Prof. Lukasz Golab

MSCI 446: Introduction to Machine Learning

Project Proposal: Amsterdam Airbnb Price Predictor

Team 12
Aditi Lohtia - 20754943
Liliana Paroski - 20788970
Aurchon Datta - 20780084
Paloma Tonigussi - 20765573

Part 1 - The general goal of your project (obtaining insights from data towards solving your problem).

When travelling somewhere new, people often turn to popular accommodation rental sites like Airbnb to find a place to stay. Airbnb is a rental site where users can find an array of accommodations, from houseboats to treehouse cabins and everything in between. Airbnb offers not only a lot of diversity when it comes to the type of stay, but the duration of the stay as well. It can be used to find a simple apartment for a single night, or a house to rent for several months. This flexibility has made the platform skyrocket in popularity over the past decade and is why many choose to stay in Airbnb, as opposed to traditional hotels.

There are two types of users on Airbnb, the 'potential guest' who is on the platform searching for accommodation and the 'renter' who is renting out their property. As someone searching the site for somewhere to stay, the process is relatively straightforward. Users fill out search preferences and then explore the different listings available that meet these preferences. For example, someone would indicate where they would like to stay, the number of nights, and then the number of people. Users can then further refine the results by filtering for many different attributes such as property type, number of beds, or even little details like if the home has a fireplace.

On the flip side, as a renter looking to list a property on Airbnb, this can be a much more confusing process. Figuring out a reasonable rate to charge for an Airbnb accommodation can be a challenging process. There are multiple aspects to consider when looking to list properties on Airbnb such as nightly rates, cleaning fees, damage deposits, insurance, cancellation rates, and so much more. Currently, renters are often left to look at what other properties are charging on the platform and make an educated guess as to what they should charge.

This guesswork can be stressful as there are consequences to both overpricing and underpricing the rental. If the rental property is overpriced, this may turn potential guests away as they are likely to find similar stays at a more reasonable price, which puts the renter out of a sale. Now if the stay is underpriced, they are losing out on potential revenue that they could have made. This uncertainty amongst renters of what to fairly charge for their Airbnb stay is the problem our team is looking to tackle.

The goal of our project is to help people who want to post their property on Airbnb understand how much they should charge per night, based on the characteristics of their property. Our group has thus decided to take the route of obtaining insights from the data we have found online to solve our problem. Our end goal of creating a model that will predict how much someone should list their property for on Airbnb based on certain features, we will start by cleaning the data we have found and uncovering potential patterns from the data through various data analytics techniques. This will allow us to further explore what features have the greatest impact on the nightly rate of a stay. We will then implement an appropriate model and ensure that the results drawn from the model are communicated in a clear and concise way, so this tool is easy to use and understand. After all this, we hope to alleviate some of the stress new Airbnb renters may feel when setting up their rental property on the platform.

Part 2 - The specific problem you chose to study: justify your choice. Why is it important in practice? You may cite research papers, white papers, or business reports to support your claims.

In this project, we want to provide Airbnb owners with an accurate rental estimate of their property. For new renters of the app (and even current ones), it can be fairly difficult to determine which characteristics of a property can increase or decrease its value, as well as other factors to take into consideration before listing a place. It can also take a long time to truly understand the property's value, and we hope to alleviate this stress with this estimator. We decided to use a dataset for Amsterdam as it is a very touristy area, and we are curious to learn more about the pricing scheme and demand in such a city.

According to iGMS, there are several steps to ensure an Airbnb property is profitable. The first main point is the location, as areas with high-density tourist attractions would have higher demand all year round. In our project, we chose Airbnb properties in Amsterdam, the Netherlands, which is a hot tourist spot year-round. Some neighbourhoods in Amsterdam, however, may be worth more than others depending on their distance to nearby attractions. The property type and amenities are also crucial for owners when deciding to price, as units with more rooms will cost more than smaller places. "Great amenities such as a swimming pool, hot tub, or breathtaking views can significantly bump up your pricing" (iGMS).

In terms of the pricing strategies, iGMS indicates there are five main points to consider: nightly rate, cleaning fee, security deposit, extra charges, and discounts. The nightly rate is the base rate per night for guests to stay, which can vary depending on how an owner may want to charge i.e. charge per guest or charge for the entire place. The number of rooms, beds, and bathrooms may also affect the nightly rate. We account for the price of each property per night, the number of guests it can accommodate, beds, bedrooms, bathrooms, and bed type, which can help better estimate the nightly rate.

