our system presents a compelling business opportunity amidst fierce competition.

II. Related Work

In recent years, the application of machine learning technologies in the hospitality industry has seen significant advancement, particularly in enhancing personalized guest experiences through recommendation systems. Previous studies and existing systems often focus on collaborative and content-based filtering methods, much like the approach described in the proposed system for Airbnb.

One such example is the system implemented by Expedia, which utilizes machine learning to provide personalized travel recommendations (Melián-González et al., 2019). Their model analyzes historical booking data and user preferences to suggest hotels and travel options tailored to individual tastes. Similarly, Booking.com employs machine learning algorithms to optimize search results based on user interactions and feedback, thereby improving customer satisfaction and engagement (Levi, 2018).

Academic research has also contributed to this field. A study by Xie et al. (2016) highlighted the effectiveness of hybrid recommendation systems that combine collaborative and content-based filtering. This research found that hybrid systems are particularly adept at addressing the cold start problem common in collaborative filtering, by integrating content-based attributes from the onset.

Furthermore, leveraging large-scale data processing technologies like PySpark in cloud environments, such as Google Cloud or AWS, has proven beneficial. Zhang et al. (2020) demonstrated the scalability of PySpark for processing extensive datasets in real-time, which is critical for dynamic and responsive recommendation systems in high-demand sectors like hospitality.

In addition, the use of advanced analytics and data mining techniques to derive insights from user data has been extensively documented. For instance, Gavurova et al. (2021) explored how data-driven decision-making in the hospitality sector could lead to optimized pricing strategies and enhanced market competitiveness.

However, while these systems provide a robust framework for understanding user preferences and behavior, challenges remain. Issues such as data privacy, the accuracy of recommendations in the face of sparse data, and the need for real-

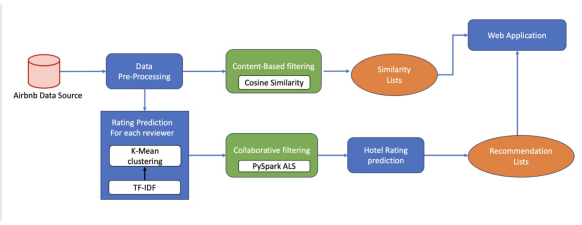

Fig. 1. This figure shows framework of the hotel recommendation system.

time processing capabilities are areas that continue to require innovative solutions.

The proposed system aims to build upon these foundational works by integrating state-of-the-art machine learning models with scalable cloud infrastructure, thereby addressing some of these ongoing challenges. By doing so, it seeks to offer a novel contribution to the field by enhancing the precision of recommendations and thereby, improving the overall user experience in the hospitality industry.

This body of related work underscores the importance of technological advancements in recommendation systems and sets a solid foundation for further innovation in personalizing guest experiences in the hospitality sector.

III. Methods

The Figure 1 shows the framework of the hotel recommendation system.

A. Data Pre-processing

1) Exploratory Data Analysis: Our exploratory data analysis provides insightful delineation of listings' geographical spread and users' booking predilections.

- **Geographic Distribution of Listings:** As depicted in Figure (2), our listings data encompass a broad geographic scope, encompassing major urban centers across the United States. The dispersion of listings is particularly dense in metropolitan areas, such as the West Coast cities of Los Angeles and San Francisco, as well as East Coast cities like New York and Boston. This spatial analysis indicates a strong presence in both coastal and inland cities, reflecting diverse user demographics and a wide-ranging market reach. The distribution suggests varied user preferences, potentially influenced by factors like urban tourism, business travel, and regional attractions.

- **Consolidated Booking Insights:** As depicted in Figure (3), availability trends point to a user inclination towards weekend stays and seasonal booking patterns, while a notable decline in listings availability beyond a six-month forecast reflects host practices of short-term listing horizons.

The data informs two potential enhancements for the recommendation system: prioritization of listings with high weekend availability and prompts for hosts to update listings, improving long-term booking viability.

