DS 810: Big Data & Artificial Intelligence

Predicting High Booking Airbnbs: Maximizing Return on Investment

Shawn Bedard, Pegah Karimi, Lisa Olsson,
Amie Rowland, Hannah Wirth, Sam Woodward

We, the undersigned, certify that the report submitted is our own original work; all authors participated in the work in a substantive way; all authors have seen and approved the report as submitted; the text, images, illustrations, and other items included in the manuscript do not carry any infringement/plagiarism issue upon any existing copyrighted materials.

Table of Contents

Executive Summary

"Predicting High Booking Airbnbs: Maximizing Return on Investment" uses predictive methods to classify Airbnbs in the Nashville, Tennessee market as high booking and subsequently uses that classification to calculate the property's return on investment. The importance of this project was to be able to provide recommendations to an investor interested in building a portfolio of properties and determine what variables influence high booking rates in Airbnbs.

After running a series of classification models, it was found that high booking Airbnbs in Nashville could be predicted with 88% accuracy. Then, after identifying which properties were predicted as high booking, the estimated cost of these properties was evaluated to calculate the expected ROI and determine which Airbnbs could be the most lucrative to our investor. In addition to the classification model analysis, a panel data analysis was conducted to determine the impact of Airbnb-related variables on high booking rates.

The classification model and return on investment analysis resulted in a total recommendation of 22 properties for consideration by the investor, at an estimated investment of $7.7 million. Additionally, further recommendations such as buying properties centrally located in downtown Nashville and various host behaviors to increase the likelihood of achieving a high booking (subsequently resulting in a high return on investment) were made based on the model analyses.

Introduction

The main purpose of this project is to explain why certain Airbnbs achieve a high booking rate while others do not which ultimately will help an investor decide what properties to invest in based on their likelihood of becoming a high booking Airbnb. After considering different market characteristics, opportunities, and risks, this project decided to focus on Nashville, Tennessee.

A recent article published February 3, 2023, had a headline of "Nashville named one of the best cities in US to open an Airbnb in new report" (Gerasimenko, 2023). Another article published by WZTV Nashville reported that Airbnb hosts in Nashville collectively made $260 million in 2022 which is about $24,000 per host (Keller, 2023). With its warm weather and variety of activities and attractions, Nashville is a popular city for tourists including college students on spring break,

country music lovers, bachelorette parties, and group vacations. When compared with Airbnb-filled cities such as New York City or Los Angeles, booming cities such as Nashville appear to be a lucrative option for investors.

In addition to the external research on the city of Nashville as an Airbnb investment location, a brief survey of individuals that have vacationed in Nashville was conducted to learn more about why a tourist might want to choose to book an Airbnb in Nashville.

> *"My friends and I visited Nashville for our senior year spring break trip to experience a new part of the world for all of us. We loved the atmosphere the nightlife had to offer and the constant live music. One of our favorite hobbies during the trip was keeping track of how many live bands we saw perform on a given night, every night had a minimum of 12." - Jenson Scott, '22*

> *"For spring break a couple of my friends and I booked an Airbnb about half a mile away from Broadway Street. During the day we visited various tourist attractions nearby, such as the Johnny Cash Museum and the Country Music Hall of Fame, and at night we visited the bars to experience the nightlife. The whole experience was surreal, and our apartment was beautiful; it was a one-of-a-kind experience, and I wouldn't change a single part of it!" - Jack Sharpe, '22*

Project Objective

The primary objective of the project is to provide recommendations to a property investor on which potential listings in Nashville, Tennessee will generate the highest return on investment (ROI) when run as an Airbnb. The objective function is to maximize the investor's ROI on an investment. To calculate ROI and understand what impacts ROI, the project will also predict whether an Airbnb achieves a high booking rate. The booking rate will determine how much an Airbnb will generate based on how many people book a stay at a particular Airbnb.

The project focuses on the city of Nashville to realistically fit within the constraints of the investor's budgeting, time, and management of investments, as well as to capitalize on the claims that Nashville is one of the most profitable cities for Airbnbs. Some benefits and opportunities of focusing on the specificity of choosing one city are the ability to achieve in-depth domain knowledge about the market and the minimal barriers to entry (minimal regulations on Airbnbs). On the other hand, a major risk to focusing investments to only Nashville is the risk of being too specific and the Airbnb not performing as well as expected.

To address the primary objective, the project uses a predictive model to predict a property's probability of having a high booking rate while controlling for the variables that are impacted by how a host runs its Airbnb and the property's features. Based on the classification of a property, the return on investment will be calculated based on the Airbnb's price per night and the cost of the property. In addition to the classification analysis, a panel data analysis will be conducted to investigate the relationship between high booking Airbnbs and host and property variables.

The goal of the investor is to obtain a high return on investment and to minimize "lemons", or properties that turn out to be unsatisfactory even if they were classified as having potential to be high booking.

