NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

BC2407: Analytics II: Advanced Predictive Techniques

Airbnb Booking Destination Prediction Model

Main Document

Seminar Group 1 Team 2

Poh Kang Yu	U1720388D
Ooi Jia Xuan	U1710203J
Shen Weixian	U1510418K
Tu Anqi	U1622399F

Content Page

Case Overview

This project aims to develop an analytics-based solution to increase users' booking rate and reduce the average days between account creation and booking for Airbnb, through providing a personalized experience to customer by predicting their country destination and urgency status. This is especially important in today's highly competitive e-commerce space where consumers are facing information overload and expect more personalized content. Besides, the saturated markets and large percentage of inactive users faced by Airbnb makes it important to further engage its registered users to maintain its growth.

A dataset contains the users' demographics, behaviours and web session records is used for the analysis. Four models - logistic regression, Multivariate adaptive regression spline (MARS), decision tree and random forest - are built for each of the prediction problem of the country destination and the urgency status of users. The most important factor and its effects on the prediction outcome will be analyzed. All models will also be compared based on three criterion - performance, interpretability and efficiency, and then a final model will be selected for each prediction problem based on the identified most important criteria for this business case.

Based on the data exploration and model analysis, several recommendations can be made to Airbnb to achieve greater level of personalization of user experience. The recommendations can be split into three categories: All users' blanket approach, segmented users' targeted marketing strategies and finally a strategy for further improvements of the prediction model for Top-5 country destination. A SWOT Analysis (Strengths, Weaknesses Opportunities, Threats) and financial feasibility are then performed on the recommendations to evaluate the effectiveness in maintaining Airbnb's growth.

Business Background

Airbnb's background & profile

Founded in 2008, Airbnb is an online accommodation and experiences booking platform which has shook up the travel industry. One of the pioneers of the online peer-to-peer business model, Airbnb uses a community-based, two-sided online platform to connect renters with potential guests. Individuals can list their properties or search for properties to rent on its website, and in more recent years, experiences such as classes and tours can similarly be booked as well. In addition to these, Airbnb also generates travel content on a separate blog as a content marketing strategy for its experiences.

Airbnb's revenue comes mainly from the service fee it charges its hosts and guests; currently it charges guests 3% of the booking amount and hosts 20% of the booking amount as its service fee (Nath, 2018). Considered as one of the few profitable "unicorn" companies, Airbnb has over 4 million listings in 81,000 cities worldwide, serving over 150 million guests since its founding (Airbnb, 2019).

Airbnb's largest market to date is in the United States (US), which accounted for 38.4 million users in 2018 (Conway, 2018). Out of the 50 to 65 million visits Airbnb's website gets each month, more than half of the traffic comes from the US, making the US market Airbnb's stronghold. Trailing shortly behind the US is France, which is the second-largest market, and hosts the city with the most number of Airbnb listings. While Airbnb has a number of listings in Asia, Africa and Oceania, the bulk of its customer base still lies in America and Europe (Bishop, 2017).

Analytics in Airbnb

Analytics plays a central role in Airbnb's key business areas, where it is used to analyse business trends and facilitate data-driven decision-making through the continuous process of creating new hypotheses, testing new ideas and improving on their existing ones.

Resources-wise, currently Airbnb has over 100 data scientists in total, and all cross-functional teams in Airbnb have at least one data scientist (Airbnb, 2018). An ecosystem of models supports Airbnb's operations, including algorithms to predict host preferences, customer loyalty and determine search ranking for searches by users. User data is actively collected and stored through the logging of user searches and guests and hosts interactions in a complex data infrastructure stack, which is shared across all its models (Mayfield et. al., 2016).

The largest application of analytics in Airbnb is its search ranking algorithm, which returns a personalised list of the highest quality listings within a radius of a user's search. This model is a neural network that ranks listings based on metrics such as a location relevance signal and encoded listing features, as well as user data to return listings that best meet each user's unique needs (Charkov et. al., 2013). Apart from this, Airbnb has also introduced a neighborhood matching system on its application in 2016, which takes in a user's preferences via a survey and matches its user-contributed neighborhood guides to users (eHotelier, 2016). Through personalised search capabilities and content, Airbnb hopes to capture the interest of potential users quickly to convert them into customers.

Current issues faced

While Airbnb's usage statistics are still growing, usage is slowing, which has led to the slowdown in revenue growth (Appendix 1), causing worries over the sustainability of Airbnb's growth. Usage by both hosts and guests have been increasing at a slower rate, evidenced by the decrease in the average available listings per month and decrease in growth of the number of nights booked. Additionally, market research company eMarketer lowered its 2018 forecast for Airbnb's usage rate within the US by 5 million guests, down from its earlier projection of 43.2 million users (Graham, 2018).

This comes amidst an impending Initial Public Offering (IPO), which puts additional pressure on Airbnb to perform well, given that investors believe the trend in guest arrivals is a key indicator of growth (Bosa & Salinas, 2019). With a valuation of $38 billion in March 2019, Airbnb is valued higher than its public competitors such as Expedia and Marriott, which demands Airbnb to justify this figure to investors through its earnings (Carson, 2018). Evidently, Airbnb cannot afford to slow down now.

According to a report from Morgan Stanley Research (Newsdesk, 2018), this slowdown is largely due to a saturated market, inactive users and increased competition from online travel agencies (Ting, 2018).

1.1 Saturated Market

Airbnb's success can be largely credited to its strong brand awareness, which has been the main focus of its growth strategies since its founding (Zheng, 2017). However, Airbnb's market penetration in mature markets like the US has reached its peak, with the awareness of the brand plateauing at 86% in recent years (Newsdesk, 2018). Awareness is the first and largest funnel out of the 3 main stages in the Sales funnel (Figure 1), which describes how out of the total number of consumers who are aware of Airbnb's brand, a percentage of them will consider using Airbnb, of which a percentage will actually make the decision to book.

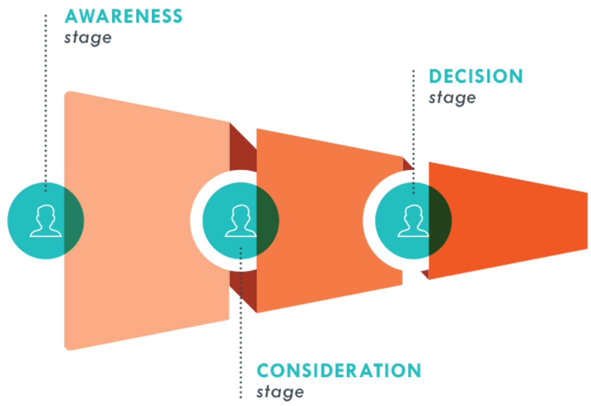


Figure 1: The Sales Funnel (Duggan-Herd, n.d.)

