AirBnB New User Bookings - Capstone Project Report

Vijay Gopalakrishnan

2016

Problem Definition

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. Airbnb's registered users engage on the site and explore the featured accomodations across destinations before making a booking. Newly registered users do not have any booking history for Airbnb to personalize recommendations. However, by accurately predicting whether a new user will book their first travel experience based on their early activity, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand. This task is to predict the likelihood of a new user to complete a booking.

Data about the problem

This problem was posted by Airbnb on Kaggle as an open competition –> https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings). For a list of US users, Airbnb provided the following data.

- Demographic and user account data
  - id: user id,date_account_created, timestamp_first_active, date_first_booking, gender, age, signup_method, signup_flow, language, affiliate_channel, affiliate_provider, first_affiliate_tracked, signup_app, first_device_type, first_browser, country_destination
- Web session log data for users
  - action, action_type, action_detail, device_type, secs_elapsed
- Summary statistics on destination countries, user's age group and gender.

Data is provided seperately for the Test and Training groups. The Test and Training data sets have been separated based on the user's first activity. Training data set has all the users with first activity date prior to July 2014 while the test data set has all those starting July 2014. The target variable, destination country, has been provided for the Training data set to build the prediction model. There are 12 possible destination countries for the users that complete their first booking: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'.

Solution Outline

The problem to predict one of the 12 destination outcomes is a classification one. However, I have considered to simplify the problem into a basic logistic regression to predict whether an user will complete their first booking or not (yes/no) and further refine and fine tune the predictiveness of this model.

About the Data

Here is the complete description of the Airbnb data files.

File descriptions

- train_users.csv - the training set of users
- test_users.csv - the test set of users
    - id: user id
    - date_account_created: the date of account creation
    - timestamp_first_active: timestamp of the first activity, note that it can be earlier than + date_account_created or date_first_booking because a user can search before signing up
    - date_first_booking: date of first booking
    - gender with values male, female, other, unknown
    - age is a continuous variable
    - signup_method with values basic, facebook and google
    - signup_flow: the page a user came to signup up from with continuous values
    - language: international language preference
    - affiliate_channel: what kind of paid marketing with values like direct, sem-brand, sem-non brand, api, seo etc.
    - affiliate_provider: where the marketing is e.g. direct, google, bing, yahoo craigslist, other etc.
    - first_affiliate_tracked: whats the first marketing the user interacted with before the signing up
    - signup_app with values Android, iOS, Moweb, Web
    - first_device_type like Desktop (Mac, Windows, other), ipad, iphone, Android phone etc
    - first_browser values like Chrome, Safari, Firefox, IE, etc
    - country_destination: this is the target variable to predict
- sessions.csv - web sessions log for users
    - user_id: to be joined with the column ‘id’ in users table
    - action
    - action_type
    - action_detail
    - device_type
    - secs_elapsed
- countries.csv - summary statistics of destination countries in this dataset and their locations
- age_gender_bkts.csv - summary statistics of users’ age group, gender, country of destination
- sample_submission.csv - correct format for submitting your predictions

Exploratory Analysis of Airbnb data

First, let’s us import all the data files into R dataframes

Let’s take a look at the data `structure` and samples of the data.

The `countries` dataset (10 observations, 7 variables) provides geographical coordinates of each country and w.r.to US (since the users are based in USA). The language levenshtein distance provides the closeness (smaller the number the closer to US English) of the destination country language with respect to US English.

The `sessions` dataset (10.6M rows, 6 variables) provides all the user interaction events or actions. Note that there are about 360 different user actions and 156 different action details. This may need to be aggregated into a few logical groups to capture aggregated user interaction times for a given user id. Otherwise, we will end up with one too many session interaction data features for prediction thus posing system limitations to run the model and also complicating the model.

The `age_gender_bckts` (420 rows, 5 variables) dataset consists of demographic summary for each country.

The `train_users` (213K rows, 16 variables) is the training dataset with the destination country while `test_users` (62K rows, 15 variables) is a similar dataset without the destination country. The session information could be aggregated at each user level in the training dataset. The destination demographics could also be attached to each destination country in the training dataset.

Training dataset has NA values for Age and Gender

From the above, we observe that the training data set 'train_users' has NA or Unknown values for Age and Gender. From the user sign up flow for Airbnb, Gender is not mandatory. Further more, in the profile setting, the options for Gender are only 'Male', 'Female' and 'Other'. Let us dig in further.

```
levels(train_users$gender)
```

```
## [1] "-unknown-" "FEMALE"    "MALE"      "OTHER"
```

```
summary(train_users$age)
```

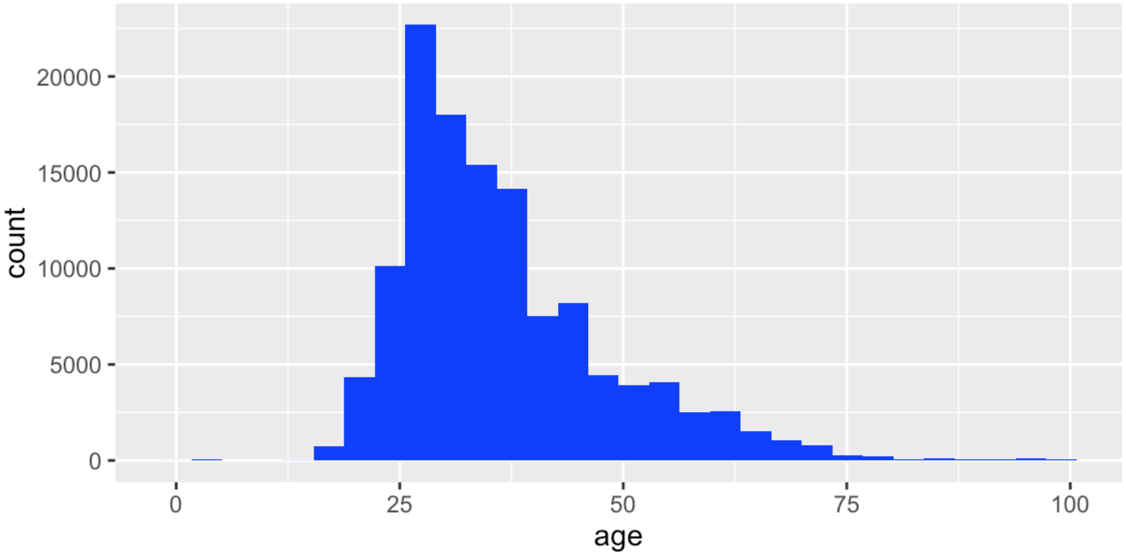
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    1.00   28.00   34.00   49.67   43.00 2014.00   87990
```

There are some erroneous Age values. We need to clean up the data to eliminate them.

