15.071 Analytics Edge **Fall 2019**
Final project report

Airbnb First Booking Prediction

Where will a new guest book their first travel experience?

Team members

Killian Farrell
Suzana Iacob
Brandy Piao
Alexandru Socolov

TA

Renbo Zhao

Instructor

Alexandre Jacquillat

Contents

1 Problem and Relevance

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. This vast space of decisions makes it difficult for users to navigate their search and for Airbnb to guide it. Especially new users might be simply overwhelmed. Therefore, our goal is to predict where users will book their *first* travel experience. By doing this, Airbnb can recommend relevant content to assist the booking process and thereby increase the likelihood and speed of booking.

Online booking services like Airbnb have rich knowledge in recommendation systems. Classical methods rely on individual users being *rich* data points. In the given case, we are presented with a 'cold start' problem. Having little to no information about new users poses a difficult question for the sales and marketing taskforce of Airbnb. A purposeful prediction model would help close bookings of first-time customers and concentrate the company's resources on the customers who are likely to proceed with a booking. As acquiring new customers is the hardest problem, registered users are a valuable group that should be addressed and guided towards an initial booking.

2 Analytics Framework

2.1 Data

The data is supplied by Airbnb and consists of 100,000 new users based in the US. The dependent variable is the destination country with 12 possible levels: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL', 'DE', 'AU', 'NDF' (no destination found), and 'other'. The the independent variables available are age, gender, first device and browser used, how the user reached Airbnb (e.g. via Google), length of the web session, etc. (total predictors = 18, see Appendix 5.2 for full list).

2.2 Exploratory Analysis

We begin the analysis by considering which countries are often booked by users. The dataset is highly unbalanced. For example, some rarely-booked countries appear in less than 0.3% of 120,000 observations. The US is the most common destination, representing 70% of the booked destinations. This is sensible due to the US-based dataset, but at the same time making the prediction problem particularly challenging.

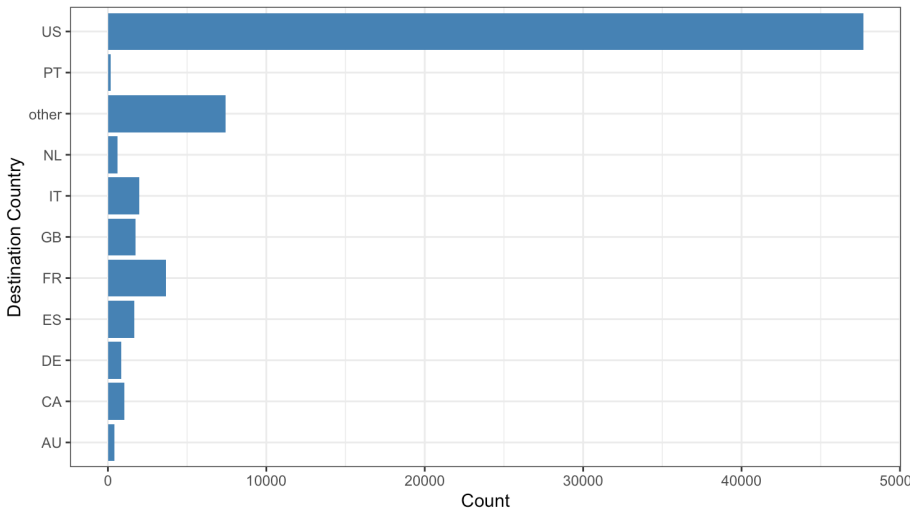

Figure 1: Counties Booked in our dataset

We inspect the Airbnb user-base, attempting to find significant differences among users that could be strong predictors for bookings. We considered the age of users and the device types they used, which are useful when deciding targeted advertisements if these features prove important. When exploring pairwise relationships between booking and factors such as age, we see no strong correlation, and the distribution of booking by age appears by-modal, peaking at 30 and 65 (see Appendix 5.1 Figure 6).

