

**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

BC2407: Analytics II: Advanced Predictive Techniques

Airbnb Booking Destination Prediction Model

Solution Manual

Seminar Group 1 Team 2

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

Content Page

1. Business Problem	4
1.1 Business Problem Statement	4
1.2 Business Outcome Measures and Target	4
2. Analytical Problem	4
2.1 Analytical Problem Statement	4
2.2 Analytics Performance Measures and Targets	5
3. Data Preparation	5
3.1 Data Sources and Data Dictionary	5
3.2 Data Cleaning	6
3.2.1 Outliers	6
3.2.2 Missing data	6
3.2.3 Categories with Low Frequencies	6
3.3 Feature Engineering	6
3.3.1 Extract Information from Datetimes	6
3.3.2 Extract Device and System Information from Device Type	6
3.3.3 Summarize Statistics for Session Data	7
3.4 Factorization	7
3.5 Data Splitting	7
4. Data Exploration	7
4.1 Destination Distribution	7
4.2 User Demographic	7
4.3 Account Creation and Booking Day	7
4.4 Peak Season	8
4.5 Country Destination and Month	8
5. Modelling	8
5.1 Predict Country Destination	8
5.1.1 Linear/Logistic Regression	8
5.1.2 Multivariate Adaptive Regression Splines (MARS)	9
5.1.3 Decision Tree	9
5.1.4 Random Forest	9
5.2 Predict Urgency Status	9
5.2.1 Linear/Logistic Regression	9
5.2.2 Multivariate Adaptive Regression Splines (MARS)	10
5.2.3 Decision Tree	10
5.2.4 Random Forest	10
6. Model Evaluation	11
7. Recommendations	11

7.1 Recommendations for All New Users	11
7.1.1 Personalize Recommendations	11
7.1.1.1 Validate KPI	12
7.1.2 Festivals and Peak seasons increase marketing	12
7.2 Recommendations for Different User Segmentation	12
7.2.1 Recommendation for Urgent Users (Predicted to book within ≥ 2 Days)	12
7.2.2 Recommendation for Non-Urgent Users (Book after < 2 Days)	13
7.3 Recommendations to Further Improve Prediction Models	13
7.3.1 Collect More Data	13
8. Feasibility	14
8.1 SWOT	14
8.2 Financial Feasibility	14
9. Limitations and Further Research	15
9.1 Data Collection	15
9.2 Research Scope	15
9.3 Model Improvement	15
10. Conclusion	15
11. References	16
Appendix 1: Age Outliers	17
Appendix 2: Distribution of Destination Countries	18
Appendix 3: Distribution of User Demographics	18
Appendix 4: Distribution of Duration between Account Creation and First Booking	20
Appendix 5: Peak Season	20
Appendix 6: Association between Country Destination and Month	21
Appendix 7: Important Features for Models Predicting Country Destination	22
Appendix 8: Important Features for Models Predicting Urgency Status	25
Appendix 9: Logistic Regression Value for Country Destination	28
Appendix 10: Logistic Regression Value for Urgency	30
Appendix 11: Airbnb Website and Mobile App	32
Appendix 11: Airbnb Website and Mobile App	34
Appendix 12: Website Snippet after Personalisation	35
Appendix 13: Voting boxes Examples	36
Appendix 14: Feasibility Analysis	37

1. Business Problem

1.1 Business Problem Statement

The business problem is that Airbnb must improve the booking rate of new users through personalised marketing in order to sustain its growth.

1.2 Business Outcome Measures and Target

Two business measures can be used to gauge the solution for the problem - booking rate and average time to first booking.

KPI 1: Overall Airbnb Customer Booking Rate	
Business Target: Booking rate of new users > 60%	
KPI Explanation	This KPI refers to the percentage of registered users who book a listing on the Airbnb website . This can be obtained through the total number of booking and total number of accounts created. The current booking rate from our data is 40% and the aim is to boost the booking rate by half of the current percentage.
Relevance and Impact	Airbnb's main revenue stream is the booking service fees it charges from its users. As such, the booking rate is directly related with Airbnb's revenue growth rates. Personalised marketing aims to increase the number of bookings by enticing users to book with tailored content and promotions. Thus, achieving this KPI means that the personalisation solutions have successfully made a significant contribution to helping Airbnb sustain its growth rates.

KPI 2: Average Time to First Booking	
Business Target: Average Booking Day < 10 Days	
KPI Explanation	This KPI refers to the average days a user takes to book after account creation. This can be obtained through tracking of users' booking history. From our data, most users book within 10 days, thus we aim to further reduce the time.
Relevance and Impact	This rating criteria feedbacks to Airbnb on whether their personalization strategy is successful in attracting users to book within a shorter time period. The impact of achieving this KPI would mean that Airbnb has a higher chance of capturing booking revenue from new customers compared to losing them to competitors that helped them navigate to the listing of their choice before Airbnb does.

2. Analytical Problem

2.1 Analytical Problem Statement

Tailoring the destination country recommendations to each user's preferences prevents information overload and captures their interest, increasing user engagement. Segmenting user based on urgency

status allows implementation of personalized marketing strategy, which leads to higher booking rates.

2.2 Analytics Performance Measures and Targets

The following key analytical performance indicators (KAPI) are recommended to evaluate the performance for the analytical problem.

KAPI 1: Top-5 Accuracy	
Target: Top-5 Accuracy (> 90%)	
KAPI Explanation	The Top-5 Accuracy measures the probability that any of the model's top 5 country predictions (based on probability) matches the user's dream destination country.
Relevance and Impact	In the personalisation solution, Airbnb will recommend 5 countries to users in the 'Recommended for you' section on its homepage. As long as one of it matches the user's ideal destination, the recommendation is considered as successful and personalized because it caters to the user's preference. Predicting 5 countries will lead to a higher prediction accuracy than randomly selecting 5 countries out of all the possible destination countries. Thus, evaluating accuracy with Top-5 approach is the most appropriate and suitable.

KAPI 2: Urgency Status Accuracy	
Target: Urgency Status (> 75%)	
KAPI Explanation	The urgency status accuracy measures the percentage of accurately predicted outcome of whether the user needs to book urgently. A user is defined as urgent if they make a booking within 2 days after account creation. A non-urgent user will make the booking after 2 days.
Relevance and Impact	When segmenting a user into urgent or non-urgent, a higher percentage of correct predictions implies more precise segmentation. In order for the targeted marketing solution to be effectively implemented, the segmentation of users has to be sufficiently accurate.

3. Data Preparation

3.1 Data Sources and Data Dictionary

The target variable for predicting user's dream destination is destination country which is a categorical variable. The target variable for predicting user's urgency status is also a categorical variable, however it is not explicitly present in the data. It can be obtained from the difference between the account creation date and first booking date. If the difference is less than or equal to 2 days, the user is urgent. Otherwise, the user is considered as not urgent.

3.2 Data Cleaning

3.2.1 Outliers

The boxplot of age distribution in Appendix 1 shows there are some outliers for 'age' column. For ages with values greater than 1900, it is assumed that the user put the year of birth instead of age. It is fixed by subtracting the given year from the current year (for this dataset it was 2015) to get the age of the user. The boxplot of age distribution after fixing those outliers with outstanding high values, shows there are still some outliers with invalid values, ie. ages less than 15 or more than 100. For these values, they are considered as incorrect inputs and thus set to be NA.

3.2.2 Missing data

The age column has 42.3% missing values. It is assumed that users did not provide their age. To deal with this, due to the large percentage of missing values, an age bucket is created, from 15 to 95 with an interval of 5. Then all missing values are grouped into the 'unknown' bucket.

The column 'date_first_booking' has 61.4% missing values. The assumption is that it is missing because the user has never booked before and 'date_first_booking' will only be known if the user has booked a trip. Besides, this column should not be included for the purpose of training a model as it contains unknown information at the time of making the prediction.

The 'first_affiliate_tracked' contains 2.89% missing values, which are replaced by 'untracked', the most common category, with the assumption that these values are missing because of being untracked.

The missing values for 'gender' column are represented by '-unknown-'. As these also provide information that the user does not choose to put the gender information, these values are kept.

3.2.3 Categories with Low Frequencies

There are several categories including 'affiliate_provider', 'age_bkt', 'language', 'first_affiliate_tracked', 'first_browser' and 'signup_flow' which have some levels with frequency lower than 0.1%. All low-frequency levels are converted to 'others'.

3.3 Feature Engineering

3.3.1 Extract Information from Datetimes

With the assumption that public holidays also affects users' decision making process, US holidays data is added to create columns to indicate number of days from account creation date to the next holiday.

3.3.2 Extract Device and System Information from Device Type

Two new columns are formed by mapping the device type to the device and OS respectively.

3.3.3 Summarize Statistics for Session Data

The same can be done for the action detail and device column.

3.4 Factorization

The month features are automatically recognized as integer by R. R set the features as nominal variables, they need to be factored as categorical variables.

3.5 Data Splitting

The library caTools is used for the splitting, with a seed number 2019. After splitting, the train set contains 15679 rows and the test set contains 6720 rows.

4. Data Exploration

4.1 Destination Distribution

The distribution of country destination (Appendix 2) shows that most of registered users (62%) have no destination country and have never booked yet, reflecting a low booking rate of 38%. This might be due to users who do not have a country destination in mind yet or become inactive after registration. For those users who have made their first booking, the most popular country is US (17%), followed by US, France, Italy, Spain and Canada. A possible reason why US is the most popular destination country is because the dataset only contains users from the US who might be more inclined towards domestic travel.

4.2 User Demographic

Most users (43.5%) have the age as 'unknown' (Appendix 3). They might not be willing to disclose their personal information to Airbnb. For people who have disclosed their age, it is found that the major user group of Airbnb are people aged in their 20s and 30s. This might be because people from this age group are more comfortable using technology-based services, compared to older age group who are more conservative. Besides, Airbnb's advantage over traditional hotels is its affordable price and unique experience, which is valued by people in their 20s and 30s.

Almost half of users (45.4%) have their gender as 'unknown', as they might be unwilling to share their gender. Additionally, there is slightly more female users than male users, but the difference between the gender of the users is not significant.

Most users use English, followed by Chinese, French and Spanish. This is because the dataset only includes users from US, where the majority of the population speaks English.

4.3 Account Creation and Booking Day

The distribution of duration between account creation and booking day (Appendix 4) shows most people make their bookings within 10 days of account creation. For this group of people, it is highly likely that they already have a country destination in mind when creating an Airbnb account.

4.4 Peak Season

The midyear period always experiences a surge in bookings (Appendix 5). This summer rush begins in around June, as schools and universities end their semesters and families and college students head onto their summer vacations. Another contributing factor that most popular travelling countries for US people have the warmest and most enjoyable weather from June to August due to their geolocations.

4.5 Country Destination and Month

The 10 association rules in Appendix 6 have the highest lift. The first rule will be used to illustrate the concept of support, confidence and lift. The confidence of 8.19 implies that 8.19% of users who book in December travel to Australia. The support of 0.46 shows there are 0.46% of users who book in December and travel to Australia. The lift of 2.1 shows that users are 2.1 more likely to travel to Australia if they book in December, compared to what can be expected if booking month and country are not associated. This is probably due to Australia's location in the southern hemisphere, where the warmest weather and peak tourist season falls between December and February.

Conversely, summer occurs from June to August in the Northern Hemisphere, thus the popularity of Spain, Canada and France around June to August is likely to be due to the ideal weather in summer and travellers can enjoy long days of sunshine. Germany is also more popular in October, probably due to the Oktoberfest event in Munich, which is acclaimed as the biggest beer festival in the world. Likewise, Netherland is more popular in May and Italy in April due to the spring tulip season and the grand Easter celebrations in each country respectively.

5. Modelling

5.1 Predict Country Destination

5.1.1 Linear/Logistic Regression

Logistic regression is more appropriate here because it can solve classification problem while linear regression is suited for regression problem. Using 'multinom' function of R, a full multinomial logistic regression model trained on all variables was built first, with a Top 5 Accuracy of 0.835. This 'multinom' algorithm fits a multinomial log-linear model via neural networks. Then 'step' is applied to choose the model with the best feature combination in a stepwise algorithm through forward feature selection, improving the Top 5 Accuracy to 0.862. The second model with the optimal feature subset is shown as below:

```
multinom(formula = country_destination ~ language + age_bkt + signup_app + gender + first_device +  
          date_account_created_dayofyear + first_os, data = train)
```


As reflected from the accuracy score, feature elimination improves the model performance. The coefficients shows language, age and affiliate channel are most important factors (Appendix 9).