```
train_users_new <- train_users[which(train_users$age <= 100),]
```

Let us plot some histograms of the variables

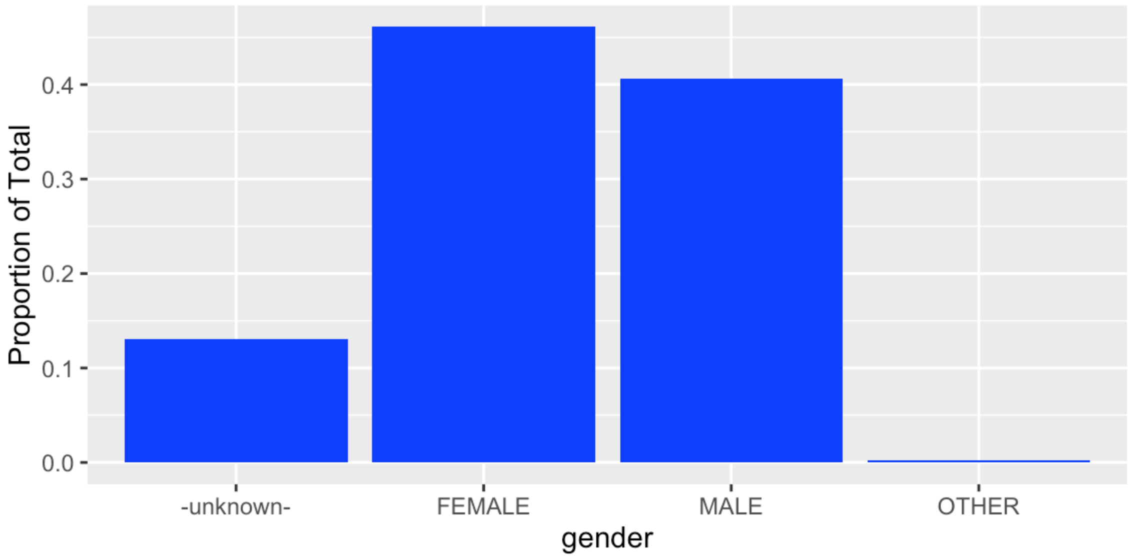
```
library(ggplot2)
ggplot(train_users_new, aes(x=age)) +
  geom_histogram(fill = 'blue')
```



Age of majority of the users range between 25-50.

Distribution of Users by Gender

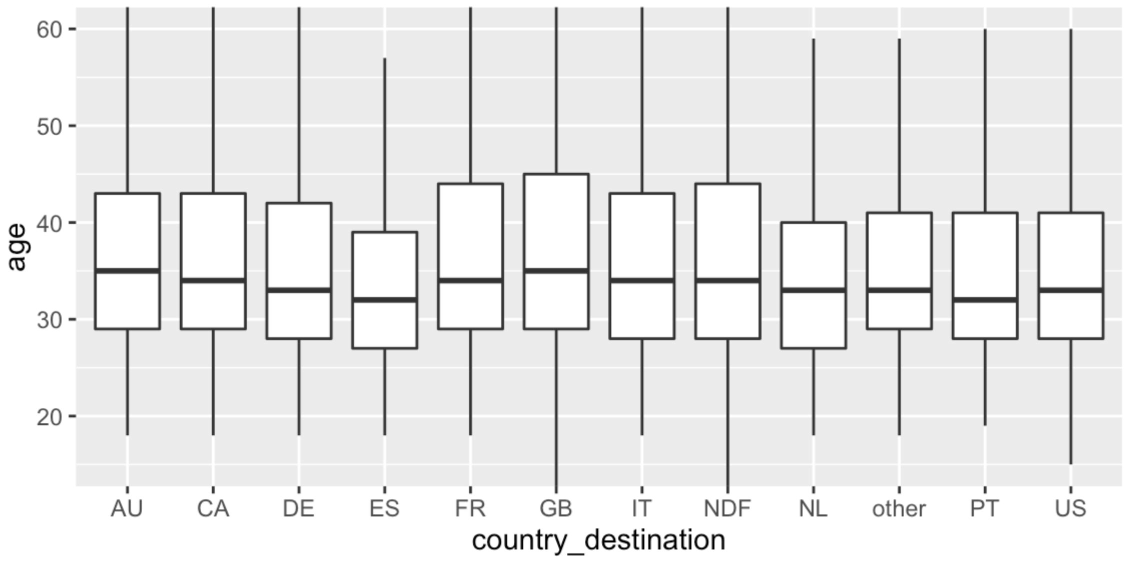
```
# proportion of gender
ggplot(train_users_new, aes(x=gender)) +
  geom_bar(fill = "blue",aes(y=..count../sum(..count..))) + labs(y = "Proportion of Tota
l")
```



Both Male and Female are similar in proportion (~ 40-45%) and ~13% have not specified any Gender.

Let us plot Age by Country Destination using a Box Plot option

```
ggplot(train_users_new, aes(x=country_destination, y=age, fill=age)) + geom_boxplot(outl
ier.shape = NA) +
  coord_cartesian(ylim = c(15, 60))
```



There is no destination country that is particularly attractive to a certain age range.

Create a book (Yes/No) flag

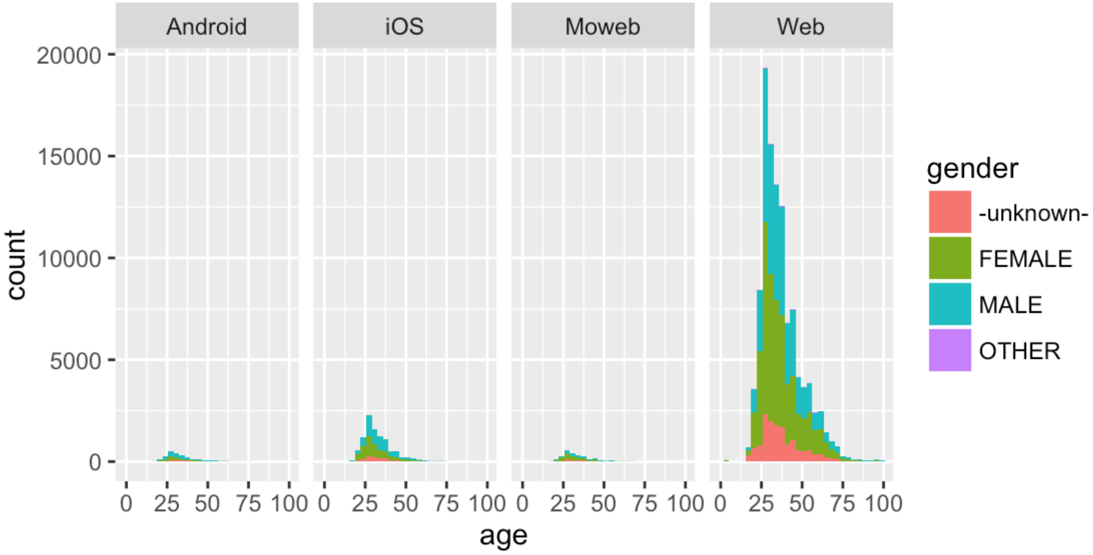
Let us create a flag to indicate whether the user completed a booking or not.

```
train_users_new$book_yorn[train_users_new$country_destination == "NDF"] <- "0"
train_users_new$book_yorn[train_users_new$country_destination != "NDF"] <- "1"
train_users_new$book_yorn <- factor(train_users_new$book_yorn)
```

More plots.

