**UC Davis Graduate School of Management**
**BAX 452 Winter Quarter, 2023**

# Powering rental bookings for Airbnb's biggest competitor: VRBO

---

## Leveraging machine learning to drive business impact

**Submitted By:**
**Yuto Takeda, Tamalika Basu, Shulang (Simon) Ning**

**Table of Contents**

1. Executive Summary 2

2. Introduction of the Domain 3

3. How does the industry tackle the problem traditionally? 3

4. Our Analysis 4

5. Recommendations & Next Steps 10

6. Conclusion 10

7. Appendix 11

**Executive Summary**

Operating successfully in the online short-term rental industry involves providing a pleasant experience to both, the users looking to book properties and the property owners renting their resources. To augment the given use case in hand, we undertook the project to solve it in two halves. Firstly, we gathered data from VRBO.com, one of Airbnb's largest competitors, to identify various attributes that determine the pricing for properties in well-known US cities like San Francisco, New York, Chicago, and many more. This part of the project where we identify, gather, and prepare the data required for our analysis has already been completed in BAX 422. And now, as part of this report, we will detail the second task, where we utilize the collected data to further perform two essential sub-tasks using machine learning algorithms. The first sub-task is to develop algorithms that will analyze the pricing trends on VRBO.com[1], and then use the models to predict rental property prices based on the factors that will best describe a relationship with 'price'. The second sub-task is to use the algorithms to offer customer-centric recommendations to property owners based on the reviews on their property, enabling them to improve their property listings for better profitability and increased booking chances. In this report, we will delve into our discoveries from machine learning model results to guide business owners in this industry.

Our goal through this project is to enhance user experience, increase engagement, and supplement usage of short-term rentals like VRBO, among customers by providing fairly determined prices and promoting a more transparent listing that can ultimately create business value. The insights obtained from this project will be useful for home rental business owners, allowing them to make well-informed decisions about prices and amenities that can optimize their properties for the market. Furthermore, since the project concentrates on popular tourist destinations in the US, it is particularly relevant to the hospitality sector, providing stakeholders with valuable insights into consumer behavior and preferences. Overall, this project has the potential to contribute to the growth and evolution of the short-term rental market, while also providing significant value to users and industry stakeholders.

**Introduction of the Domain**

The vacation rental industry is expanding rapidly and has the potential for significant innovation and disruption. The emergence of platforms like Airbnb and VRBO has transformed the way people travel and find accommodations, making it more convenient and economical for travelers to find comfortable and affordable lodging while also providing property owners with new avenues to generate revenue. According to Allied² market research, the global vacation rental market size reached $91.2 billion in 2021 and is projected to grow to $315 billion by 2031. Additionally, Statistica² reports that there were over 1,000 distinct properties available in San Francisco alone in April 2022, indicating a potential user base of over 62.99 million by 2027.

Despite this growth, the industry faces several challenges such as fluctuating demand, varying pricing strategies, and the need to provide personalized experiences for customers. By leveraging data (as done in BAX 422) and employing machine learning techniques (as will be done in BAX 452), we can assist property owners in optimizing their rental properties for the market and improving the rental experience for tenants. The insights gleaned from this project can aid industry stakeholders in better understanding consumer behavior and preferences, resulting in more efficient and effective rental markets. Therefore, this field offers an exciting and challenging opportunity for data-driven innovation and value creation.

**How does the industry tackle the mentioned problems traditionally?**

Per our knowledge of this domain, to tackle the two business problems in hand, here is how traditionally, the rental industry handled them:

Price prediction: Property owners and rental managers traditionally use their own experience and market knowledge to set prices for their properties. They would often rely on competitive analysis, comparing their prices to other similar properties in the same area. However, this method was not always accurate and could lead to overpricing or underpricing of properties. In recent years, there has been increasing use of data analysis tools and software to help property managers set prices based on real-time market data

and demand trends. Therefore, with our machine learning models, we will be able to exactly identify the essential and contributing factors to determine the price for a property, thereby helping the owners set a strategically transparent and analytically driven price for their property

Customer reviews: Traditionally, property owners relied on their experience and intuition to improve their properties based on customer feedback. However, with the growth of online booking platforms, customer reviews have become more important than ever in improving property listings, therefore it is crucial to consider and carefully analyze these reviews to understand the ongoing sentiments. Many rental businesses now actively encourage customers to leave reviews and ratings, which are displayed prominently on their online listings. Not only it is helpful for property owners in the way that they can use this feedback to make targeted improvements and adjustments to their properties but these reviews also help advertise the credibility of a property to other prospective customers. Because of this growing focus on customer reviews, the owners should use an automated sentiment analysis tool to help them identify common themes and issues across customer reviews, giving them the opportunity to improve their listing