5.1.2 Multivariate Adaptive Regression Splines (MARS)

A MARS model is built to predict the country destination with a Top 5 Accuracy of 0.895, using the 'earth' package. Because this is a classification problem and a categorical response is expected, the argument `glm=list(family=binomial)` must be added as shown below:

```
mars <- earth(country_destination~., nfold=10, data=train, glm=list(family=binomial), degree=2)
```

The *evimp* functions can be used to obtain the most important factors. The *evimp* function uses *nsubsets*, *rss* (Residual sum of squares) and *gvc* (generalized cross validation) for estimating variable importance. For the *nsubset* criterion, variables that are included in more subsets are considered as more important. For the *rss* criterion, variables which cause larger net decrease in RSS are considered more important. For the *gcv* criterion, variables which increases the GCV a lot are considered less important. As shown in Appendix 7, Language is the most important factor, followed by day of account creation age. For language: 1) Users using Catalan are 0.994 more likely to travel to Spain 2) Users using Italian are 0.887 more likely to travel to Italy 3) Users using French are 0.783 more likely to travel to France. 4) Users using German are 0.452 more likely to travel to Germany. For day of account creation: 1) At the start of the year, the earlier the user creates account, the more likely that he/she travels to Australia. 2) At the end of the year, the later the user creates account, the more likely that he/she travels to Australia. For age, people aged more than 65, they are 0.237 more likely to travel within the home country.

5.1.3 Decision Tree

Firstly, the decision tree is grown to the max with 'minsplit' set as 2. It is then pruned using optimal CP value that results in lowest CV error. As expected, the fully grown decision tree was overfitted, as it achieves a Top 5 accuracy of 0.99 on the train set, but a Top 5 Accuracy of 0.763 on the test set. In contrast, the pruned decision tree performs comparatively for both train and test set, achieving a Top 5 accuracy of 0.867 and 0.852 on the train and test set respectively. As shown in Appendix 7, the most important factors are related to the language, followed by marketing channel and age.

5.1.4 Random Forest

A random forest model is built to predict the country destination with a Top 5 Accuracy of 0.909, using the 'randomForest' package. Though also tree-based, it performs much better than the decision tree predictor. This is because random forest grows many trees with sampled rows and sampled columns, and then combine the results from many diverse models. As shown in Appendix 7, the most important factors are day of account creation, language and age.

5.2 Predict Urgency Status

5.2.1 Linear/Logistic Regression

Logistic regression is more appropriate here because it can solve classification problem. Using *glm* function of R, a full generalized linear model trained on all variables was built first, with a Accuracy of 0.655. This 'glm' algorithm fits a multinomial log-linear model via neural networks. Then 'step' is applied to choose the model with the best feature combination in a stepwise algorithm through forward feature selection, improving the Accuracy to 0.705. The second model with the optimal feature subset is shown as below:

```
glm(formula = urgent ~ action_search_results + action_ajax_refresh_subtotal + signup_app +  
    gender + action_ask_question + action_similar_listings + age_bkt + signup_method  
    date_account_created_dayofyear + first_browser + first_os + first_device +  
    action_wishlist_content_update, family = binomial, data = train)
```

The accuracy score shows that the feature elimination improves the performance of the logistic regression. The coefficients shows the counts of actions of refreshing subtotal cost, viewing search result and age are most important factors (Appendix 10).

5.2.2 Multivariate Adaptive Regression Splines (MARS)

The MARS model achieves an accuracy of 0.852. Similarly, because this is a classification problem and a categorical response is expected, the argument *glm=list(family=binomial)* must be added:

```
mars <- earth(urgent ~., nfold=10, data=train, glm=list(family=binomial), degree=2)
```

Counts of actions of refreshing subtotal cost, viewing search result and checking similar listings are most important factors:

- When the user refreshes the subtotal after changing trip characteristics for one time, the probability that he/she is urgent to book increases by 0.521. For every refreshing the user performs, the probability that he/she is urgent to book increases by 0.068.
- When the user views the search result for one time, the probability that he/she is urgent to book increases by 0.441. For every search result viewing, the probability that he/she is urgent to book increases by 0.045.
- When the user checks similar listings for one time, the probability that he/she is urgent to book increases by 0.385. For every checking of similar listings, the probability that he/she is urgent to book increases by 0.01.

5.2.3 Decision Tree

As expected, the fully grown decision tree was overfitted, as it achieves an accuracy of 0.99 on the train set, but an accuracy of 0.765 on the test set. In contrast, the pruned decision tree performs comparatively for both train and test set, achieving a Top 5 accuracy of 0.826 and 0.817 on the train

and test set respectively. As shown in Appendix 8, the most important factors are related to the user's actions of refreshing subtotal cost, viewing search result and checking similar listings.

5.2.4 Random Forest

A Random Forest model was built to predict the country destination with an accuracy of 0.835, using the 'randomForest' package. The most important factors are related to the user's actions of refreshing subtotal cost, viewing search result and checking similar listings.

6. Model Evaluation

In terms of model performance, random forest and MARS performs the best for each of the prediction problem (Appendix 9). Regarding efficiency, the linear regression trained on selected features always has the shortest training time while MARS has the longest training time, followed by random forest. For interpretability, all models are able to provide the feature importance, as discussed in the previous section. However, while the coefficients of logistic regression, the knot values of MARS and decision paths of decision trees can be used to explain how the respective model produces the prediction result and subsequently interpret the effect of each factor, the random forest is a black box model.

In this business context, the model accuracy is of the highest priority as the utmost goal is to provide the right content to the right user, instead of making sense of why a user wants to go a certain destination or why a user needs to make a booking urgently. The factors of those decisions (in this dataset) are not controllable by Airbnb and thus of less importance to interpret. The training time of models are also not a concern as they still within the same scale. Thus, random forest is considered as the best choice for country destination prediction and MARS is considered as the best candidate for urgency status prediction. However, if Airbnb emphasizes on the interpretability of the results, or data of factors that can be influenced by Airbnb (eg. promo code) are made available, MARS, which achieves a similar performance but offers much more interpretability, would be a preferred choice.

7. Recommendations

7.1 Recommendations for All New Users

7.1.1 Personalize Recommendations

At first glance, the reader should notice a mixture of listings from different countries displayed on the website's main page. This gives an unpersonalised experience by including unnecessary distractions like all types of country listings that the users might have no interest in. According to Taylor (2014), removing irrelevant content improves conversion rate.

With the relevant prediction model techniques demonstrated in this report, data from Airbnb users can be used to predict the top 5 country destinations the user would most likely make their first

booking in. After visiting the websites noted in question 7.1a, the following few observations on the potential opportunities and issues can be identified:

- 1) “Recommended for you” section not curated and personalised (Appendix 11)
- 2) Variety of experiences and homes listings from various destinations that are not related to the 5 recommended destinations in part 1
- 3) Blog content not curated and could be shown on the main page

The solution for question 7.1b is that the **website** should include the predicted 5 countries in the “Recommended for you” section so that users who do not know where they want to travel to will be inclined to book with Airbnb. (refer to Appendix 12 for screenshot of website after personalisation)

The prediction model can also be incorporated into Airbnb’s **mobile application**. From their current application home page layout (Appendix 11), the prediction model’s results can be included in a top 5 recommended destination section for the app user.

Other than the top 5 country destination section, Airbnb can utilise their website and app to push out **blog content material** related to these predictions to entice the users to book any predicted destination. A depiction of this recommendation specifically for Airbnb’s website is depicted in the figure in Appendix 12.

Feasibility of this idea is largely dependent on the Airbnb marketing team. They have to update their content and keep creating new attractive content to place as the centerpiece of each predicted country’s position on the website or app.

This recommendation will personalise homes, experiences and blog content for every user and increase the possibility of them to find a booking they like with the right country displayed. With the personalisation, distractions will be minimised and thus improve Airbnb’s booking rate.

7.1.1.1 Validate KPI

After all improvements have been implemented, after a year, the analysis should be run again to get the new booking rate. It should increase to more than 60%.

7.1.2 Festivals and Peak seasons increase marketing

Using the findings from data exploration in section 4.4, the student should identify peak and off-peak seasons in Airbnb’s sales. A possible recommendation using this insight would be to increase marketing and send relevant email content to entice users to book during peak seasons. For off-peak seasons, Airbnb could send out time-sensitive promotions and discounts to attract more booking.

From their association rules analysis in section 4.5, the student should discover that more users make bookings to specific country destinations in certain months. We assume the month of the first day of booking is the month they visit the country. After the student researches about each country and their public events, they should notice a trend in these events held in the countries compared to the increase in booking during these months. Hence, Airbnb can send related events email marketing

content according to the user's predicted top 5 country destinations to entice them to make a booking. By incorporating the prediction model in the digital marketing strategy, it reduces spam emails while ensuring that the content sent out will be for locations the users are interested in.

7.2 Recommendations for Different User Segmentation

7.2.1 Recommendation for Urgent Users (Predicted to book within ≥ 2 Days)

According to Hyken (2018), offering convenience to your customers makes them less likely to make their purchase with competitors instead. Hence, after classifying users as "Urgent", Airbnb should use their search history, streamline the most appropriate listings in the country the user searched for and send an email to them with the listings displayed clearly. The ability to make customers feel like Airbnb understands their needs, coupled with the convenience of having suitable listings handed to them will increase the chance of "Urgent" users booking.

7.2.2 Recommendation for Non-Urgent Users (Book after < 2 Days)

After predicting the "Non-urgent" users segment, the most useful method to increase their booking rate would be to create urgency for them to make their booking with Airbnb. According to Loo (2017), 60% of travellers would consider an impulse trip if they receive a good hotel or flight deal. Hence, Airbnb should use a different marketing strategy compared to the "Urgent" users by sending these group of users timed promotions and discounts. By injecting urgency into "non-urgent" users, this will increase the rate of booking a listing with Airbnb.

7.3 Recommendations to Further Improve Prediction Models

7.3.1 Collect More Data

Airbnb only requires new users to provide personal information like email address, name and birthday on sign up. From the data visualisation in section 4.2, 43.5% of users did not specify their age and 45.4% of users did not provide gender. A possible solution for Airbnb to motivate users to input such data would be to provide incentives for users who provide more personal data (for example, their job, country they are living at now) during signup. The incentives could come in the form of Airbnb rewards system where they would get extra rewards points in their account. The rewards could range from redemption of one free night stay or vouchers for experiences and restaurants. As 63% of millennial consumers agree that they are willing to share data with companies that send personalised offers and discounts (Altexsoft, n.d.), providing incentives would be an appropriate motivator for Airbnb to collect more data. According to the MARS model insights in section 5.2, the most important factor that affects the prediction model is demographics ("age", "gender", "language"), hence continuously collecting more demographic data is important for progressively improving the model.

Simple voting boxes could be added into the website and app to get trends and insights from fun and casual questions (refer to **Appendix 13** for depiction of voting boxes). An example of questions in these voting boxes could be "Are you an adventure lover or indoor enthusiast?". Users would likely to

click their preference to find out what the Airbnb community's average results are. With the votes, Airbnb can better understand the user's preferences to give better personalised listings on the website and via emails. Getting more demographic data will also improve the prediction model's accuracy.

Feasibility of this recommendation is largely dependent on the country's privacy and data laws that might limit Airbnb's ability to collect more demographic data. For example, the EU's General Data Protection Regulation that requires organisations to comply with data minimisation by only collecting minimal personal data that it needs to process to achieve their processing purposes.

After finding out feature importance of all models built in section 5, the readers realise all the actions of the users on the website are important in improving the accuracy of all models and continually feeding more "actions" data into the model will help to improve the model's accuracy even more.

8. Feasibility

8.1 SWOT

Here is a summary of the SWOT (Strength, Weakness, Opportunities and Threat) analysis of the different action strategy of the recommendations. Refer to Appendix 14 for more details.

Strengths	Weaknesses
<ul style="list-style-type: none"> ● Top 5 - algorithm has a matching rate of 93%. ● Peak Season - Able to obtain what are the influential festivals and peak season easily by exploring the dataset. ● Urgency Status - Model has a accuracy of 85 percentage 	<ul style="list-style-type: none"> ● Top 5 - Feasibility study on their database is required on how to implement the system (Presto, Druid and Airpal) ● Peak Season - Yearly review to identify what are the possible seasons and Reliance on customer allowing airbnb to send email to them ● Urgency Status - Effectiveness of this strategy is highly dependent on the accuracy of the model.
Opportunities	Threats
<ul style="list-style-type: none"> ● Top 5 - Potential advertising revenue from recommended listing ● Peak Season - Potential advertising revenue from marketing the content. ● Urgency Status - Model has a accuracy of 85 percentage 	<ul style="list-style-type: none"> ● Top 5 - Performance with larger dataset would change, yearly evaluation required. ● Peak Season - Wrongly identified season may lead to a wrong strategy applied ● Urgency Status - Potential area of conflict with data protection of many countries and conversion from non-urgent to urgent user might make the whole airbnb experience worse off

8.2 Financial Feasibility

The overall financial feasibility translates to a overall revenue increase of 15%. Refer to appendix 15 for more details.

Considerations	Estimated(\$)
Additional Cost	501 million
Additional Revenue	910 million
Total	409 million

9. Limitations and Further Research

9.1 Data Collection

Other important information such as search history and users' interests are not included in the training dataset, but they are likely to be highly relevant in the prediction of destination country. The performance of our model is thus limited due to the lack of such information in our dataset. However, this limitation will be addressed by the "collecting more data" recommendation in section 7.3.

9.2 Research Scope

In terms of specification level, the specificity of the prediction can be narrowed down from country to city. The majority of users chose to travel domestically in the US but there are many cities to choose from within the US itself. By predicting and recommending the dream city for each user, greater personalization can be achieved, which encourages users to make bookings with Airbnb.

