# The Airbnb Revenue Model: Analyzing KPI Accomplishments

Sajid Hussain Rafi Ahamed, Shahid Shakil, Siddharth S, Hailey Thanki and Tapan Pradyot

***Abstract*— Airbnb has proliferated over the past several years, with millions of tourists using the service. The hotel industry loses approximately $450 million in direct revenue per year to Airbnb. We collected, analyzed the Airbnb listings in the Phoenix market, and built an artificial intelligence system that predicts the property's availability based on various amenities, locations, and rental prices on Airbnb. The Random Forest model performed very well with an accuracy of 92.05%, and we identified the Airbnb earnings, date of the month, number of bathrooms, and property capacity as the key features that helped the model predict with high efficacy.**

## I. INTRODUCTION

Airbnb became one of the most popular sharing economy platforms within a few years after it was founded in 2009 [1]. Over 150 million worldwide users have booked over one billion stays, with at least six guests checking into an Airbnb listing every second, and 400,000 companies directly engage with Airbnb to manage the travel for their employees. Hosts have earned well over $110 billion with their listings.

Central Reservation System used in the hotel industry centralizes reservations, distribution, rates, and inventory in real-time. Our study aims to help the individuals in the hospitality business by predicting accommodation availability during the renters' target time frame. For Airbnb, which makes money from service fees from bookings charged to guests and hosts, their customer base typically seeks accommodations with a homey feel that hotels cannot provide, while most hosts want to rent out their homes to supplement their income. Our goal is to predict the availability of these lodgings using machine learning and their identify key performance indicator.

## II. LITERATURE REVIEW

To predict the availability of an Airbnb property shortly, it is crucial to know the reasons which lead to the large number of bookings accommodations provided by different hosts Li and Yang et al.[2]. It is crucial to analyze if the impact is based on location or the structure of the other businesses Moon, Hyoungeun, et al. [3]. The fast pacing improvements have resulted in a significant increase in the clash between Airbnb and various other businesses. There is a compulsion to improve and diminish the expenses to keep the flow of visitors attracted to themselves. To match the steps with the business of Airbnb, it is essential to have a proper understanding of the business model and numerous segments. The business models should be changed from time to time to maintain growth and development [4]. The Airbnb host should also be aware of what is expected from the property as compared to the hotels [2]. Various attributes of the listing affect the prices of Airbnb, which makes it essential to have a look at the relation between several attributes and how they differ while reflecting the price. It is also essential to match the needs of the visitors to make them revisit the place [5] The host thereby can decide the attributes which help the community grow more by keeping the check on the prices as well [4].

Previously, Yu and Wu's [5] tried to implement a real estate availability and price prediction using feature importance analysis along with linear regression, SVR, and Random Forest regression. For our primary goal, to predict growth and availability, we reviewed Oskar et al. [7], which is a purported successor to Facebook Prophet to provide revenue forecasting at scale. To help with driving our data analysis, the methods provided by David et al. [12] will be helpful in our study. Lastly, [9] will help us with picking the necessary Statistical testing methodology we require to extract critical actionable insights from our dataset.

## III. HYPOTHESIS AND GOAL

We expect to answer the following with our project:

- What data attribute would you eliminate from your study because it is untrustworthy or could be a block rather than a real key that aids in the prediction?
- What is an effective method for estimating unit occupancy and revenue?
- Which month does it appear to be the most profitable?
- What is the revenue difference between various types of rooms on offer?
- Which machine learning algorithm has the highest accuracy in predicting hotel availability?

## IV. DATA

### A. About the dataset

The dataset used has daily availability and pricings for AirBNB listings in the Phoenix market from 4/1/18 to 5/31/18. The two data tables provided as flat CSV files are structured as scraped_listings.csv and scraped_data.csv. We have 1,048,576 observations in the dataset.