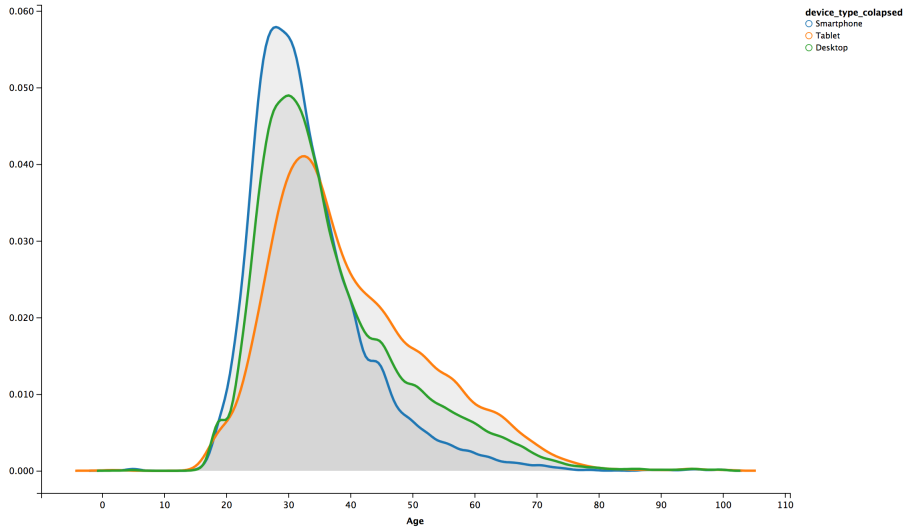

Figure 2: Age vs Devicetype

We performed clustering in order to find patterns of the distribution of user characteristics and get a spatial interpretation of what our user base look like. K-means clustering revealed 7 groups (found by scree plot, see Appendix 5.1 Figure 8).

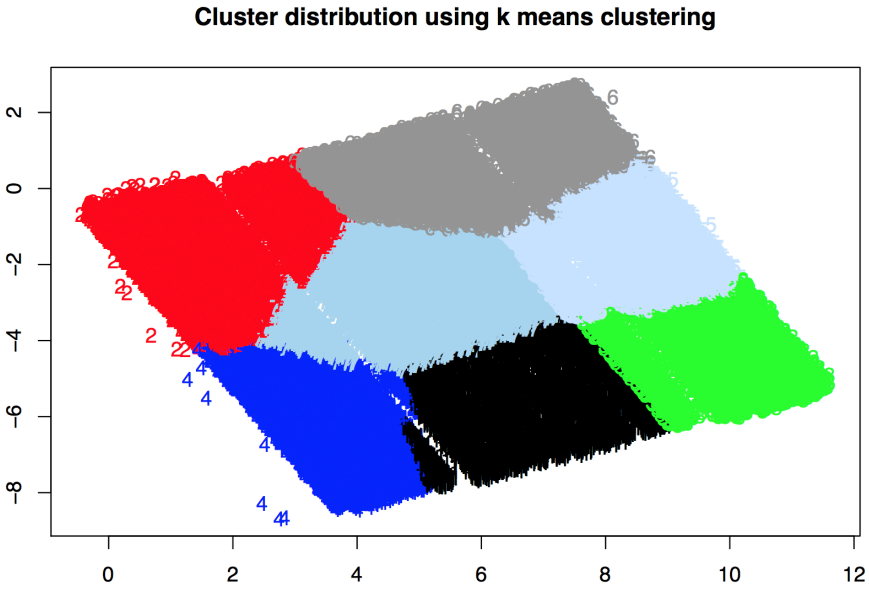

Figure 3: Clustering using all variables

Apparently, users appear to be pretty well clustered. This leads us to perform feature engineering, in order to put users into interpretable groups and identify different booking habits between them.

2.3 Feature Engineering

In addition to the overall prediction, we also want to highlight the effect of age groups, as age can tell us a lot of stories about one's purchasing power and new information (e.g., people in working age going abroad due to better financial state). Therefore, we created a new column called 'age generation' discretely distinguishing people in their 20s, 30s, 40s, and 50s.

In addition, we categorized people based on the devices they used to create their account, into 'Desktop', 'Smartphone', and 'Tablet'. Then, we further divided each device type into 'Android' and 'iPhone', 'Windows Desktop' etc.