**Our analysis of each of the business problems**

**First, predicting price based on essential features for each property**

- **Approach**

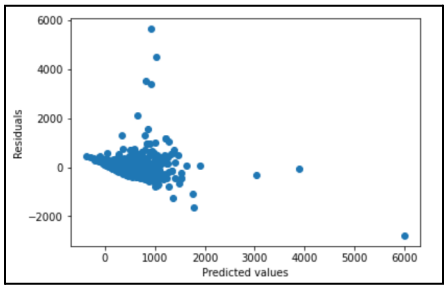
  To predict prices using the different features extracted for each property, we applied an ensemble of models (linear regression, cross-validation + lasso regression, random forest) to obtain a good price prediction. The detailed steps are as follows:

- We started with performing EDA on our dataset to check for the presence of null values, and missing values and also standardized the relevant columns for better interpretation. First, we removed the columns not needed for the price analysis, such as: '_id', 'amenities_text', 'facilitates_text', 'reviews_text', 'name', 'vrbo_near', 'vrbo_text', as few of these would also contribute in the sentiment analysis task later. Second, we observed a very high percentage of missing values in the data, therefore it was important to treat them rather than remove them.

Mentioned below are details on how we handled these missing values and why, (note that: a description of all the fields is present in Appendix[3]):

- ❖ No. of Bathrooms: Replaced nulls with '0' as no bathroom information was listed
- ❖ No. of Baths: Replaced nulls with '0' as no bathroom information was listed
- ❖ No. of Beds: Replaced nulls with '1' as another column called *'no of bedrooms'* had no null values, therefore we assumed a presence of at least one bed per bedroom
- ❖ Area Square Ft: Replaced nulls with '300' as per the guidelines from International Residential Code, no property built will be lesser than 300 square feet typically
- ❖ No. of Images: Replaced nulls with '1' as one image is a mandated requirement for a property to be listed
- ❖ Star Rating: Created a dummy class for Star Rating, and for nulls, we assigned a new dummy class called '99'

- We then split out the dataset into train and test subsets with an 80% - 20% ratio, as we will work with supervised learning, therefore predictions would need to be compared against unseen data
- Next, to start with our model exploration, we applied a linear regression model with 'VRBO price' as the response variable and all other variables in the dataset as explanatory. Here, VRBO price refers to the price of a property for 1 night, 2 adults, 0 children per room basis. However, on observing the residuals, we realize that the normality assumption is invalid for this data set. Below is an image of the residual plot:




- Therefore, to perform feature selection, we apply a CV Lasso method to reduce the number of features and include only the most influential ones. Using CV, we aimed to obtain the optimal

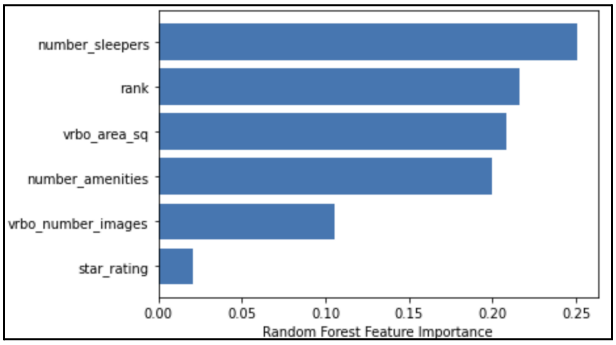
regularisation parameter (λ), which is then fed in a Lasso regression to obtain 6 non-zero variables, out of the 21 variables, indicating that these six are critical for price prediction