With Airbnb's brand awareness reaching saturation point, correspondingly the number of users passed on to the consideration and decision stages will be affected, ultimately affecting the number of bookings made. If Airbnb does not rethink its strategies to focus on the "Consideration" and "Decision" stages, its growth in sales is bound to weaken.

1.2 Inactive Users

Since Airbnb's launch in 2008, the company has significantly grown its user base to approximately 150 million users today (Much, 2019). However, a report from RJ Metrics showed that 85% of Airbnb's user base are inactive without any bookings or listings (Carr, 2012) and this number has been increasing which led to a slowdown in the number of 'nights booked' over the recent years (Spinks, 2017). Despite the expanding user base, there is a rising trend of user inactivity, which is represented by the widening gap between the number of users in the Awareness and Consideration stages on the Sales funnel. An analysis by behavioral analytics specialist CleverTap suggests that this could largely be due to loopholes in the onboarding process for new users - on average, a travel app loses 64% of new users after 30 days, rising to 78% after 60 days, and 82% after 90 days (Fox, 2018).

1.3 Fierce Competition

While Airbnb started off with home-sharing, it has been expanding into different sub-sectors of the online travel market in recent years. In 2014, the company added travel activities and tours, branded as "experiences", to its product catalogue. With the acquisition of Luxury Retreats for $200 million in 2017, Airbnb moved into the luxury market as well (CB Insights, 2017). Later on, in February 2018, executives announced that they were planning for the greater integration of hotels to power long-term growth (Ting, 2018). With its expansion into these markets, the scope of competition for Airbnb is widened (Figure 1).

COMPETITOR SCOPE – Travel Tech Market Map

Predicted to reach $167.9 billion in 2019 (PR Newswire, 2016), the rapidly growing vacation rental market has also attracted the attention of leading online travel agencies such as Expedia Group and Priceline Group, which have responded to this market opportunity by venturing into the home-sharing market themselves (Ahmed & Hook, 2017). Technology plays a central role in these companies and they have extensive technological and financial resources to power their development. Leveraging on their resources and expertise, they are continuously innovating and pushing out new initiatives, with a strong emphasis on their personalisation efforts. The increased overlap in market scope pits Airbnb directly against these large incumbent firms, leading analysts to believe that Airbnb should be especially worried about competing with online travel agencies (Ting, 2017).

Additionally, there has been a surge in the number of online-based travel startups which offer similar services to Airbnb, such as Flipkey and HomeToGo. As of now, there are 34 such firms competing with Airbnb for market share in the same space (Appendix 2). This surge has largely been credited to the growth of the online travel market and the fairly low barriers to entry, given the low upfront investment requirement and availability of digital infrastructure (Clark, 2018). As these startups are based off the same home-sharing concept that Airbnb is based on, Airbnb needs to differentiate itself through providing superior service.

The fierce competition means that substitutes are becoming stronger and more numerous, and as Airbnb is becoming more substitutable, this puts a damper on growth rates.

Business Problem

Approach

In its infancy, Airbnb's strategies focused on building brand awareness, which allowed Airbnb to increase its brand presence quickly and grow exponentially. While that built a strong foundation for Airbnb's future success, these strategies need to be refocused to suit Airbnb's current needs now that it has already become "mainstream". Linking back to the Sales funnel (Figure 1), as the Awareness stage is fully optimised, the bottleneck of increasing revenue growth now lies in the Consideration stage, which means that Airbnb should work on how it can better convince customers to book with Airbnb.

This case will focus on the conversion of new users to customers. The dataset used in this case records the activities of new users to Airbnb, including whether or not a booking has been made and which country is the booking in. Through developing a model to accurately predict of the first booking country of new users, Airbnb can use the model's predictions to tailor its marketing to user preferences. While it is not a cure-all, personalised marketing has the potential to engage users effectively and lift sales significantly, both of which are key for Airbnb to sustain its current growth rates.

Significance of first booking

The general lack of robust data on new users makes focusing on new users as opposed to all users seem counterintuitive however this exactly makes analysing new users more valuable: Competitors are not developing personalised marketing for their new users yet, which gives Airbnb a headstart in targeting new users.

The time between account creation and first booking is also the time in which many new users become inactive users due to insufficient engagement. This signals that there is a loophole in the onboarding process that Airbnb needs to work on.

However, if the user is successfully converted into a customer, they are more likely to become a recurring customer. A report by RJ Metrics suggests that the likelihood of a user booking an Airbnb listing follows an exponential upward trend: Out of Airbnb's total user base who books a stay, more than 22% of these customers will book a second stay. Once a user has booked 5 or more stays with Airbnb, there is a 50% probability that the user will book again (Carr, 2012). Given that a large proportion of users are lost before their first booking, and a new customer acquisition is likely to lead to a recurring customer, it is paramount for Airbnb to develop strategies to get users to make their first booking.

Importance of personalisation and how it can be done

Personalisation is hailed by marketers for its ability to increase conversion rates and customer loyalty - according to research by Epsilon Marketing, 80% of consumers are more likely to buy from a company which personalizes the customer experience and 49% of consumers who receive personalised offers are likely to become repeat customers (Serrano, 2018). The benefits of personalisation make it a key component in many companies' marketing strategies, including Airbnb.

Although previously seen as a value add-on by many, personalisation is becoming more of a necessity - a recent 2017 report from Deloitte found that consumers in the travel industry, in particular millennials, expect a personalised experience as a baseline criteria (Deloitte, 2017). At the same time, many players in the travel industry have developed extensive personalisation strategies. Online travel agencies in particular, are placed at an advantage as they have the full oversight over a customer's booking process, from flights to accommodations and activities. This translates to rich customer data which they use to analyse the big picture regarding motivations to travel and develop personalisation strategies to target customers at each stage of their travelling process. Thus in order to increase its market share, Airbnb needs to offer a level of personalisation that not just fulfills the expectations of consumers but goes above competitors' standards as well.

Every personalisation strategy can be split into 3 main stages: Data Collection, Predicting Customers' Preferences and Tailoring Recommendations. Airbnb has an established data ecosystem, thus the main areas the Airbnb should work on are prediction of preferences and tailoring recommendations to the predictions.

Usually, a combination of various analytical models such as neural networks and regressions are used to predict the outcome variable or extract important features, for example. Subsequently, these predictions can be applied to tailor marketing efforts and to guide customer choices. In this case, the main prediction outcome is the top 5 destinations of new users.

User segmentation

In general, there are 2 types of new users: those who know where they want to go and those who are still undecided. The first type tends to go straight to searching for listings and experiences in their destination country but are easily lost to competitors with better prices and discounts. On the other hand, the second type may require more guidance to help them narrow down the choices to their preferred destination, but tend to have higher bounce rates. With the top 5 model's predictions, Airbnb can target both groups effectively through displaying content that match user interests and tailoring its marketing channels and promotion mix accordingly.

Furthermore, through exploring the data, these users can be segmented based on time urgency as well. Urgent users have differing characteristics from non-urgent users, which is an area that can be explored in the personalisation of Airbnb's marketing strategy as well. Thus, a user's urgency status can be considered as a secondary prediction outcome as well.