Research Questions

Given the large scope of the project, the following research questions were defined to help guide the analysis:

1. What variables influence an Airbnb located in Nashville, Tennessee achieving a high booking rate and subsequently influence a high return on investment (ROI) for an Airbnb?
2. What recommendations can we provide to an investor on how to successfully run their Airbnb (amenities offered, host response time, etc.) and ultimately achieve a high booking rate?

The defined questions narrow the project's focus to the investor's goal of maximizing their return on investment. While there are many factors to consider when investing in properties, focusing on return on investment will provide a fair metric on whether a property is a good investment or not.

Methodology

Data

The dataset used to complete the project's analysis was a 25,000-property sample of Airbnbs containing 385,629 observations and 26,461 properties total. The dataset has a panel structure, meaning that for one Airbnb there are multiple observations over time for that property.

The variables that were selected for the analysis focused on three categories: host behavior variables, property variables, and host-controlled variables. The purpose of selecting these types of variables was to understand what behaviors and characteristics are more likely to occur with highly booked Airbnbs. Analyzing the influence of a host's behavior will provide important recommendations to an investor on how to run a successful Airbnb based on response times, booking acceptance rates, and customer feedback. The property and host-controlled variables provide information on what room types are most highly booked, what amenities are more likely to appear in highly booked Airbnbs, and what variables set by the host (price, whether there is a security deposit or not, etc.) are most highly booked.

The reference variables were selected for grouping the data by Airbnb (`id`), predicting the variable of interest (`high_booking`), and determining which Airbnbs from the original dataset are within the Nashville radius (`latitude` and `longitude`).

Reference variables	Host behavior variables	Property variables	Host-controlled variables
id	host_response_time	room_type	price
high_booking	host_response_rate	accommodates	security_deposit
latitude	host_acceptance_rate	bathrooms	cleaning_fee
longitude	host_is_superhost	bedrooms	instant_bookable
	review_scores_rating	neighbourhood_cleansed	
	review_scores_accuracy	amenities	
	review_scores_cleanliness		
	review_scores_checkin		
	review_scores_communication		
	review_scores_location		
	review_scores_value		

Assumptions for Analysis

To build a classification model to classify a property as having a high booking rate or not, a number of assumptions had to be made.

1. We assume that all host-related variables are fixed, meaning that if an Airbnb changes owners at any point, the account and its host attributes (response time, super host, reviews) remain attached to the Airbnb itself.
2. We assume that if any value in the `security_deposit` and `cleaning_fee` columns are NA, their value is 0 (as in the security deposit and cleaning fee are $0).
3. We assume that the median property price for Nashville houses or apartments can be inferred by the number of bedrooms and is sufficient to model the cost of the property (as in, we are not accounting for maintenance,

renovations, etc.). The Rocket Homes' Nashville Housing Market Report was used for data on median housing prices (Rocket Homes, 2023).

4. We assume that the occupancy rate of an Airbnb (what percentage of the year the Airbnb is occupied and therefore is being paid for) can be inferred from the `high_booking` variable. If an Airbnb is classified as 1 (having a high booking rate), the Airbnb has an occupancy rate of 75%. If an Airbnb is classified as 0 (not having a high booking rate), the Airbnb has an occupancy rate of 25%. This assumption was estimated using values below and above 36.9%, the average occupancy rate of an Airbnb in Nashville (Average Airbnb Occupancy Rates by City [2022], n.d.).
5. We assume that the total yearly revenue generated by a property can be calculated by multiplying an Airbnb's nightly price by the occupancy rate multiplied by 365 days.

$$Yearly\ revenue\ =\ Nightly\ price\ \times (Occupancy\ rate \times 365)$$

Data Cleaning and Preparation

Location

Our initial dataset included Airbnbs from areas across the US. After we identified Nashville, Tennessee as the emerging market of interest, we filtered the dataset to extract only those Airbnbs in Nashville. To achieve this, we could not perform a straightforward selection of city and state, as there were missing values in those columns. Instead, we used the latitude and longitude columns by filtering the data to include only those points within a specific radius of Nashville, Tennessee, which has coordinates (36.174465,-86.767960). We utilized the Leaflet library to assist with this, which allowed us to map the 9301 Nashville observations and confirm that our selection included our desired subset of the data.

Preliminary Data Cleaning

The preliminary data cleaning process involved the following procedures:

1. Removing dollar sign and comma characters from the `price`, `security_deposit`, and `cleaning_fee` columns and converting to a numeric data type to use columns for calculations.
2. Removing the percent signs from the `host_response_rate` and `host_acceptance_rate` columns, converting the column to a numeric data type, and dividing the numeric value by 100 to obtain a numeric percentage value.
3. Converting TRUE/FALSE values in `host_is_superhost` and `instant_bookable` columns to dummy 1/0 values.
4. Cleaning the `amenities` column.
    a. Removing unnecessary characters from list of amenities.
    b. Expanding the list of amenities to separate each unique amenity into one row.
    c. Pivoting the list of amenities grouped by Airbnb into unique columns by amenity and creating dummy 1/0 variables (1 = Airbnb has amenity, 0 = Airbnb does not have amenity).
    d. Summarizing and counting the number of unique Airbnbs that contain each amenity to determine the most found amenities in Nashville Airbnbs.
    e. Selecting the top amenities that were property-specific and excluding amenities that were not inherently permanent (i.e., shampoo, conditioner, etc.).
    f. Combining columns with similar amenity names – for example, the air conditioning and central air conditioning columns were combined.
        i. Amenities included: Air conditioning, heating, kitchen, parking on premises, private entrance, patio or balcony, street parking, grill, and pool.