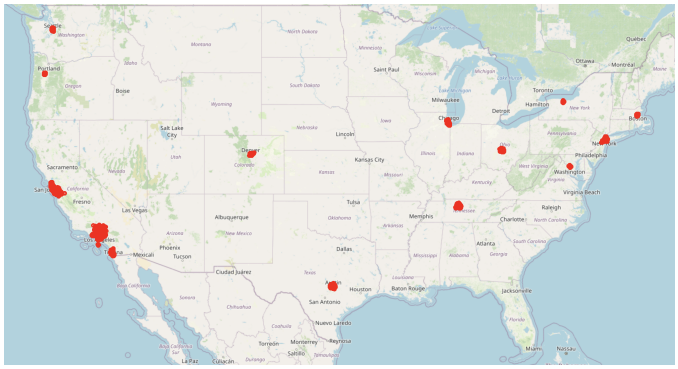

Fig. 2. A Geographic Overview of our listings

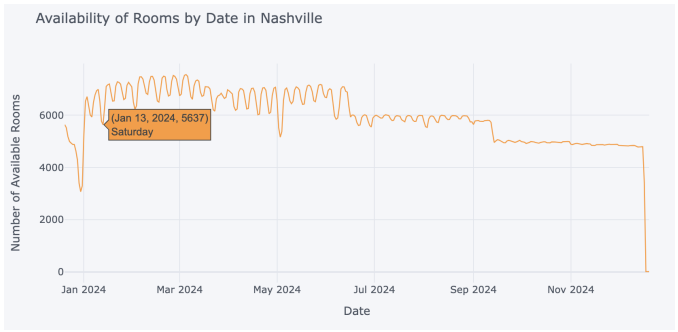

Fig. 3. Availability of rooms by Date in Nashville

2) Features Engineering: During the feature engineering stage, our goal was to refine our dataset to improve the predictive capabilities of our Airbnb recommendation system. We applied several techniques to transform raw data into informative features that could better capture the nuances influencing user preferences.

- **Incorporating Location-Based Features:** Recognizing the pivotal role of geographic location in accommodation selection, we introduced a "location rating" feature. This novel feature was engineered using latitude and longitude coordinates to quantify the desirability of a listing's location. Using the Haversine formula, we calculated each listing's distance to the central business district of Nashville, a proxy for the city's downtown area. We normalized these distances to a 0-1 scale to create a uniform measure of location attractiveness across all listings.
- **Addressing Price Variation Across Regions:** Our exploratory analysis uncovered substantial regional variations in average listing prices, with some locales, such as Hawaii, exhibiting averages above 400USD, and others, like Rochester, averaging near 100USD. To normalize this disparity, we segmented listing prices into quartiles, assigning each listing to a category based on its relative price position within its city or region. These categories were then encoded using one-hot encoding, allowing our recommendation model to discern price tiers without being skewed by the wide range of absolute price values.

In this section, we outline the methodologies employed in developing the rating prediction model for user reviews. The process involves two main stages: unsupervised learning for feature extraction and clustering, and supervised learning using collaborative filtering and content-based filtering techniques.

B. Unsupervised Learning Method

1) Feature Extraction: To convert textual comments from reviewers into a structured form suitable for modeling, we employed the Term Frequency-Inverse Document Frequency (TF-IDF) technique. TF-IDF assigns weights to words based on their frequency in a document relative to their frequency in the entire corpus.

2) Clustering Algorithm: We utilized the K-Means clustering algorithm to group the TF-IDF transformed comments into distinct clusters. Each cluster corresponds to a rating on a scale from 1 to 5, allowing for the identification of patterns and sentiments within the textual data.

C. Modelling Approaches

1) Collaborative Filtering: We explored collaborative filtering techniques, which leverage user-item interactions to make predictions.

- **User-Item Collaborative Filtering:** We implemented the ALS (Alternating Least Squares) model in Apache Spark to predict ratings by learning latent factors for both users and items.
- **User-User Collaborative Filtering:** Predicting a user's preferences based on similar users, we calculated the similarity between users using cosine similarity or Pearson correlation.
- **Item-Item Collaborative Filtering:** Focusing on relationships between items, we computed item similarities using cosine similarity and utilized weighted averages of ratings for recommendation.

