Developing Airbnb Price Matching Tool Using Various Machine Learning Models

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Vaidic Naik

Spring, 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Briana Morrison, Department of Computer Science

Developing Airbnb Price Matching Tool Using Various Machine Learning Models

CS4991 Capstone Report, 2023

Vaidic Naik

Computer Science

The University of Virginia

School of Engineering and Applied Sciences

Charlottesville, VA USA

vn3qg@virginia.edu

ABSTRACT

Capital One's Shopping Travel Team needed to develop an end-to-end pipeline that would serve as a price-matching tool for the vacation rental property service, Airbnb. The pipeline used various machine learning models to compare different features between a given Airbnb property and other vacation rental properties on Capital One's partner services. The compared features included size (maximum occupancy, bedrooms, bathrooms), amenities, name, description and distance. The machine learning models used to compare these features were k-nearest neighbors, natural language processing, and density-adjusted geospatial distance. Our final pipeline was 500% more accurate than the original model, and it made more meaningful predictions. We also explored further enhancements to our project, such as tracking user click-through data, to further fine-tune our model.

1. INTRODUCTION

The vacation rental industry has seen exponential growth over the past decade, with Airbnb being a major player in this space. Vacation rental properties provide a convenient and affordable alternative to traditional hotels, resulting in a significant increase in demand. However, one of the key challenges faced by vacation rental property services like Airbnb is the difficulty in pricing properties competitively in a constantly changing market. Without the ability to adjust prices to match market demand, vacation rental services risk losing out on potential customers or underpricing their properties, resulting in lower profits. Therefore, there is a need for a price-matching tool that can accurately compare the prices of different vacation rental properties and recommend optimal pricing strategies.

To address this challenge, we present an end-to-end pipeline that uses machine learning models to compare different features between a given Airbnb property and other vacation rental properties on Capital One's partner services. The compared features include size (maximum occupancy, bedrooms, bathrooms), amenities, name, description and distance. The machine learning models used in the pipeline include k-nearest neighbors, natural

language processing, and density-adjusted geospatial distance.

K-nearest neighbors is a machine learning algorithm that is used for classification and regression. It works by finding the k-nearest data points in a dataset and then using their labels or values to predict the label or value of a new data point. Natural language processing (NLP) is a subfield of computer science that focuses on the interaction between computers and humans in natural language. NLP techniques are used in our pipeline to analyze and compare the textual descriptions of vacation rental properties. Density-adjusted geospatial distance is a method that is used to measure the similarity between two points in space. It takes into account the density of points in the surrounding area to produce more accurate distance calculations.

By presenting the development and evaluation of the price-matching tool, our aim is to provide insights into its accuracy, effectiveness, and limitations. In addition, we explore further enhancements to the tool, such as tracking user click-through data to further fine-tune our model. Our approach can help vacation rental property services like Airbnb make more accurate and informed pricing decisions, ultimately leading to better profits and customer satisfaction.

2. RELATED WORKS

A review of scholarly sources that focus on machine learning-based solutions for problems related to Airbnb and other short-term rental platforms reveals several similar works that highlight the varied applications of machine learning in this context. For instance, Garrido, et al. (2021) conducted a comparative study of Airbnb pricing tools using machine learning models like k-nearest neighbors, random forest, and gradient boosting [1]. The study aimed to predict Airbnb prices using various features such as location, room type, and number of bedrooms. Although similar in nature to other works, the study differed in the specific features used to predict Airbnb prices.

Jiang, et al. (2020) present a machine learning-based approach to matching Airbnb listings with hotel rooms [2]. The authors used several features such as location, price, and amenities to create a similarity score between Airbnb listings and hotel rooms. This score was used to identify the closest hotel rooms to each Airbnb listing. Although similar to other works, the specific features used in this study differed from those used in other works.

In addition, Wang, et al. (2019) proposed a machine learning-based approach to predicting Airbnb property rental prices [3]. Their model used features such as location, number of bedrooms, and amenities to make predictions. The authors used a variety of machine learning models, including decision trees, random forests, and support vector regression, to train their model. Although the three works share the common thread of predicting Airbnb prices using machine learning, their approaches, models, and features differ.

The review of similar works highlights the diverse range of applications for machine learning in the context of short-term rental platforms like Airbnb [1, 2, 3]. While there are similarities in using machine learning to solve problems related to Airbnb, the specific approaches, models, and features used in each work differ. Insights gained from the review can help to improve future projects in this field by providing a better understanding of the potential applications of machine learning and the specific factors that affect short-term rental management. Machine learning can provide a better guest experience and help

hosts make better decisions, ultimately, maximizing their revenue.

3. PROCESS DESIGN

Our pipeline relied on a very basic architecture. This is because our pipeline was a proof-of-concept project, so we were not worried about connecting our code to a front end or the client. The theoretical final product of our project would work like the current Capital One shopping extension.

3.1 System Requirements and Overview

Our system requirements were given to us at the beginning of our project. Our Airbnb matching pipeline had to be compact and efficient. We could not have it take up too much space because it would theoretically be installed as an extension on a web browser. And it had to find matches quickly since we wanted the extension to work in real time as the user browses through Airbnb listings. We were not given any strict restrictions as to which programming languages or libraries to work with. We just had to keep in mind that our pipeline was able to read from the testing data that was stored in a hive database.