Furthermore, only the first booking history of each user is analyzed. If the entire booking history of all Airbnb users is made available for analysis, the research scope can be extended to predicting the next country destination even for those people who have made a booking. This personalize the experience of existing users as well, which helps to boost customer retainment rates in addition to acquisition rates. This will improve the customer loyalty and help to sustain Airbnb's market share.

9.3 Model Improvement

It is recommended to train more rigorous and advanced models such as Neural Network and XGBoost, a gradient boosting algorithm, which might produce even better results than random forest. Though these models are black box models that are difficult to interpret, the priority for solving this business problem is still the accuracy of the model, making them suitable candidates for future research.

10. Conclusion

In conclusion, data is valuable in conducting analytics and building models for Airbnb to provide better user experience by personalisation. With accurate and efficient user travel destination and urgency status prediction models, Airbnb can develop personalized and targeted marketing strategies for each user, increasing booking rates while avoiding wasted marketing resources, thus leading to greater revenue growth.

Random forest and MARS are the most preferred models for each of the prediction problems, based on their performance. However the model still needs to be upgraded and modified in the long run to ensure sustained performance. Feedback, customer action data and customer demographic data needs to be continuously collected to measure performance and improve the models if necessary.

11. References

Coldwell, J. (2001). *Characteristics of a Good Customer Satisfaction Survey*. New Delhi, DL: Tata McGraw-Hill , pp. 193 - 199.

Hyken, S. (2018, August 21). Convenience: The Next Competitive Differentiator. Retrieved from <https://retailminded.com/convenience-the-next-competitive-differentiator/>

Loo, J. (2017, November). The future of travel: New consumer behavior and the technology giving it flight. Retrieved from <https://www.thinkwithgoogle.com/marketing-resources/new-consumer-travel-assistance/>

Taylor, M. (2014). How Creating a Sense of Urgency Helped Me Increase Sales By 332%. *Conversion XL*. Retrieved from <https://conversionxl.com/blog/creating-urgency/>

Kutschera, S. (2018, July 4). Travel statistics to know about in 2018 and 2019. Retrieved from <https://www.treksoft.com/en/blog/65-travel-tourism-statistics-for-2019>

Patrick Veenhoff (2018, July 26). Successfully Implementing the AirBnB of Agile Corporate Learning. Retrieved from <https://www.linkedin.com/pulse/successfully-implementing-airbnb-agile-corporate-patrick-veenhoff/>

Pam Neely (2018, July 26). 5 Things That Make People Click On Content. Retrieved from <https://whatagraph.com/blog/articles/5-things-that-make-people-click-on-content>

Statista (2018). Number of Airbnb users in the United States from 2016 to 2022 (in millions). Retrieved from <https://whatagraph.com/blog/articles/5-things-that-make-people-click-on-content>

Forbes (2018). What Technology Stack Does Airbnb Use? Retrieved from <https://www.forbes.com/sites/quora/2018/02/20/what-technology-stack-does-airbnb-use/#2e74c8fd4025>

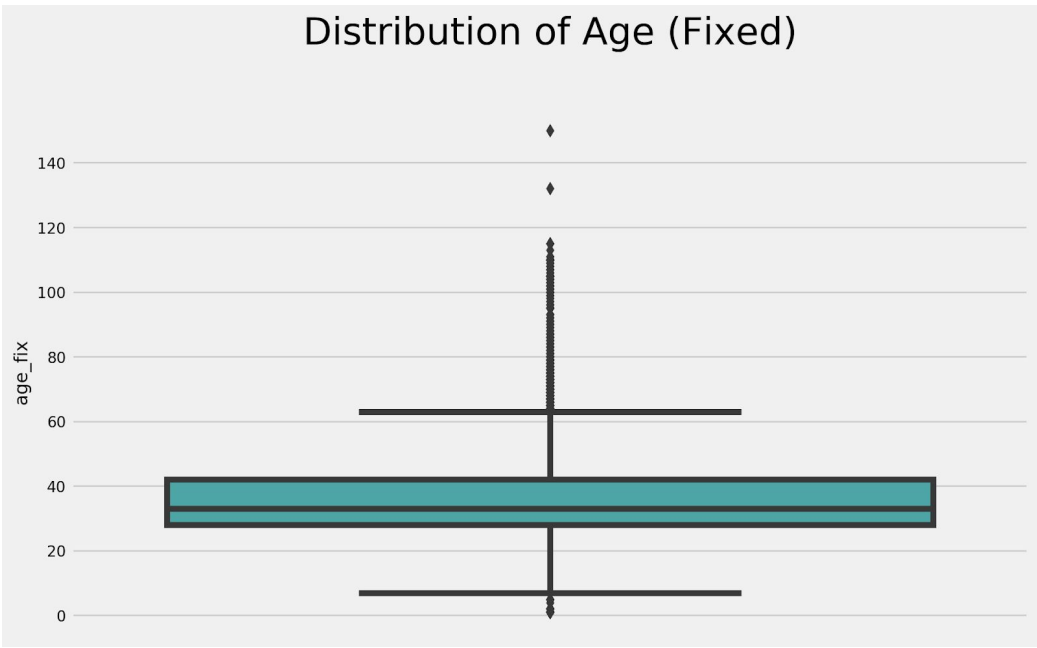
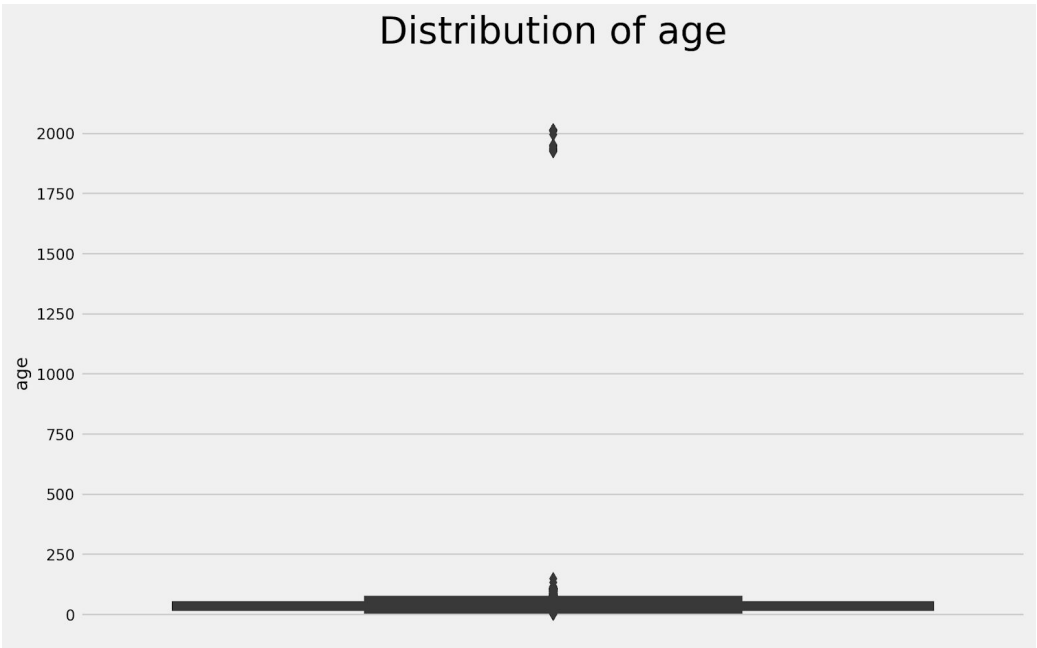
Rejoiner (2018). The Amazon Recommendations Secret to Selling More Online Retrieved from <http://rejoiner.com/resources/amazon-recommendations-secret-selling-online/>

Businessinsider (2018). Airbnb made \$93 million in profit on \$2.6 billion in revenue Retrieved from <https://www.businessinsider.sg/airbnb-profit-revenue-2018-2/?r=US&IR=T>

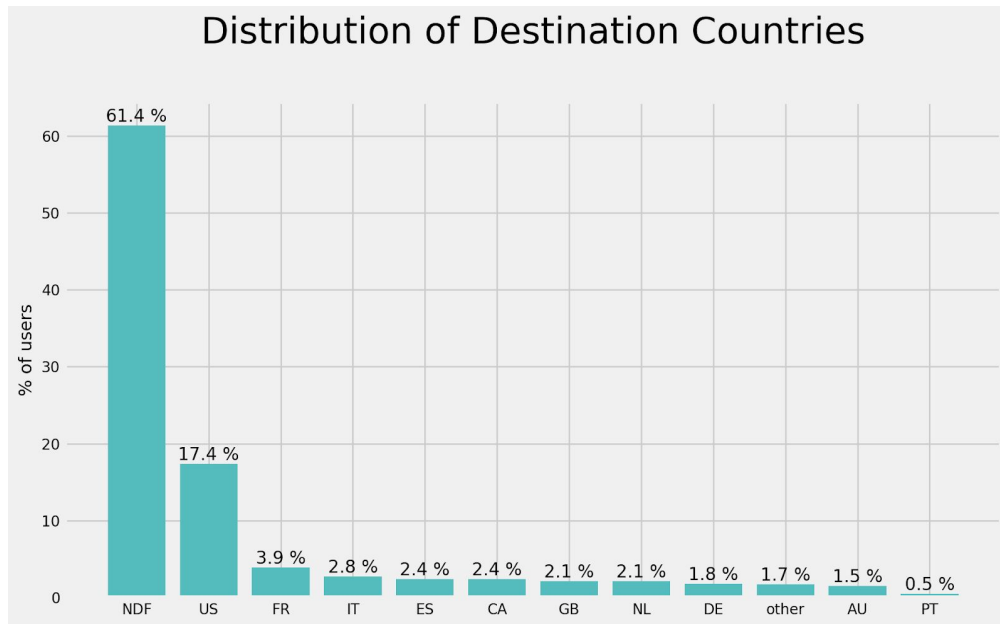
VoucherCloud (2015). What Science Says About Discounts, Promotions and Free Offers. Retrieved from https://www.huffpost.com/entry/what-science-says-about-discounts_b_8511224

Verticalmeasures (2018). How Much Does Real Content Marketing Cost? Retrieved from <https://www.verticalmeasures.com/blog/digital-marketing/how-much-does-content-marketing-cost/>

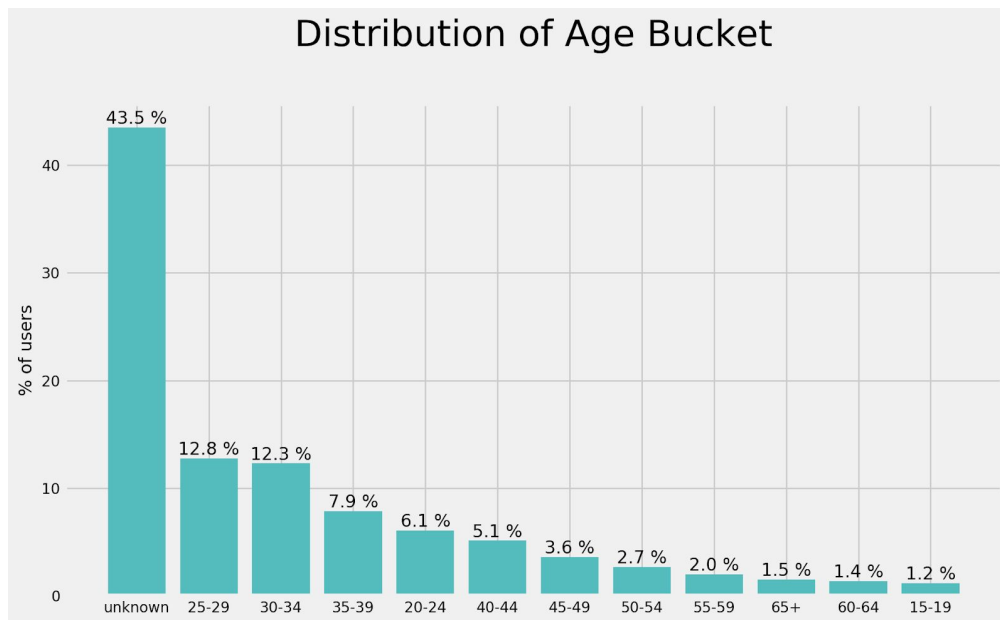
Appendix 1: Age Outliers



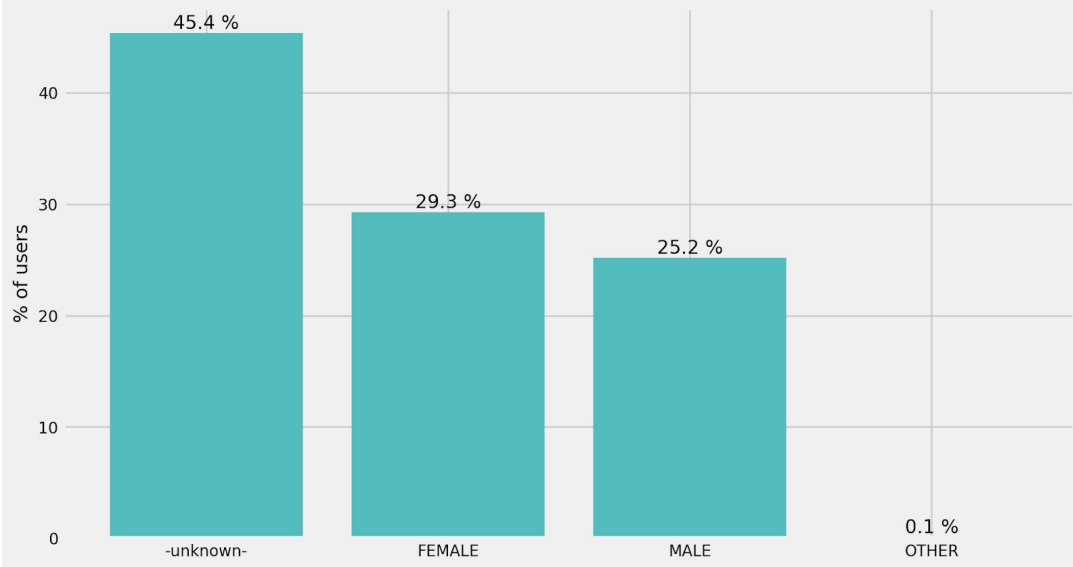
Appendix 2: Distribution of Destination Countries