In addition to the physical aspects of the property, other details are also crucial to ensure that an owner will be able to rent and make a profit off of their Airbnb. According to Price Labs, "the key to great service is an immediate and appropriate response. Ensure that you are keeping the guests on top of all their requests." Listings that do not have great host response rates may result in fewer bookings, and the price will no longer play as big of a factor in booking rates. Our project utilizes information about how long a host has responded i.e. within an hour, a few hours, a day, or a few days or more. It also utilizes the host's response rate, which according to Airbnb is "the percentage of new inquiries and reservation requests you responded to (by either accepting/pre-approving or declining) within 24 hours in the past 30 days." A host can improve their response rate by replying to new inquiries, accepting/declining reservation requests, or pre-approving/declining trip requests.

When viewing a property to rent for a weekend, guests place a high priority on the reviews of a listing. According to a study conducted by Boston University's School of Hospitality Administration, Airbnb guests place "value on the sociability, trustworthiness, and friendliness of their Airbnb hosts and the experience," as well as high-rated reviews to ensure that the place is safe, clean, and the host is communicative. We account for different metrics related to reviews such as scores on cleanliness, location, communication and check-in. There are several categories in which guests can review an Airbnb listing, which all contribute to the overall rating of the property. Properties that have higher reviews sometimes have higher prices, as customers feel more trustworthy of the unit and are willing to spend more. Properties that are new will not have reviews, and thus may be cheaper initially until they accumulate enough reviews to increase the rental price.

There are a variety of characteristics that are necessary to take into consideration when uploading a property on Airbnb for rent, and our data consists of several traits that can help a renter determine what rental price is best for their unit. We hope that this predictor will allow owners to maximize their monthly profit based on the current listings, as well as help them save time when finalizing the rental price.

Part 3 - Dataset description: describe your datasets and include links if publicly available. Include information such as: the number of rows, a few sample rows, column names and datatypes.

To find the dataset the group spent some time searching through Kaggle and other sites to find a dataset relating to Airbnb listings that had relevant data but didn't have a model attached to it. The dataset we chose was taken from Kaggle (https://tinyurl.com/team12Airbnb) and describes the Amsterdam Airbnb market with 33 columns and 7,834 rows worth of data. Its columns are as follows: host_id, host_name, host_since_year, host_since_anniversary, id, neighbourhood_cleansed, city, state, zipcode, country, latitude, longitude, property_type, room_type, accommodates, bathrooms, bedrooms, beds, bed_type, price, guests_inlcuded, extra_people, minimum_nights, host_response_time, host_response_rate, number_of_reviews, review_scores_rating, review_scores_accuracy, reviews_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, and review_scores_value.

When listing a property on Airbnb, location is a strong factor in determining the listing price. In this dataset, there are multiple columns that describe the location such as neighbourhood_cleansed, city, state, zipcode, country, latitude, and longitude. The nicer the neighbourhood, the more chances are the host will be able to list their property at a higher price, even if their interior is not as nice as others in the neighbourhood.

Now although the importance of the property's location can trump the interior of the property, it is still an important attribute of the prediction model. The dataset has seven attributes, property_type, room_type, accommodates, bathrooms, bedrooms, beds, and bed_type, pertaining to the interior of a listing which would help determine the listing price. For example, if we keep all attributes constant besides the number of beds, the listing could change drastically.

Other key attributes include number_of_reviews, review_scores_rating, review_scores_accuracy, reviews_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, and review_scores_value, which all describe the ratings that were given from the guests. In general, the better the ratings a listing achieves, the higher potential for the listing to increase its list price. Although this is not always the case, we have decided to use these columns as attributes to help determine the price.

Table 1 below lists all the columns (transposed to rows for clarity since there are many columns) and 4 rows of data (transposed to columns). The rightmost column shows the data types.

Column	**Sample Data**				**Data Type**
host_id	3718	21669	42725	56142	Categorical
host_name	Britta	Mark	Marco	Joan	Categorical
host_since_year	2008	2009	2009	2009	Categorical
host_since_anniversary	19-Oct	15-Jun	01-Oct	20-Nov	Categorical
id	103026	8061	933385	1003865	Categorical
neighbourhood_cleansed	De Baarsjes - Oud-West	De Baarsjes - Oud-West	De Baarsjes - Oud-West	De Baarsjes - Oud-West	Categorical
city	Amsterdam	Amsterdam	Amsterdam	Amsterdam	Categorical
state	Noord-Holland	Noord-Holland	North Holland	North Holland	Categorical
zipcode	1053	1056 TM	1053	1053 LB	Categorical
country	Netherlands	Netherlands	Netherlands	Netherlands	Categorical