Lastly, we created three new columns from the date the account was created: day, month, and year. This is because we believed that booking displays a temporal relationship (as the Airbnb platform progressed the number of bookings grew). Yet at the same time bookings are likely to be seasonal (e.g. more bookings around holidays) hence we expect the month to be and important contributor to our models.

2.4 Methodology

As observed in 2.2, the dataset has highly unbalanced classes. Indeed, 45.4% of all users have not booked and and 38.7% have made a booking in the US. As such, initial machine learning tools tested, i.e. Logistic Regression, CART, Random Forest and Gradient Boosted Machines, were predicting only the two biggest classes - no booking and US. To address this problem, we split the classification task into three consecutive steps:

- Step 1: Predicting whether a user will book
- Step 2: Given a booking is predicted to occur, classifying whether the country of destination is US or not
- Step 3: Predicting which country the user will book.

At each of the three steps, we apply Logistic Regression, CART, Random Forest and Gradient Boosted Machines and record the main metrics: accuracy and AUC. We obtain moderate results for all 3 steps due to the high imbalance in the dataset, e.g. for US/non-US a baseline model predicting the most common destination will have 0.7 accuracy. Therefore we fit additional models such as ensembles, experimented with balancing the dataset by up-/down-sampling, and fitting multi-class predictions to improve upon the results.

The table below summarises the results. For step 3 - Elsewhere we report the results of models predicting which country *other than the US* the user is most likely to book, and the models predict one country per user. As part of step 3, we created additional models to predict all countries (including the US) and to predict multiple countries per user. These resutls are detailed in the sections below.

3 Results

Table 1: Performance of CART, RF and GBM models across all steps

	Booking / no booking				US / non-US		Elsewhere	
Model	Accuracy	AUC	TPR	FPR	Accuracy	AUC	Accuracy	AUC
Logit	0.632	0.678	0.621	0.354	0.709	0.558	–	–
CART	0.633	0.623	0.657	0.396	0.693	0.551	0.539	0.587
RF	0.635	0.664	0.726	0.477	0.709	0.525	0.633	0.611
GBM	0.600	0.673	0.531	0.762	0.709	0.562	0.633	0.657

3.1 Step 1: Predicting whether a user will book

This section focuses on the prediction of a binary outcome, whether there is a booking made by the user.

We use Logistic Regression, CART, Random Forest (RF) and Gradient Boosted Machines (GBM) models, with cross validation. Across all models, the most important variables are age and sign-up method, device type, as well as month and day of the account creation. We interpret the time dependence as seasonality, age and device type dependence as distinct customer behaviour groups, e.g. younger travelers are more likely to book on mobile devices. All the insights align with business sense.

3.1.1 Best-performing model and extension

Judging by the AUC, the best-performing model is Logistic Regression (Logit) using a LASSO-regularizer with an out-of-sample AUC of 0.678 (Table 1). We also fit Optimal

Classification Trees, with a similar performance (accuracy 0.645 and AUC 0.65).

Given the difficulty of the task and limited features available, this is a promising result. It is very interesting to see logit as the best performing model, although it is known that GBM do not always perform the best. Overall, this robust model has 33 variables (all factor variables where encoded binary). We predict more false negatives than false positives, which is beneficial if we have a limited budget to advertise only to users are most certain will book.

Motivated by search for a better predictive performance, we combine the class predicted by the Gradient Boosted Machines with the training data-set and feed it to the Logistic Regression. We observe this ensemble configuration to provide an out-of-sample AUC performance of 0.677, which is marginally better than with GBM alone.

3.2 Step 2: Predicting whether the destination is US

In order to predict if a user will book a **US/non-US** destination we start by applying CART, Random Forest, and Boosting. Note that this is done on the subset of data which contains the users that actually booked a trip. Due to the high unbalance of the data, we tried adding a loss matrix to CART (penalizing the classification of a US trip) with different penalty terms, achieving an AUC of 0.529, which is worse than CART without modifying the loss matrix, i.e. AUC of 0.551.