1. Business Problem

1.1 Business Problem Statement

Question: Based on the business background and current problems faced by Airbnb as described above, formulate the business problem statement.

1.2 Business Outcome Measures and Target

Question: Setting clear Key Performance Indicators (KPIs) is crucial to measure the success of a solution and understanding which aspects work well and which aspects are underperforming. The SMART framework denotes the key criterias for crafting a KPI: a KPI should be Specific, Measurable, Attainable, Relevant and Time-Bound (Nichols, 2018). The overall success of personalising marketing for new users can be measured with a two-dimensional approach: **Booking rate** measures the direct impact of the solution on revenue while **time from account creation to first booking** measures the efficiency of the solution. Construct the two KPIs that can be used for this business problem.

2. Analytical Problem

2.1 Analytical Problem Statement

Question: Based on the business problem and the outcome measures, the analytical problem statement is how to recommend the right destination countries to the right user and how to accurately segment user according to their booking urgency. Explain how solving each of the analytical problem can achieve the business targets.

2.3 Analytics Performance Measures and Targets

Question: Similar to the KPIs set for measuring business outcome, Key Analytical Performance Indicators (KAPIs) should be set to quantify the success of the models built. The overall success for the main prediction is measured by the **accuracy of the top 5 destination country prediction**, while the **accuracy in determining a user's urgency status** measures the success of the secondary prediction. The specific percentages which define a "good" accuracy can be obtained through exploring the data. Based on the analytical problems, choose suitable analytical performance indicators to evaluate the performance of the analysis.

3. Data Preparation

3.1 Data Dictionary

Question: According to the data dictionary in Appendix 3 and the analytical problems, what is the targeted variable for each of the analytical problem? Is it continuous or categorical?

3.2 Data Cleaning

As the dataset clearly reflects the real situations in the travel accommodation industry, it is also clearly subjected to a lot of real world noises, including incorrectly stated values and missing values. Thus, data cleaning must be performed before the modelling.

3.2.1 Outliers

Question: Outliers might be signs of erroneous entries and could affect the performance of analysis and modelling. In particular, their presence is especially common in continuous variables. Thus, it is important to check the existence of outliers in the dataset and find ways to fix their values, either by dropping the row or recompute a reasonable value. Boxplot is a recommended way to identify outliers. Identify outliers for this data set, fix the values and provide your assumptions if any.

3.2.2 Missing data

Question: Real world data is always susceptible to missing values. While the existence of a few missing values can be fixed by computing reasonable values to fill in, the occurrence of many missing values poses a larger challenge. Sometimes, a missing value might provide extra information with the very fact that it is missing. Another challenge associated with missing values is that the missing values might be represented in different forms. Identify columns with missing values for this data set and the percentage of missing values, provide a way to fix the values and give your assumptions if any.

3.2.3 Categories with Low Frequencies

Question: If a categorical column contains levels that have very low frequency (eg. less than 1%), it might be of insignificant statistical values. To produce more general analysis, avoid overfitting for modelling and speed up computations, all rare values are recommended to be replaced with 'other'. Identify the categorical columns with low-frequency levels and convert them to 'others' if any.

3.3 Feature Engineering

Apart from the explicit information presented in the dataset, there are still some potentially useful but implicit information which we can extract from existing columns, such as the date columns, device type column, and web records data. The newly engineered features could be helpful for both data analysis and model building.

3.3.1 Extract Information from Datetimes

Question: Many useful information can be extracted from from the datetime columns such as month and weekday. Days difference between account creation date and first booking date should also be computed to provide extra information and converted to the urgency status target variable. Apart from the above two techniques, propose one more way to generate information from the datetime columns and state your assumption.

3.3.2 Extract Device and System Information from Device Type

Question: The device type column contains both the device and os information. Extract those information with the mapping in Appendix 4.

3.3.3 Summarize Statistics for Session Data

Question: The action column in web records have hundreds of categories. However, most categories appear with extremely low frequency. Find those actions with frequency higher than 1% and obtain the counts of and total time spent on each action for each user. Are there any other columns that you can perform similar transformation?

3.4 Factorization

Question: Some categorical columns might be recognized as numbers and thus treated as continuous features. It is always a good practice to check the type of each column after importing and ensure that it matches the appropriate column type. Factorization is the recommended way to perform the transformation. Identify columns that need to be factorized and perform the factorization.

3.5 Data Splitting

Question: It is always a good practice to check the model's performance on a different dataset that is it trained on. Split the cleaned data into train set (70%) and test set (30%), where train set is used for training and test set for evaluation.The splitting must be done in a stratified fashion on the target variable ('country_destination' and 'urgency_status'), such that the train and test subsets have the same proportions of class labels as the original dataset and thus could reflect the original distribution. Find a suitable library for the splitting and ensure the results can be reproduced by setting a random seed number.

4. Data Exploration

4.1 Destination Distribution

Question: For categorical target variable, barplot can help visualize the the distribution of levels and identify the most frequent category. For Airbnb, knowing the the popularity of each destination among its users could help the business better understand market demand. Use a barplot to show the distribution of the country destination. What are the observations and what could be the possible explanation?

4.2 User Demographic

Question: It is important for Airbnb to know the demographics of its users, so as to identify the main user group and cater to their interest and needs. Visualize the distribution of demographic features of users, such as age, gender and language. What are the observations and what could be the possible explanation?

4.3 Account Creation and Booking Day

Question: The duration between account creation and first booking is an important performance indicator for the business. Use a histogram to identify the number of days between account creation and bookings for most users.

4.4 Peak Season

Question: It is useful for business to identify the peak and off season in order to forecast demand to plan marketing strategy. Use a time series plot to identify the seasonal trend for users’ booking.

4.5 Country Destination and Month

Question: Association rule analysis can discover interesting relations between variables in large databases. This can help the business find patterns in the dataset without depending on target variable. There are three important information for each rule - support, confidence and lift. Support measures the applicability of the rule, whereas Confidence measures the strength of the rule in situations where it is applicable. Lift measures how useful the rule is, in the context of the existing situation. For this dataset, association rule analysis can be used to identify useful relationship between destination and month. Convert the data into transactional format, and perform association rule analysis to find any useful relationship between country destination and booking month. Explain the meaning of support, confidence and lift in this business content.

5. Modelling

As the goal is to predict for the top 5 destination country for newly registered users and predict whether the user is urgent to make a booking, the predictive models must be able to perform classification prediction and output the probability for each category. A few predictive models fulfill these requirements and thus can be implemented on the preprocessed dataset using R, including Regression, Multivariate Adaptive Regression Splines (MARS), Decision Tree and Random Forest. For both analytical problems, all four models will be built. Several techniques to improve the performance of each model will also be discussed. Furthermore, the feature importance indicated by each model could also be helpful in identifying useful information to be collected from user for better prediction result.