Imputation

After the preliminary data cleaning, the dataset contained a total of 19,548 NA values. To run classification models, the options to accommodate the NA values are to drop the NA values or impute values. Given that the subset of Nashville properties was already a smaller sample (9,301 observations), the decision to impute values was made.

The first simple imputation was to impute 0 for any rows that had a security deposit or cleaning fee as NA.

For the remaining numerical variables, the predictive mean matching (PMM) imputation method was used to fill in the missing values. PMM calculates a predicted value of the NA value based on a set of donors that have predicted values closest to the missing entry. A donor is randomly selected from the set and the donor's value is imputed for the NA value (Van Buuren, n.d.). For this project, a set of five donors was specified with a maximum of fifty iterations. The PMM method appeared to be the best option for imputation given that NA values for `price` should be inferred from similarly attributed Airbnbs.

Training and Testing Sets

The cleaned dataset was split into training and testing sets grouped by Airbnb (`id` column) to account for the panel data structure. The training and testing sets include all observations of an Airbnb and do not exclude any instances of an Airbnb. The training and testing set split was 70% training and 30% testing.

Return on Investment (ROI) Calculation

Based on the assumptions listed above, ROI was calculated with the following steps.

1. *Determining the occupancy rate of each Airbnb*

The occupancy rate was determined based on the `high_booking` column. To eliminate occupancy rates of 0% or 100%, a 75% rate was applied to Airbnbs with high booking rates and a 25% rate for Airbnbs without high booking rates. Since there are multiple observations of one unique Airbnb, the average of all occupancy rates at any point in time of one Airbnb (i.e., if the Airbnb was classified as high booking at one time and not high booking at another) was taken to find the average occupancy rate per Airbnb.

2. *Calculate the number of days out of the year that an Airbnb is occupied.*

To determine the number of days out of the year that an Airbnb is occupied, the average occupancy rate per year was multiplied by 365 and rounded to the nearest integer.

3. *Calculate yearly revenue generated by an Airbnb.*

Yearly revenue was simply calculated by multiplying the `price` column by the number of days out of the year an Airbnb is occupied.

4. *Determining the total cost associated with an Airbnb.*

To determine the cost of investing in an Airbnb, the number of bedrooms associated with a property was used to determine what price the property would have been bought at. This data was pulled from a 2023 report from Rocket Homes.