In addition, we also performed Logistic Regression with a LASSO-regularizer and performed cross-validation. This gave us an out-of-sample AUC of 0.558, which is almost at par with the best-performing GBM method, which gave us an AUC of 0.562.

This part of the analysis sheds a light on interesting variable influences regarding the prediction of a non-US stay, such as: speaking Italian positively influences the probability of booking non-US, speaking Chinese negatively influences the probability, a the sign-up through an affiliate link from an email marketing campaign or a website sign-up increases probabilities. Overall, this robust model has 47 variables (all factor variables where encoded binary).

3.3 Step 3: Predicting which country the user will book

3.3.1 Prediction individual country

Now that we have a dataset that contains all booked trips to non-US destinations, we would like to know **where exactly** people would go. Similarly, we perform CART, Random Forest and Boosted Trees to predict specific destinations. For Step 3, the AUC is calculated using the multi-classification ROC, while TPR and FPR are unavailable. We also experiment with ensemble methods and balancing the dataset by up-/down-sampling. To guide our business decision, we use the list of countries a person is likely to go to.

Out of all the models, GBM performs the best, with an out-of-sample AUC of 0.657. From the feature importance matrix, we can also see that gender plays an important role, and that the channel and device used to create their account are also important (Figure 3).

Before going deeper into our results, we have come to realize that if Airbnb is going to

recommend travel destinations to its users, it should not recommend only one destination, but a list of possible countries. In this way, Airbnb not only can give users a wider range of options, but can potentially increase their click-through rate. Therefore, our final model became a multi-class prediction using our best-performing method and a threshold probability which determines which countries to include in a recommendation.

3.3.2 Multi-class prediction

In this prediction, we used GBM, which is our best performing model so far, and we set our probability threshold to be 5%. This means our model will include this country if it is at least 5% confident that the user will book it (naturally the first country will have a probability of above 50% and for the other countries it decreases significantly). The majority of the nodes includes 2 or 3 countries (Figure 4). This allows us to reach an accuracy score of 87%, i.e. the right country of destination is included in our prediction in 87 out of 100 cases. Varying the 5% threshold, the managers will be able to get the level of accuracy they deem optimal.

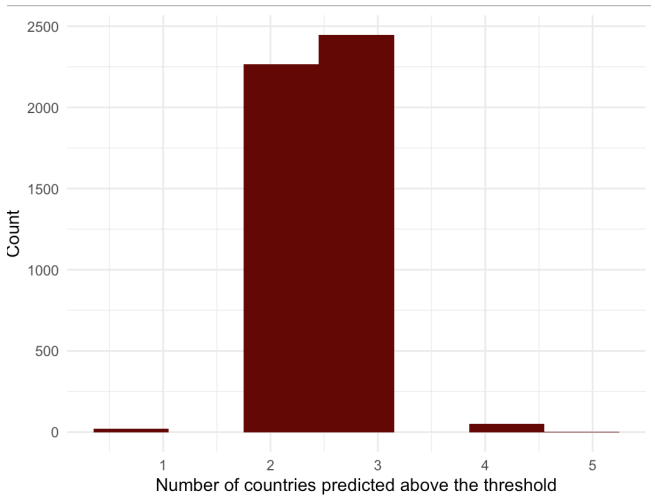

Figure 4: Histogram of number of countries predicted per user

As we can see, the number of countries in each multi-class prediction is two or three in the majority of cases. Thus, it appears feasible for Airbnb to recommend them to corresponding users, be it on their website or via ads. Moreover, along with the country predictions, we obtain a probability of how confident the model is in making such predictions.

4 Managerial Implications

The goal of our project is to help Airbnb predict the first booking of new users. This is valuable for recommending travel destinations and running targeted ads. Such a campaign is likely to increase the speed and chances of a first booking. Quickly capturing value from customers is key in stabilizing revenue streams and binding customers by offering an effortless first booking experience. The way we utilized the data and approached our analysis gives us an opportunity to look at both *whether* a new user is likely to make a

booking, so when to show ads, and *where* the user is most likely to go, so what country to recommend. At the same time, information about new users is sparse and we have to acknowledge this in our predictions. Thanks to our design of recommendations on websites and in targeted ads, we can make robuster promotion decisions.