5.1 Predict Country Destination

For the section, models for predicting top 5 country destination for users are to be built.

5.1.1 Linear/Logistic Regression

Question: Choose between linear regression and logistic regression for this prediction problem and explain why it is more appropriate. An effective way to improve the model performance is to perform stepwise algorithm to perform backward feature elimination or forward feature selection. This is because some features in the dataset might be insignificant in making the prediction, thus requiring feature selection to be performed to find the best feature subset. Find the optimal feature combination using forward feature selection with the *step* function and compare the model performance before and after feature selection. Find the most important 3 factors.

5.1.2 Multivariate Adaptive Regression Splines (MARS)

Question: MARS is an extension to linear regression that captures nonlinearities and interactions between variables. Build a MARS model with the earth package in R and provide the R code to show the settings of parameters such that it suits the prediction problem. Compare its performance to that of the previous model. Based on the knot values below, find important factors affecting user's country destination and explain how each factor affect the destination.

	AU	**CA**	**DE**	**ES**	**FR**	**GB**	**IT**	**NL**	**other**	**PT**	**US**
(Intercept)	0.0080	0.0289	-0.0007	-0.0078	-0.0091	0.0443	-0.0074	0.0203	0.0635	-0.0019	0.8620
languagefr	-0.0033	-0.0319	-0.0041	0.0179	0.7834	-0.0276	0.0043	-0.0250	-0.0622	0.0023	-0.6538
languageit	-0.0117	-0.0360	0.0104	0.0216	-0.0047	-0.0571	0.8874	-0.0248	-0.0519	-0.0018	-0.7313
languageca	-0.0116	-0.0356	-0.0040	0.9937	-0.0173	-0.0563	-0.0102	-0.0246	-0.0660	-0.0015	-0.7667
h(date_account_created_dayofyear - 322)	0.0146	0.0075	0.0108	0.0046	0.0073	0.0022	0.0034	0.0051	0.0007	0.0026	-0.0050
h(55 - date_account_	0.0123	0.0035	0.0014	0.0076	0.0041	0.0082	0.0013	0.0051	0.0007	0.0026	-0.0050

created_dayofyear)											
age_bkt65+	-0.0007	0.0026	-0.0038	0.0096	0.0187	0.0088	0.0298	0.0009	0.0082	0.0057	0.2371
languagede	0.0231	-0.0360	0.4514	0.0107	0.0694	-0.0567	0.0433	-0.0250	-0.0655	0.0162	-0.4311

5.1.3 Decision Tree

Question: Decision trees implicitly perform feature selection and are very intuitive and easy to explain. However, without proper pruning or limiting tree growth, they tend to overfit the training data, making them poor predictors. One common technique is to grow the tree to the maximum first, find the optimal complexity parameter (i.e. a specified penalty cost for model complexity) value, then prune the tree with the optimal CP value. Compare the performance of the full tree and pruned tree on both train and test set and explain the observation. Obtain the variable importance of the pruned tree. (Though it is always recommended to visualize the tree for interpretation, we are not going to visualize and interpret in details in this case because of the large size of the tree.)

5.1.4 Random Forest

Question: Random Forest is a powerful ensembling machine learning algorithm which works by creating multiple decision trees and then combining the output generated by each of the decision trees. Build a random forest model with the randomForest package in R. Compare the result with that of the decision tree and explain the observation. Obtain the importance of each variable.

5.2 Predict Urgency Status

For the section, models for predicting the users' urgency status are to be built. Note here that a user is classified as urgent if he or she books within 2 days after account creation.

5.2.1 Linear/Logistic Regression

Question: Choose between linear regression and logistic regression for this prediction problem and explain why it is more appropriate. Find the optimal feature combination using forward feature selection with the *step* function and compare the model performance before and after feature selection. What other information can you get from the feature combination?

5.2.2 Multivariate Adaptive Regression Splines (MARS)

Question: Build a MARS model with the earth package in R and provide the R code to show the settings of parameters such that it suits the prediction problem. Compare its performance to that of the previous model. Find important factors affecting user's country destination and explain how each factor affect the destination. Based on the knot values below, find important factors affecting user's country destination and explain how each factor affect the destination.

	Yes
(Intercept)	1.3415

h(1-action_ajax_refresh_subtotal)	-0.5214
h(1-action_search_results)	-0.4416
h(action_similar_listings-1)	-0.2265
h(1-action_similar_listings)	-0.3847
h(action_ask_question-1)	-0.2159
h(1-action_ask_question)	-0.3643
h(action_ask_question-2)	0.2210
h(action_search_results-8)	0.0455
h(action_ajax_refresh_subtotal-8)	0.0685
signup_appiOS	-0.1042
h(action_similar_listings-9)	0.2364
age_bktunknown	0.0627
h(action_ask_question-9)	0.2398
h(action_search_results-3)	0.2028
h(action_search_results-2)	-0.1986
h(action_ajax_refresh_subtotal-9)*h(1-action_search_results)	0.2638
h(9-action_ajax_refresh_subtotal)*h(1-action_search_results)	0.0111
h(1-action_similar_listings)*h(action_search_results-9)	0.2258
h(1-action_similar_listings)*h(9-action_search_results)	0.0073
h(action_ask_question-9)*h(1-action_ajax_refresh_subtotal)	0.3957
h(9-action_ask_question)*h(1-action_ajax_refresh_subtotal)	0.0117
genderFEMALE	-0.0430
h(1-action_ajax_refresh_subtotal)*h(action_search_results-9)	0.2263
h(action_ajax_refresh_subtotal-2)	-0.2002
h(action_ajax_refresh_subtotal-3)	0.2013
h(1-action_ask_question)*h(action_search_results-9)	0.1852
h(1-action_ask_question)*h(9-action_search_results)	0.0053
h(action_similar_listings-9)*h(1-action_ajax_refresh_subtotal)	0.3685
h(9-action_similar_listings)*h(1-action_ajax_refresh_subtotal)	0.0077
h(1-action_ask_question)*h(action_ajax_refresh_subtotal-9)	0.1664

5.2.3 Decision Tree

Question: Compare the performance of the full tree and pruned tree on both train and test set and explain the observation. Obtain the variable importance of the pruned tree.

5.2.4 Random Forest

Question: Build a random forest model with the randomForest package in R. Compare the result with that of the decision tree and explain the observation. Obtain the importance of each variable.

6. Model Evaluation

Question: Three criterion are commonly used to evaluate the models - prediction performance, interpretability and efficiency.

1) **Prediction performance** reflects how different the predicted values are from the actual values. For both models, it is reflected by the accuracy rate. Prediction performance measures the effectiveness of the model and how well it can predict the real values. For example, for the first analytical problem, the higher the top 5 accuracy for the country destination prediction model, the more likely that the predicted dream country to be recommended will match the user's interest. For the second analytical problem, the higher the accuracy for the urgency status prediction model, the higher probability that Airbnb can segment the user to the correct group and implement marketing strategy that is more effective to the user.