2) Content-based Filtering: We employed content-based filtering to recommend items based on their features and user preferences.

- Utilizing item features, we recommended items similar to those a user liked in the past.
- We measured feature similarity between items using cosine similarity to provide personalized recommendations.

D. Considerations

Throughout the modeling process, several considerations were taken into account:

- **Hybrid Approaches:** The potential benefits of combining collaborative filtering and content-based filtering were explored to enhance recommendation accuracy and diversity.
- **Evaluation Metrics:** We defined appropriate evaluation metrics, such as RMSE (Root Mean Square Error) and precision-recall curves, to assess model performance.
- **Scalability:** Considering the scalability of the chosen methods, we ensured compatibility with large datasets, utilizing techniques like ALS in Apache Spark for distributed computing.
- **Model Interpretability:** Transparency and interpretability of the models were prioritized to enhance user trust and understanding of the recommendation process.

By integrating unsupervised learning for feature extraction with collaborative and content-based filtering techniques, we aimed to develop a comprehensive rating prediction model capable of providing accurate and personalized recommendations for user reviews.

IV. Experiments

A. Data

Our analysis is based on the Inside Airbnb dataset, a rich compilation encompassing listings, calendar data, and reviews across 19 major US cities. Initially, the dataset was organized into separate folders for each city, containing files like 'Austin listings.csv', 'Austin calendar.csv', and 'Austin reviews.csv'.

To streamline our analysis, we transformed and aggregated this data into three master CSV files:

'master listings.csv': Consolidating detailed listings information such as amenities, pricing, host details, and property characteristics from all 19 cities. 'master calendar.csv': Combining calendar data detailing availability and pricing changes over time across all locations. 'master reviews.csv': Aggregating comprehensive review data, offering qualitative insights from guest experiences in different cities.

B. Evaluation Metrics

We evaluated the performance of our recommendation system using the following metrics:

- **Root Mean Square Error (RMSE):** RMSE measures the average magnitude of the errors between predicted and actual ratings. Lower RMSE values indicate better accuracy in rating prediction. Our system achieved an RMSE of 0.3741.
- **Mean Absolute Error (MAE):** MAE calculates the average absolute errors between predicted and actual ratings. Similar to RMSE, lower MAE values signify better accuracy. Our system achieved an MAE of 0.1978.
- **Precision:** Precision measures the proportion of true positive predictions (relevant items recommended) among all recommended items. Higher precision indicates fewer false positives in recommendations. Our system achieved a precision of 0.9851.
- **Recall:** Recall measures the proportion of true positive predictions among all relevant items in the dataset. Higher recall indicates fewer false negatives in recommendations. Our system achieved a recall of 0.9925.
- **F1-score:** F1-score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance, considering both precision and recall. Our system achieved an F1-score of 0.9888.

V. Results

In this section, we present the results obtained from the collaborative filtering and content-based filtering approaches for predicting hotel ratings.

A. Collaborative Filtering

We first employed collaborative filtering techniques, including Alternating Least Squares (ALS) and Singular Value Decomposition (SVD), to predict hotel ratings based on user-item interactions.

1) Model Performance: Table I summarizes the performance of different collaborative filtering models in terms of Root Mean Square Error (RMSE). We observed significant improvements in prediction accuracy through fine-tuning the ALS model parameters, achieving an RMSE of 0.89 compared to the initial RMSE of 1.99. Additionally, the SVD approach yielded an RMSE of 1.2470, providing competitive results.

Model	Parameters	RMSE
ALS	(maxIter=5, rank=10, regParam=0.01)	1.99
Finetuned ALS	(maxIter=5, rank=10, regParam=0.1)	0.89
SVD		1.2470

TABLE I
COMPARISON OF RMSE FOR DIFFERENT PARAMETERS AND MODELS.