4.1 When to show ads

During the first stage of our analysis, we predict whether there is a booking made. Given by the feature importance matrix, we see that sign-up flow is one of the most critical aspects of the user decision. This result also matches our business intuition because signing up is the first encounter that a user has with the product. The design of it and the way you integrate advertisement into it can have a major impact on how users make a decision. Our boosted tree model indicates that for sign-up flows 3 and 10, the user is most likely to book. These indices correspond to specific websites that users came from to sign up with Airbnb. Therefore–on Airbnb's end–they can de-anonymize this information and investigate how these sites capture users that are likely to book.

In addition, we also found that Facebook referrals increased the likelihood of actually making a booking. So Airbnb can investigate whether investing in Facebook advertising is more effective than other channels. In addition, working with influencers on Facebook's websites (including Instagram) might additionally boost the results of acquiring users that will also book.

4.2 What to recommend

At the final stage our analysis, we managed to help answer the key objective of our project: where would first-time users book their trip. Once we can predict possible destinations of each user, we can recommend 2-3 places using the multi-class model. Eventually, the interface that users see could look as presented in Figure 5. If the model predicts that the user is most likely to book a trip in the US, the background picture will be set as some destination in the US (e.g. Oregon). Then it would show other three travel destination countries in tabs, in order of their probabilities (e.g. France, Italy, and Spain).

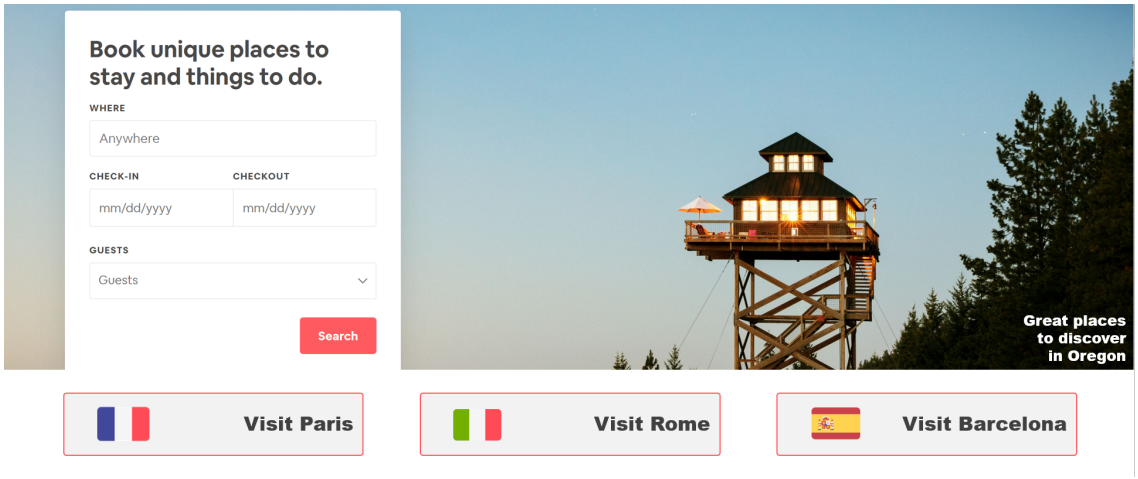

Figure 5: Sample interface of Airbnb using our model

The probabilities obtained with the multi-class model provide us with a range of insights. Thus, given budget constraints, we can show the ads we are more sure about. A composition of ads may depict different countries to the user, proportional to the probabilities obtained. Giving a purposefully designed menu of options guides the user and gives us the possibility to not only rely on a single point prediction.