- Next, using this set of features, we use a Random Forest algorithm, which will be highly robust and efficient in predicting prices for properties with the reduced feature set. In order to fine-tune the hyperparameters of the random forest, we used the CV grid search method to optimize the model inputs. Here is a snapshot of the grid used:

```
# Random Forest and Grid Search

# Define the parameter grid for GridSearchCV
param_grid = {
    'n_estimators': [100, 200, 300, 400, 500],
    'max_depth': [None, 10, 20, 30, 40, 50],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

- Finally, using the optimal parameters, we use our *'Test data'* to analyze the performance of the model. We obtained an out-of-sample R square of 48%. Additionally, below is the comparison of the six features, as gauged by random forest




- **Result & Interpretations**

Based on our analysis, the features that seem to be most influential to predict prices are: No. of Amenities, No. of Sleepers, Star Rating, Rank, Area Square Feet, No. of Images. Further, we can conclude that these features attribute to explaining almost ~50% of the variation in Price data

**Second, analyzing the topics with negative comments among the lower-rated properties.**

- **Approach**

  The primary objective of this task is to analyze the reviews of properties with low ratings to identify potential reasons for their low ratings. To accomplish this, we employed sentiment analysis to classify the reviews into positive, negative, or neutral using the "Twitter-RoBERTa-Base-Sentiment[4]" transformer. Additionally, we developed a Latent Dirichlet Allocation (LDA)[5] model to gain insights into the negative reviews' underlying topics. The following are the detailed steps we took in our analysis.

  **Labeling the reviews**

- We first transformed the dataset. Since different reviews in the data are concatenated with a pipe symbol "|||" under each property, we then transform those data into a list type
- Nextly, we select the subset of the dataset. Since we need the value of "star_rating" and "reviews_text", we only select the properties which have both values. We use almost 30% of the whole dataset (743/2503)
- In this step, we load the "Twitter-RoBERTa-Base-Sentiment" transformers, which is a pre-trained model. Then, we labeled the reviews for each property into "Positive" "Negative" or "Neutral". LABEL_0 indicates "Negative", LABEL_1 indicates "Neutral", and LABEL_2 indicates "Positive"
- Then, we divide these properties into two groups. One is the group with the higher rating and the other is the group with the lower rating. The definition of a highly-rated group is when star_rating is higher than 3, and the definition of a low-rated group is when star_rating is 3 or less.
- In this step, we aggregated the label of all of the reviews for each group. The result of the review proportion in both groups are in fig1, 2 in the result section

  **LDA (Identity what topic are important in these negative comment)**

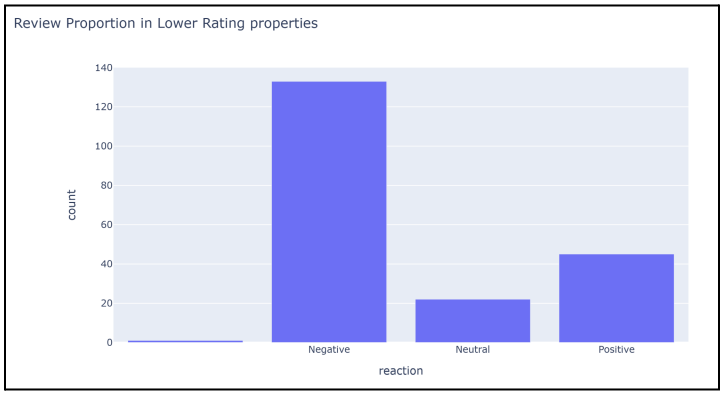
- First, we only select the reviews labeled as negative in the lower rating group

- Next, we removed the unnecessary specific characters for the LDA model such as ”,\.!?”, and any stopwords in the NLTK module.
- Finally, using these dictionaries of words, we build the parallelized LDA model. We set the number of topics as 10. The result is in fig 3 in the next section.

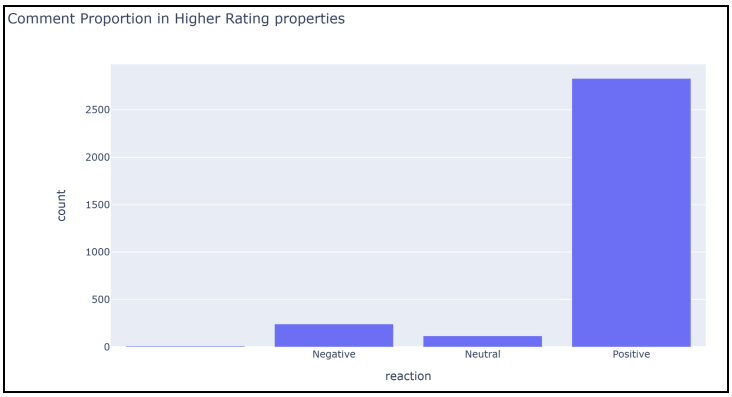
- **Result & Interpretations**

  The result of the labeling are as follows:

  (fig1) showing the counts by labeling of review in the group of lower rating properties




  (fig2) showing the counts by labeling of review in the group of higher rating properties




  (fig3) The result of LDA model

| Topic Num | Words |
|---|---|
| 1 | [room,property,us,back,get,hotel,check,dirty,outside,pictures] |
| 2 | [vrbo,stay,check,morning,place,us,owner,property,room,like] |
| 3 | [stay,hotel,also,check,floor,dirty,room,broken,rooms,like] |

| | |
|---|---|
| 4 | [house,property,day,vrbo,stay,back,apartment,like,check,owner] |
| 5 | [stay,place,vrbo,room,owner,book,horrible,booked,days,hotel] |
| 6 | [room,place,get,two,never,dirty,us,time,check,stay] |
| 7 | [room,check,get,back,property,us,hotel,owner,place,vrbo] |
| 8 | [room,place,property,never,vrbo,us,hotel,called,night,people] |
| 9 | [property,stay,room,place,like,owner,get,check,even,night] |
| 10 | [room,property,get,check,us,unit,dirty,hotel,good,services] |

(fig4) The word cloud of the data used in the LDA model


From the result of the first two visualizations, we can see that the proportion of negative reviews is large in the lower rating group. On the other hand, the proportion of negative reviews is quite low and one of the positive reviews is large in the higher rating group.