```
ggplot(train_users_new, aes(x=age, fill = gender))+
  geom_histogram() + facet_grid(.~signup_app)
```

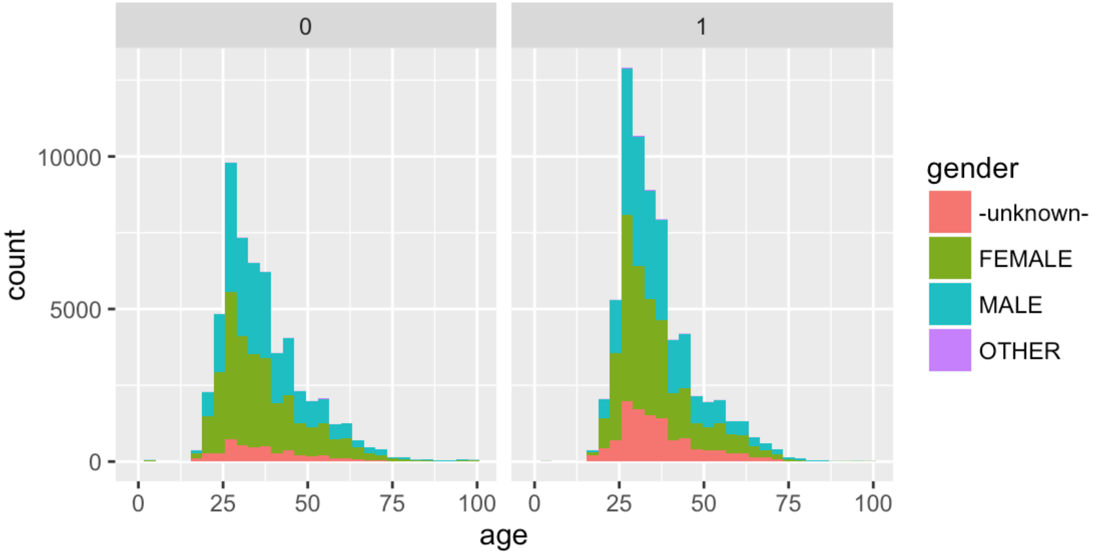
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Most of the users are web based with a few iOS mobile users. For the web users, it will be worthwhile to check the browser type.

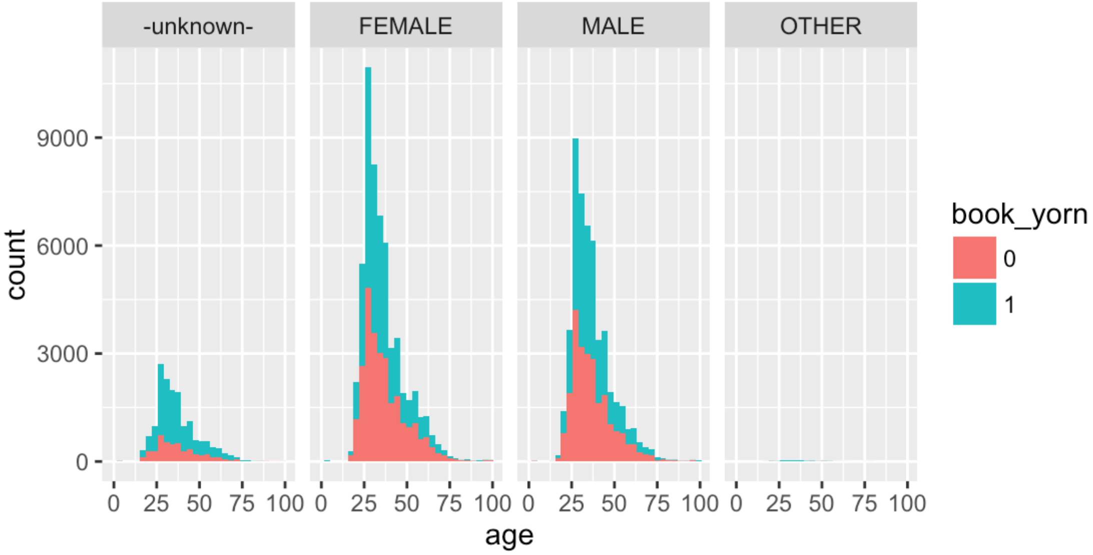
```
ggplot(train_users_new, aes(x=age, fill = gender))+
  geom_histogram() + facet_grid(.~book_yorn)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



