RSM456
BDAP: Final Report
Professor Hemant Sangwan

Group Members:

Name	Utorid
Po Shan Sandra Kan	1003103447
Julien Bernardo	1003515671
Michael Gonzaga	1006630488
Pelin Mamaoglu	1002972483
Siddhant Banerjee	1002976942

Executive summary

Airbnb is an online marketplace for booking or offering room & house rentals and tourism experiences. We use an Extreme Gradient Boosting (XGB) classifier model to predict, based on account information, a user's first booking destination. With these predictions we can help Airbnb deploy targeted marketing on Airbnb's website to entice newly registered users to book a first stay by customizing each user's homepage with their predicted three most likely first destinations with a 92.31% accuracy rate. We believe that accurately predicting first-stay destinations will increase revenues through higher customer attraction and repeat booking rates, which will further lead to increased return on investment with more efficient marketing. Furthermore, since Facebook sign up was the most important feature in our model, assuming the relation is positive with booking a first destination, we recommend Airbnb to advertise through Facebook. Before implementing the predictive based marketing, Airbnb should employ A/B testing to determine what marketing platforms are most effective. In doing so, Airbnb can have a risk-free approach to implementing our predictive model in their marketing strategy while the potential rewards could drastically improve customer attraction and marketing efficiency.

Key Problem

What factors, if any, are predictive of where an Airbnb's user's first stay will be? Can we accurately predict where a customer will go for their first AirBnB stay?

Methodology

I. Preliminary analysis, data manipulation, descriptive analysis

After cleaning our data, we used descriptive statistics and data visualization to know more AirBnB users and how they behave. Since most of our dummy variables don't provide us with valuable insights when calculated standard deviation, mean, variance and range, we decided to discard these statistics for this part of the project. We included the most relevant variables: age brackets, gender, age, and destinations.

A. **Gender distribution among users:** refer to Graph 1

Our graph illustrates the gender distribution of the sample in percentages, showing if any segment dominates the sample. It can be seen that females are the bigger segment with 53.07% of the data, leaving 46.64% to male. This means that the slight majority of users are females. The remaining 0.29% is the "others" tab, which is insignificant due to the extremely small size. Since this is a categorical dummy variable, disregarding the descriptive statistics such as mean, median, standard deviation and skew are irrelevant and insignificant.

B. **Most popular destinations:** refer to Graph 2

By graphing the data on the count of destinations preferred by each user, we can see that the US dominates the choice of first stays with 70% of the sample. The second most preferred destination is "other" by 11%, meaning a combination of countries that weren't significantly sized enough to have their own dummy variable. After "others", all the countries with a dummy variable such as France and Italy are below 5%, showing the extreme demand of visitors to pick the US as their first destination. Once again, since this is a categorical variable converted into a dummy, descriptive statistics weren't relevant or significant. This chart provides a good insight into understanding patterns in the data, however, it lacks more details such as the user preference of destination by gender and age which would help us better reach our objective of providing better first-stay suggestions to increase retention and adoption rates thus profits.

C. **Age distribution:** refer to Graph 3

Age is a continuous variable which provides us with insightful statistics. Graph 3 is a histogram of the age variable, illustrating a clear positive skew, where the mean age of 36 is higher than the median age of 33. This means that 50% of users are younger than 33, creating a big emphasis for us to target the younger audiences in our marketing efforts when we are suggesting destinations to first time Airbnb users. Even though this is an important insight for us to focus our targeting efforts to the younger audience, it is relevant for us to know which location to suggest to these users in each age group.

D. **Destination preference based on age group:** refer to Graph 4

To be able to get better insights on user preference by segments, we first created Graph 4 showing the frequency of destinations chosen with a percentage of users in specific age brackets within each location. We can easily see that in each age bracket, the US is the most chosen destination. Among each age group, age groups 18-29 and 30-39 are the dominant age brackets representing more than 25% each, interested in the US as a first destination, which is more than 20% compared to other destinations. Apart from the US, these age groups also dominate the users in each specific destination with a significant difference among all age brackets.

For our objective of increasing user adoption and retention rates by making more accurate predictions of first stay destinations, these are good insights because Airbnb can put a greater emphasis on advertising American destinations to each age group, knowing the significant likeliness of a user to be interested in a destination in the US than others.