2) **Interpretability** measures the level to which humans are able to comprehend and explain why certain predictions have been made by the model. A model is said to be more interpretable than another model if its decisions are easier for a human to understand than that of the other model. In some cases, it is not required to understand the process of how a decision is made, but rather it is enough to know that the predictive performance is good. But in other cases, knowing how the model derived its decisions can help the business learn more about the problem, the data and the reasons why a model might fail.

The need for interpretability arises from when a correct prediction only partially solves the problem. For example, if the business wants to know how each factor influences the target variable and use this information to change the target variable through modifying the contributing factor, a correct prediction by the model is definitely insufficient. Interpretability could also improve the reliability and acceptance of the model, and make it easier to find any potential bias and debug the model.

Some common methods to interpret a model includes importance of factors and how each factor affect the prediction result. Knowing factor importance aids future feature engineering and data collection while factor influence helps generate useful insights from the data for the business.

3) **Efficiency** of the model largely depends on the resources and time that is expected to be devoted to the model training and maintaining phase. Some complicated models trained on extremely large dataset could take days to train and extra efforts to monitor and maintain. In the case that the prediction must be made in a short time, efficiency will be of significant importance.

Evaluate all models based on the training results for three criterion. Are there any interesting observations? Select the best model and explain why it is the best (ie. which criteria do you use for the justification and why this criteria is the most important).

7. Recommendation

7.1 Recommendations for All Users (All Users who have not made first booking)

7.1.1 a) Training the model is not a standalone solution and the model prediction results should be incorporated into Airbnb's business strategies. Visit airbnb's website (www.airbnb.com) and their blog content (https://blog.atairbnb.com). Put yourselves in the shoes of a potential traveller who is browsing through this main page. What are some possible issues that might limit the success of you finding a space that you want to book? In addition, what do you think Airbnb could do to better utilise their interesting blog posts that are write-ups on various country's events and attractions?

7.1.1 b) With the issues in mind, discuss methods that Airbnb can use to effectively include the model's predictions so that their customers can have a personalized experience when using their service. (Hint: Airbnb website, blog posts and marketing email content). An example could be the main Airbnb website's "Recommended for you" section or blog content (refer to Appendix 5).

7.1.1 c) Imagine that your recommendation in part a) is being used by Airbnb, how would you recommend the company to measure their KPIs?

7.1.2 a) Based on the findings in section 4.3, identify the relationship between the day of account creation and booking day of an Airbnb user. What can you recommend to Airbnb based on your insights?

7.1.2 b) Refer to section 4.4 and 4.5, what are some insights you have gathered? How should Airbnb use your findings to aid with their marketing strategy to increase their booking rate?

7.2 Recommendations for Different User Segmentation

7.2.1) After the model predicts a user to be "Urgent", what do you think is the main focus and needs of this group of customers? Suggest how you would target your marketing strategy to capture this group of users according to their historical data and actions collected since the date they first made their account.

7.2.2) After the model predicts a user to be "Non-Urgent", what do you think will entice such a user book with Airbnb? Suggest how Airbnb should target this group of users with a different marketing strategy to attract "Non-urgent" users to make their first booking with Airbnb.

7.3 Recommendations to Further Improve Prediction Models

7.3.1) Take a look at the new user sign-up page for Airbnb (refer to Appendix 6) and the findings from the activity in section 4.2. What other types of data might be relevant and useful? Propose ways for Airbnb to collect more data from new users.

8. Feasibility

8.1 SWOT

Question: For each of the recommendation listed, analyse the effectiveness of the action plan on the targeted user group with SWOT.

8.2 Financial Feasibility

Question: Provide the cost estimation for each of the recommendations and assess its financial feasibility.

9. Limitations and Further Research

9.1 Data Collection

Question: Are the performance of models limited by the amount of data? Propose some other attributes that will be useful.

9.2 Research Scope

Question: The current data only contains users from USA, thus limiting the research scope. As Airbnb has over 150 million users around the world and offers listings in over 191 countries, it is feasible to expand the research beyond customers booking from USA. The research expansion can be done so that analysis can derive useful business insights, such as difference in user preferences or demographics in different regions, and more data can be used to train the model. This would mean that the research scope expansion could potentially improve the model performance. Propose two more ways to expand the research scope, in terms of specification level and user coverage.

9.3 Model Improvement

Question: The model accuracy could be further improved. For example, for random forest, model selection has not been performed due to hardware constraint. The number of trees to grow, the percentage of rows and columns for each subtree, the maximum depth of each subtree all potentially affect the model performance. Though it outperforms all other models, changing the default model setting might further improve the accuracy of the final predictions. Identify one more limitation of the current models and propose ways for improvement.

10. Conclusion

Question: What is the best performing model? What is its significance in helping the business?

What are its impacts in the long run?

References

Carr, A. (2012, 27 April). *Airbnb By The Numbers: 85% Of Users Inactive?* Fast Company. Retrieved from https://www.fastcompany.com/1835654/airbnb-numbers-85-users-inactive

eHotelier. (2016, April 21). Airbnb updates app to focus on personalisation. Retrieved from https://insights.ehotelier.com/global-news/2016/04/21/airbnb-updates-app-focus-personalisation

Duggan-Herd, T. (n.d.) Why You Need to Personalize Your Sales Funnel (and How to Do It). Retrieved from https://www.campaigncreators.com/blog/why-you-need-to-personalize-your-sales-funnel-and-how-to-do-it

Konar, R. and Kumar, J. (2016). Do the Traditional Travel Agents have any Future? An Empirical Analysis of Indonesian Travelers' Attitude. Retrieved from https://www.researchgate.net/publication/303434192_Do_the_Traditional_Travel_Agents_have_any_Future_An_Empirical_Analysis_of_Indonesian_Travelers'_Attitude

Macaulay, T. (2017, August 15) How Expedia uses data science to personalise the customer experience. Retrieved from https://www.computerworlduk.com/data/how-hotelscom-uses-data-science-personalise-customer-experience-3662447/

Much Needed. (n.d.). *Airbnb by the Numbers: Usage, Demographics, and Revenue Growth*. Retrieved from https://muchneeded.com/airbnb-statistics/

Newsdesk, A. (2018, November 14). *Airbnb's growth slowing in the face of competition from Booking and Expedia*. The association of serviced apartment providers. Retrieved from https://hub.theasap.org.uk/airbnbs-growth-slowing-in-the-face-of-competition-from-booking-and-expedia/

Spinks, R. (2017, December 3). *Using Airbnb Isn't Fun Anymore.* Quartz. Retrieved from https://qz.com/quartzy/1132556/how-i-fell-out-of-love-with-airbnb/

Ting, D. (2018, November 14). *Airbnb's Growth Is Slowing Amid Increasing Competition From Booking and Expedia: Report.* Skift. Retrieved from https://skift.com/2018/11/14/airbnbs-growth-is-slowing-amid-increasing-competition-from-booking-and-expedia/