5 Appendix

5.1 Figures from Exploratory Analysis

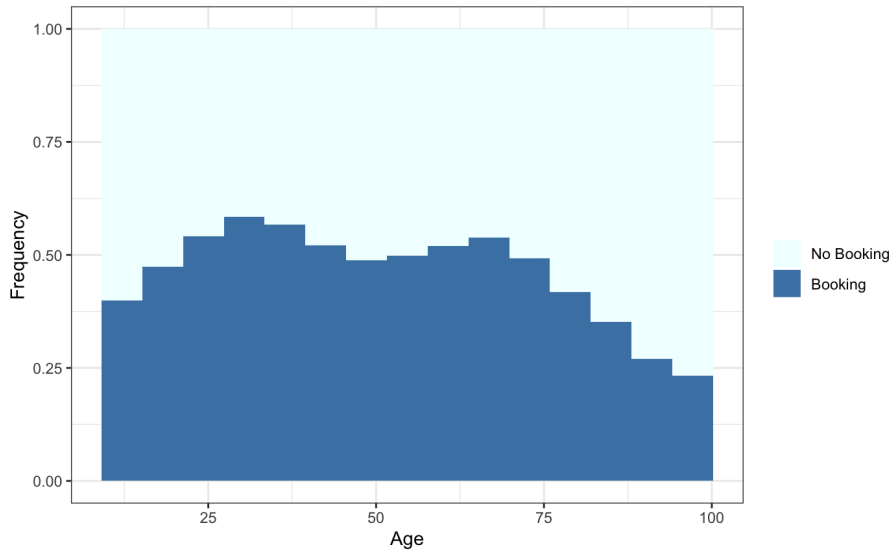

Figure 6: Booking rate vs age

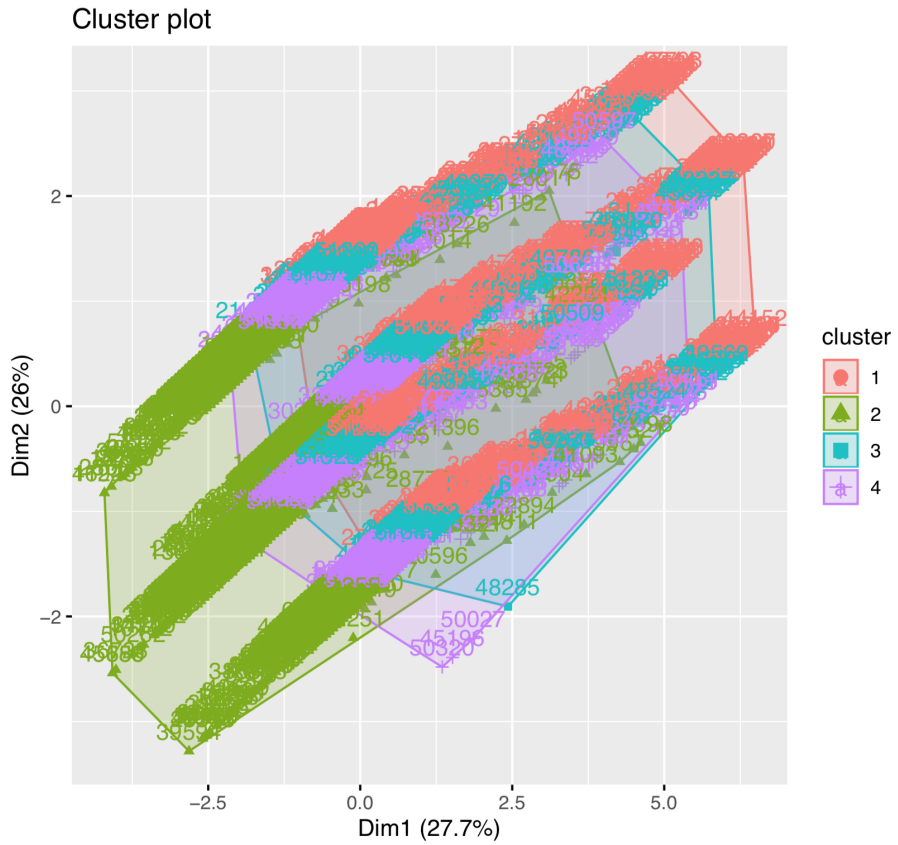

Figure 7: Clustering using gender, age, and language

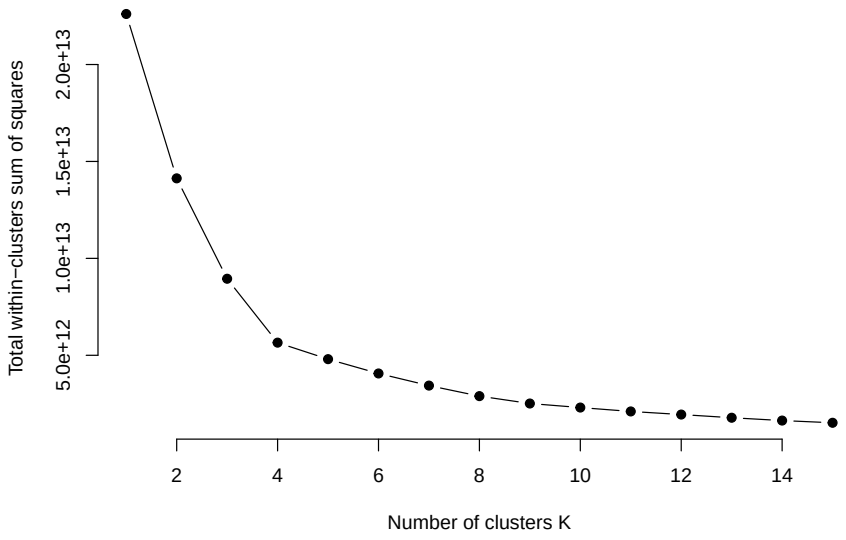

Figure 8: Selecting k for clustering

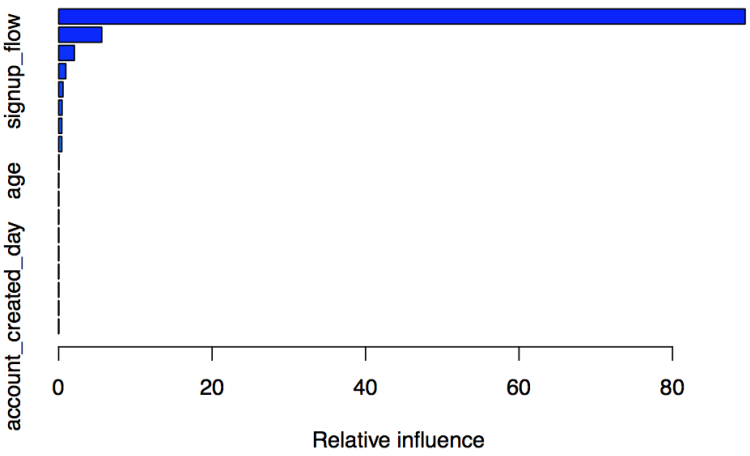