B. Content-based Filtering

Next, we employed content-based filtering techniques to predict hotel ratings based on item features and user preferences. The weighted average of ratings was utilized as the primary mechanism for recommendation.

1) Model Evaluation: Table II presents the evaluation metrics for the content-based filtering recommendation system. We achieved a low RMSE of 0.3741, indicating the model's ability to accurately predict hotel ratings. Moreover, the Mean Absolute Error (MAE) was measured at 0.1978, demonstrating the model's precision in rating prediction. The precision, recall, and F1-score were also calculated, showing high values of 0.9851, 0.9925, and 0.9888, respectively, reflecting the system's effectiveness in recommending relevant items to users.

Metric	Value
RMSE	0.3741
MAE	0.1978
Precision	0.9851
Recall	0.9925
F1-score	0.9888

TABLE II
EVALUATION METRICS FOR THE RECOMMENDATION SYSTEM.

Overall, the collaborative filtering and content-based filtering approaches exhibited promising results in predicting hotel ratings, showcasing the effectiveness of the proposed models in providing personalized recommendations to users based on their preferences and historical interactions.

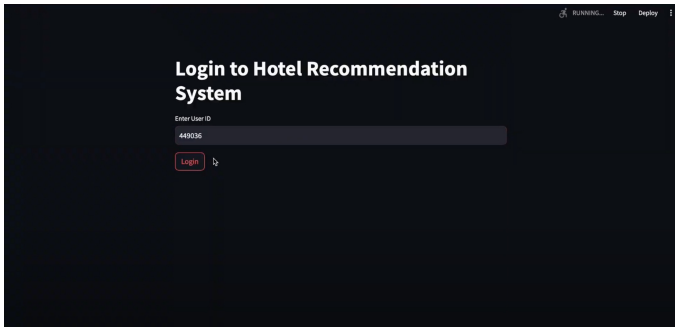

Fig. 4. Login Page

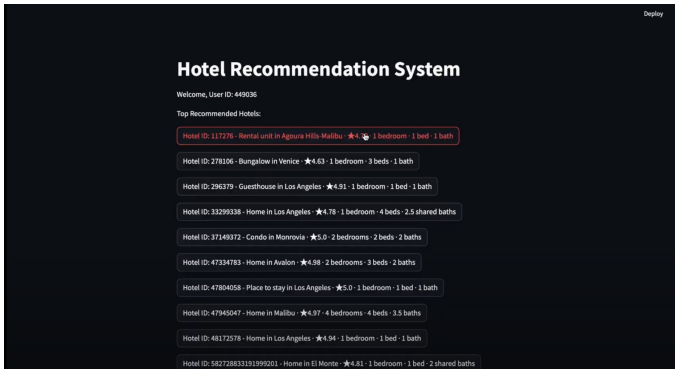

Fig. 5. Recommendation Page

C. Web App

To implement the functionality described above, we utilized Streamlit, a powerful framework for building interactive web applications in Python. Below is the structure of the web app:

1) Login Page: The login page in Figure (4) allows users to enter their user ID to access personalized recommendations.

2) Recommendation Page: The recommendation page in Figure (5) displays the top 10 recommended hotels for the logged-in user.

3) Map Page: The map page in Figure (6) shows the recommended hotels on a map, allowing users to visualize their locations.

4) Hotel Details and Similar Hotels Page: This page in Figure (7) displays the details of a selected hotel and the top 5 similar hotels.

By combining these pages and functionalities, we created a seamless user experience for exploring and interacting with hotel recommendations. The web app provides personalized recommendations based on user input and allows users to explore recommended hotels on a map, view hotel details, and discover similar hotels of interest.