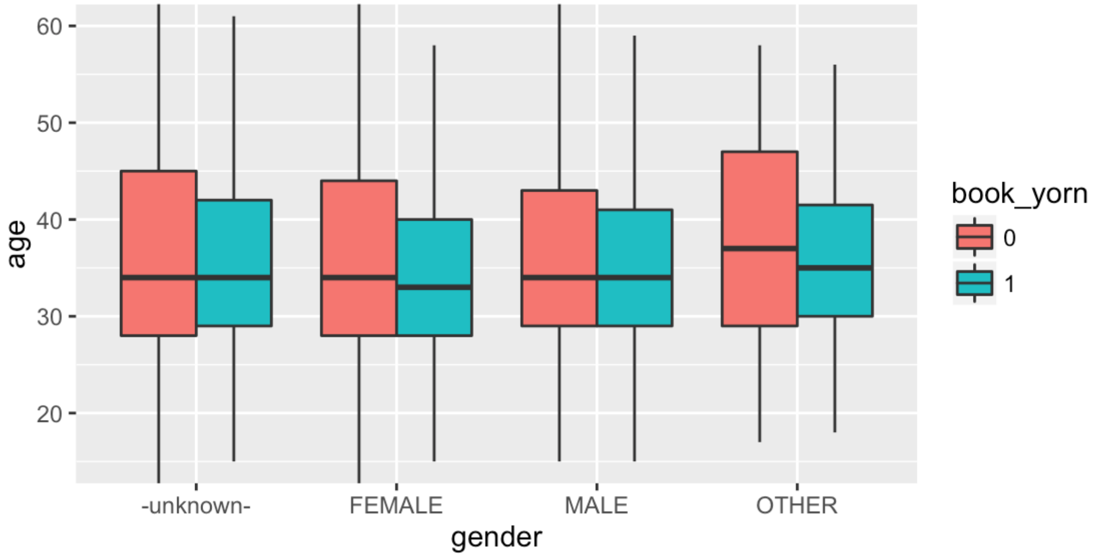
```
ggplot(train_users_new, aes(x=age, fill = book_yorn))+
  geom_histogram() + facet_grid(.~gender)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Irrespective of Gender and booking outcome, the age of the users fall between 30 and 40 with similar average age. (Refer Chart below)

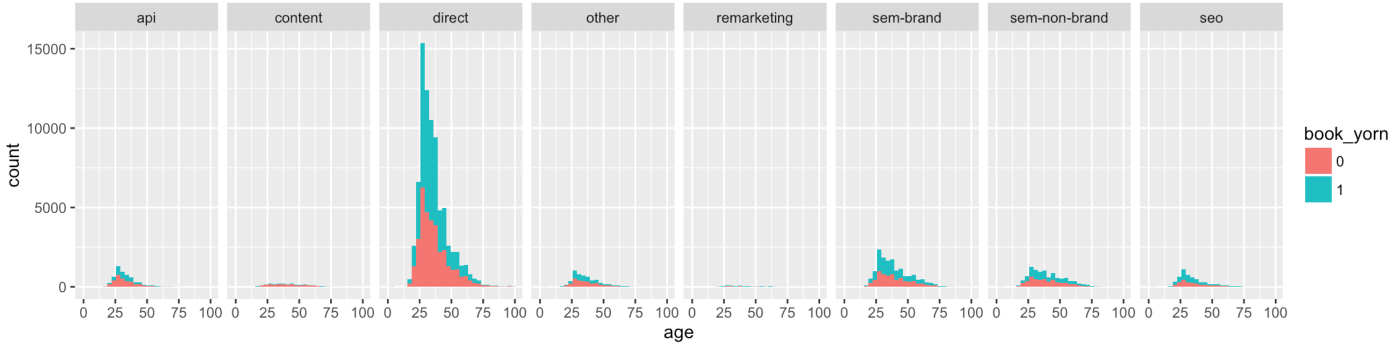
```
ggplot(train_users_new, aes(x=gender, y=age, fill=book_yorn)) + geom_boxplot(outlier.sha
pe = NA) +
  coord_cartesian(ylim = c(15, 60))
```



Airbnb users under consideration are primarily through direct affliate channel and cannot be attributed to specific few marketing channels. (Refer Chart below)

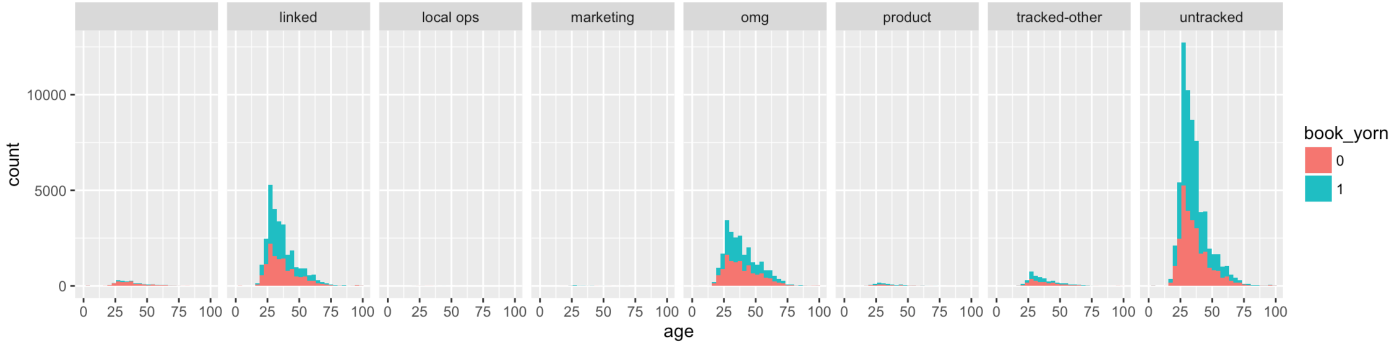
```
ggplot(train_users_new, aes(x=age, fill = book_yorn))+
  geom_histogram() + facet_grid(.~affiliate_channel)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(train_users_new, aes(x=age, fill = book_yorn))+
  geom_histogram() + facet_grid(.~first_affiliate_tracked)
```

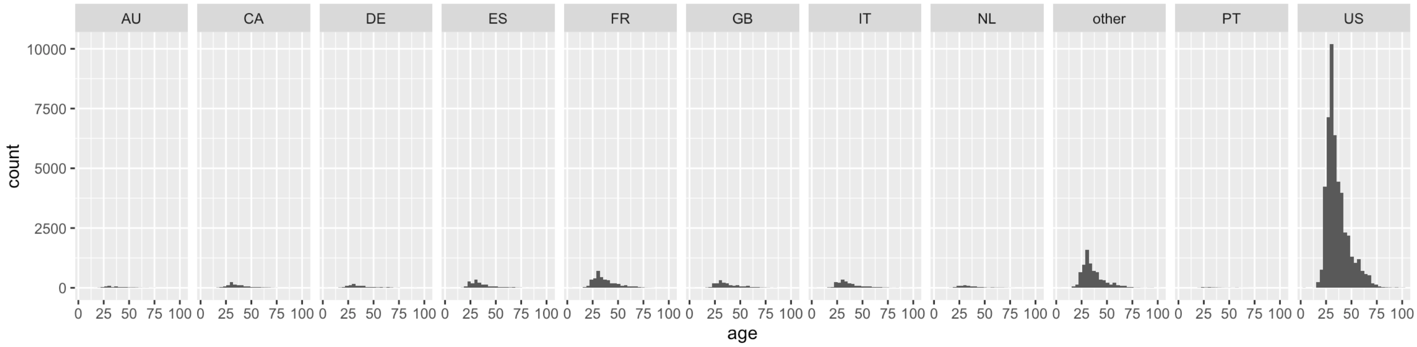
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



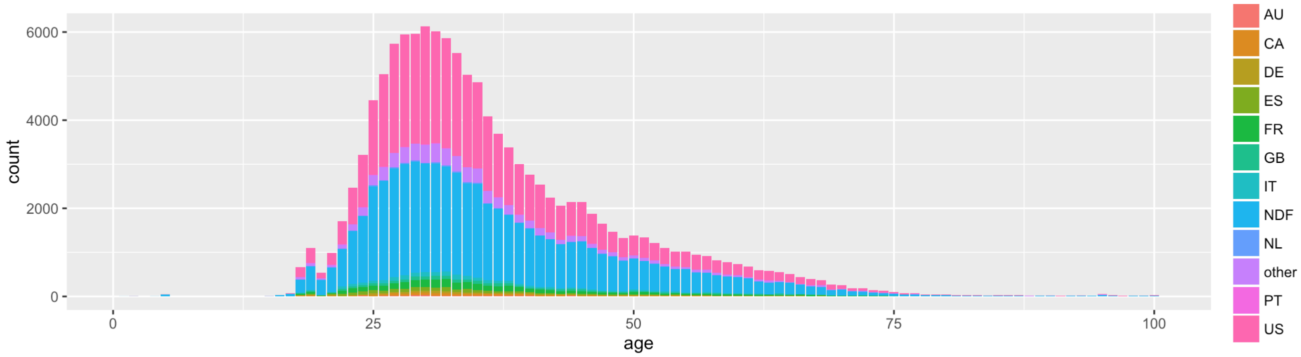
For the given data set, the booked destination country is skewed towards US.