Number of bedrooms	Price (in USD)
0	0
1	324,900
2	330,500

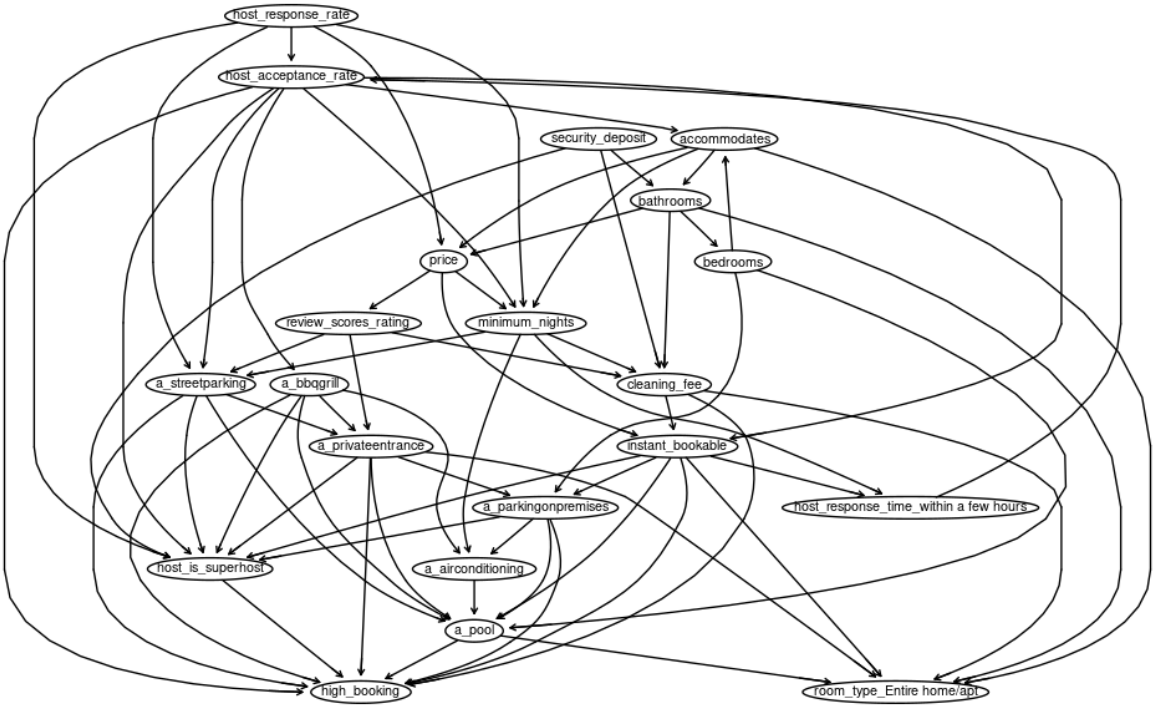
3	444,200
4	649,900
5+	1,200,000

5. *Calculating return on investment (ROI)*

The final step, calculating ROI, was done by dividing yearly revenue by the property cost.

Exploratory Data Analysis

Narrowing down to Nashville, it was important to first understand the variables in context to the city itself. Using PC algorithm for causal discovery as it ignores confounders allowing for a simplistic understanding of the relationships between various variables of interest, the following PDAG was created:



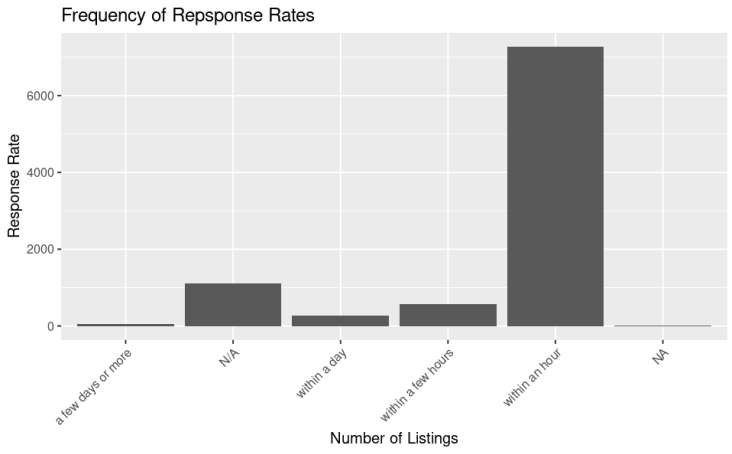
rcode

Figure 1

Based on the PC algorithm results, the high booking appears to be directly affected by the host acceptance rate, host being a super host, having a grill, street parking, private entrances, a pool, parking on premises, having a cleaning fee and being instantly bookable. There are many other variables that relate or have a causal link to those mentioned prior. There are a few links that do not appear accurate based on logical inference, like the host response rate causing the minimum nights of the stay, or the instant bookable affecting if there is parking on premises. These causal links can be ignored with domain knowledge or can be inferred that there is another hidden effect between the two variables. For the variables that appear to have causation, they can be further explored through exploration.

Host response rate is a variable that has relationships with many other variables in this data set. This is a percentage between 0 and 1. What is interesting is that there is not a direct connection between the host responding withing a few

hours. There is an indirect connection with host acceptance rate in between. Below shows the distribution of response rates, showing that within an hour has most responses, whereas within a few hours does not hold as much of a presence – this could help explain the indirect link between the two variables.

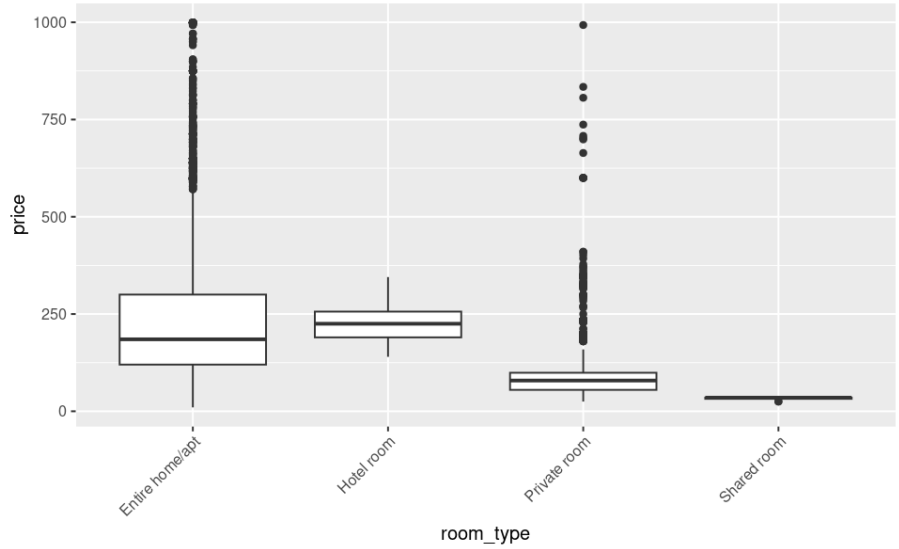


rcode

Being a super host appears to directly affect the ability for high bookings. Of all the listings in Nashville, there is almost an even split of super hosts and non-super hosts. For each Airbnb, a host can become a super host over time, so this value could change per Airbnb, but there are many hosts that have not become a super host at the time of listing.