VI. Future Work

To further enhance the performance and comprehensiveness of our Airbnb recommendation system, incorporating a broader range of data types and sources is essential. Expanding the dataset to include more user interactions, behavioral metrics, and demographic information could provide a deeper

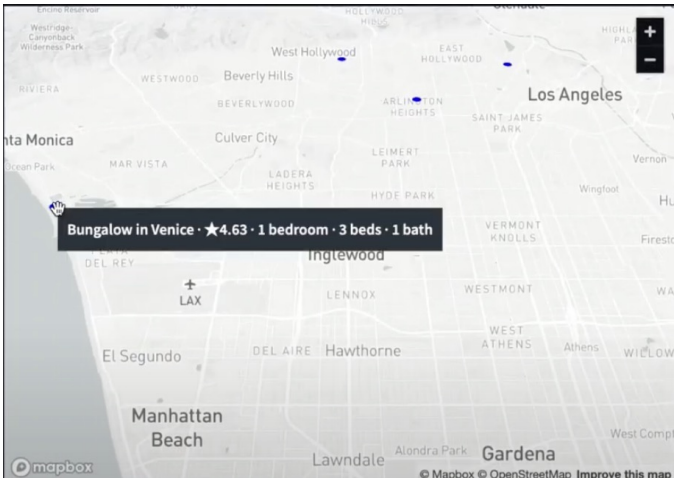

Fig. 6. Map Page

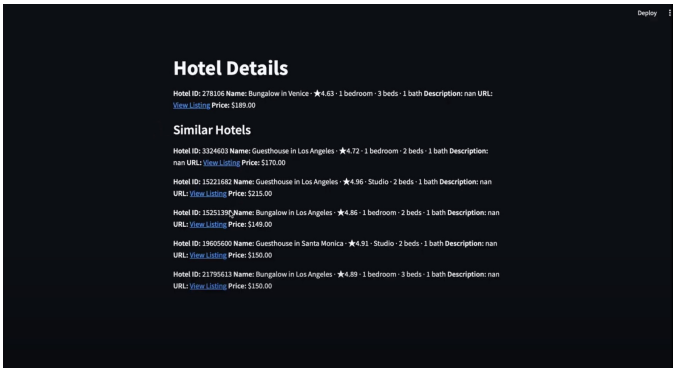

Fig. 7. Hotel Details and Similar Hotels Page

understanding of customer preferences and improve the accuracy of our predictions. Additionally, integrating data from social media platforms could offer insights into current trends and preferences, allowing the system to adapt to changing consumer behaviors dynamically.

In terms of algorithmic development, further exploration and comparison of various machine learning models will be crucial. Future studies should investigate the efficacy of neural networks, which are well-suited for handling large and complex datasets with non-linear relationships. Probabilistic models, which can manage uncertainty and variability in user data effectively, also merit consideration. Moreover, developing hybrid models that combine the strengths of multiple machine learning techniques could potentially address limitations found in single-model approaches.

Lastly, continuous model evaluation and comparison against current benchmarks in the industry will ensure that the system remains state-of-the-art. Implementing a robust framework for ongoing testing and updates will be vital to maintain relevance and effectiveness in the highly competitive hospitality market. These efforts will not only refine our system but also significantly contribute to the broader field of personalized recommendation systems.

VII. Acknowledgements

This work is supported by Vanderbilt University and Data Science Institute. The project is supervised by Professor Dana Zhang and serves as the final deliverable for DS 5460-01 Big Data Scaling in the Spring 2024 semester.

References

[1] ChatGPT, response to author query. OpenAI [Online]. https://chat.openai.com/ (accessed April 16th, 2024).
[2] Melián-González, S., et al. (2019). Personalized travel recommendation model using machine learning. Journal of Travel Research.
[3] Levi, A. (2018). Enhancing customer experience with machine learning at Booking.com. Hospitality Technology.
[4] Xie, K. L., et al. (2016). Examining the effectiveness of hybrid recommendation systems for enhancing user experiences. Journal of Hospitality Marketing Management.
[5] Zhang, Y., et al. (2020). Scalability of large-scale data processing in cloud environments with Apache Spark. Journal of Cloud Computing.
[6] Gavurova, B., et al. (2021). Data-driven decision-making in hospitality management for optimized pricing strategies. International Journal of Hospitality Management.