```
ggplot(subset(train_users_new,country_destination %in% c("AU", "CA", "US", "DE", "ES",
"FR", "GB", "IT", "NL", "other", "PT" )), aes(x=age))+
  geom_histogram() + facet_grid(.~country_destination)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

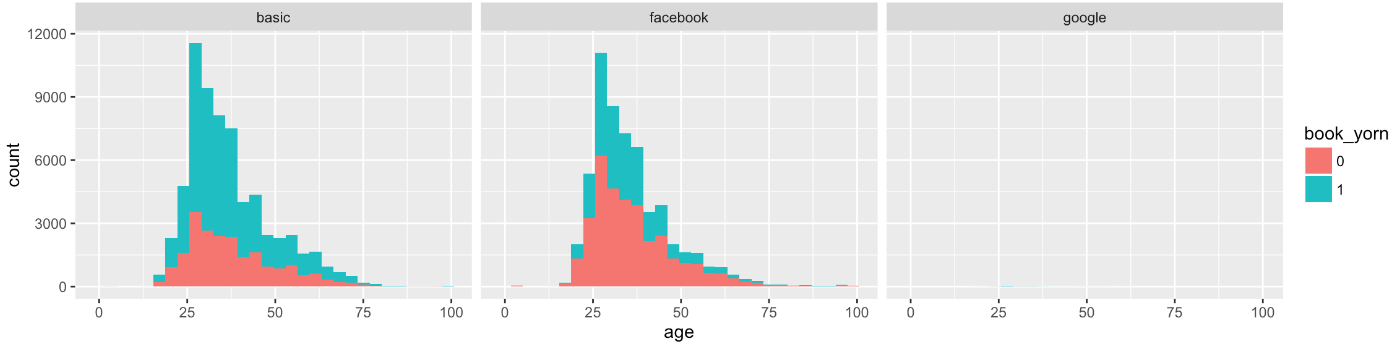


```
ggplot(train_users_new, aes(x=age, fill = country_destination ))+
  geom_bar()
```



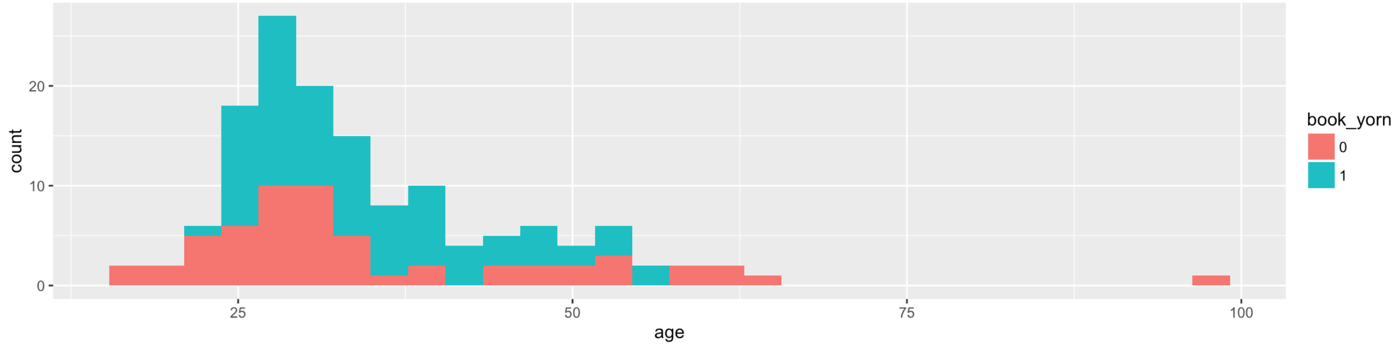
```
ggplot(train_users_new, aes(x=age, fill = book_yorn))+
  geom_histogram() + facet_grid(.~signup_method)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(subset(train_users_new, signup_method %in% c("Google", "google")), aes(x=age, fill = book_yorn))+ geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Summary

From the above charts, following observations can be made.

- Most of the user sign ups have been through Web with ios at a distant second. Browser type could be a strong predictor variable
- Difference in age distribution between those that book and not book is not quite apparent. However, Age and Gender could be highly predictive of booking outcome and should be explored by including them in the model
- From the user sign up flow for Airbnb, specifying Gender is not mandatory. This explains why many users have Unknown Gender. Furthermore, in the profile setting, the options for Gender are ‘Male’, ‘Female’ and ‘Other’
- Direct channel seems to be the most efficient in originating users. Channels may be a key factor in attracting a certain age group of users. Users originated through API channel are predominantly under 50 years of age compared to say direct or SEM channels
- Most of the booking destinations is within US with ‘other’ at a distant second. Low volume for the non US countries in the training dataset may make the prediction of the non US destination challenging
- The chart on signup method suggests that quite a number of user signups through Facebook marketing comapred to Google. It is surprising that Google generates very few sign ups. It will be interesting to see how the signup method determines first booking (yes/no). Sign up method is clearly an interesting predictive feature to include in our model.

Can the browser type can be a good indicator of whether the user books or not?? Let us check.

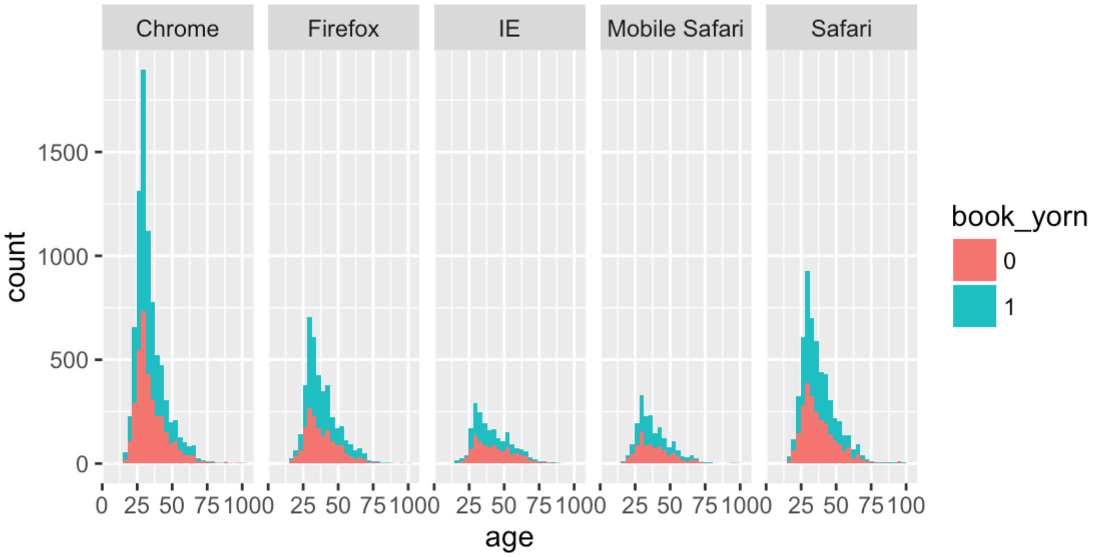
```
summary(train_users_new$first_browser)
table(train_users_new$first_browser)
```

The top 5 browsers are Chrome, Firefox, IE, Mobile Safari and Safari. Let us create a plot for those by gender and age.

```
library(dplyr)
#filter(countries, destination_language == "eng", language_levenshtein_distance == 0.0)
data_browser <- filter(train_users_new, first_browser == c("Chrome","Firefox", "IE", "Mo
bile Safari", "Safari"))
#head(data_browser)
```

```
ggplot(data_browser, aes(x=age, fill = book_yorn)) +
  geom_histogram() + facet_grid(.~first_browser)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Chrome and Safari are the most used browser types with the most traffic. There is no explicit evidence that any particular browser makes a difference to the booking outcome. However, age and browser type can influence the booking outcome and we should consider this in our model as a predictor variable.