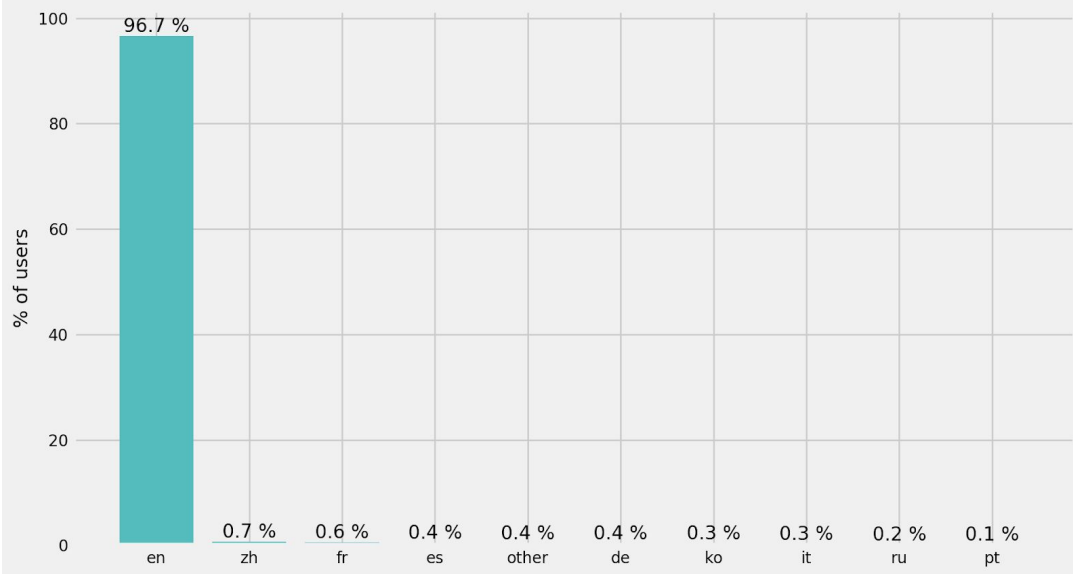
Appendix 3: Distribution of User Demographics



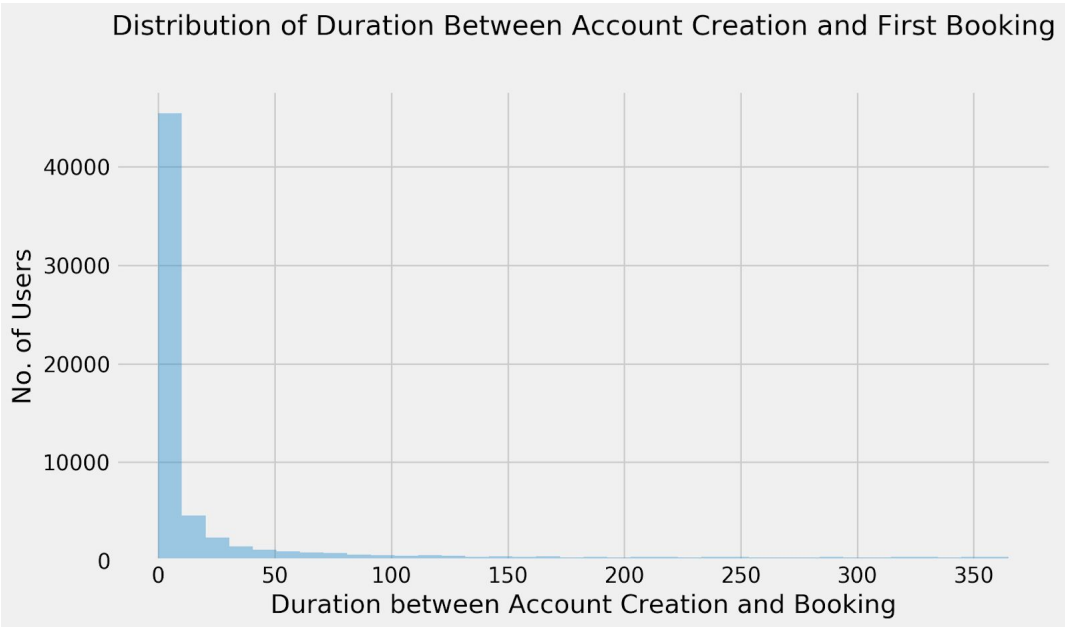
Distribution of Gender



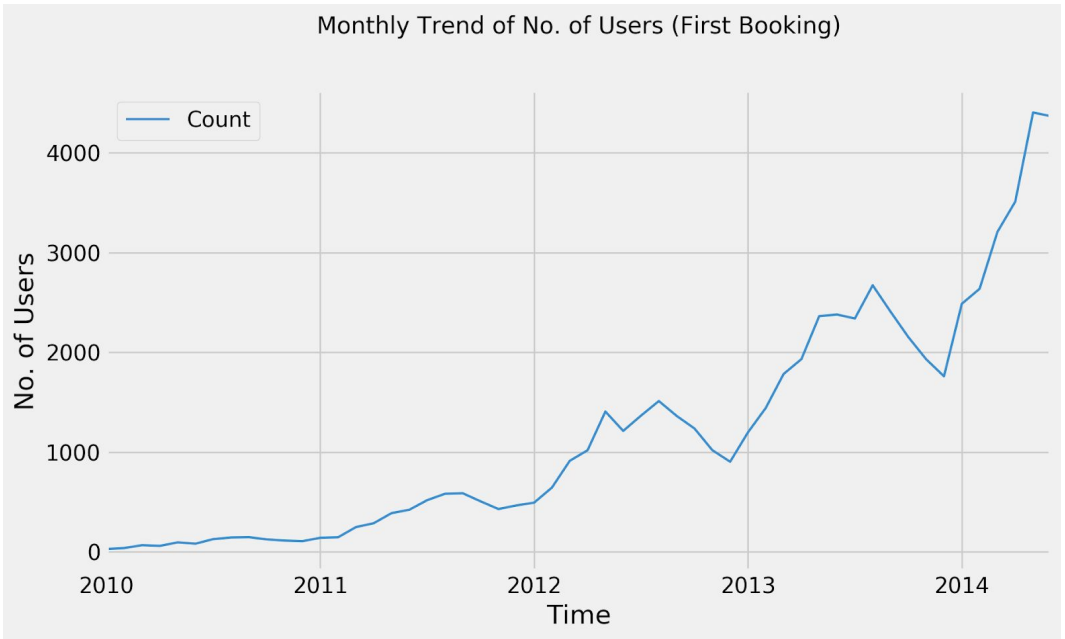
Distribution of Language



Appendix 4: Distribution of Duration between Account Creation and First Booking



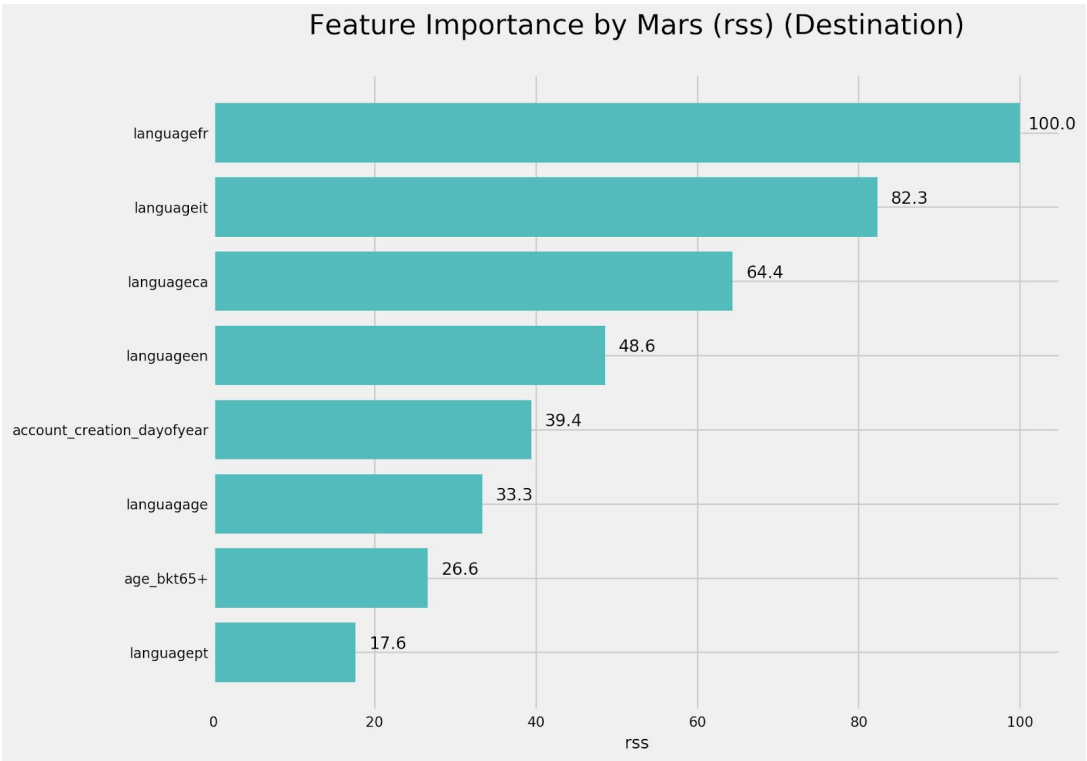
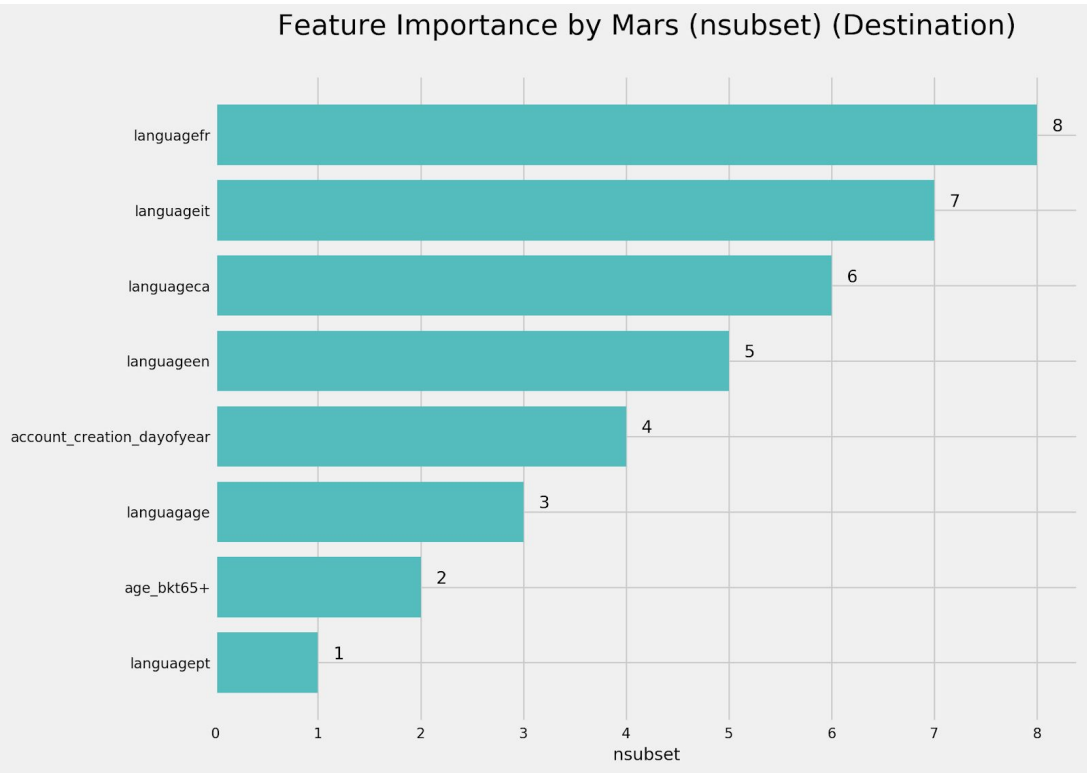
Appendix 5: Peak Season