Figure 9: Boosted Trees Feature Importance for Booking/No booking Prediction

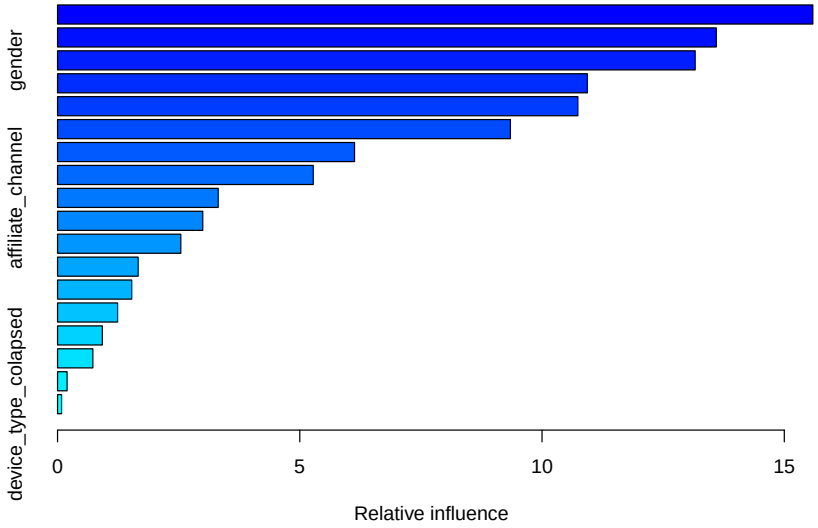

Figure 10: Boosted Trees Feature Importance for Multi-class Prediction

5.2 Independent Variables

- Gender
- Sign-up method
- Sign-up flow
- Sign-up app
- Language
- Affiliate channel