### B. Features in the dataset

scraped_listings.csv has 1 row for each 'scraping_id' with various information about that listing. This file has the following features:

- scraping_id – ID key
- listing – URL link to the Airbnb posting
- city – Name of city within the Phoenix market
- lon – Longitude of the unit rented

- lat – Latitude of the unit rented
- mapped_location – A google maps URL of the location
- name – Posting name
- capacity – Number of people is said to accommodate
- bathrooms – Number of bathrooms at unit
- bedrooms – Number of bedrooms at unit
- has_pool – 1 if unit has pool listed; 0 if does not
- cleaning_fee – The amount in dollars to cover cleaning
- i_superhost – 1 if the host is a superhost; 0 if not
- host_name – Name of host

scraped_data.csv has the availability and price of all listings for April to May 2018. This file has the following features:

- scraping_id – ID key
- as_of_date – Date the information was scraped
- date – Date of the night to be booked
- price – Price in dollars of the night
- available – 1 if unit is available to be booked; 0 if not

### C. Exploratory Data Analysis

We did extensive exploratory data analysis to understand the trends underlying the dataset. First, we began with visualizing the target variable, i.e., the available column. We observe from Figure 1 that the dataset is well balanced with an almost equal number of available and unavailable listings.

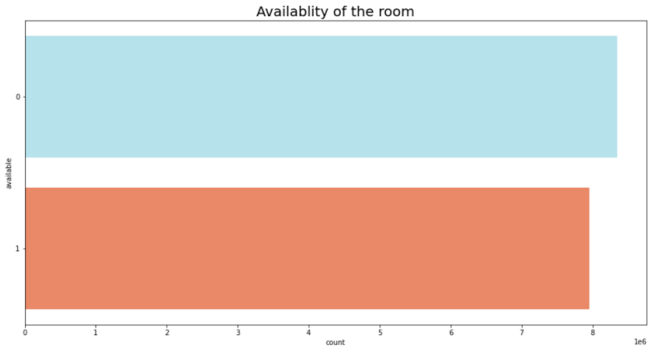


Fig. 1. Count plot of the Target Variable

We wanted to know how the prices of the listing vary through various months and dates. In Figure 2, we see that the prices have increased drastically in May compared to April. This may be because many people move around taking a vacation at the start of summer. Hence the prices inflate with the increase in the demand.

For a property to attract people, it should provide various amenities to make their stay more amiable. Next, we analyzed the various facilities that the listings provided in the Phoenix area.

From Figure 3, we can depict that more than 6000 properties out of 8445 have pools. This shows that Phoenix's climate is a little harsh during summer. People will prefer Airbnb to have pools. Another essential feature that the customer pays attention to is the number of bedrooms, and we found that most of the properties in this area mainly have about 2-4 bedrooms. Surprisingly, there are properties with 0

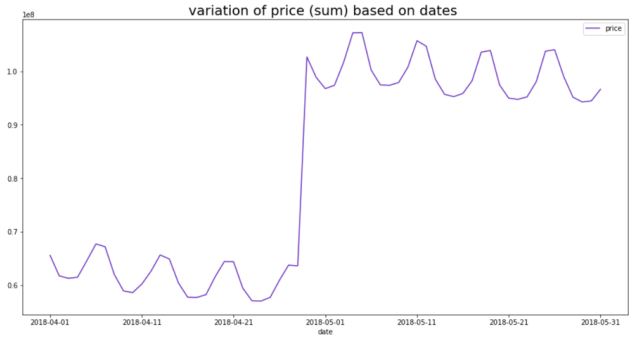


Fig. 2. Variation of price of the listing with respect to dates

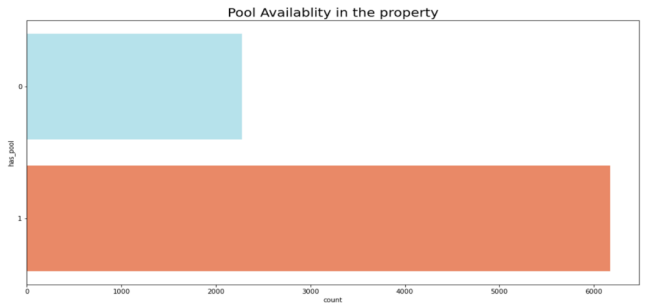


Fig. 3. Count plot of the number of the properties with pools

bedrooms, and null values are present in this column which needs to be taken care of during data cleaning.