Mayfield, J., Puttaswamy, K., Jagadish, S. and Long, K. (2016, February 23). *Data Infrastructure at Airbnb.* Retrieved from https://medium.com/airbnb-engineering/data-infrastructure-at-airbnb-8adfb34f169c

Charkov, M., Newman, R. and Overgoor, J. (2013, May 2). *Location Relevance at Airbnb.* Retrieved from https://medium.com/airbnb-engineering/location-relevance-at-airbnb-12c004247b07

Ting, D. (2018, February 24). *Airbnb Struggles to Balance Its Past and Future.* Retrieved from https://skift.com/2018/02/24/airbnb-wants-to-evolve-while-staying-true-to-its-roots/

Bishop, J. (2017, April 5). You'll Never Guess Which City Has The Most Airbnb Properties. Retrieved from https://www.forbes.com/sites/bishopjordan/2017/04/05/airbnb-paris-london-nyc-la-barcelona-rome-infographic/#fd347bf4127b

Conway, T. (2018, July 2). *Airbnb Competitors – Who Are They and What's the Difference?* Retrieved from https://blog.pillow.com/airbnb-competitors-who-are-they-and-whats-the-difference/

Airbnb. (2018, March 23). *Sharing More About the Technology That Powers Airbnb.* Retrieved from https://press.airbnb.com/sharing-more-about-the-technology-that-powers-airbnb/

Nath, T. (2018, June 8). *How Airbnb Makes Money.* Retrieved from https://www.investopedia.com/articles/investing/112414/how-airbnb-makes-money.asp

Airbnb. (n.d.). *About Us.* Retrieved from https://press.airbnb.com/about-us/

eHotelier. (2016, April 21). *Airbnb updates app to focus on personalisation.* Retrieved from https://insights.ehotelier.com/global-news/2016/04/21/airbnb-updates-app-focus-personalisation/

Clark, K. (2018). *Travel startups are taking off.* Retrieved from https://techcrunch.com/2018/11/11/travel-startups-are-taking-off/

Ting, D. (2017, November 15). *Airbnb Growth Story Has a Plot Twist — A Saturation Point.* Retrieved from https://skift.com/2017/11/15/airbnb-growth-story-has-a-plot-twist-a-saturation-point/

Carr, A. (2012, April 27). *Airbnb By The Numbers: 85% Of Users Inactive?* Retrieved from https://www.fastcompany.com/1835654/airbnb-numbers-85-users-inactive

Graham, J. (2018, June 6). *Airbnb's growth curbed by by safety concerns, competition: researcher.* Retrieved from https://www.usatoday.com/story/tech/talkingtech/2018/06/06/airbnbs-growth-curbed-safety-concerns-competition-researcher/675024002/

Zheng, Z. (2017, August 4). How does Airbnb increase brand awareness in China in 2017 ? Retrieved from https://medium.com/@zoeyzheng/how-does-airbnb-increase-brand-awareness-in-china-in-2017-e06e70ca3e76

Bosa, D. and Salinas, S. (2019, January 15). *Airbnb says it's been profitable for two years straight as it heads for IPO.* Retrieved from https://www.cnbc.com/2019/01/15/airbnb-sustains-profit-as-it-heads-toward-ipo.html

CB Insights. (n.d.) *Airbnb Strategy Teardown.* Retrieved from https://app.cbinsights.com/research/////airbnb-strategy-teardown-expert-intelligence#future https://www.ft.com/content/f132390c-f90f-11e6-9516-2d969e0d3b65

Carson, B. (2018, October 3). *Old Unicorn, New Tricks: Airbnb Has A Sky-High Valuation. Here's Its Audacious Plan To Earn It.* Retrieved from https://www.forbes.com/sites/bizcarson/2018/10/03/old-unicorn-new-tricks-airbnb-has-a-sky-high-valuation-heres-its-audacious-plan-to-earn-it/#2ad955106fa3

Carey, M. (2018, March 16). *Airbnb Is Finally Mainstream, But CEO Brian Chesky Wants More.* Retrieved from https://www.cntraveler.com/story/airbnb-is-finally-mainstream-but-ceo-brian-chesky-wants-more

Serrano, S. (2018, September 28). *Definitive Guide: Personalization Statistics to Guide Strategy & Tactics.* Retrieved from https://www.barilliance.com/impact-website-personalization-conversion-rates/#tab-con-3

Deloitte. (2017). *2017 travel and hospitality industry outlook.* Retrieved from https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consumer-business/us-cb-2017-travel-hospitality-industry-outlook.pdf

Nichols, R. (2018, September 13). *How to Define and Track your KPIs.* Retrieved from https://www.abtasty.com/blog/key-performance-indicator/

Landis, T. (n.d.) Customer Retention Marketing vs. Customer Acquisition Marketing. Retrieved from https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/

Fox, L. (2018, August 1). *More than half of travel apps deleted after a month.* Retrieved from https://www.phocuswire.com/clevertap-travel-apps-benchmark

Appendix 1: Airbnb’s Falling Revenue Growth Rate

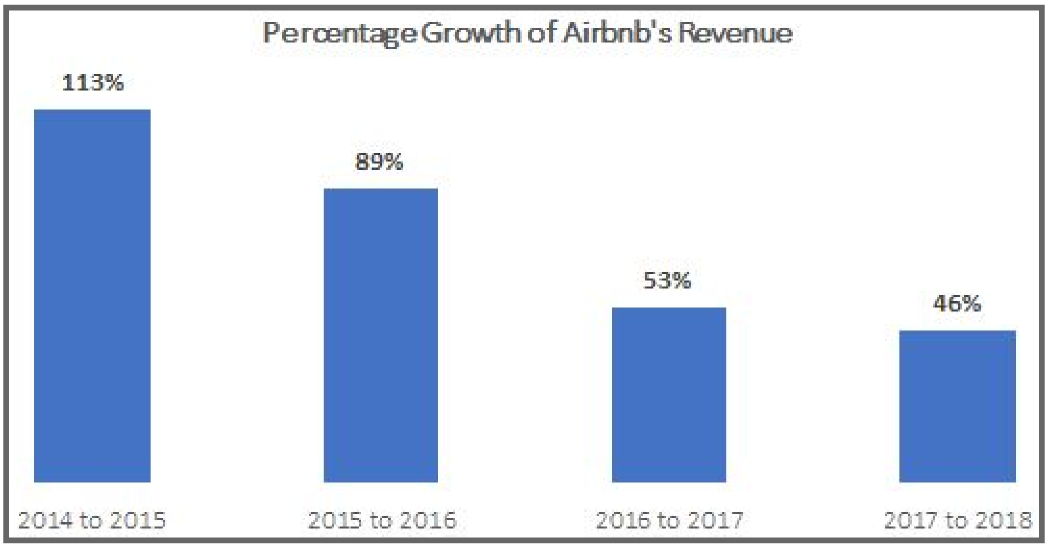


Figure 1: Percentage change in Airbnb’s Revenue from 2014 to 2018

As Airbnb is not a publicly listed company, they do not file their annual reports for investors and the public. Hence, the 2014 to 2018 revenue numbers are obtained from different sources. After calculations, the percentage change in Airbnb’s Revenue was found to be falling over the years.