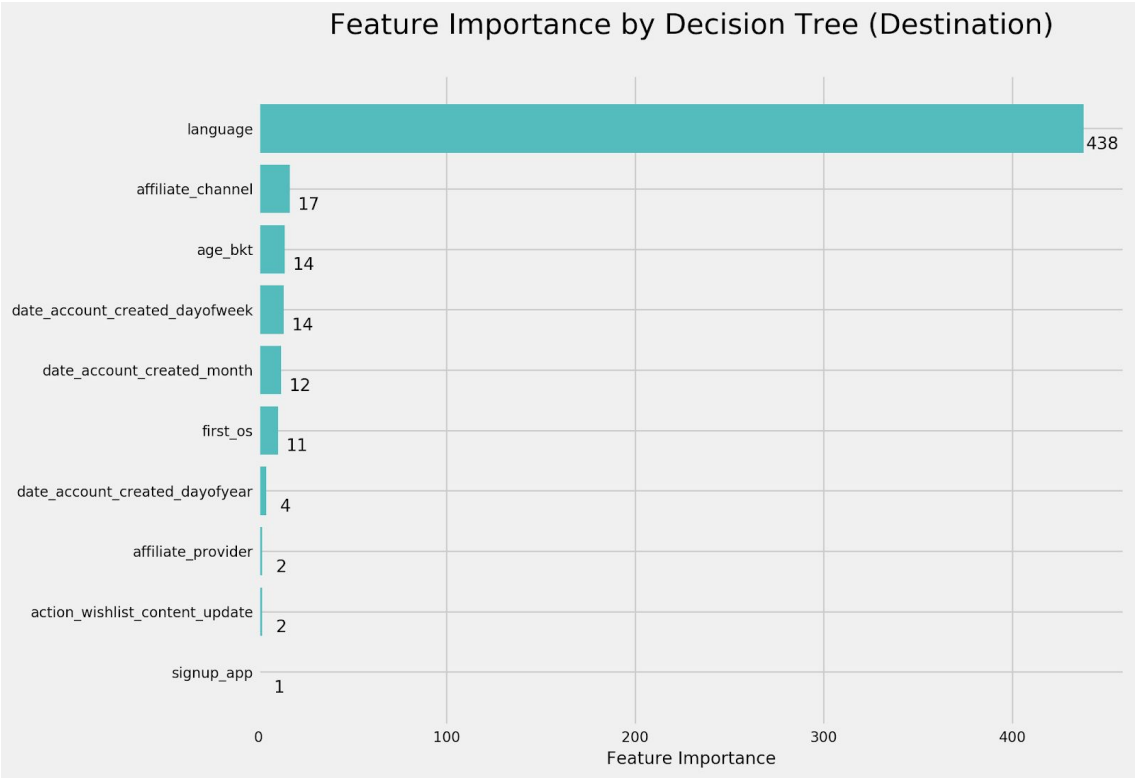
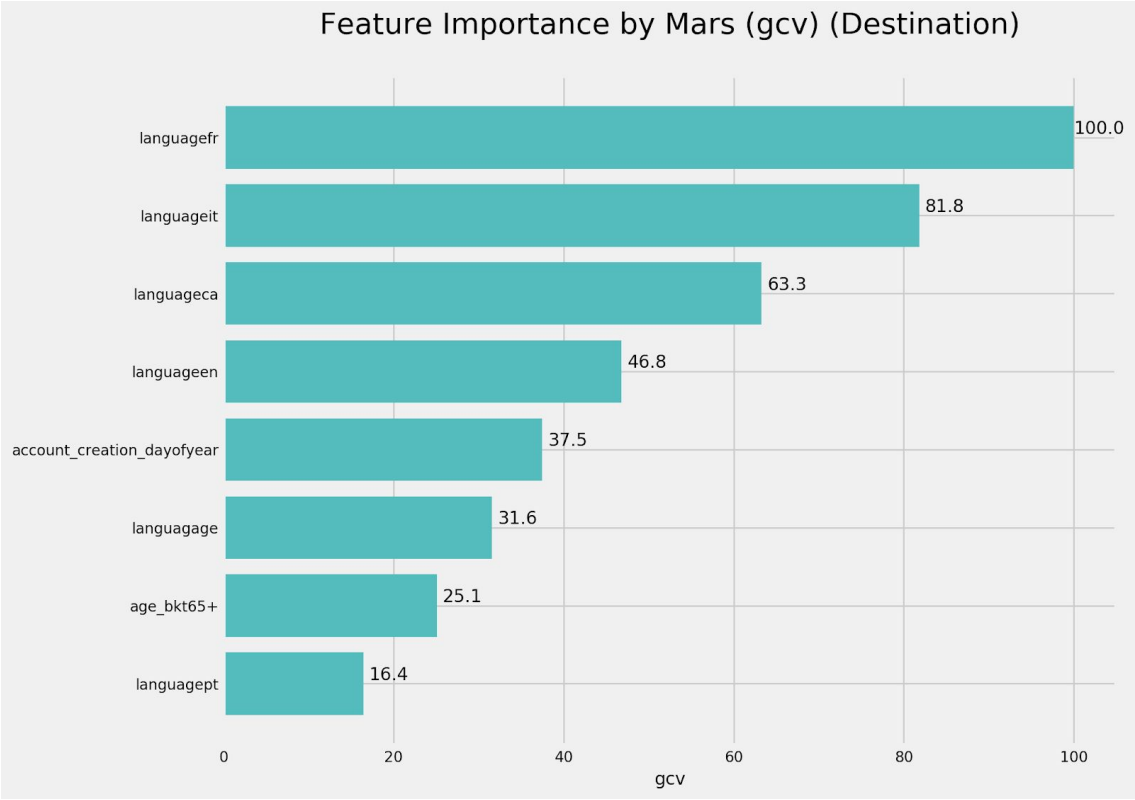


Appendix 6: Association between Country Destination and Month

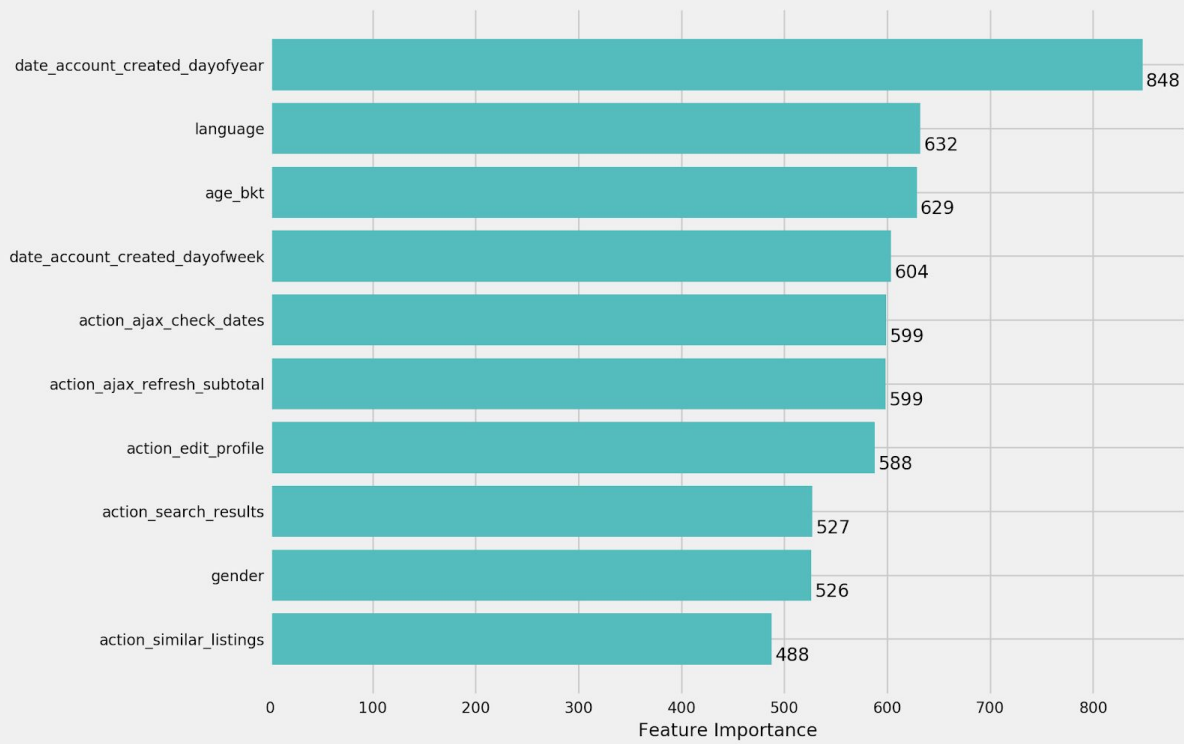
Confidence (%)	Support (%)	Lift	Rule
8.19	0.46	2.1	December → Australia
6.57	0.49	1.43	October → Germany
7.09	0.59	1.29	May → Netherland
9.05	0.92	1.26	April → Italy
7.88	0.66	1.26	July → Spain
7.72	0.95	1.25	June → Canada
7.67	0.64	1.24	July → Canada
11.61	1.47	1.24	July → France
7.57	0.62	1.22	August → Canada
7.69	0.95	1.22	June → Spain

Appendix 7: Important Features for Models Predicting Country Destination

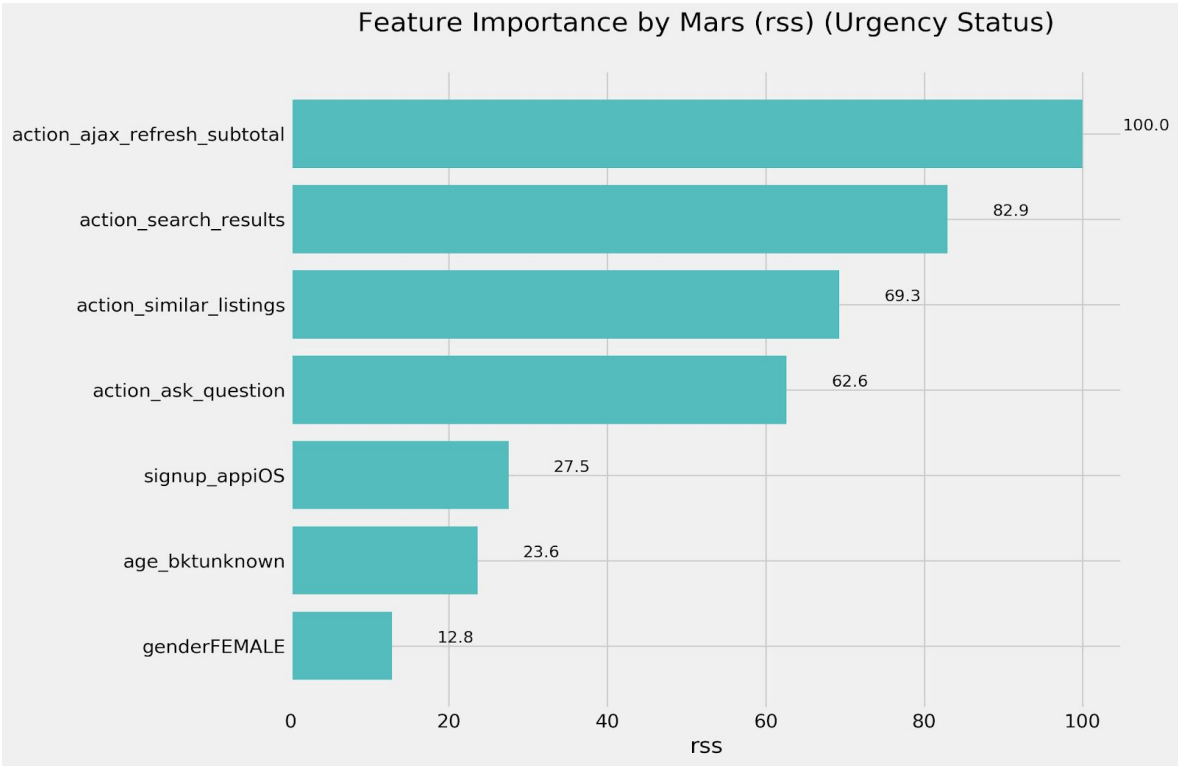
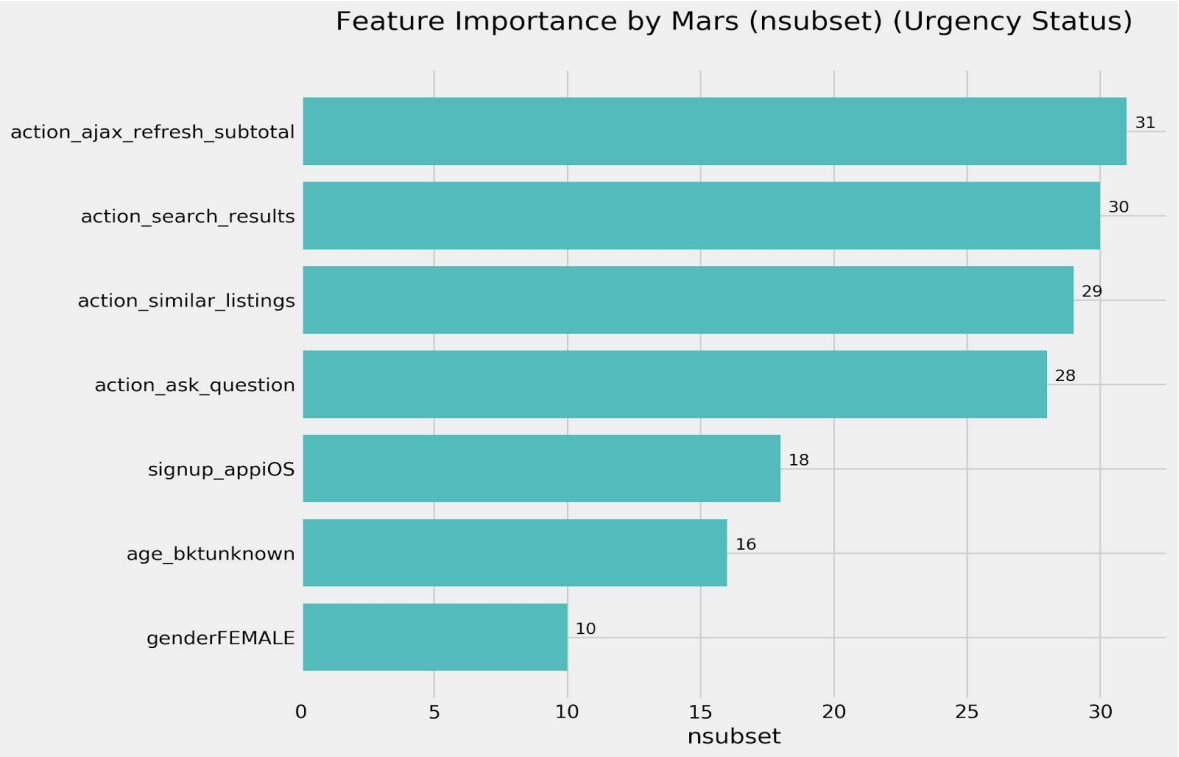