Let us try a basic regression model with all the features so far.

```
airbnb_fb = glm(book_yorn ~ gender+age+signup_method+signup_flow+language+affiliate_chan
nel+affiliate_provider+first_affiliate_tracked+signup_app+first_device_type+first_browse
r, data = train_new_2, family = binomial())
summary(airbnb_fb)
```

The above identifies a few variables as significant predictors. Adding more feasibly predictive variables results in non convergence due to data leakage and the model does not yield any meaningful result. So, we can start with all the significant features to get a workable model that we can improve upon.

```
airbnb_fb = glm(book_yorn ~ gender+age+signup_method+signup_flow+affiliate_channel+first
_affiliate_tracked+signup_app+first_device_type, data = train_new_2, family =
binomial())
summary(airbnb_fb)
```

```
## 
## Call:
## glm(formula = book_yorn ~ gender + age + signup_method + signup_flow + 
##     affiliate_channel + first_affiliate_tracked + signup_app + 
##     first_device_type, family = binomial(), data = train_new_2)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.1182  -1.0841   0.7597   0.9757   2.6601  
## 
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -1.255926   0.152623  -8.229  < 2e-16
## genderFEMALE                         -0.382840   0.020668 -18.523  < 2e-16
## genderMALE                           -0.344722   0.021054 -16.373  < 2e-16
## genderOTHER                           0.116154   0.151708   0.766  0.44389
## age                                  -0.008295   0.000527 -15.739  < 2e-16
## signup_methodfacebook                -1.003652   0.012957 -77.462  < 2e-16
## signup_methodgoogle                   0.509114   0.182691   2.787  0.00532
## signup_flow                           0.029125   0.002543  11.454  < 2e-16
## affiliate_channelcontent             -1.602908   0.072660 -22.060  < 2e-16
## affiliate_channeldirect              -0.375106   0.041980  -8.935  < 2e-16
## affiliate_channelother               -0.525924   0.053237  -9.879  < 2e-16
## affiliate_channelremarketing         -0.586392   0.095621  -6.132 8.65e-10
## affiliate_channelsem-brand           -0.347945   0.047667  -7.300 2.89e-13
## affiliate_channelsem-non-brand       -0.391937   0.049478  -7.921 2.35e-15
## affiliate_channelseo                 -0.312512   0.051623  -6.054 1.42e-09
## first_affiliate_trackedlinked         1.379312   0.096742  14.258  < 2e-16
## first_affiliate_trackedlocal ops      0.640710   0.478572   1.339  0.18064
## first_affiliate_trackedmarketing      1.768388   0.232636   7.602 2.93e-14
## first_affiliate_trackedomg            1.228647   0.097699  12.576  < 2e-16
## first_affiliate_trackedproduct        1.109735   0.120739   9.191  < 2e-16
## first_affiliate_trackedtracked-other  1.222112   0.102606  11.911  < 2e-16
## first_affiliate_trackeduntracked      1.530192   0.095656  15.997  < 2e-16
## signup_appiOS                         0.218002   0.078149   2.790  0.00528
## signup_appMoweb                       0.682477   0.083311   8.192 2.57e-16
## signup_appWeb                         1.050876   0.093660  11.220  < 2e-16
## first_device_typeAndroid Tablet       0.221756   0.106357   2.085  0.03707
## first_device_typeDesktop (Other)      0.554801   0.103533   5.359 8.38e-08
## first_device_typeiPad                 0.404434   0.072612   5.570 2.55e-08
## first_device_typeiPhone               0.177212   0.072804   2.434  0.01493
## first_device_typeMac Desktop          0.614264   0.069156   8.882  < 2e-16
## first_device_typeOther/Unknown        0.269986   0.082808   3.260  0.00111
## first_device_typeSmartPhone (Other)  -0.084293   0.341820  -0.247  0.80522
## first_device_typeWindows Desktop      0.445162   0.069310   6.423 1.34e-10
##                                         
## (Intercept)                          ***
## genderFEMALE                         ***
## genderMALE                           ***
## genderOTHER                             
## age                                  ***
## signup_methodfacebook                ***
## signup_methodgoogle                  ** 
## signup_flow                          ***
## affiliate_channelcontent             ***
## affiliate_channeldirect              ***
## affiliate_channelother               ***
## affiliate_channelremarketing         ***
## affiliate_channelsem-brand           ***
## affiliate_channelsem-non-brand       ***
## affiliate_channelseo                 ***
## first_affiliate_trackedlinked        ***
## first_affiliate_trackedlocal ops
## first_affiliate_trackedmarketing     ***
## first_affiliate_trackedomg           ***
## first_affiliate_trackedproduct       ***
## first_affiliate_trackedtracked-other ***
## first_affiliate_trackeduntracked     ***
## signup_appiOS                        **
## signup_appMoweb                      ***
## signup_appWeb                        ***
## first_device_typeAndroid Tablet      *
## first_device_typeDesktop (Other)     ***
## first_device_typeiPad                ***
## first_device_typeiPhone              *
## first_device_typeMac Desktop         ***
## first_device_typeOther/Unknown       **
## first_device_typeSmartPhone (Other)
## first_device_typeWindows Desktop     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 169635  on 123115  degrees of freedom
## Residual deviance: 157161  on 123083  degrees of freedom
## AIC: 157227
##
## Number of Fisher Scoring iterations: 4
```

Baseline Accuracy

If we predict everybody is going to book, based on the training dataset, our baseline accuracy will be 54.6%. We need to improve this accuracy with our logistic regression model.

Let us look at variables with negative and positive correlation relation with the probability of booking outcome that we want to predict. It should be noted from our charts earlier that some of the sub segments have very little or miniscule volume and the following correlations may not hold strong on a test population.

Variables that are negatively correlated to p(booking = 1) are Gender, age and affiliate channel (the kind of paid marketing). Only when the Gender is Other which is a very small proportion, the user is more likely to complete a first booking. It is likely that more of the users that do not book would have specified Male/Female gender and the age.

Variables that are positively correlated to p(booking =1) are first affiliate tracked (first marketing the user interacted with before the signing up), signup_flow (page the user came to sign up from), signup app and first device type. When the user sign up through web/web on mobile, use desktop (mac/windows/other) as first device, or come through affiliate marketing interaction, the likelihood of them booking increases.

However, signup method is mixed in correlation with the booking outcome. Signup method Google is positively correlated whereas Facebook is negatively correlated.

```
predict_train_fb <- predict(airbnb_fb, type ="response")
# statistical summary of our predictions
summary(predict_train_fb)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.02907 0.42110 0.57550 0.54590 0.68430 0.89390
```

```
tapply(predict_train_fb, train_new_2$book_yorn, mean)
```

```
##         0         1 
## 0.4925545 0.5903013
```

If the model works, then we should see higher probability for predicting actual bookings. This is indeed true from the above average prediction for each of the true booking outcomes.