TABLE – R CODE

Price is an important variable for both hosts and guests. Hosts need to optimize the price to attract guests based on the type of stay and features, but also need to ensure they are pricing competitively to help make a profit. In the CPDAG, price is one of the variables that has a bunch of relationships causing it as well as it is having an impact on other variables. One of the features that price has an indirect relationship with is the type of room being offered. Although many other factors may contribute to price, the price of an Airbnb may help indicate what type of rental it is. Comparing different rental types to price is seen in the figure below. With entire homes and apartments having the most listings, there is also a greater range in pricing, though the average price is just below that of a hotel room. Private and shared rooms are less frequent in Nashville but have a much lower average price. It is interesting to note that private rooms have a lot of out outliers in pricing – this could be due to data entry error, or other confounding effects that we see in the PDAG.



R code

Models

A series of classification models were used to investigate how accurately an Airbnb can be classified as high booking. While some of the models chosen for analysis are considered black-box methods and do not explain variables and their relationships, they are still relevant to the analysis of determining what variables influence an Airbnb's booking rate. Additionally, they provide a reference point for the results of the subsequent panel data analysis.

The metric used to compare the performance of each measure was measuring the rate of false positives. The rate of false positives will be calculated using the following formula:

$$Rate\ of\ false\ positives\ =\ 1 - Specificity$$

Given that the goal of the investor is to minimize the number of properties that fail to achieve a high booking rate even when they classified as such, it is preferable to achieve a low false positive rate.

Decision Tree

A decision tree model was used in the analysis for its simplicity, interpretability, and its ability to predict categorical or binary outcomes. Given that `high_booking` is a binary (0/1) variable, the decision tree was a good place to start to understand what variables might influence high booking rates. The model initially starts with all the defined Airbnb variables and recursively partitions the data to create its branches and nodes. At each node, the algorithm selects a variable and creates a split point that best separates the data into 1 (high booking) and 0 (not high booking). A downside of utilizing a decision model tree is that they tend to over fit the training data making it difficult to interpret new incoming data. Additionally, a decision tree also struggles when attempting to create a split point within continuous data, potentially leading to skewed results. The model structure of the classification of `high_booking` and the decisions of each decision node were visualized for interpretation and can be referred to in the appendix (Figure xxx). The decision tree determined the deciding variables as `host_response_time`, `neighborhood_cleansed` (specific neighborhood districts), `price`, `a_bbqgrill`, `security_deposit`, and `host_is_superhost`. The results from the decision tree model when classifying Airbnb units as a high booking property has an accuracy of 85.52% and a false positive rate of 0.949%.

Random Forest

A random forest model was created for its ability to use all the variables selected in our analysis while being robust to outliers (i.e., Airbnb properties at high or low-price points) within the data. The model functions by selecting training data through bootstrap sampling which introduces variation and enables the created trees to be more diverse. The

construction of the trees involves a process called recursive binary splitting which involves selecting a feature and creating a split point for each node. The prediction of the random forest model is made by aggregating the results of the individual trees created, allowing the model to be better suited for handling new incoming data. A benefit of using the random forest model for predicting high booking Airbnbs is the ability to add more data on recently listed Airbnbs with minimal effort. Some downsides of the random forest model are that its decision-making process is not easily interpretable and can include noisy features that alter its performance. The model created for the Airbnb analysis predicted the `high_booking` variable and was built using all the variables selected for analysis. The importance parameter (`TRUE`) enables the model to compute and assess the importance of each Airbnb variable. The proximity (`TRUE`) of the model refers to the model's utilization of the proximity matrix of the data which helps the model identify similar instances for its classification clusters. For the final two parameters, `ntrees` (1000) represents the number of individual trees built in the model and `mtry` (10) represents the number of predictor variables used at each split of the decision tree. The results of the random forest model being able to identify an Airbnb unit as a high booking unit has an accuracy of 83.67% and a false positive rate of 1.793%. While there is a loss of interpretability when using a random forest model, the results were still useful when deciding which classification model to use for final analysis.

Logistic Regression

Logistic regression was used to model high bookings because of its ease of set up, efficiency in training and testing, and interpretability, assuming linearity between the dependent and independent variables. The first logistic regression model that used all the variables in our selected dataset. This model classified Airbnbs as high booking with an 86.47% accuracy and a 2.2% false positive rate. The second logistic regression model used variables that the 'feglm' function of 'alpaca' library found not to have multicollinearity issue. These were chosen by manually removing variables one by one until they did not violate the multicollinearity assumption. This model classified high bookings with an 88.02% accuracy and a 0.8% false positive rate. The third and final logistic regression model used parameters based on the subsequent panel analysis's findings. The `feglm` function of 'fixest' package automatically removed variables or specific factors of the variables from the model that caused issues with multicollinearity – the remaining variables from the `feglm` panel analysis were used in this logistic regression model. The third model classified with an 85.61% accuracy and a 1.8% false positive rate.