latitude	52.36938767	52.371207	52.36761407	52.36675578	Numerical
longitude	4.866972319	4.857291017	4.866895471	4.871953549	Numerical
property_type	Apartment	Apartment	Apartment	Apartment	Categorical
room_type	Entire home/apt	Entire home/apt	Private room	Entire home/apt	Categorical
accommodates	4	3	2	4	Numerical
bathrooms	1	1	1	1	Numerical
bedrooms	1	2	1	1	Numerical
beds	1	2	2	2	Numerical
bed_type	Real Bed	Real Bed	Real Bed	Real Bed	Categorical
price	95	95	82	110	Numerical
guests_included	2	1	2	2	Numerical
extra_people	25	0	0	10	Numerical
minimum_nights	3	3	2	5	Numerical
host_response_time	within a few hours	within a day	within a few hours	within a few hours	Categorical
host_response_rate	1	0.75	1	1	Numerical
number_of_reviews	15	2	19	0	Numerical
review_scores_rating	92	100	98		Numerical
review_scores_accuracy	9	10	10		Numerical
review_scores_cleanliness	9	10	10		Numerical
review_scores_checkin	10	10	10		Numerical
review_scores_communication	10	10	10		Numerical
review_scores_location	9	10	10		Numerical
review_scores_value	9	10	10		Numerical

Table 1: Columns, Sample Data and Data Types from our dataset

After exploring the data further we discovered that many Airbnb renters are listing their properties for over $100 (*Figure 1)*, which we want to explore further to determine why they can list their place for so much and what are the key attributes that allow them to do this. Also worth mentioning, this dataset was previously cleaned so most columns do not have many blanks, but the review columns (number_of_reviews, review_scores_rating, review_scores_accuracy, reviews_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, and review_scores_value) contain many blanks, with review_scores_communication having the highest about at 1,713 blanks, as not all listing have to have been reviewed. Since we are using these as features, we will need to make sure to remove these blanks when we are testing the model.

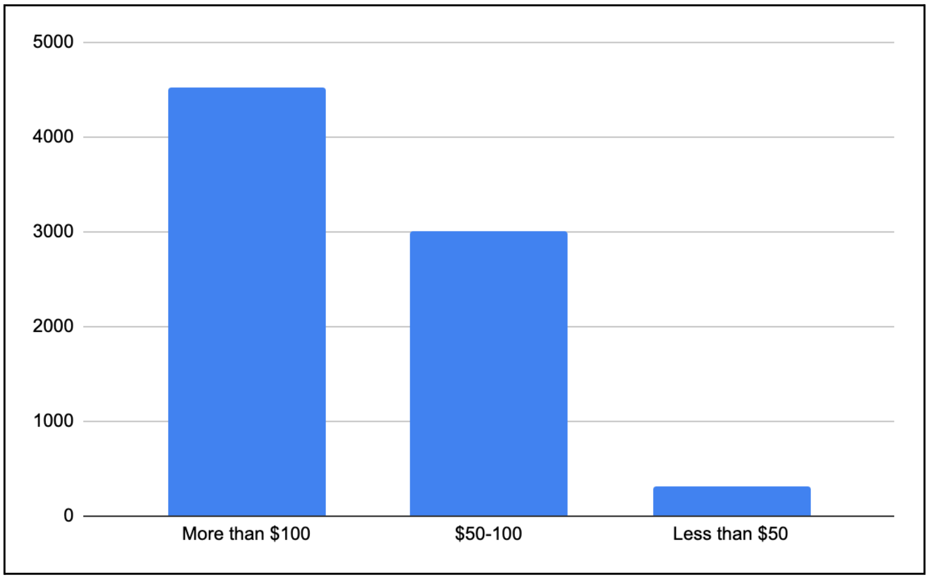

Figure 1: Group price of a listing

Part 4 - Describe the machine learning that you wish to do: what are the features, what is the class label, etc. Justify why your datasets are appropriate for training a model to solve your problem.

We intend to perform supervised machine learning, training a linear regression model to predict Airbnb prices. Supervised machine learning involves the use of features, which are the inputs of the model and will be mapped to a class label, which is the output of the model. The model is trained on a set of training data to learn how to predict the value of the class label.

Afterwards, test data can be used as input and the model will give a value for the class label output. Within the context of our project, this involves certain columns of the Airbnb dataset as our features and the predicted price of the Airbnb as the class label.