Year	Revenue	Percentage Growth
2014[a]	$423,000,000	
2015[b]	$900,000,000	113%
2016[c]	$1,700,000,000	89%
2017[d]	$2,600,000,000	53%
2018[e]	$3,800,000,000	46%

Table 1: Airbnb’s Yearly Revenue and Percentage Growth 2014 - 2018

(a) https://www.cbinsights.com/research/airbnb-hospitality-industry-valuation-breakdown/
(b) http://fortune.com/2015/06/17/airbnb-valuation-revenue/
(c) https://pitchbook.com/news/articles/airbnb-expects-28b-in-2017-revenue-85b-by-2020
(d) https://www.businessinsider.sg/airbnb-profit-revenue-2018-2/?r=US&IR=T
(e) https://www.forbes.com/sites/greatspeculations/2018/05/11/as-a-rare-profitable-unicorn-airbnb-appears-to-be-worth-at-least-38-billion/#6456bcf62741

Appendix 2: Competition from Startups

☐	Last Funding	Competitor	Stage	Total Funding	Location
☐	12/12/2018	Airbnb	Secondary Market - III	$4,398.13M	California
☐	4/1/2019	Oyo Rooms	Series E - III	$1,552.65M	India
☐	2/27/2019	StayList	Seed VC	$0.36M	Japan
☐	2/14/2019	NightSwapping	Acquired	$4.82M	France
☐	2/5/2019	2nd Address	Series C	$22.76M	California
☐	1/2/2019	Luxstay	Bridge	$5.5M	Vietnam
☐	10/10/2018	XiaoZhu	Series F	$500M	China
☐	9/27/2018	Wimdu	Dead	$90M	Germany
☐	9/5/2018	Knok Exchange	Acquired	$0.64M	Spain
☐	8/23/2018	Sonder	Series C	$130.05M	California
☐	6/28/2018	Starcity	Angel	$20.37M	California
☐	3/9/2018	HubHaus	Series A	$11.4M	California
☐	2/1/2018	Apartum	Unattributed	$2.66M	Spain
☐	12/21/2017	Common	Series C	$65.05M	New York
☐	10/11/2017	TuJia	Series E	$763.7M	China
☐	7/31/2017	Love Home Swap	Acquired	$4.28M	United Kingdom
☐	7/29/2017	MyTwinPlace	Acquired	$1.04M	Spain
☐	3/8/2017	HomeExchange	Series B	$40.86M	France
☐	4/28/2016	HouseTrip	Acquired	$59.7M	United Kingdom
☐	4/6/2016	Onefinestay	Acquired	$80.7M	United Kingdom
☐	11/5/2015	HomeAway	Acq - P2P	$495.5M	Texas
☐	10/31/2015	Rent Like a Champion	Angel	$0.2M	Illinois
☐	1/28/2015	Sykes Cottages	Private Equity	$82M	United Kingdom
☐	7/16/2014	Tutt.ru	Seed	$0.02M	Russian Federation
☐	6/11/2014	Homestay Technologies	Series B	$7.13M	Ireland
☐	2/26/2014	TripVillas	Seed VC	$0.5M	Singapore
☐	5/28/2013	SleepOut	Seed VC	$0.2M	Mauritius
☐	11/23/2012	Lezu365	Angel		China
☐	8/22/2012	CouchSurfing	Series B	$22.6M	California
☐	8/8/2012	MorningCroissant	Seed VC	$0.51M	France
☐	8/6/2012	iStopOver	Acquired	$3M	Canada
☐	6/20/2012	TravelRent	Series A	$2M	Russian Federation
☐	1/12/2012	9flats	Series B	$10M	Germany
☐	11/20/2006	VRBO	Acquired		California
☐		Vacation Rentals Online			Nevada

CB Insights, 2019
https://app.cbinsights.com/profiles/i/b4wO/competitors

Appendix 3: Data Sources and Data Dictionary

This two datasets used for analysis and creation of the prediction model can be obtained from Kaggle - https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data. The first dataset contains the demographics and behaviours of real-life customers who signed up on Airbnb as a member but may or may not have made their first booking. The second dataset contains the web session records for those users and is ordered by user id and then chronologically. There are 12 possible outcomes of the destination country for each user: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'. All the users in this dataset are from the USA, who visit Airbnb from 2009/03/19 to 2015/06/28.

The data dictionary for the columns in the users and session records dataset is documented below.

User Demographics and Behaviour

Name	**Type**	**Definition**
Variables related to the user's data		
id	categorical	User Id
gender	categorical	Gender of the user ('-unknown-', 'MALE', 'FEMALE', 'OTHER')
age	continuous	Age of the user
language	categorical	International language preference code of the user ('en', 'fr', 'de', 'es', 'it', 'pt', 'zh', 'ko', 'ja', 'ru', 'pl', 'el', 'sv', 'nl', 'hu', 'da', 'id', 'fi', 'no', 'tr', 'th', 'cs', 'hr', 'ca', 'is')
Country_destination	categorical	The country where the customer's first booking is made in. This is the target variable you are to predict ('US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'other' and 'NDF' (no destination found). * 'NDF' is different from 'other' because 'other' means there was a booking, but is to a country not included in the list, while 'NDF' means there was not a booking.
Variables related to the user's behaviour		

signup_method	categorical	The signup method by the user. (basic, facebook, google)
signup_app	categorical	The application that the user used to sign up on (IOS, Android, Moweb, Web)
signup_flow	categorical	The index of the page a user came to signup up from (Range from 0 to 25)
first_device_type	categorical	The type and operating system of the device used by customers when accessing Airbnb's website. ('Mac Desktop', 'Windows Desktop', 'iPhone', 'Other/Unknown', 'Desktop (Other)', 'Android Tablet', 'iPad', 'Android Phone', 'SmartPhone (Other)')
first_browser	categorical	The browser type that was used by the customer to visit Airbnb's website and content for the first time. ('Chrome', 'IE', 'Firefox', 'Safari', '-unknown-', 'Mobile Safari', 'Chrome Mobile', 'RockMelt', 'Chromium', 'Android Browser', 'AOL Explorer', 'Palm Pre web browser', 'Mobile Firefox', 'Opera', 'TenFourFox', 'IE Mobile', 'Apple Mail', 'Silk', 'Camino', 'Arora', 'BlackBerry Browser', 'SeaMonkey', 'Iron', 'Sogou Explorer', 'IceWeasel', 'Opera Mini', 'SiteKiosk', 'Maxthon', 'Kindle Browser', 'CoolNovo', 'Conkeror', 'wOSBrowser', 'Google Earth', 'Crazy Browser', 'Mozilla', 'OmniWeb', 'PS Vita browser', 'NetNewsWire', 'CometBird', 'Comodo Dragon', 'Flock', 'Pale Moon', 'Avant Browser', 'Opera Mobile', 'Yandex.Browser', 'TheWorld Browser', 'SlimBrowser', 'Epic', 'Stainless', 'Googlebot', 'Outlook 2007', 'IceDragon')
first_affiliate_tracked	categorical	The first marketing the user interacted with before the signing up ('untracked', 'omg', nan, 'linked', 'tracked-other', 'product', 'marketing', 'local ops')
affiliate_channel	categorical	The kind of paid marketing channel that attracts the user. ('direct', 'seo', 'other', 'sem-non-brand', 'content', 'sem-brand', 'remarketing', 'api')