K-Nearest-Neighbors (KNN)

The K-Nearest Neighbors (KNN) model is a non-parametric model that does not rely on the actual variables in the dataset but simply relies on the value given to the K parameter. The algorithm focuses on grouping the Airbnb properties based on k number of neighbors by comparing their labels and values to make the classification prediction. The distance metric used in a KNN model determines the closeness or similarity between data points. The model used the Euclidian distance metric which calculates the straight-line distance between two points. A benefit of the model is its ability to capture potential nonlinear relationships between `high_booking` and other Airbnb variables. The KNN model can also easily adapt to new Airbnb data by grouping it to similar points. A drawback of the KNN model is that it can have deteriorating effectiveness when the dimensionality of the data increases and may impact the accuracy of the distance-based calculation. KNN models are also sensitive to outliers and noisy data which may make the Airbnb model ineffective if there are Airbnb properties with extreme high or low prices. The results for the K-Nearest Neighbors model when classifying Airbnb listings as high booking are an accuracy of 77.9% and a false positive rate of 7.595%. The poor performance of the KNN model in comparison to the other classification models may be attributed to the high dimensionality and outliers of the data.

Summary of Classification Model Results

Model	Accuracy	Specificity	False Positive Rate

Decision Tree	85.52%	99.051%	0.949%
Random Forest	83.67%	98.207%	1.793%
Logistic Regression 1	86.47%	97.31%	2.69%
Logistic Regression 2	88.02%	99.156%	0.844%
Logistic Regression 3	85.61%	97.78%	2.22%
KNN	77.90%	92.405%	7.595%

Panel Data Analysis

The Airbnb data set consisted of panel structured data, with Airbnbs having multiple records for different dates. To account for this structure and ensure the effects of the variables are measured properly, a panel analysis was performed accounting for the within effects of each Airbnb over time. By modeling each variable separately to high bookings and controlling for id time and time for most variables, and just time for those variables that would be controlled for by the id. Using a generalized linear model, each effect was measured for magnitude and significance. These results were then compared to an ordinary least squares model to ensure effect magnitude and significance were similar for interpretation as the effects in OLS can be interpreted in the same scale of the high booking. Figure XX shows the variables that had significant positive and negative effects as well as similar GLM and OLS results. There are a few other variables that were not significant but that were deemed important to examine the effect size, these have been included in the results.

Variable	feglm effect	feglm se	feols effect	feols se
Host Acceptance Rate	-1.03642.	0.586306	-0.11857**	0.040801
Host is Superhost	0.15503	0.270266	0.001866	0.013941
Security Deposit	0.001153***	0.000246	0.000111**	4.1e-05
Cleaning Fee	0.003225***	0.000827	0.000228.	0.000128
Instant Bookable	-1.20818*	0.488091	-0.052304 .	0.027448
District 12	2.561181**	0.974722	0.304715***	0.057997
District 14	3.876016***	1.011528	0.603755***	0.055633
District 15	2.511984*	1.016996	0.282609***	0.036692
District 19	0.846321	1.015281	0.032049	0.031227
Total Amenities	-0.193024***	0.017397	-0.023543***	0.002018
Pool	0.677627***	0.061189	0.093418***	0.008485
Grill	-0.577805***	0.05853	-0.577805***	0.05853
Air conditioning	1.08934***	0.389654	0.106152***	0.02681
Parking on Premises	0.60005***	0.107474	0.060187***	0.009061
Private Entrance	-0.869761***	0.072891	-0.109853***	0.00764

It was found that the neighborhood districts had the largest positive significant effects on high booking. Having an Airbnb in neighborhood districts 14, 12, and 15 (east of downtown) contributes to an Airbnb being a high booking by 60%, 30%, and 28% respectively. It was interesting to see that district 19 (downtown area) did not have a significant effect, though it does show that it is a positive effect which would make sense as people tend to gravitate there to be close to attractions. The next largest positive effect size was for amenities of a pool and air-conditioning. These make sense because of the warmer climate of Nashville, especially in the summer. Being a super host also had a very small effect and was insignificant. Being a superhost is not a significant variable in predicting whether a listing will have a high booking.

The largest negative effects came from having a grill (-58%) followed by the host acceptance rate (-12%). Having a grill does seem to be surprising that it would have a negative effect, but being a vacationer in Nashville, most will be exploring the food scene so having a grill may not be as important. The host acceptance rate was another surprising negative effect as it would be thought that as the host accepts more high booking would also increase. In attempting to check for collider bias and if there were effects needing to be controlled for by adding variables to the model and the fixed effects, there was nothing that significantly changed the effect. This would need to be further tested to check this effect.