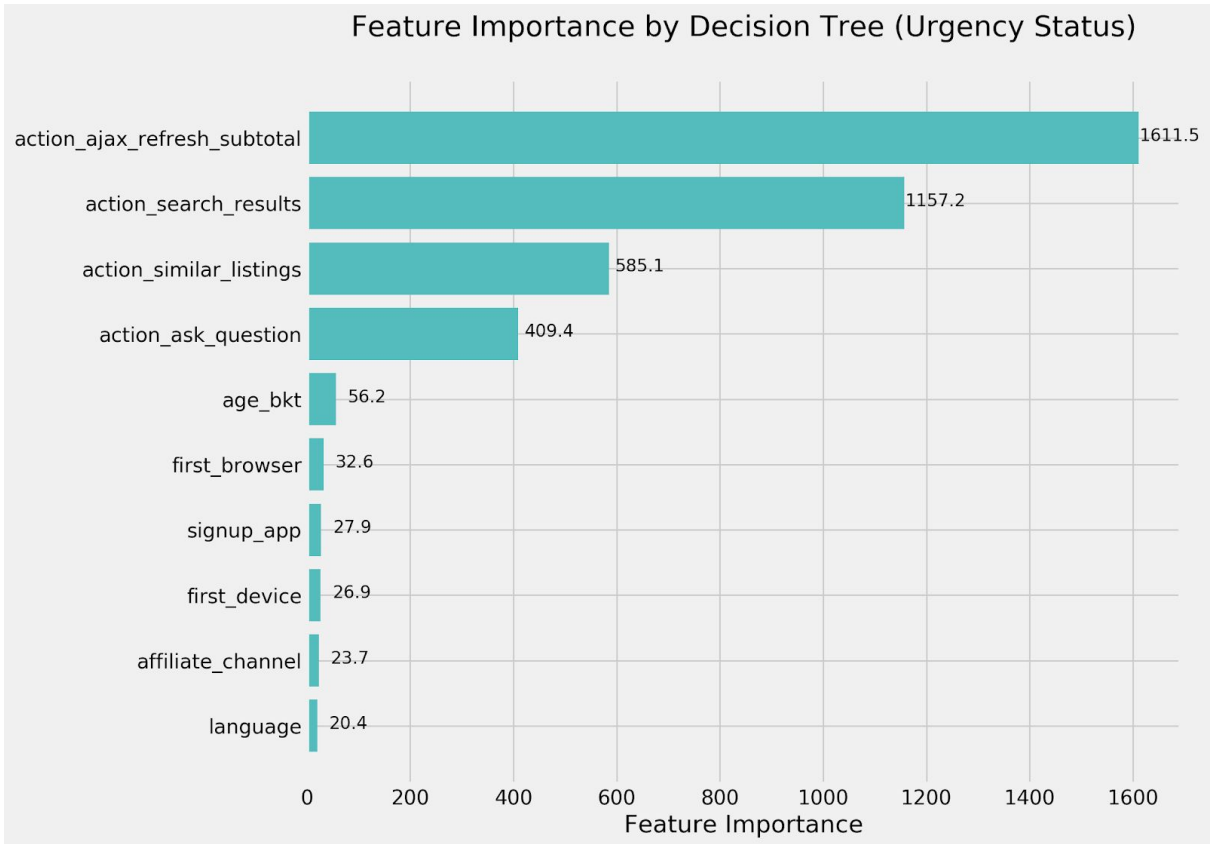
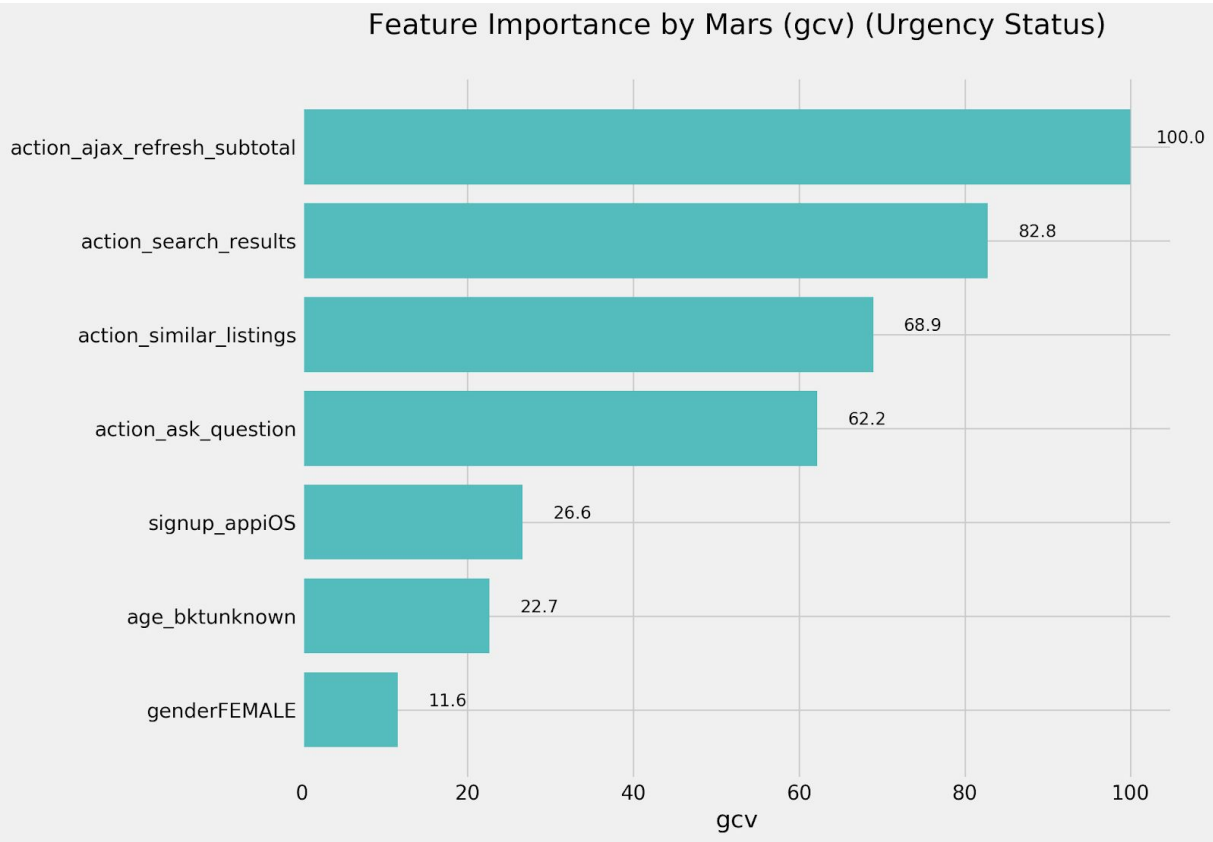


Feature Importance by Random Forest (Destination)

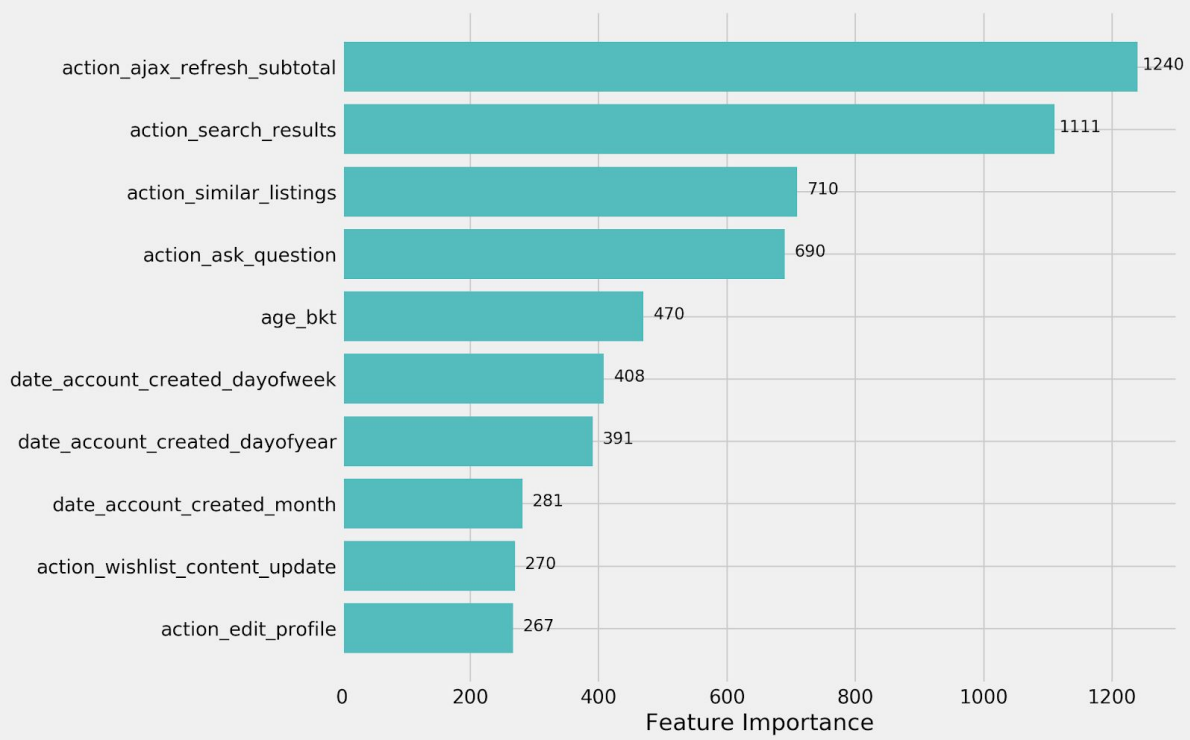


Appendix 8: Important Features for Models Predicting Urgency Status





Feature Importance by Random Forest (Urgency Status)