3.2 Selecting Features

In order to compare the properties, we had to decide on our own which features to focus on. We decided this by seeing what information we had on both the Airbnb listings and our other rental properties. The features that were common between the two ended up being max occupancy, number of bathrooms, number of bedrooms, Wi-Fi, kitchen, laundry, parking, pets allowed, distance/location, name of property, and property description. We used these features to determine how similar two rental properties were.

3.3 Integrating Machine Learning Models

Once we determined which features we were going to focus on to determine how similar the two properties were, we had to decide which Machine Learning (ML) models we would use to compare the features. We knew we had to use some type of Natural Language Processing (NLP) when comparing the names and descriptions. Since the descriptions of the property contained more important information, used a more robust NLP model to properly analyze the descriptions. We decided to train our own model using doc2Vec library, the trained library was fine-tuned to detect keywords in the descriptions of properties. This gave us insight into information about the property that was not possible with the other features.

We opted for a pre-trained opensource NLP model for the names since they were much simpler to compare. The library we used for comparing names was spacy. The main difference between the two NLP models is that spacy is pretrained tool that generates a similarity score between two words or short phrases, while doc2vec creates a numeric vector for each word and generates a model that can make a more nuanced comparison.

We decided to lump together a couple of our features and use the same ML model to compare them. We combined max occupancy, number of bathrooms and number of bedrooms into one group called property size. Similarly, Wi-Fi, kitchen, laundry, parking, and pets allowed were combined to form property amenities. For both property size and amenities, we used k nearest neighbor as the machine learning model. Using an unsupervised model ensured that our pipeline ran as efficiently as possible.

For our final feature distance, we used a unique approach rather than just comparing distance. When analyzing our distance, we decide to consider the concentration of

Airbnb properties in the area. For instance, the fact that two properties are a mile away would be a lot less significant in a busy city compared to a rural area with few Airbnb properties. We got the “density” of an area by looking at the number of properties in a 10-mile radius. The geospatial distance was divided by the density to produce a density adjusted distance.

3.5 Final Aggregate Scoring

The final function of our pipeline was to be able to generate a similarity score between an Airbnb property and any other vacation rental property. For this we had to take some weighted average of all the individual scores calculated for our features. Most of our individual scores were already standardized so they were between 0-1 except for a density adjusted distance. To scale the distances, we took the furthest property and set it to 0 and normalized everything else.
With these five scores we tool a weighted average to generate a final score. In our testing we tweaked the weights so that they best fit our own definition of what makes the two properties similar. We then sorted the five most similar properties based off the score.

4. RESULTS

To determine the accuracy of the similarity score generated by our pipeline, we needed a baseline to compare. We did this by doing a general survey of our coworkers to generate their own similarity score out of five between three Airbnb properties and accompanying properties. We then compared our scores to the surveyed scores to the recommendations generated by the initial filters. We found that our pipeline was able to make more accurate predictions by a factor of 500% when compared to the original filters.

5. CONCLUSION

Our internship experience working on the price-matching tool project with Capital One's Shopping Travel Team was a resounding success. We were able to develop an end-to-end pipeline that utilized various machine learning models to compare different features between Airbnb properties and other vacation rental properties on Capital One's partner services. Our pipeline accurately predicted the similarity score between two properties, providing valuable insights into optimal pricing strategies for vacation rental property services like Airbnb.

The pipeline's performance was evaluated by comparing it to a baseline of similarity scores generated by human judgment. Our pipeline significantly outperformed the original filters, making more meaningful predictions with a 500% increase in accuracy. Our pipeline's use of k-nearest neighbors, natural language processing, and density-adjusted geospatial distance allowed for a more nuanced and accurate analysis of the compared features, highlighting the potential of machine learning in the vacation rental industry.

Our project demonstrates the usefulness of machine learning-based solutions to complex problems and emphasizes the need for continued research and development in this area. The insights gained from our internship experience provide valuable contributions to the broader field of machine learning and can help inform future projects aimed at enhancing vacation rental property services like Airbnb.

6. FUTURE WORK

In terms of future work there are a couple of things that can be done moving forward with our project. We could work to push our pipeline into production and design a proper front end for our tool. This would also require us to integrate our work with another similar intern teams project. This

team worked to develop the tool that scraped through capital ones partners to find multiple listings of the same property to find the cheapest offer. This would work with our similarity matching pipeline to generate a full price matching tool.

Another aspect that could be improved upon is the assignment of weights for our final aggregate score. Not all users are looking for the same features. By analyzing user click-through data, we could update our pipeline to generate better-suited matches for individual users. For example, if we notice that a user has only looked at Airbnb's with a max occupancy of 8 and nothing else is common, the pipeline would adjust to prioritize the size category. This personalization would make our price matching tool more marketable and ultimately more profitable for Capital One Shopping.

REFERENCES

[1] F. J. Garrido, J. B. Roca, J. M. Luna, and A. Damas, (2021). "Comparative study of Airbnb pricing tools using machine learning techniques," in Proceedings of the 9th International Conference on Industrial Engineering and Operations Management, 2021, pp. 2312-2322.

[2] L. Jiang, Y. Liu, and S. Zhang, (2020). "Matching Airbnb Listings with Hotel Rooms Using Machine Learning," in Proceedings of the 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 2020, pp. 100-105.

[3] Y. Wang, C. Xu, Y. Li, and L. Liu, (2019). "Airbnb Property Rental Price Prediction Using Machine Learning," in Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), 2019, pp. 1393-1395