From the result of the LDA model, it might be hard to label the name for each topic. However, we can describe in topic number 1, there are “dirty”, “pictures”, “property”, and “room” which represent this topic. Moreover, in topic number 10, there are “dirty”, “room”, “unit”, “services”, and “property”.

Furthermore, in topic number 3, there are “dirty”, “room”, “broken”, and “floor”.

From these topics, it might be possible to interpret that rooms and properties are dirty, and floors are broken, and these facts might differ from the pictures.

**Recommendations and Next Steps**

As per the analysis performed, we have the following recommendations:

- In order to determine analytically justified prices for the properties, the owner should channel their primary focus towards the attributes that are most essential for establishing prices, which are: No. of Amenities, No. of Sleepers, Star Rating, Rank, Area Square Feet, No. of Images. Clearly, customers are attracted to more on-premises amenities available like a swimming pool, barbeque stand, etc, and are drawn towards properties that feature more images of different locations of the accommodation. Customers also highly value the sleeping space, and size of the property available for them and their guests. And ultimately, star ratings will always be essential for a property to be recognized.
- Additionally, our price prediction analysis did not consider a few critical factors such as the real estate rate prices of the location, ongoing market prices, any seasonal aspect that may affect prices, and weights on the popularity of famous landmarks around the property. We believe that if these factors are accounted for, the price predictions will be more accurate
- Finally, the cleanliness of a property is considered the most important factor for the owners to not get a low evaluation from the users. It is also considered important to make sure that the properties or amenities are not broken. As an important point, in this analysis, the topic model was built based on the lower-rated reviews. However, since negative reviews may differ depending on the price range of the property, we believe that the topic can be further clarified in the future by dividing the targets a little more.

**Conclusion**

In this report, we present an analysis using VRBO.com, aimed at enhancing user experience and promoting the popularity of properties among customers. Our objectives are to provide fair prices and ensure a more transparent listing to create business value. To achieve this, we developed a combination of models to predict the price of a property based on essential features, enabling owners to set prices that are

analytically driven. Our approach involved the use of the lasso regression model to select key variables and the random forest model to predict prices. In addition, we conducted sentiment analysis to identify factors that influence lower ratings and built a topic model to uncover the key themes and topics affecting these ratings.

In conclusion, our findings have the potential to contribute significantly to the growth and evolution of the short-term rental market. By providing transparent pricing information and promoting fair competition, we can enhance the user experience and promote customer engagement.

**Appendix**

1. Link to VRBO.com: https://www.vrbo.com/. Some facts about VRBO:

   VRBO stands for Vacation Rentals By Owner and was founded in 1995 by husband and wife team, David and Lynn Clouse, who wanted to create a platform for property owners to advertise their vacation homes. VRBO was acquired by HomeAway in 2006, and then by Expedia Group in 2015. It now operates as a part of Vrbo, a global vacation rental online marketplace.

   VRBO has over 2 million vacation rentals in more than 190 countries around the world, making it one of the largest vacation rental platforms. The site allows property owners to list and rent out their entire homes, apartments, cabins, and villas, as well as individual rooms in shared spaces across a range of unique and unusual vacation rentals, such as treehouses, houseboats, and even castles.

   VRBO's upcoming potential has stirred the establishment of the market leader Airbnb. Travellers and customers are increasingly using VRBO as a replacement, and therefore there seems to be a huge untapped market laying with VRBO to explore. Refer to the link here to see a comparison shown between the 2 sites: https://travelfreak.com/airbnb-vs-vrbo/

2. References to data stated in "Introduction":
   https://www.grandviewresearch.com/industry-analysis/vacation-rental-market

   https://www.igms.com/vacation-rental-sites/#

   statista.com/outlook/mmo/travel-tourism/vacation-rentals/united-state