E. **Destination preference based on gender:** refer to Graph 5

Now that we know that Airbnb's user base has slightly more females, we want to know whether or not there is gender bias when it comes to choosing the first destination. Looking at Graph 5, we can see that there is also a slight dominance of females at 37.5% for the US as first destination and males at 32.5%. For the rest of the destinations, gender distribution is close to even. However, this graph doesn't include other variables and only categorises preference based on gender, which might not be accurate to discard gender as an important variable after considering other factors such as age, platform used when logging in and so on. This tells us that our marketing and prediction model should still test gender's importance since it seems that gender might play a significant role in determining the user's first destination.

F. **Gender based on age group:** refer to Graph 6

So far, we have seen that gender distribution was close to even based on destination and user base, but there could be gender distribution discrepancies among age brackets. Looking at Graph 6, we see that in the age bracket 18-29 the majority of users are male while in the age bracket 30-39 the majority are female. We can interpret this as males tend to be younger than females in Airbnb's user base which could lead to market segmentation based on both age and gender allowing for a better prediction of a user's first destination.

Takeaway

From our preliminary analysis we can conclude that most Airbnb users go to the US as their first destination and that gender doesn't seem to play a major role when deciding first destinations, but our insights may not be accurate without the model building. Furthermore, we now know that segmenting users by age and gender allows for a clearer picture for categorizing new users by first destination.

II. Models building and evaluation

A. Model Building: Extreme Gradient Boosting

We chose to use Extreme Gradient Boosting (XGBoost) as a classifier to predict the customer's decision. We chose this because our dataset is sufficiently large and complex. XGBoost is an iterative algorithm that incorporates decision trees into a similar logic as simple gradient boosting. The result is a powerful, optimizing algorithm that can be used for both regression and classification. Our model is trying to understand based on available data, what a new client's first destination is going to be with a probability attached to each possible outcome which is, in this case, countries.

B. **Model Evaluation:** refer to Graph 7

The model is evaluated by the accuracy score. The "*soft probability*" option of XGboost ranks, for each client, the most likely first destination to the least likely. We calculated the accuracy of the model based on whether or not the actual first destination was among the top "x" most likely destinations from our prediction, where "x" is an arbitrary integer.

This model analyzes all available variables on a AirBnB user's profile and provides us with their most likely destination(s). One observation is that the accuracy score increases at a diminishing rate as we incorporate more destinations into the model as recommendation to clients.

Accuracy of predicting the most probable destination: ***63.29%***
Accuracy of predicting the top 2 most probable destinations: ***87.55%***
Accuracy of predicting the top 3 most probable destinations: ***92.31%***
Accuracy of predicting the top 4 most probable destinations: ***94.53%***
Accuracy of predicting the top 5 most probable destinations: ***95.89%***

C. **Feature importance:** refer to Graph 8

The most important feature is signing up through Facebook, this could mean that Facebook signups are a group of people that prefer a specific first destination, or it could also say alot about whether or not they will book with AirBnB. Also, in contrast to our findings in our preliminary analysis, gender seems to be one of the most important factors. Unfortunately, with a multi-class classification problem, we cannot extract whether the feature has a positive or negative influence for each possible destination, however, with assumptions, we can use the most important features in our marketing plan when we are deciding on the platforms and marketing content to consider when advertising to potential users in specific categories.

III. Insights and summarizing results

The sign-up method of using Facebook is the most important feature. We assume that people who are on social media platforms are more likely to sign into Airbnb, because they are already interested in following trends such as using popular platforms like Facebook, and are more likely to be open to more current trends, one of the most significant one being booking a vacation stay in a house or apartment rather than the traditional hotel. With the millennials between the ages of 18 to 29 being the biggest segment, this assumption based on our data seems to be valid and can be used to make important decisions on the marketing plan.

To make accurate recommendations on the Airbnb website while avoiding overwhelming the consumers, Airbnb should recommend top three destinations to every user based on the customers' sign up method, age, gender and location. There is a 92.31% rate at which the actual first destination is in the top three recommended destinations. Providing more recommendations only increases the accuracy by an insignificant amount but also it would also risk overwhelming and confusing the customer that may lead them to not book at all.

Implementation plan & impact evaluation

To test the impact of our destination prediction model and Facebook advertisement, Airbnb could A/B test it, expose a small sample of their customers to the predicted destination and compare with the control sample to see change of the rate at which website viewers convert to clients by booking their first destination with Airbnb.