affiliate_provider	categorical	The kind of paid marketing provider that attracts the user. ('direct', 'google', 'other', 'craigslist', 'facebook', 'vast', 'bing', 'meetup', 'facebook-open-graph', 'email-marketing', 'yahoo', 'padmapper', 'gsp', 'wayn', 'naver', 'baidu', 'yandex', 'daum')
Variables related to date		
date_account_created	datetime	The date of account creation (Range from 2010-06-28 to 2014-06-30)
timestamp_first_active	datetime	The timestamp of the first activity. It can be earlier than date_account_created or date_first_booking because a user can search before signing up (Rang from 2009-03-19 to 2014-06-30)
date_first_booking	datetime	The date of first booking (Range from 2010-08-02 to 2015-06-28)

Web session records

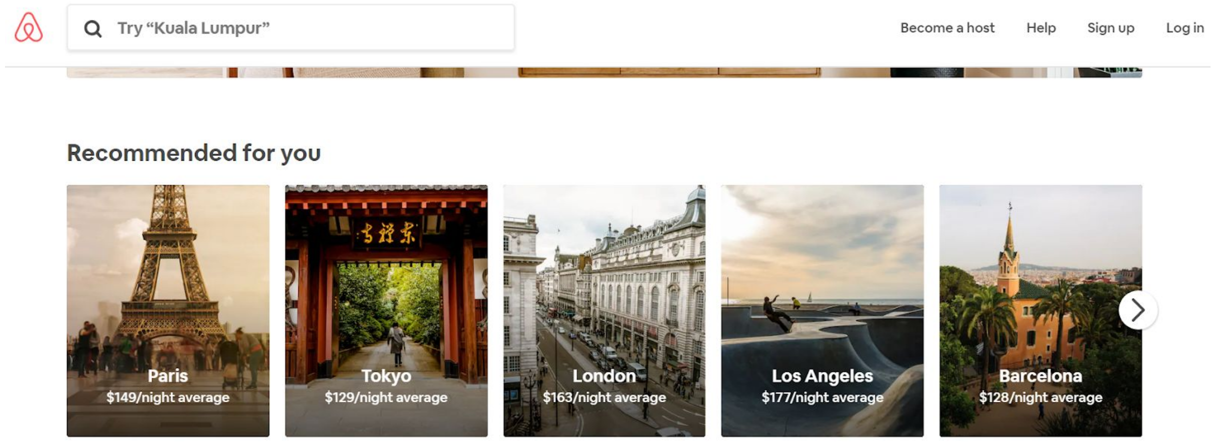
Variables related to sessions		
user_id	categorical	User Id (to be joined with the column 'id' in users table)
action	categorical	A general description of the action performed by the user
action_type	categorical	The type of the action performed by the user
action_detail	categorical	The detail of the action performed by the user
device_type	categorical	The type of device used by the user when performing the action
secs_elapsed	continuous	The number of seconds between actions were recorded.

Appendix 4: Device Type Mapping

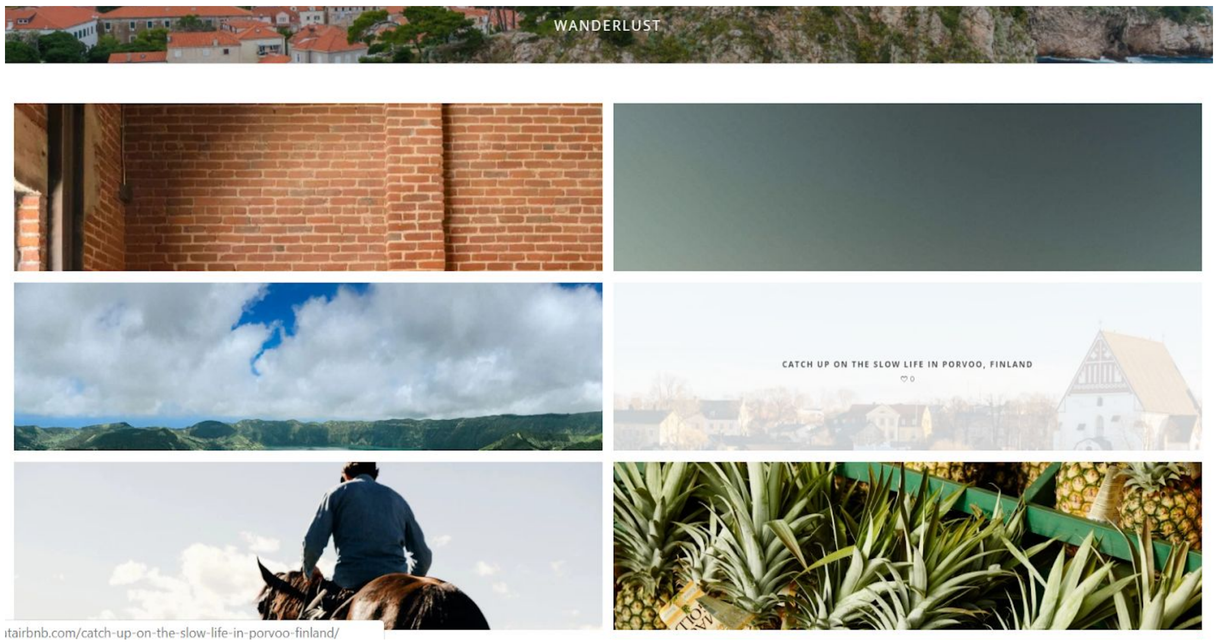
	Device	**OS**
Mac Desktop	Desktop	MacOS
Windows Desktop	Desktop	Windows
iPhone	Phone	MacOS
iPad	Tablet	MacOS
Other/Unknown	Others	Others
Android Phone	Phone	Android
Android Tablet	Tablet	Android
Desktop (Other)	Desktop	Others
SmartPhone (Other)	Phone	Others

Appendix 5: Airbnb Website and Blog Content

Airbnb Website



Airbnb Blog Content



Appendix 6: Airbnb Sign up Page