Sensitivity, Specificity and ROC Curve of the above Logistic Regression Model

Let us use a threshold value of 0.6 to assess the prediction of the model.

```
table(train_new_2$book_yorn, predict_train_fb > 0.6)
```

```
##    
##     FALSE  TRUE
##   0 37830 18075
##   1 27672 39539
```

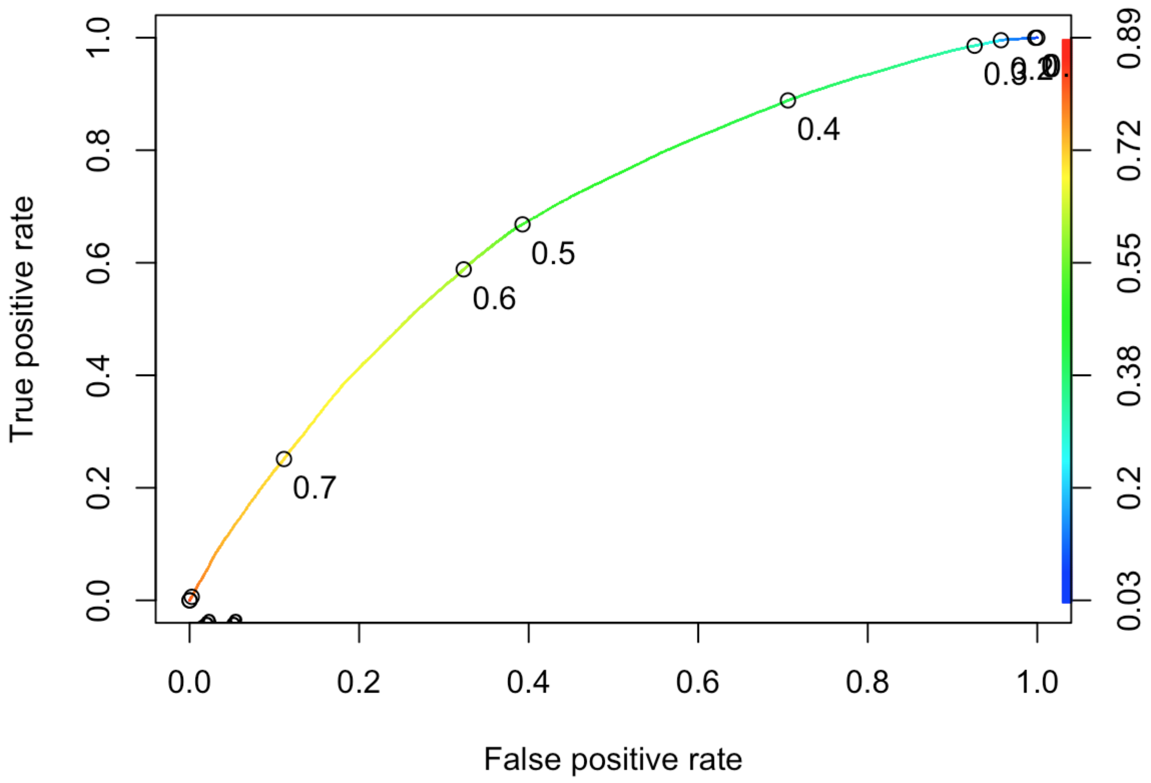
Using the above, we get Sensitivity or True positive rate = 0.5993 and Specificity or True Negative rate = 0.6637. By decreasing the threshold to 0.5, sensitivity increases to 0.6996 and specificity decreases to 0.5767

ROC Curve

ROC curve helps us to determine which threshold value to pick to mazimize True Positive rate while keeping the False positive rate to the minimum.

```
ROCRpred_fb <- prediction(predict_train_fb, train_new_2$book_yorn)
ROCRperf_fb <- performance(ROCRpred_fb, "tpr", "fpr")
```

```
plot(ROCRperf_fb, colorize = TRUE)
plot(ROCRperf_fb, colorize = TRUE, print.cutoffs.at = seq(0,1,0.1), text.adj = c(-0.2,
1.7))
```



From the above ROC curve, a threshold value of 0.6 will result in a TP rate of 0.6 (better than our baseline of 54.6%) and a FP rate of ~0.35. Moving eitherside of the threshold does not seem to offer a better tradeoff. The above ROC curve suggests that at a threshold of 0.6, we will be able to predict bookings 60% of the times while erring on non bookings 35%of the time. The model is not making a huge difference in terms of predictiveness.

Features from Users session (website interaction time) Data

Airbnb has provided User session data that can be used in our modeling. Time spent interacting on the site is likely to be a good predictor of the propensity to complete a booking. We could incorporate the time spent by each user for various actions in our training data set to build the model

```
airbnb_fbs = glm(book_yorn ~ gender+age+signup_method+signup_flow+signup_app+first_devic
e_type+booking_request+click+data+message_post+partner_callback+submit+view
                , data = train_new_3, family = binomial())

summary(airbnb_fbs)
```

```
## 
## Call:
## glm(formula = book_yorn ~ gender + age + signup_method + signup_flow + 
##     signup_app + first_device_type + booking_request + click + 
##     data + message_post + partner_callback + submit + view, family = binomial(), 
##     data = train_new_3)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.7952   0.4014   0.5486   0.6314   1.3397  
## 
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          1.570e+01  4.251e+02   0.037 0.970545
## genderFEMALE                         5.146e-01  2.104e-01   2.446 0.014431
## genderMALE                           1.344e-01  1.819e-01   0.739 0.459911
## genderOTHER                          1.399e+01  7.258e+02   0.019 0.984621
## age                                 -7.234e-03  7.586e-03  -0.954 0.340318
## signup_methodfacebook               -3.156e-01  2.940e-01  -1.073 0.283061
## signup_flow                         -2.655e-02  4.859e-02  -0.546 0.584815
## signup_appiOS                       -1.364e+01  4.251e+02  -0.032 0.974408
## signup_appMoweb                     -1.424e+01  4.251e+02  -0.033 0.973282
## signup_appWeb                       -1.414e+01  4.251e+02  -0.033 0.973463
## first_device_typeAndroid Tablet      6.380e-01  1.386e+00   0.460 0.645304
## first_device_typeDesktop (Other)     1.711e-01  1.080e+00   0.158 0.874079
## first_device_typeiPad               -1.425e-01  9.417e-01  -0.151 0.879705
## first_device_typeiPhone             -6.544e-01  8.972e-01  -0.729 0.465798
## first_device_typeMac Desktop         1.963e-01  8.822e-01   0.223 0.823901
## first_device_typeOther/Unknown       3.301e-01  1.376e+00   0.240 0.810431
## first_device_typeSmartPhone (Other)  1.489e+01  1.455e+03   0.010 0.991836
## first_device_typeWindows Desktop     8.892e-02  8.843e-01   0.101 0.919898
## booking_request                      2.004e-05  2.872e-05   0.698 0.485433
## click                               -1.725e-05  8.282e-06  -2.083 0.037280
## data                                 2.410e-05  2.223e-05   1.084 0.278375
## message_post                         6.143e-05  1.652e-05   3.718 0.000201
## partner_callback                     2.971e-05  8.337e-05   0.356 0.721608
## submit                               4.129e-06  9.982e-06   0.414 0.679155
## view                                -3.147e-06  3.689e-06  -0.853 0.393535
##                                        
## (Intercept)                            
## genderFEMALE                        *  
## genderMALE                             
## genderOTHER                            
## age                                    
## signup_methodfacebook                  
## signup_flow                            
## signup_appiOS                          
## signup_appMoweb                        
## signup_appWeb                          
## first_device_typeAndroid Tablet        
## first_device_typeDesktop (Other)       
## first_device_typeiPad                  
## first_device_typeiPhone                
## first_device_typeMac Desktop
## first_device_typeOther/Unknown
## first_device_typeSmartPhone (Other)
## first_device_typeWindows Desktop
## booking_request
## click                               *
## data
## message_post                        ***
## partner_callback
## submit
## view
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1165.9  on 1336  degrees of freedom
## Residual deviance: 1116.9  on 1312  degrees of freedom
##   (121779 observations deleted due to missingness)
## AIC: 1166.9
##
## Number of Fisher Scoring iterations: 14
```

The above model suggests that three features 1) Gender - female 2) Time from Click interaction and 3) Messages Posted are the most significant in predicting the booking. However, removing observations with missing session data results in only 1337 observations(out of total 123K). The sessions data has a lot of NA's and in the above form is not adding value to the predictiveness of the model.