Appendix 9: Logistic Regression Value for Country Destination

	CA	DE	ES	FR	GB	IT	NL	other	PT	US
(Intercept)	21.948	0.231	0.289	0.235	1.149	0.196	1.745	10.319	0.436	223.739
languageca	0.673	13.237	85.428	2.791	0.205	2.889	0.651	0.056	14.120	0.000
languageede	0.014	120.222	7.768	17.164	0.007	19.413	0.019	0.004	3.607	0.128
languageen	0.737	14.436	72.783	73.388	0.764	63.259	0.575	0.662	5.221	0.697
languagees	0.028	0.400	186.179	89.331	0.635	67.261	0.758	0.300	0.414	0.476
languagefr	0.094	0.389	40.251	631.762	0.510	16.110	0.005	0.072	4.104	0.133
languageit	0.014	47.100	136.338	38.915	0.009	361.429	0.011	0.848	1.409	0.124
languageja	1.430	0.811	242.259	0.550	0.851	0.656	0.149	1.053	0.631	1.327
languageko	0.556	0.611	0.260	227.107	0.935	139.591	0.049	2.335	0.610	0.852
languageother	0.040	57.358	0.525	24.451	0.026	0.484	5.320	1.120	29.136	0.387
languagept	0.626	0.974	0.741	1.089	0.295	0.925	0.574	0.301	943.945	0.001
languageru	0.052	65.625	103.262	29.599	0.032	133.281	0.076	0.021	0.506	0.645
languagezh	0.103	3.857	0.230	14.838	0.065	5.390	0.006	0.272	0.168	0.253
date_account_created_dayof year	1.806	1.001	1.001	1.000	1.002	1.000	1.004	0.999	1.001	1.000
age_bkt20-24	0.851	4.958	2.919	2.000	4.854	2.435	5.394	2.122	1.461	3.305
age_bkt25-29	0.803	2.024	0.740	0.961	1.290	0.875	1.847	1.518	0.332	1.981
age_bkt30-34	1.843	3.360	1.418	1.806	2.510	1.644	3.912	2.688	0.552	3.555
age_bkt35-39	0.846	1.469	0.524	0.739	1.111	0.705	1.018	1.267	0.260	1.493
age_bkt40-44	0.770	1.238	0.308	0.940	0.937	0.591	0.909	0.724	0.353	1.183
age_bkt45-49	0.690	1.562	0.555	1.080	1.791	0.932	1.269	0.980	0.221	1.407
age_bkt50-54	0.322	1.550	0.488	0.713	1.670	0.778	0.812	0.572	0.445	0.896
age_bkt55-59	0.779	3.459	0.813	1.594	2.698	1.113	0.830	1.519	0.981	1.893
age_bkt60-64	2.144	1.884	2.501	2.411	5.744	1.880	5.129	2.115	0.171	2.920
age_bkt65+	0.559	1.461	1.932	2.869	4.191	1.982	2.251	3.014	0.734	2.311
age_bktunkno wn	0.593	1.365	0.860	0.925	1.432	1.011	1.397	1.070	1.183	1.099
signup_appiOS	0.795	1.497	0.298	2.821	10.470	2.229	4.071	2.012	1.158	1.164
signup_appMo web	0.657	0.770	0.143	0.966	2.418	0.716	2.633	0.986	0.721	0.953

signup_appWeb	0.398	0.612	0.271	1.614	4.140	0.953	1.564	1.218	0.569	0.570
genderFEMALE	0.802	0.916	1.094	0.949	0.982	0.912	0.792	0.739	1.892	0.891
genderMALE	0.324	0.577	0.405	0.316	0.277	0.272	0.365	0.345	0.679	0.344
genderOTHER	8.767	4.936	0.098	2.071	3.820	0.012	4.291	1.495	0.773	2.043
first_deviceOthers	0.059	0.084	4.754	0.529	1.311	0.281	3.998	1.026	1.082	0.324
first_devicePhone	0.495	0.272	1.068	0.747	0.674	0.669	0.408	1.319	0.372	1.042
first_deviceTablet	0.576	0.724	0.935	0.669	0.740	0.888	0.875	0.764	0.773	0.709
first_osMacOS	0.374	0.439	1.245	0.410	0.766	0.563	0.517	0.653	0.414	0.478
first_osOthers	3.146	1.722	0.061	0.913	0.833	0.926	0.044	1.171	0.286	1.147
first_osWindows	0.627	0.505	1.656	0.538	0.937	0.790	0.519	1.037	1.060	0.730
first_affiliate_trackedomg	1.045	1.064	0.687	0.995	1.365	1.289	1.113	0.816	0.657	0.913
first_affiliate_trackedother	0.837	0.025	0.233	0.631	0.475	0.606	0.706	0.322	3.632	0.566
first_affiliate_trackedtracked-other	5.370	1.507	0.932	1.679	2.164	1.334	4.716	0.943	2.074	2.133
first_affiliate_trackeduntracked	1.459	1.279	0.948	1.360	1.501	1.476	1.445	0.996	0.950	1.159

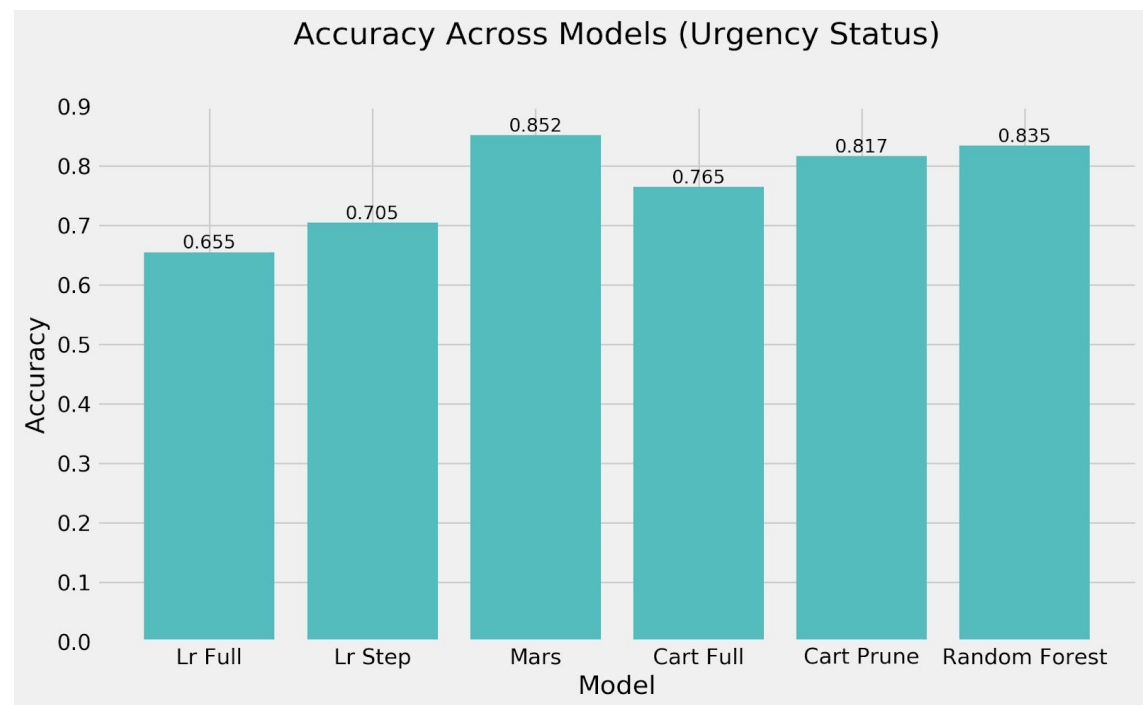
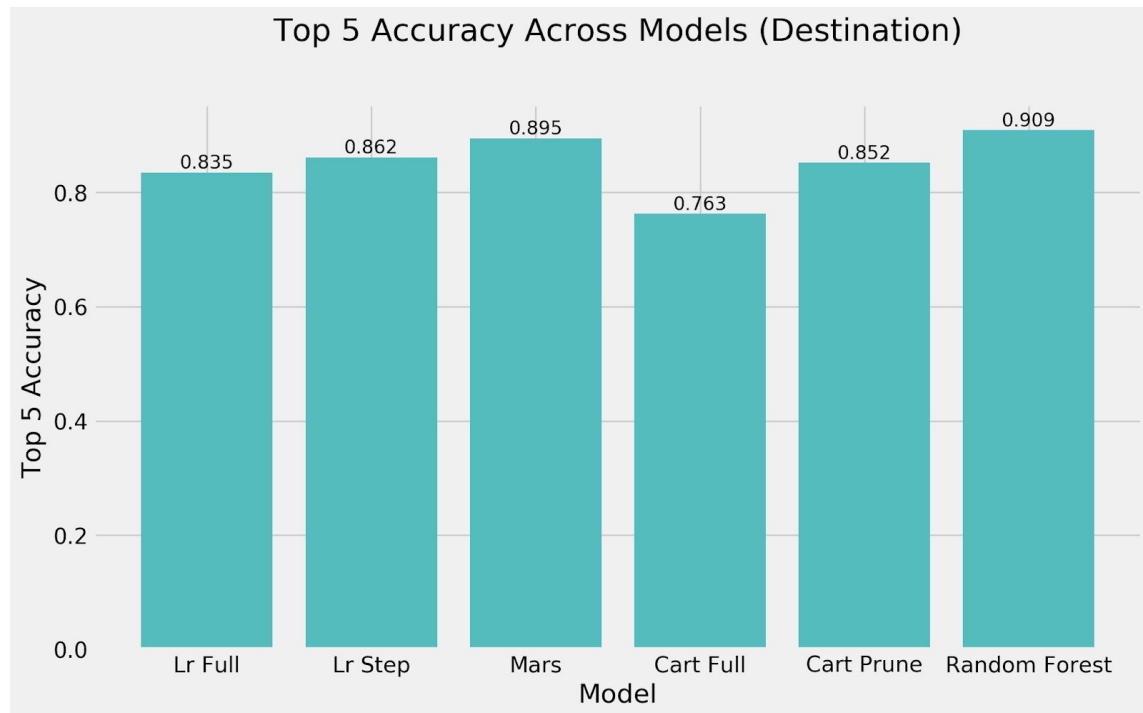
Appendix 10: Logistic Regression Value for Urgency

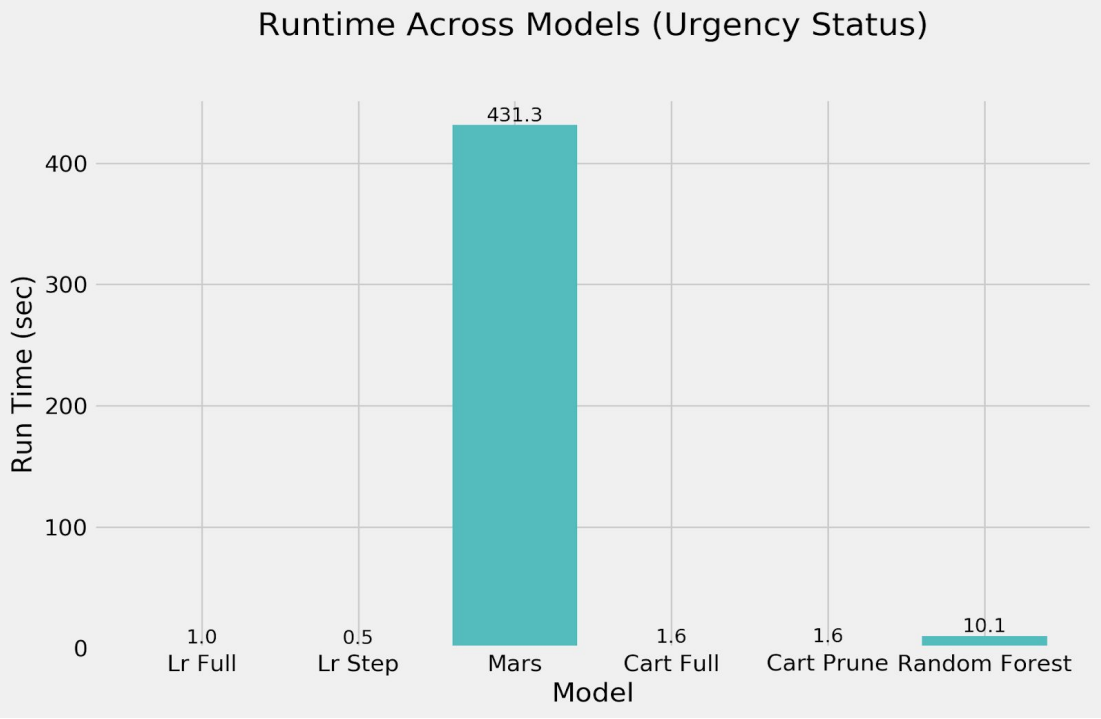
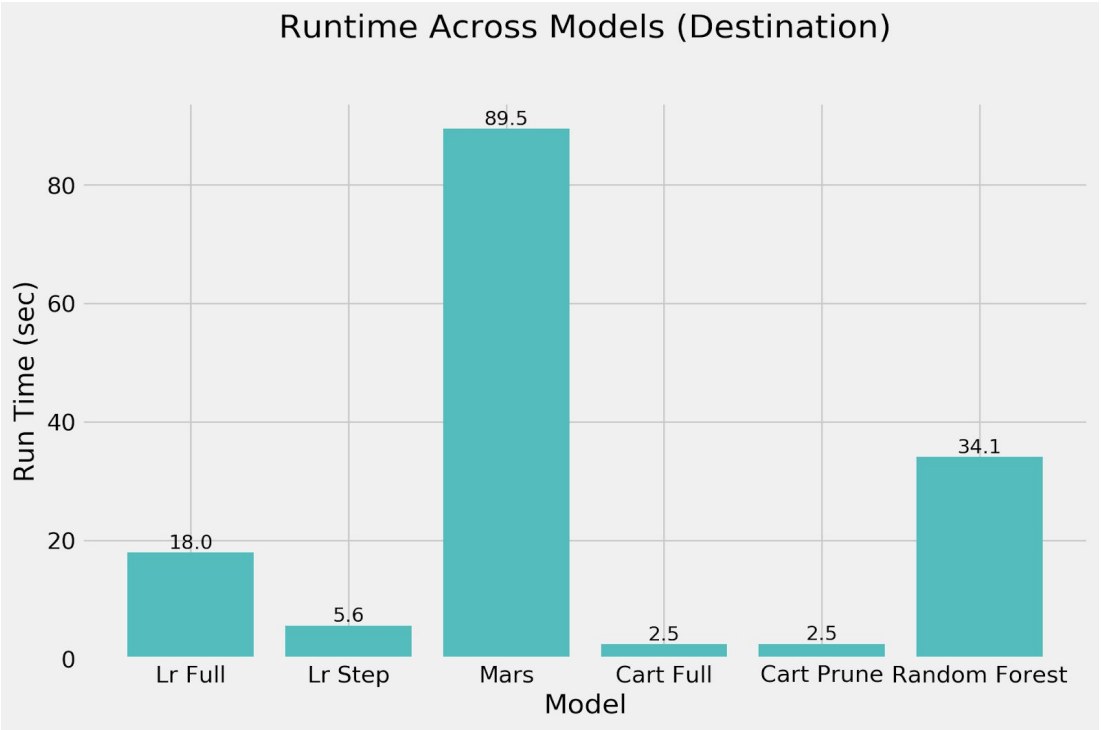
	Coefficient
(Intercept)	0.1209217255
action_search_results	2.113235539
action_ajax_refresh_subtotal	2.114783166
signup_appiOS	0.9040160399
signup_appMoweb	1.378331207
signup_appWeb	1.780216763
genderFEMALE	0.7455601642
genderMALE	0.9554778371
genderOTHER	0.8543996732
action_ask_question	1.059781233
action_similar_listings	1.058175846
age_bkt20-24	1.19564333
age_bkt25-29	1.272319803
age_bkt30-34	1.296165431
age_bkt35-39	1.270223611
age_bkt40-44	1.234168905
age_bkt45-49	1.250751354
age_bkt50-54	1.0085489
age_bkt55-59	1.483801502
age_bkt60-64	1.384876528
age_bkt65+	1.609881743
age_bktunknown	1.860832412
affiliate_channelcontent	0.3933461562
affiliate_channeldirect	0.8907608077
affiliate_channelother	1.34482904
affiliate_channelremarketing	0.5997260162
affiliate_channelsem-brand	0.8746700978
affiliate_channelsem-non-brand	0.8480047729
affiliate_channelseo	1.156467654