- Affiliate provider
- First affiliate tracked
- First device type
- First browser
- Age
- Age generation (engineered)
- Age quantile (engineered)
- Millennials (engineered)
- Device type (engineered)
- Year in which the account was created (engineered)
- Month in which the account was created (engineered)
- Day in which the account was created (engineered)

5.3 Non-Zero Variables of Logistic Regressions with LASSO-regularizer

Listing 1: Non-Zero LASSO-Logit Coefficients for Booking/No-Booking

(Intercept)	3.474866e+01
gender-unknown-	3.388997e-01
genderOTHER	2.088883e-02
age	-3.842107e-03
signup_methodfacebook	-1.000815e+00
signup_flow1	-1.142483e+00
signup_flow2	1.121943e-01
signup_flow3	6.532696e-01
signup_flow23	-5.664504e-03
languageen	3.485090e-01
languageit	-8.707444e-02
affiliate_channelcontent	-1.032063e+00
affiliate_channelother	-1.463688e-01
affiliate_channelremarketing	-1.690925e-02
affiliate_providerfacebook-open-graph	-2.729912e-01
affiliate_providermeetup	-2.361923e-01
affiliate_provideryahoo	-6.650859e-04
first_affiliate_trackedomg	-7.901567e-02
first_affiliate_trackedproduct	-6.430732e-03
first_affiliate_trackedtracked-other	-6.805537e-02
first_affiliate_trackeduntracked	1.083573e-01
signup_appMoweb	1.723651e-01
signup_appWeb	5.097542e-02
first_device_typeMac Desktop	1.631242e-01
first_device_typeOther/Unknown	-1.583290e-01
first_browserChrome	1.559980e-01
first_browserFirefox	1.318452e-01
age_generationNon-millennials	-1.885206e-01
age_quantilyonger than 20	-3.313960e-01
device_type_colapsedOther	-4.603679e-05
device_type_colapsedSmartphone	-1.076829e-01
account_created_year	-1.709644e-02
account_created_month	-3.950223e-03

Listing 2: Non-Zero LASSO-Logit Coefficients for US/Non-US

(Intercept)	-1.000020792
genderMALE	0.022383065
genderOTHER	0.151771636
signup_flow3	0.002160220
signup_flow8	0.637648734
signup_flow16	1.203597161
signup_flow21	0.258554376
signup_flow24	-0.235258605
languageel	0.011474755
languageen	-0.059937117
languageit	0.253245584
languagezh	-0.288024875
affiliate_channelcontent	0.310250369
affiliate_channelother	-0.468775151
affiliate_channelremarketing	0.099109392
affiliate_channelsem-non-brand	0.158143367
affiliate_providercraigslist	-0.110030168
affiliate_provideremail-marketing	0.279313441
affiliate_providerfacebook	0.214285951
affiliate_providerfacebook-open-graph	0.530745726
affiliate_providerother	-0.113422092
first_affiliate_trackedtracked-other	-0.100084505
first_affiliate_trackeduntracked	0.001151430
signup_appMoweb	-0.106574911
signup_appWeb	0.129671622
first_device_typeiPad	0.069665820
first_browserAvant Browser	0.074721793
first_browserChrome	-0.003830385
first_browserKindle Browser	0.024883982
first_browserSafari	0.014001317
first_browserSeaMonkey	1.155216892
age_generationNon-millennials	0.153366108
age_quantil30-40	0.026402000
age_quantil40-50	-0.066986625
age_quantilolder than 60	0.009149131
age_quantilyonger than 20	0.153486534
millenials1	-0.032225655
device_type_colapsedSmartphone	-0.058840094
account_created_year2011	0.051324868
account_created_year2012	0.057329185
account_created_year2014	-0.069648626
account_created_month2	0.033896597
account_created_month5	0.083752703
account_created_month6	0.102267065
account_created_month7	-0.004267462
account_created_month10	-0.084534207
account_created_month12	-0.094569992