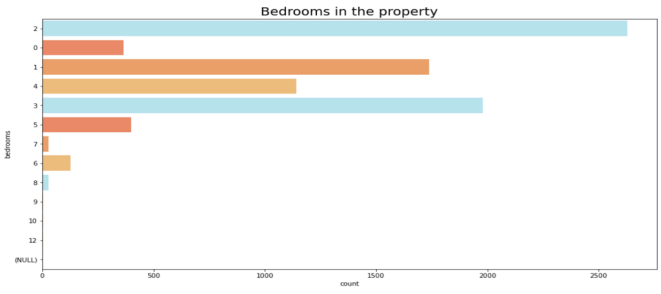


Fig. 4. Number of bedrooms available in the properties

With the different number of bedrooms, we found that the cleaning prices of prices increase with more bedrooms as expected. Also, we found certain outliers in the prices of the properties having more than eight bedrooms.

With the feature of longitude and latitudes of the listings, we plotted their location based on their coordinates. We observed that the properties are well spread out through Phoenix city, indicating that customers have options to choose from based on their requirements.

Overall, with EDA, we learned about the trends in the data, anomalies, and null values that needed to be handled during data cleaning.

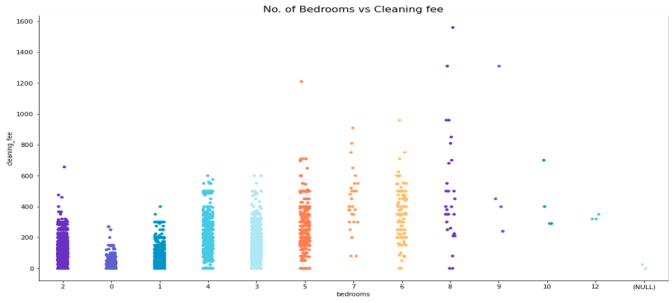


Fig. 5. Trend of cleaning prices v/s number of bedroom

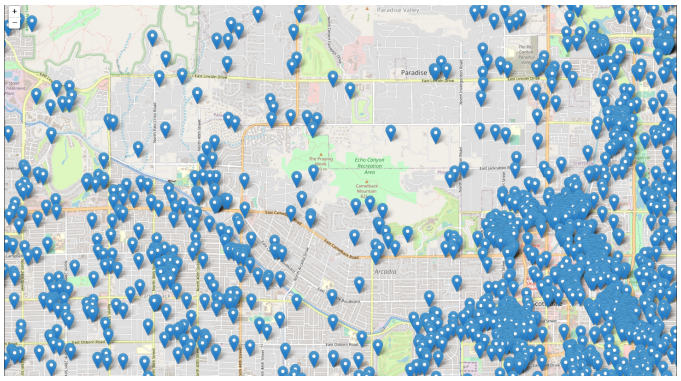

Fig. 6. Trend of cleaning prices v/s number of bedroom

## V. Methodology

### A. Data Preprocessing

Amenities, price, and date of the listing are the columns that would most likely lead to greater insight into whether or not a user's listing is a "block." By block, we define it as a user intentionally setting the listing to unavailable, but no one paid for the stay (revenue = 0, availability = 0). Thus, an excellent approach to assume occupancy and revenue involved processing a new revenue feature based on Airbnb's two-sided fees for guests and additional fee slabs for whether the host is a super host or not. This feature helped with our modeling phase.

Though some of the names are duplicated, their property features are different, so we have retained them in our analysis. This might provide insight into why this feature would not help predict price. The host's name tends to be listed, and sometimes it might add superfluous words. This column is a candidate for removal since most of the description in the name is added in the other column features.

We could, in theory, lemmatize this column and add X number of columns that correspond to certain words in the description. We could create a one-hot embedding for these words and use it as a potential feature in price prediction. The columns of data we excluded are listings URL, listing names, scarping ID, scarping dates, mapped location, and hostnames.

We've determined the following columns to be of significant importance to the model.