date_account_created_dayofyear	1.001772029
first_browserChrome	1.33509821
first_browserFirefox	1.506357297
first_browserIE	1.281683037
first_browserOther	1.516162273
first_browserSafari	1.43231061
first_osMacOS	1.143710237
first_osOthers	1.993047166
first_osWindows	1.281056639
first_deviceOthers	0.7545323158
first_devicePhone	1.052908899
first_deviceTablet	0.7666894963
signup_methodfacebook	0.8778717407
signup_methodgoogle	1.10608416
action_wishlist_content_update	0.9940675531

Appendix 11: Airbnb Website and Mobile App

The plots below shows the performance and training time for all models that have been built.

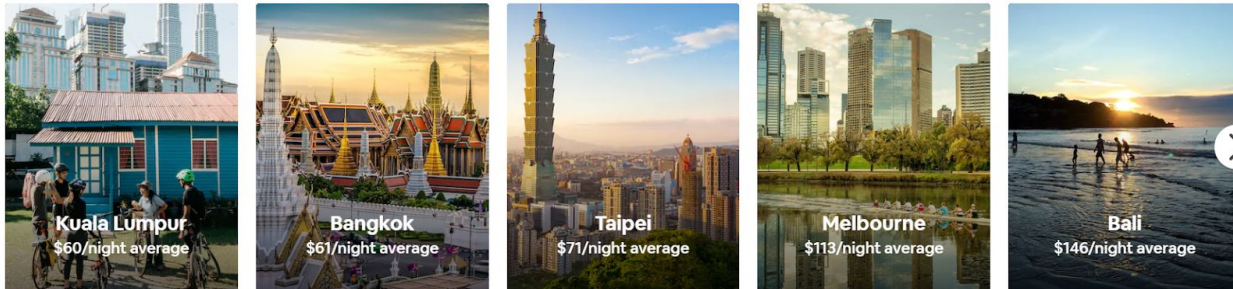




Appendix 11: Airbnb Website and Mobile App

Airbnb's website layout

Recommended for you



Airbnb's Mobile Application Front Page Layout

Homes around the world



ENTIRE FLAT · BOSA
**apart.terrace
overlooking the river**
\$91 per night
★★★★★ 205



CAVE · SANTORINI
Hector Cave House
\$425 per night
★★★★★ 226 · Superhost



ENTIRE VILLA · SINGARAJA
**180° VIEW, PRIVATE
POOL VILLA..**
\$175 per night
★★★★★ 198 · Superhost



PLUS VERIFIED · MENAGGIO
**Romantic, Lakeside
Home with Views of La...**
\$183 per night
★★★★★ 215

Show all (2000+)



Appendix 12: Website Snippet after Personalisation

Recommended For You

Country #1

Country #2

Country #3

Country #4

Country #5

#Country 1 related Blog Post

Blog posts related to the top 5 predicted countries initially only located in a separate Airbnb blog website. This is to capture the attention and pique interest for the country. Examples of the blog posts could be “Short Weekend Getaways Ideas” targeted at users that have US in their top predicted countries.

Top Home Listings for Country #1

Home #1

Home #2

Top Experiences for Country #1

Experience #1

Experience #2

Experience #3

Home #3

Recommended for you




Copenhagen
\$167/night average



Berlin
\$96/night average



Prague
\$94/night average



London
\$165/night average



Los Angeles
\$178/night average

Copenhagen Inspiration, Homes and Experiences



Step into Spring

Find the best Cherry Blossom viewing spots, discover colorful spring festivals, and enjoy the beauty of the blossoming season

Explore More



BOAT RIDE
Hey Captain, let's sail!
\$85 SGD per person
4.81★ (545)



CRAFT CLASS
Knitting in the nordic
Hygge tradition
\$74 SGD per person
4.93★ (29)



BIKE RIDE
Copenhagen Discovery On
Stylish Bike
\$85 SGD per person
4.93★ (14)



BEER TASTING
BeerWalks with
Copenhagen Locals
\$53 SGD per person - 2 hours -
Drinks included



ENTIRE LOFT - 1 BED
★CityCentre
Penthouse.PrivateTerrace.25min...
\$226 SGD per night - Free cancellation
★★★★ 404 - Superhost



ENTIRE APARTMENT - 7 BEDS
Penthouse, 4-5 rooms + 5
balconies
\$382 SGD per night - Free cancellation
★★★★ 290



ENTIRE APARTMENT - 1 BED
Charming apartment in the heart
of Copenhagen
\$195 SGD per night - Free cancellation
★★★★ 246

Appendix 13: Voting boxes Examples

To obtain more general information from users to better predict urgency status and target them with the right marketing strategy, small voting boxes could be included.

What defines me best when I am on vacation?

ADVENTURE JUNKIE

INDOOR LOVER

What defines me best when I am on vacation?

ADVENTURE JUNKIE

54%

INDOOR LOVER

46%

Local's eat rule! Bring on all the unusual dishes and stories behind them

TRUE

FALSE

Local's eat rule! Bring on all the unusual dishes and stories behind them

TRUE

78%

FALSE

22%

Appendix 14: Feasibility Analysis

Personalization Recommendations

Strengths	Weaknesses
<ul style="list-style-type: none">• Top 5 algorithm has a matching rate of 93%.• High match percentage would influence user to click and see more details about the listing (Neely).• Able to implement with the Airbnb's current Agile practices (Veenhoff).• Blanket approach to all users.	<ul style="list-style-type: none">• 45.6 million user using airbnb services (statista)• 45.6 million dataset would take a long time and a huge amount of resources to generate the model• Feasibility study on their database is required on how to implement the system (Presto, Druid and Airpal)• Unable to interpret the randomforest model
Opportunities	Threats
<ul style="list-style-type: none">• Potential increase in booking rates• Potential advertising revenue from recommended listing	<ul style="list-style-type: none">• Performance with larger dataset would change.• While randomforest currently is the best model, it might not be the best model on a larger dataset.• Factors such as imbalanced data, normality of dataset might skew the result as the dataset grows

Festivals and Peak seasons increase marketing

Strengths	Weaknesses
<ul style="list-style-type: none">• Able to obtain what are the influential festivals and peak season easily by exploring the dataset. <p>Peak season:</p> <ul style="list-style-type: none">• Content market would be able to get more website traffic and usage of the application. <p>Off Peak season:</p> <ul style="list-style-type: none">• A promotional discount, when applied	<ul style="list-style-type: none">• Yearly review to identify what are the possible seasons• Reliance on customer allowing airbnb to send email to them <p>Peak season:</p> <ul style="list-style-type: none">• Content marketing would only be effective if the right content is written. (Altimeter)• High reliance on the quality of content to attract more users.

<p>with the correct financial strategy would be able to boost revenue. (ipsos)</p> <ul style="list-style-type: none"> Furthermore, typically around 51% of user would be influenced by promotion deals. (huffpost) Able to control and limit the usage of discounts 	<p>Off Peak season:</p> <ul style="list-style-type: none"> An in-depth promotional strategy needs to be implemented, taking into account factors such as timing of sales, pricing strategy and place of sales(which country the user is from). Potential financial burden.
Opportunities	Threats
<p>Peak season:</p> <ul style="list-style-type: none"> Potential advertising revenue from marketing the content. Potential collaboration with airbnb host for them to market their places themselves with guidance from airbnb <p>Off Peak season:</p> <ul style="list-style-type: none"> Potential collaboration with airbnb host to offer the discount rate to ease financial burden on airbnb 	<ul style="list-style-type: none"> Wrongly identified season may lead to a wrong strategy applied <p>Peak season:</p> <ul style="list-style-type: none"> Fierce competitors such as Expedia Group and Priceline Group have existing content marketing efforts (blog.advertising.expedia) <p>Off Peak season:</p> <ul style="list-style-type: none"> Competitors might retaliate back with the same strategy, leading to minimal effect of promotional discount.

Recommendation for Urgent Users and Non-Urgent Users

Strengths	Weaknesses
<ul style="list-style-type: none"> Model has a accuracy of 85 percentage Strategy is highly scalable. Able to decide who is urgent and non urgent according to internal benchmarks <p>Urgent Users:</p> <ul style="list-style-type: none"> Sending a customized email allows for airbnb to have a more personalized experience. <p>Non-Urgent Users:</p> <ul style="list-style-type: none"> Converting them to be urgent user help airbnb sales and increase user base. 	<p>Effectiveness of this strategy is dependent on the accuracy of the model.</p> <p>Urgent Users:</p> <ul style="list-style-type: none"> Using the search history of user to send out content listing may be inaccurate. Reliance on customer allowing airbnb to send email to them <p>Non-Urgent Users:</p> <ul style="list-style-type: none"> An in-depth needs to be implemented, taking into account factors such as timing of sales, pricing strategy and place of sales(which country the user is from). Potential financial burden.

Opportunities	Threats
<p>Urgent Users:</p> <ul style="list-style-type: none"> Another model can be used to improve the accurate of the content of the email for the airbnb users. <p>Non-Urgent Users:</p> <ul style="list-style-type: none"> Potential collaboration with airbnb host to offer the discount rate to ease financial burden on airbnb 	<p>Urgent Users:</p> <ul style="list-style-type: none"> Potential area of conflict with data protection of many countries <p>Non-Urgent Users:</p> <ul style="list-style-type: none"> Conversion from non-urgent to urgent user might make the whole airbnb experience worse off. This is because user may feel rushed into using airbnb.

Financial Feasibility

Implementation Process	Functions	Estimated Cost
<p>(1) Adding new codes for the prediction model into the existing underlying codes of the website to incorporate the function.</p> <p>(2) Overall testing of all newly introduced functions to ensure the functionality and the authenticity of the output results.</p>	Coding for the New Functionality + Testing of Whole Package	6,000 (Giover, 2017)
(3) Redesigning the web pages to ensure the accessibility of the new function to the website users, and the results obtained from the prediction model would be made accessible to all portal users, being presented in a considerably clear, pronounced and aesthetically pleasing manner.	Android, IOS and Webpage Redesigning	7,650 (estimatemyapp)
(4) Blog content material content generation to suit the needs of the user. Assume that a content creator makes 1 post every day.	Generation of content to market to users.	5,000 x 5 per month Verticalmeasures (2018)
<p>(4) Promotion and discount. Airbnb collects from the host a total of 20%. Assuming that the promotional discount is capped at 5%</p> <p>Urgency Model - 60 percent of user will be offered promotional discount and 57 percentage of them will take it out. VoucherCloud (2015)</p> <p>Peak season model - 9 month off - peak to all and 57 percent of them taking up. VoucherCloud (2015)</p>	<p>Airbnb Revenue was 2.6 billion in 2017(Businessinsider, 2017)</p> <p>$2.6 \text{ billion} * 60\% * 57\% * 25\% \text{ (promo of 5\% only)} + 2.6 \text{ billion} * 0.75 * 57\% * 0.25$</p>	Revenue cut of up to 500 million
(4) Regular maintenance and evaluation of model's performance	Maintain Relevance of Prediction Model	93,000 Annually[#]

Implementation Process	Estimated Revenue	
Recommendation engine (rejoiner, 2017) <ul style="list-style-type: none"> - Top 5 - Recommendation for different user segments - Recommendation for peak and non-peak 	Increase in revenue of up to 35% (Rejoiner, 2018) Airbnb Revenue was 2.6 billion in 2017 (Businessinsider, 2017) 2.6 Billion * 35% = 910 million	
	TOTAL	409 million for the first year