Top three destination recommendation through:

A. Group 1: Airbnb's website
B. Group 2: Facebook
C. Group 3: Facebook and Airbnb's website
D. Control Group: None

By evaluating the impact of each form of advertising with the A/B test on customer groups, the company can then decide what marketing combination they should execute with their whole customer base. This would allow the company to eliminate their unimpactful marketing expenses while increasing customer attraction, thus increasing the ROI. In addition to the monetary benefits of increased profits and return on investment, this also leads to long term intangible benefits that drive the long term success of Airbnb such as increased market share in the accommodation industry, brand relationships, customer loyalty and customer satisfaction.

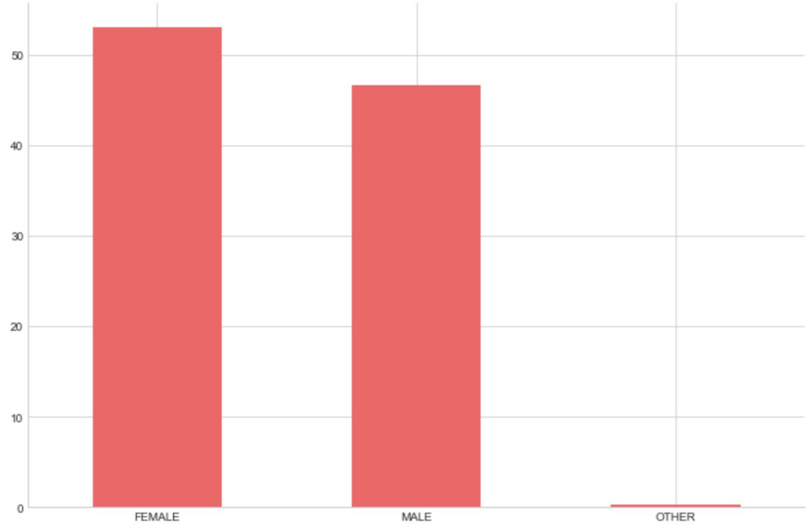
Conclusion

By understanding which destination the user finds attractive, we are able to curate a list of the top three recommended destinations to advertise to each first-time user in order to increase their likelihood of booking a first-time visit with AirBnB. We also decided that recommending three destinations hits the balance between model accuracy at 92.31% .

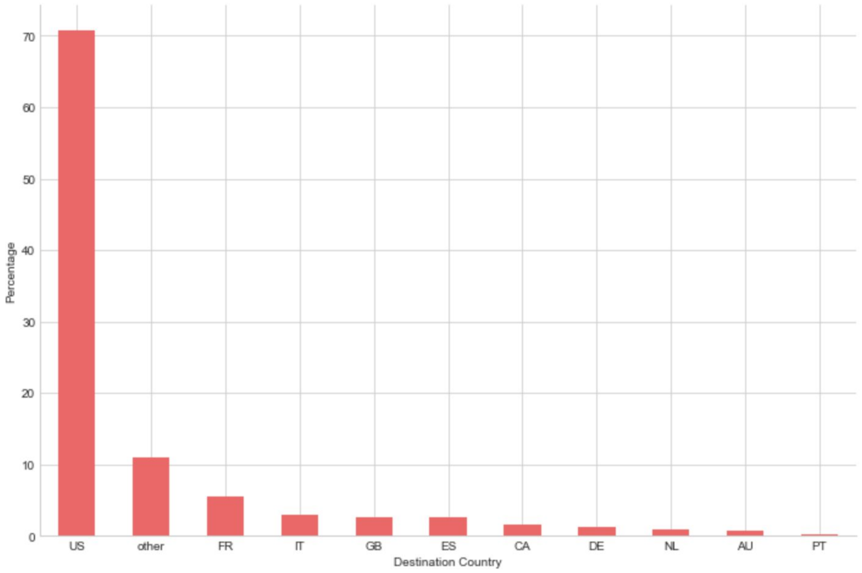
On top of recommending the top 3 destinations to each user on the Airbnb website, the company should move further in their marketing plan to maximise the customer attraction rates and ROI. Knowing the importance of Facebook sign-ups on customer attraction, Airbnb can place advertisements on Facebook to increase customer attraction rates. The brand can tailor the locations featured in the ads based on the user’s Facebook profile which includes the required information of their gender, age and location. This can be done by creatıng tailored advertisements on the Facebook platform for each specific customer profile. This would increase customer attraction, and meet the business goal of increasing return on investment. In the end, we recommend Airbnb to conduct the A/B test to see which marketing mix achieves the best result in enticing new customers to make their first booking and deploy the optimal solution.