- City: after processing the categories and removing noisy data, the city is a strong proxy for location and can help determine the price (San Francisco rent is likely higher than Orlando, Florida). So instead of doing one-hot encoding for all the cities, we just took the top 5 cities with the most listings, and one-hot encoded them.
- loc: A new feature extracted from clustering the latitude and longitude columns and cross validating with the cities column to check for appropriateness of the model.
- Capacity: great way to determine price as it is likely that larger capacity is associated with maybe square footage.
- Bathrooms: the number of bathrooms in the property, this is a common metric to determine the value of a home.
- Bedrooms: the number of bedrooms in the property, this is a common metric to determine the value of a home.
- Has_pool: Categorical variable that could determine a change in price.
- Airbnb_earning: The total price, derived from the original price column combined with cleaning fee to assess both guest total and Airbnb earnings were factored into the prediction model. The formula used is (0.14*price)+ (1.14*0.03*is_superhost*price)

Multi-collinearity can hurt any machine learning models we develop. The expected relationship between features and our target variable, available, may not hold when many features are related. This might lead to unreliable hypothesis testing results, and coefficient estimates from a linear model can be less stable. A good rule of thumb is that features above 10 in the variance inflation factor metric tend to exhibit more multi-collinearity. From initial inspection, the variance inflation factor for bathrooms, bedrooms, and capacity suggests a strong case of interdependence, as observed from the heat map below.

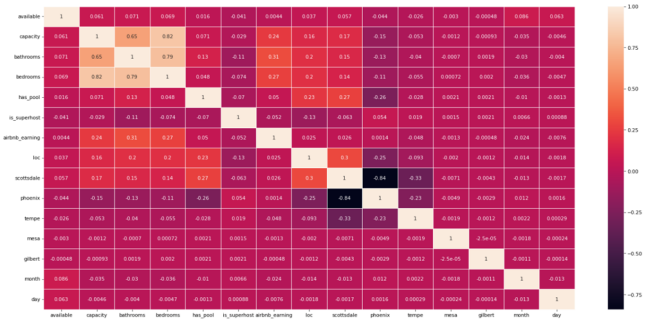

Fig. 7. Heat map of independent variables

### B. Statistical Tests and Inference

1) Chi Square tests of Independence:
The Chi-Square Test of Independence or Chi-Square Test of Association is a non-parametric test that determines whether there is an association between

categorical variables. We converted the numerical variables such as price, capacity, number of bedrooms etc. into categorical variables in order to conduct chi-square tests on different combinations of columns to understand their relationships.

Cramér's V sometimes referred to as Cramér's phi is a measure of association between two nominal variables, giving a value between 0 and 1. The following table shows how Cramer's phi or V can be interpreted.

TABLE I
INTERPRETATION OF CRAMER'S V

| Phi and Cramer's V | Interpretation |
|---|---|
| >0.25 | Very Strong |
| >0.15 | Strong |
| >0.10 | Moderate |
| >0.05 | Weak |
| >0 | No or very weak |

2) One-way ANOVA tests:
ANOVA which stands for Analysis of Variance is a statistical method for analyzing the relationship between more than two independent groups of a variable by comparing their means and their effect on the numerical dependent variable. The ANOVA method assesses the relative size of variance among group means compared to the average variance within groups.

Like the t-test, ANOVA helps you find out whether the differences between groups of data are statistically significant. It works by analyzing the levels of variance within the groups through samples taken from each of them. Put simply, ANOVA tells us if there are any statistical differences between the means of three or more independent groups.

### C. Prediction Model

We built a classification model to predict occupancy using the feature columns we listed in the pre-processing data section. The dataset was split into the train, test sets in the ratio of 90:10, and features were normalized using the Min-Max scalar. To test the performance of various Machine learning algorithms, we trained and tested Logistic Regression, KNN Classifier, Decision Tree, Random Forest, Ada Boost, and XGboost.

Finally, we also used the majority voting ensembling model. Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on the highest probability of chosen class as the output. It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is that instead of creating separate dedicated models and finding the accuracy for each of them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

## VI. RESULTS

### A. Findings from EDA

We tried to formulate certain business questions while we were approaching the problem for the first time, The EDA phase helped us put our presumptions to test which we had regarding the data.

The questions we had are

1) Which month do properties appear to generate more revenue, April? or May?