Other effects that were surprising were that are not listed in the table are that cleaning, and security fees increase the probability of a high booking. Compared to our other models this was a different effect magnitude. By controlling for price, it was found that the effect size changes to negative, which aligns more with our previous results.

Overall, the panel analysis shows that location does matter. Certain neighborhoods are more likely to get highly booked based on where there are relevant to attractions in the area. These locations are also sensitive to the types of amenities provided and those that complement the location (A/C, pool, parking) but having unnecessary amenities does not help the booking rates. It was also found that there are confounder variables that have not been controlled for that would help to better explain specific variables. There may also be collider variables that were not added to the model, and again would be necessary for getting to a better effect size.

Comparing to Overall United States Airbnb Market

In order to determine whether the classification of Airbnbs as high booking in the Nashville market is comparable to the United States Airbnb market, the highest performing classification model was run again with the original dataset of 385,629 observations. This logistic regression model had 82% accuracy and 9.8% false positivity rate, which indicates a predictive performance that does not measure up to the 88% accuracy and 0.8% false positivity rate that the model had for Nashville specifically. Thus, our model to predict high booking rate performs better for Nashville than in the overall market, which we expected because we developed the model specifically for Nashville.

Results and Findings

Return on Investment (ROI)

Our top recommendation, based on estimated ROI, is listing ID 960164, which has an estimated ROI of 31% and an investment payback period of 3.1 years. The estimated property cost is $324,900. We have compiled a list of 22 properties in Nashville, along with their respective estimated ROIs and property costs, which will be a crucial tool for investors to reference when making investment decisions. The estimated property cost, ROI, and payback time are listed in this summary. If investors can purchase the property for the estimated cost, they will know what ROI to expect. If they negotiate a bargain and purchase a property below the estimated property cost, they will know to expect a higher ROI. They can also reference properties that are more expensive than our estimations to see if the ROI is worth it. The calculations are simple to redo when exact numbers are available; however, this initial list is a great starting point for investors to identify properties of interest. Our list is short, only 22 listings, because we only used our test set to generate it. The model was used to predict high booking rates, from which we estimated the occupancy rate, yearly revenue, and finally, ROI and payback time.

A tibble: 22 × 5

id <int>	avg_roi <dbl>	avg_payback_time <dbl>	property_cost_avg <dbl>	neighbourhood <chr>
960164	0.318781163	3.136948	324900	District 19
461496	0.168667282	5.928832	324900	District 15
423245	0.164730857	6.517201	330500	District 15
752493	0.160378208	6.235261	444200	District 19
227524	0.147583872	6.775808	324900	District 1
682902	0.131117267	7.626761	324900	District 19
191699	0.078936456	12.736026	324900	District 24
326008	0.077562327	12.892857	324900	District 19
48377	0.070406279	14.203279	324900	District 6
36053	0.069446168	14.762388	649900	District 16
213187	0.064430437	16.797872	444200	District 6
818704	0.055770045	17.930773	330500	District 30
44562	0.053155825	18.812614	330500	District 21
980165	0.042243767	23.672131	324900	District 15
739556	0.042166821	23.715328	324900	District 7
786035	0.039939486	25.037879	330500	District 20
549488	0.026838001	38.114507	324900	District 6
16567	0.026297322	38.026685	324900	District 26
188925	0.025287781	39.544791	324900	District 24
313345	0.015437571	68.936457	324900	District 14
638919	0.014349030	69.691120	324900	District 19
362905	0.008925823	112.034483	324900	District 12

22 rows

Recommendations for Investor

The recommendations for the investor can be summarized as the following.

1. **Focus on centrally located Airbnbs** – Airbnbs located in District 19 and District 15 have high return on investments. Intuitively, this makes sense given that District 19 is the central downtown area of Nashville and District 15 is the neighboring district to downtown. The downtown area has multiple tourist attractions that appeal to visitors staying in Airbnbs. If an Airbnb is more centrally located, then a higher price can be charged.
2. **Optimize host-controlled Airbnb attributes and host behavior** – The results of the models revealed that it is advantageous to maintain high overall ratings to achieve a high booking rate and to make the Airbnb instant bookable. Higher overall ratings and the ability to instantly book an Airbnb have positive effects on the probability of an Airbnb achieving a high booking rate. Additionally, maintaining zero or low security deposit and cleaning fees is preferable to achieve a high booking rate. The model analysis reports a security deposit and cleaning fee having a negative effect on high booking rates, indicating that it is better to have no security deposit or cleaning fee at all or to have low values for both.

Conclusion and Discussion

Thus, using this research we were able to recommend 22 properties to the investor which total an investment of $7.7 million. These recommendations were made based on our analysis which identified which variables impact booking rate, predicted with 0.8% false positivity rate and 88% accuracy whether a property would be high booking, and estimated the ROI of high booking properties to find those most likely to be good investments. This achieves the investor's goal of avoiding "lemons" and attaining the highest possible ROI within Nashville.