Appendix

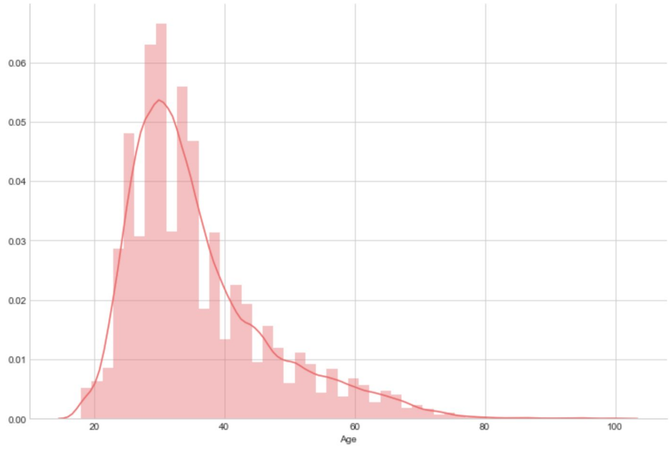
Graph 1: Gender distribution among users



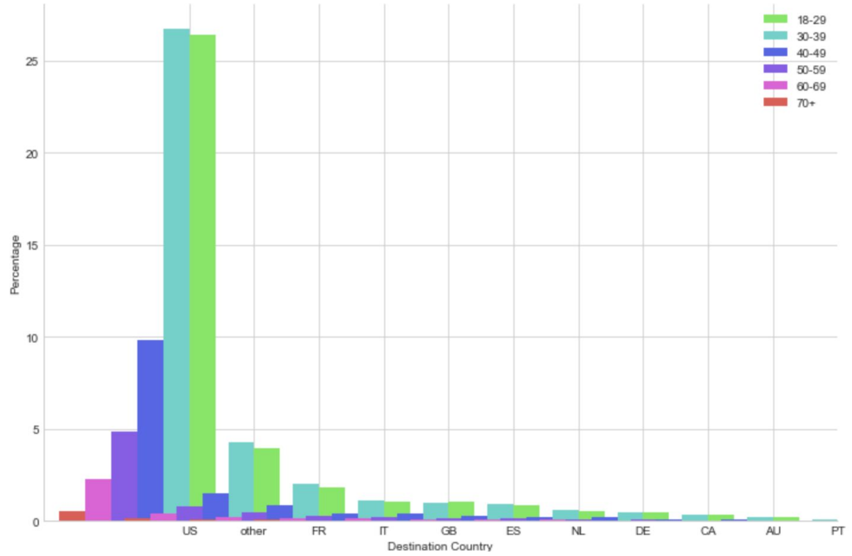
Most popular destinations: Graph 2



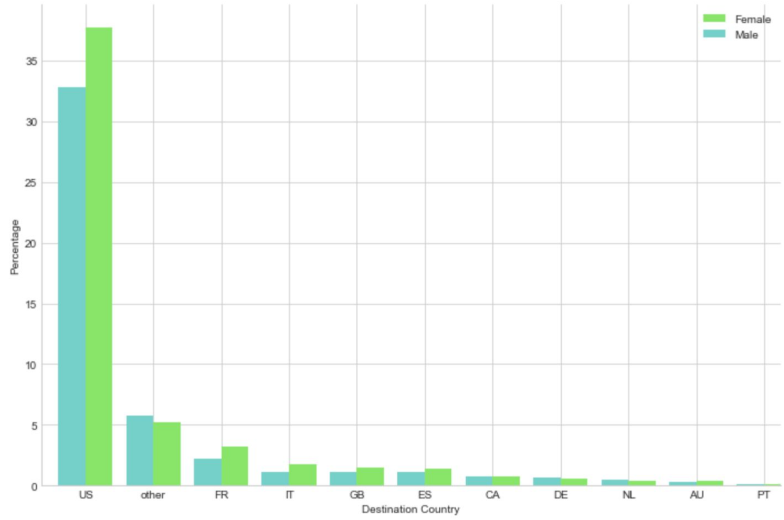
Age distribution: Graph 3



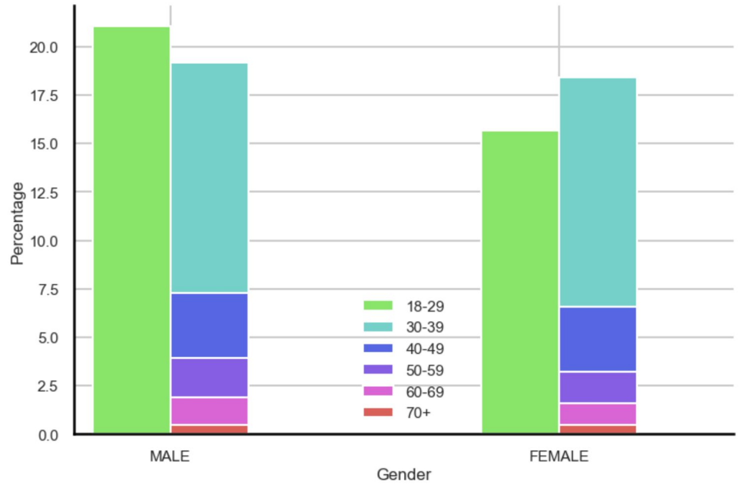
Graph 4: Destination preference based on age group



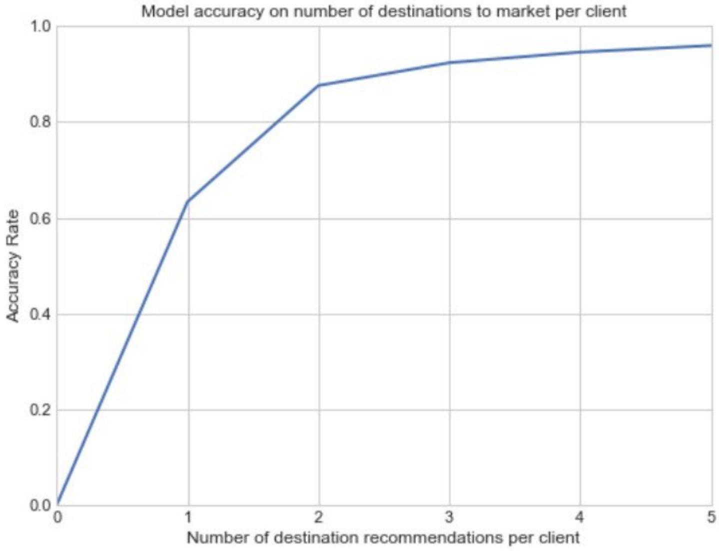
Graph 5: Destination preference based on gender



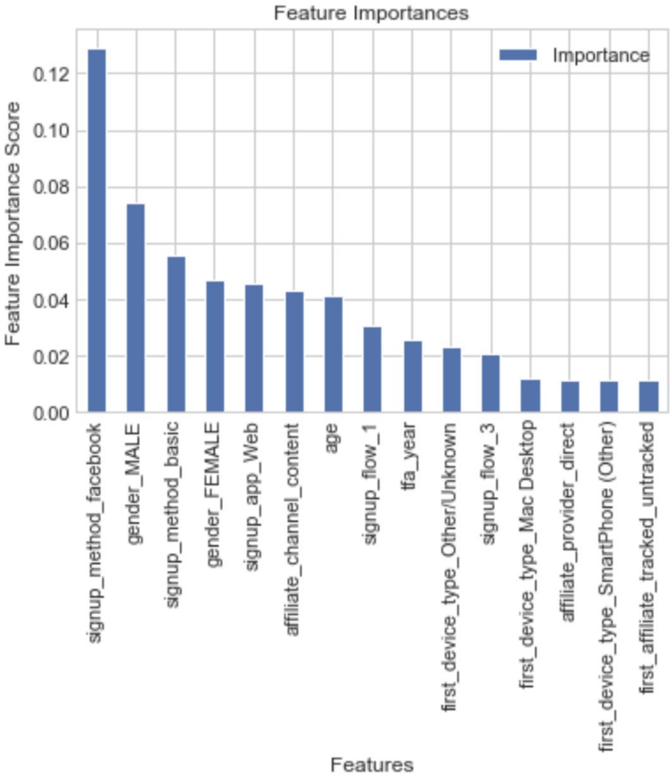
Graph 6: Gender based on age group



Graph 7: Model accuracy on number of destinations to market per client



Graph 8: Feature importance for prediction, top 15


Feature Importances
Importance
Feature Importance Score
signup_method_facebook
gender_MALE
signup_method_basic
gender_FEMALE
signup_app_Web
affiliate_channel_content
age
signup_flow_1
tfa_year
first_device_type_Other/Unknown
signup_flow_3
first_device_type_Mac Desktop
affiliate_provider_direct
first_device_type_SmartPhone (Other)
first_affiliate_tracked_untracked
Features