We identified that the revenue was more in May when compared to the month of April.This was proportionate to the bookings that we had during those months.

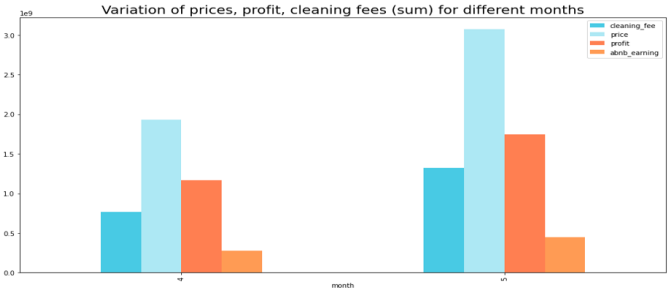


Fig. 8. Variation of prices, profit, cleaning fees for different months

2) Which month do properties potentially generate more revenue , considering listings?

The EDA suggested that the month in which the properties potentially generate more revenue based on the listings is May.This is same as when we do not try to segregate it based on the listings.

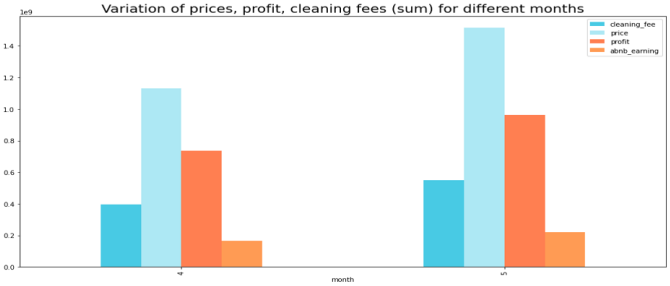


Fig. 9. Variation of prices, profit, cleaning fees for different months when it was available

3) How much revenue do places with 3 bedrooms make vs places with 2 bedrooms?

It is evident from Fig 10 visualization that the properties with 3 rooms generate more revenue than the properties with 2 rooms.This could be because people travel in a small group of family and friend and end up needing an accommodation which has 3 rooms.For the properties which have a lot of rooms they are not as frequently booked as the ones with lesser rooms.


Fig. 10. Variation of prices for 2 and 3 bedrooms

4) What are the most valuable cities for 3-bedroom or 2-bedroom apartments?

In aggregate, including all listings over all scraped dates, Scottsdale and Phoenix are the two cities with the highest potential revenue for listings with up to 6 bedrooms. The least valuable city in terms of potential revenue is Mesa. We also see that the date of bookings impacts the potential earning as well as the duration of stay which can be deduced by duplicate scraping IDs for the same listing as the prices plummet on weekday from Monday to Wednesday and then sharply rise by Friday staying stable during the weekends and showing a drop after Sunday.

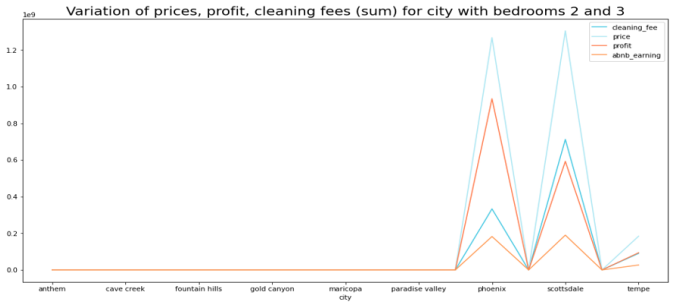


Fig. 11. Variation of prices for 2 and 3 bedrooms based on cities

5) How much of a difference does being a super host, on average, have on the price of a listing?

From these two months of data, it looks like there is more potential revenue from non_superhosts. This can be for a number of reasons, and the most likely reason is that there are more non_superhosts in the listings dataset and also the fact that being a superhost has no bearing on how much a property is listed for and what potential revenue it may generate. So, to determine which type of hosts generates the highest number of potential_revenue normalized by the number of listings over this two month period. We have 4,236,395 superhosts and 12,063,780 non_superhosts. So, to determine which type of hosts generates the highest number of potential_revenue normalized by the number of listings over this two month period, we divide and get $346 per host for non_superhosts and $193 per host for superhosts.