   https://www.igms.com/2023-vacation-rental-industry-trends/#

   https://www.lodgify.com/blog/online-vacation-rental-rates/

3. Here is the description of features used, to estimate prices:
    a. Rank: Describes the rank of the property listing under each city, based on its popularity amongst customers
    b. Name: Describes the title of the property for identification
    c. VRBO_City: Describes the city under which the property is located
    d. VRBO_Text: Describes the details about the property as updated by the owner. This section gives an overview of the uniqueness and offerings of the listing to a customer
    e. VRBO_Type: Describes the type of property offered by VRBO namely, hotel, studio, guest house, apartment, house, building, hotel suite, condo, resort, cottage, townhome, yacht, villa, estate, recreational vehicle, bungalow, cabin, houseboat, hostel, corporate apartment, mobile home, and boat
    f. Number_of_Bedrooms: Describes the number of bedrooms available in the property
    g. Star Rating: Describes the rating of the property as rated by the visitors
    h. VRBO_Near: Describes the top 6 famous tourist attractions or important landmarks around the property (typically within 0-3 miles range)
    i. VRBO_Price: Describes the price of the property based on occupancy for 1 day, 2 adults, 0 children, and 0 pets
    j. VRBO_Number_Images: Describes the number of images posted by the owner of a property. More images would indicate more confidence from the owner, in showcasing the property to ensure transparency between the owner and the customer
    k. VRBO_Area_SQ: Describes the area covered by the property in square feet
    l. Number_Beds: Describes the number of beds available in the property
    m. Number_Sleepers: Describes the sleeping capacity available in the property
    n. Number_Bathrooms: Describes the number of bathrooms available on the property
    o. Number_Baths: Describes the number of baths available on the property
    p. Number_Reviews: Contains all the reviews as shared by visitors on the property

q. Reviews_Text: Contains a concatenated format of all the reviews for a particular property, separated by '|||' to distinguish between two different reviews
r. Number_Amenities: Describes the number of amenities supported for each property, namely, microwave, fridge, room heater, hairdryer, etc.
s. Amenities_Text: Describes the type of amenities supported for each property, namely, microwave, fridge, room heater, hairdryer, etc.
t. Number_Facilities: Describes the type of facilities available namely, shower, sofa, dining table, etc.
u. Facilities_Text: Describes the number of facilities available namely, shower, sofa, dining table, etc.

4. Twitter-RoBERTa-Base-Sentiment is a pre-trained natural language processing model based on the RoBERTa architecture that has been specifically fine-tuned for sentiment analysis on Twitter data. This model is trained to understand and classify the sentiment of short text messages such as tweets, into one of three categories: positive, negative, or neutral. The pre-training of this model involves training it on a large corpus of text data, and then fine-tuning it on a smaller labeled dataset to improve its accuracy in sentiment analysis. Twitter-RoBERTa-Base-Sentiment is a popular choice for analyzing sentiment in Twitter data due to its high accuracy and ability to handle the unique language patterns and nuances present in Twitter data
5. Latent Dirichlet Allocation (LDA) is a probabilistic model used in natural language processing to identify topics in a collection of documents. The LDA model assumes that each document in the collection is a mixture of various topics and each topic is a probability distribution over a set of words. The model discovers the underlying topic structure of the corpus by identifying the distribution of words that are most likely to be associated with each topic.

   LDA works by representing each document as a bag of words, meaning that the order of the words in the document is ignored. The model then infers the distribution of topics in every

documents and the distribution of words in each topic by analyzing the co-occurrence patterns of words in the documents.

The LDA model is based on the Dirichlet distribution, which is a multivariate probability distribution used to model the distribution of probabilities in a multinomial distribution. The Dirichlet distribution is used to model the distribution of topics in the documents and the distribution of words in each topic.

LDA is widely used in various applications, including information retrieval, recommendation systems, and topic modeling. It has proven to be a useful tool for discovering hidden patterns and relationships in large collections of text data.