Our research demonstrated the importance of location, which we noticed at several different points of the analysis. One such example is our best high booking rate predictive model performing better within Nashville (the market we developed it for) than in the larger dataset, which indicates our model is customized to the area and performs better in Nashville than in the overall market. The importance of location was also demonstrated by the property which has the highest predicted ROI from our results, which is in downtown Nashville. Additionally, our panel analysis showed that a property's neighborhood is statistically significant for predicting whether that Airbnb is high booking. Considering these results together, we find that location is the most crucial factor in determining the success of an Airbnb.

Another finding from our analysis was the differences between our logistic regression and panel analysis models. These models found opposite effects for cleaning fee, security fee, and instant bookable status of the Airbnb, which suggests more investigation into potential confounders and colliders would be beneficial. Further, although we did not end up using the panel analysis model resulting from the Alpaca library, this modeling helped us identify which variables to improve in our logistic regression model which ended up being the best performing model for predicting high booking rate. We did not initially expect the development of these models to be interrelated to this extent, and we find this part of the process both instructive and fascinating.

Lastly, these findings have a few limitations. First, we are constrained by the numerous assumptions made, particularly by the external data that we used regarding the expected costs of purchasing in Nashville. If the market changes, these estimates will no longer be reliable, which would undermine our ROI. Likewise, our analysis is based on the current local policies in Nashville, which require permits prior to listing Airbnbs (Nashville, n.d.). If policies surrounding short term rentals in Nashville change, this could also impact our findings as well as the viability of investing in this location. However, despite the limitations, we believe our results are quite compelling and that there is potential for additional work in this area going forward.

Future Research

There are a variety of possibilities for future research to build upon the results from this project, including:

1. Update the expected buying price to customize per neighborhood (as currently prices are for Nashville as a whole). This would allow more accurate ROI estimation, which would improve our recommendations to the investor.
2. Expand panel analysis to investigate possible confounders and colliders, due to some unexpected effects in the panel models.
3. Acquire external data of listings on the housing market to identify possible purchases that could be converted into Airbnbs.
4. Expand analysis to other locations outside of Nashville, which would allow the investor to spend more of the budget and reduce the inherent risk of local policy change by diversifying the investments.
5. Analyze scaled data, as our analyses primarily focused on techniques using unscaled data.
6. Utilize text data analysis techniques, such as sentiment analysis or topic modeling, on the text fields in the Airbnb data.
7. Investigate the potential shift in the scale of review ratings, as suggested by Naveen in our presentation comments.

References

Average Airbnb Occupancy Rates By City [2022]. (n.d.). AllTheRooms. https://www.alltherooms.com/resources/articles/average-airbnb-occupancy-rates-by-city/

Gerasimenko, K. (2023, February 8). Nashville 2nd best city to start an Airbnb, says new report. *WZTV*. https://fox17.com/news/local/middle-tennessee-apple-google-news-inews-facebook-instagram-vacation-travel-southwest-nashville-named-2nd-best-city-to-start-an-airbnb

Keller, S. (2023, March 17). Airbnb hosts in Nashville collectively earned $260 million in 2022. *WZTV*. https://fox17.com/news/offbeat/nashville-tn-entertainment-disctrict-broadway-bars-bachelorette-weekend-nashvegas-middle-tennessee-airbnb-hosts-in-nashville-collectively-earned-260-million-davidson-county-local-news

Metropolitan Government of Nashville & Davidson County. (n.d.) Short Term Rental Property Permit Information. https://www.nashville.gov/departments/codes/short-term-rentals

Rocket Homes. (2023, April). *Nashville, Tennessee Housing Market Report April 2023 - RocketHomes*. https://www.rockethomes.com/real-estate-trends/tn/nashville

Van Buuren, S. (n.d.). *Flexible imputation of missing data*. Stef Van Buuren. Retrieved May 6, 2023, from https://stefvanbuuren.name/fimd/sec-pmm.html

Appendix

Panel Analysis

vars	notes	
host_acceptance_rate + security deposit	-0.995020 0.001119 Bic 5152	Both significant
host_acceptance_rate + security deposit + instant bookable	- 0.872244 0.001094 - 1.132253 Bic 5141	Host acceptance rate becomes insignificant – makes sense as if instant bookable, host will not need to accept -
security deposit + instant bookable	0.001119 -1.182964 Bic 5139	Interesting as would expect instant bookable to provide a higher booking as its easier and don't have to go through the hsot But could be harder as the host can not control dates and people could be booking in a way that overlaps etc
security deposit + instant bookable + host response time (na, few days, few hours, hour)	0.001383 -1.067054 -1.955469 -2.718938 -2.705110 -2.79395 Bic5147	Only na not significant Interesting how all rest have similar effect sizes and are negative would think they would be positively influencing high bookings A relationship we aren't controlling for?
security deposit + host_accpetance_rate + host response time (na, few days, few hours, hour)	0.001393 -0.874656 -1.886678 -2.698287 -2.650453 -2.773590 Bic 5158	Acceptance rate and na not significant