We can try using the Stepwise Regression (Backward and Forward) to allow the model to automatically select the best predictor variables.

```
train_new_4 <- na.omit(subset(train_new_3, select =
c(1,6,7,8,9,14,15,17,25,27,28,29,32,33,34)))
str(train_new_4)
summary(train_new_4$book_yorn)
airbnb_fb0=glm(book_yorn~1,data=train_new_4,family=binomial)
airbnb_fb1=glm(book_yorn ~ gender+age+signup_method+signup_flow+signup_app+first_device_
type+booking_request+click+data+message_post+partner_callback+submit+view,data=train_new
_4,family=binomial())

# Stepwise Regression
step(airbnb_fb0,scope=formula(airbnb_fb1), direction="forward",k=2)
step(airbnb_fb1,scope=formula(airbnb_fb0), direction="backward",k=2)
```

Both Forward and Backward regression yield the same results in terms of significant predictor variables.

```
airbnb2 <- glm(book_yorn ~ message_post + click + signup_app + gender, family =
binomial(), data = train_new_4)

predict_train <- predict(airbnb2, type ="response")
summary(predict_train)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3986  0.8121  0.8311  0.8422  0.8760  1.0000
```

```
tapply(predict_train, train_new_4$book_yorn, mean)
```

```
##         0         1 
## 0.8181524 0.8466873
```

```
table(train_new_4$book_yorn, predict_train > 0.85)
```

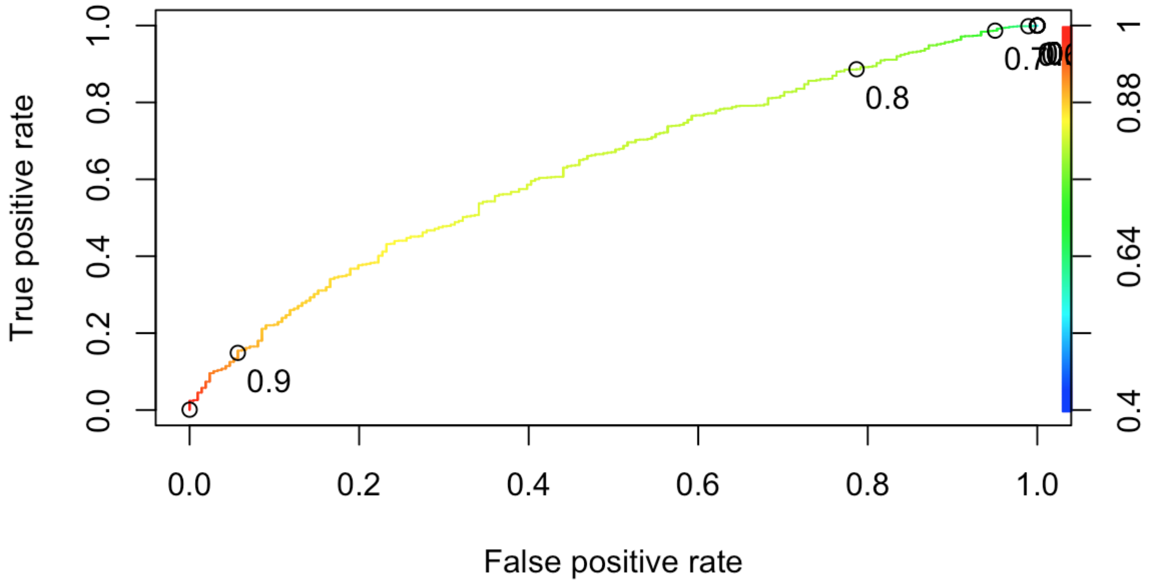
```
##    
##     FALSE TRUE
##   0   160   51
##   1   638  488
```

Sensitivity = 0.4333 and Specificity = 0.7582

```
library(ROCR)
ROCRpred <- prediction(predict_train, train_new_4$book_yorn)
ROCRperf <- performance(ROCRpred, "tpr", "fpr")
```

ROC Curve

```
plot(ROCRperf, colorize = TRUE, print.cutoffs.at = seq(0,1,0.1), text.adj = c(-0.2,
1.7))
```



From the above ROC curve, to get to a 0.6 True Positive Rate, the threshold is around 0.8 and the false positive rate increase to 0.5. The above predictive model with sessions data does not help improving the predictiveness of data mainly due to elimination of ~122K records due to lack of sessions data in the dataset.

Let us use the first model without sessions data to predict the booking for the Test data set

```
# Prepare the test data set
str(test_users)
#There are some erroneous Age values. We need to clean up the data to eliminate them.
summary(test_users$age)
test_users_new <- test_users[ which(test_users$age <= 100),]
test_users_new <- test_users[ which(test_users$age >= 16 & test_users$age <= 100),]
test_users_new <- subset(test_users_new,test_users_new$signup_method %in% c('basic','fac
ebook', 'google'))

summary(test_users_new_2$age)
summary(test_users_new)
str(test_users_new)
# Merge the sessions data with the base data
#test_new_2 <- merge(x=test_users_new, y=sessionsdataforuser, by.x = "id", by.y = "user_
id", all.x = TRUE)
#head(test_new_2)

#str(test_new_2)
#summary(test_new_2)
```

```
# apply fitted model to test sample (predicted probabilities)
predTst <- predict(airbnb_fb, test_users_new, type="response")
summary(predTst)
```

Conclusion

The following was performed to study the dataset and build a logistic regression model to predict the booking outcome of the Airbnb users.

- Exploratory analysis to identify the starting set of predictor variables
- Logistic regression model identified a few predictive variables but the ROC curve suggests that the model is not very powerful
- Stepwise regression approach, both Backward and Forward was attempted but the result does not improve
- Predicted the booking outcome for the Test data set using the Logistic Regression model from above
- Fitted a Linear Regression model on users that booked to predict the user’s interaction time (using sessions data). The obejctive here was to identify any other features that could be highly predictive of a booked user interaction time. The predictor variables were very much similar to the predictors of the logistic model above.

Scope for further refinement

- Further work may be done with more advanced feature building using additional data. For example average time between user sessions, average time between message posts, # sessions after first active date etc.

Additional data wrangling and feature building will help to achieve this.

- Subset data by weeks and see how the model works. The hypothesis is that the model may or may not be robust (exhibiting volatility) on certain weeks. This will provide an idea on how the predictiveness of the model will hold up a longer period of time.
- Other advanced modeling approaches to fit the model.