Certain columns from the dataset will not be used as features for the model as they have been deemed not to be generalizable, specifically:

- **Host_id, host_name, id:** These are unique identifiers (pieces of information to identify users) and will not help us predict prices.

- **Country, State, City:** Country and state columns always have one set value so they can be ignored. For the city column, >99% of the given data has city = ‘Amsterdam’ but <1% of the datapoints have their city column set to another city in the Netherlands. We will ignore these data points as our model is focused on Amsterdam and also there is a comparable lack of data regarding these other cities. Therefore, since we are only using datapoints where city = ‘Amsterdam’, we can ignore this column for our model.

- **Host_since_anniversary:** This is purely a given month and day in the year and will not make an impact on the model.

Our chosen dataset is appropriate for training a machine learning model as it includes different features that a potential guest might use to assess which listing they will choose. Referring back to *Part 1* and *Part 3* above, some of the core points to consider within the context of profitability/pricing strategies of an Airbnb have been included in the dataset and are categorized into four categories, general info about the listing, location, interior and host info/reviews.

The dataset’s column are split into these categories (excluding the specified columns above that are not generalizable) and will be used as the model’s features:

- **General Info**: price, guests_included, extra_people, minimum_nights
- **Location**: neighbourhood_cleansed, zipcode, latitude, longitude
- **Interior**: property_type, room_type, accommodates, bathrooms, bedrooms, beds, and bed_type
- **Host Info/Reviews**: host_since_year, host_response_time, host_response_rate, number_of_reviews, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value

The numerical features (specified in *Table 1*) will be used directly within the model while categorical features will be quantitatively mapped into the model using one-hot encoding. For example, the categorical room_type feature can have three potential values: “Entire home/apt”, “Private room” and “Shared room”. With one-hot encoding, three boolean variables will be made and for each datapoint, only one of these variables will be set to true and the others will be false to account.

These features are processed within the context of a linear regression machine learning model to output a class label output, which would be a sliding scale in terms of the final price of the unit within the Airbnb website. Least-Squares Linear regression was deemed the most appropriate form of modelling as it involves a linear relationship between multiple inputs and a continuous numerical output variable.

Depending on the results from training the model on the test data, it may be necessary to utilize piecewise linear regression. This could be the case if the data cannot be easily modelled with a singular linear model, but this is heavily dependent on the state of the data and its features after it has been processed.

Additionally, linear regression involves the concept of feature independence where each feature affects the overall class label independently from the others. If there is any amount of correlation between features of the model, this will be addressed and the features will be removed.

Lastly, our linear regression model can be improved going forward into the implementation phase of the project. One way of doing so is using LASSO regularization, which uses shrinkage towards a central point, such as the mean, in order to avoid overfitting. We will measure variables with small coefficients and remove them from the model if they do not have much of an impact on it, allowing us to enhance the accuracy of the model.

Bibliography

Arora, Mahima. “Amsterdam Airbnb Prices Dataset.” *Kaggle*, 11 Jan. 2021, https://www.kaggle.com/datasets/aroramahima1/amsterdam-airbnb-prices-dataset?fbclid=IwAR3nvgyjafMIO7Vl-bgbBfyMmc51N9X9Wq6dTK-x-0EgCg5a4zuGtFmrLfU.

Riddles, Callan. “12 Steps to a Profitable Airbnb Pricing Strategy.” *IGMS*, 18 Nov. 2022, https://www.igms.com/airbnb-pricing/.

Aishwarya. “15 Airbnb Hosting Tips to Make Your Listing Successful.” *PriceLabs*, 13 July 2022, https://hello.pricelabs.co/airbnb-hosting-tips/.

“What Are Response Rate and Response Time and How Are They Calculated? - Airbnb Help Centre.” *Airbnb*, https://www.airbnb.ca/help/article/430#section-heading-0-0.

Dogru, Tarik, and Osman Pekin. “What Do Guests Value Most in Airbnb Accommodations? an Application of the Hedonic Pricing Approach.” *Boston Hospitality Review What Do Guests Value Most in Airbnb Accommodations An Application of the Hedonic Pricing Approach Comments*, 2017, https://www.bu.edu/bhr/2017/06/07/airbnb-guest-pricing-value/

Kumar, Dinesh. “A Complete Understanding of Lasso Regression.” *Great Learning Blog: Free Resources What Matters to Shape Your Career!*, 12 Jan. 2023, https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/.