Fig. 12. Prices based on the super host

## B. Results of Chi Square Tests

The following are the results we obtained from the multiple chi-square tests we conducted on the different variable pairs.

TABLE II
RESULTS OF CHI SQUARE TESTS

| Variable Pair | Phi and Cramer's V | Correlation |
|---|---|---|
| Price & Availability | 0.1006 | Moderate |
| Capacity & Availability | 0.0589 | Weak |
| No. of Bathrooms & Availability | 0.0725 | Weak |
| No. of Bedrooms & Availability | 0.0654 | Weak |
| City & Availability | 0.0581 | Weak |
| Pool & Availability | 0.0156 | None |
| Super host & Availability | 0.0411 | None |
| Price & Capacity | 0.2515 | None |
| Price & Pool | 0.0670 | Weak |
| Price & City | 0.0825 | Weak |

The conclusions we made from the chi-square tests are:

- There is a moderate correlation between the following variables: price range of the Air BnBs and their availability.
- There was a weak correlation between the capacity of the Air BnBs and their availability, the number of bathrooms and their availability, the number of bedrooms and their availability, location and availability, the price of the Air BnB and whether it has a pool and the price and location of the AirBnBs.
- There was no correlation between their being a pool at the Air BnBs and their availability, AirBnBs' hosts being super hosts and their availability and the price of AirBnBs and their capacity.

## C. Results of Anova Tests

We used ANOVA to test a particular hypothesis. ANOVA helped us understand how our different groups respond. In this case, the different groups are availability being 0 or 1, that is, if the listing is available or unavailable. We tested whether there would be a significant difference in prices of available and unavailable Airbnbs.

The p-value obtained from the ANOVA analysis is significant ($p < 0.05$), and therefore, we concluded that there are significant differences in the mean prices of available and unavailable Airbnbs. We also conducted the same test on the capacity and the availability variables of Airbnbs. We found out that there are significant differences in the mean of the capacity of available and unavailable Airbnbs. The result obtained from the One-Way ANOVA Test on the Price and Availibility of Air BnBs is as follows:

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| available | 1.0 | 1.751672e+09 | 1.751672e+09 | 2348.231854 | 0.0 |
| Residual | 16300173.0 | 1.215917e+13 | 7.459537e+05 | NaN | NaN |

Fig. 13. ANOVA Results: price & availability

The result obtained from the One-Way ANOVA Test on the Capacity and Availibility of AirBnBs is as follows:

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| available | 1.0 | 6.226615e+05 | 622661.517755 | 61039.628336 | 0.0 |
| Residual | 16300173.0 | 1.662771e+08 | 10.200939 | NaN | NaN |

Fig. 14. ANOVA Results: capacity & availability

The conclusions we made from the ANOVA tests are as follows:

- The p value obtained from ANOVA analysis is significant ($p < 0.05$), and therefore, we conclude that there are significant differences among mean of the prices of available and unavailable AirBnBs.
- The p value obtained from ANOVA analysis is significant ($p < 0.05$), and therefore, we conclude that there are significant differences among mean of the capacity of available and unavailable AirBnBs.

### *D. Results of Predictive models*

The performance of metrics of the machine learning algorithms are shown below,

TABLE III
PERFORMANCE METRICS OF MACHINE LEARNING ALGORITHMS

| Algorithm | Accuracy | F1-score | Recall | Precision |
|---|---|---|---|---|
| Logistic Regression | 55.8 | 53 | 50 | 55 |
| KNN | 91.04 | 91 | 91 | 91 |
| Decision Tree | 92.04 | 92 | 91 | 92 |
| Random Forest | 92.05 | 92 | 91 | 92 |
| Ada boost | 92.03 | 92 | 91 | 92 |
| Gradient Boosting | 60.88 | 62 | 65 | 59 |
| XGBoost | 67.49 | 68 | 71 | 65 |
| Voting Classifier | 91.76 | 92 | 91 | 92 |

Of all the models, Random forest has the highest efficacy. We found that the airbnb earning feature we created has the highest importance for the model to make the availability prediction, followed by day of the month, capacity of property, bathrooms and bedrooms.

## VII. CONCLUSION

Based on our study, We identified that the Airbnb earning, day, capacity, number of bathrooms are the features that

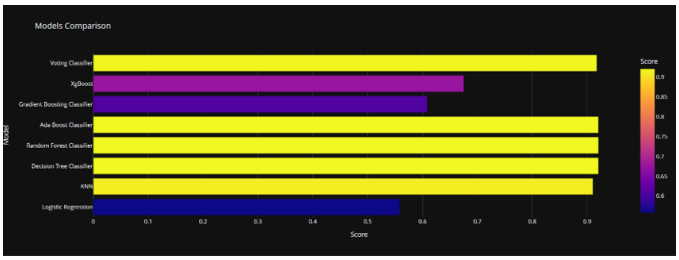


Fig. 15. Comparison of accuracy for different models

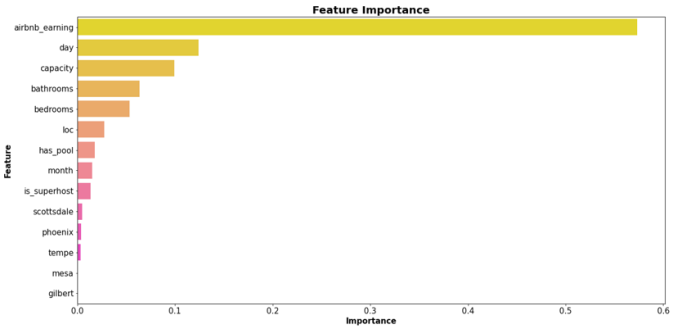


Fig. 16. Feature Importance of the Random Forest

impact the availability and price of a property.We are comfortable making the following recommendations for the Phoenix area to drive growth based on key performance indicators. A majority of listings in the Mid-Range category between $100-$1000 accounting for 90% of all listings also see the highest demand. The most popular areas for Airbnb are Scottsdale and Phoenix, with Tempe a distant third.Several statistical tests were performed to find more insights on the data. Apart from this classification model was build to predict the availability of the listing. The random forest algorithm gave the highest accuracy with 92.04%.

From our study, we conclude by stating that this predictive model will help the Airbnb search engines to predict if listing in the location be available and also for them to generate more revenue they need to have more listings with 2 or 3 bedrooms, that could accommodate a family up to 5 members.

## REFERENCES


[1] Hati, Sri Rahayu Hijrah, Tengku Ezni Balqiah, Arga Hananto, and Elevita Yuliati. "A decade of systematic literature review on Airbnb: the sharing economy from a multiple stakeholder perspective." Heliyon 7, no. 10 (2021): e08222.

[2] "Price recommendation on vacation rental websites." Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2017.

[3] "Peer-to-peer interactions: Perspectives of Airbnb guests and hosts." International Journal of Hospitality Management 77 (2019): 405-414.

[4] Sheppard, Stephen, and Andrew Udell. "Do Airbnb properties affect house prices." Williams College Department of Economics Working Papers 3.1 (2016): 43.

[5] Mao, Zhenxing, and Jiaying Lyu. "Why travelers use Airbnb again? An integrative approach to understanding travelers' repurchase intention." International Journal of Contemporary Hospitality Management (2017).

[6] https://www.kaggle.com/code/ruchi798/wids-datathon-2021-rapids-xgb-lightgbm
[7] Triebe, Oskar, Hansika Hewamalage, Polina Pilyugina, Nikolay Laptev, Christoph Bergmeir, and Ram Rajagopal. "NeuralProphet: Explainable Forecasting at Scale." arXiv preprint arXiv:2111.15397 (2021).
[8] Berthold, Michael, and David J. Hand. Intelligent data analysis. Vol. 2. Berlin: Springer, 2003.
[9] https://www.pwc.com/gx/en/audit-services/corporate-reporting/assets/pdfs/uk_kpi_guide.pdf
[10] https://www.kaggle.com/code/alexeykolobyanin/tps-nov-log-regression-with-sklearnex-17x-speedup
[11] https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/
[12] https://en.wikipedia.org/wiki/Logistic_regression