New user booking prediction

Data Mining- Project

ORIGINAL WORK STATEMENT : We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	Signature
Contact Author	Michael Antony	Michael Antony
	Sahil Bareja	Sahil Bareja
	Sarika Dhoot	Sarika Dhoot
	Roma Kaul	Roma Kaul
	Rishabh Rathod	Rishabh Rathod

INDEX

Executive Summary

For the purpose of this project, our team had chosen a dataset that we had obtained from Kaggle. This dataset contained information regarding new customer bookings for AirBnb. We aimed to employ as many methods of classification and prediction as we could so as to create a model with the best prediction accuracy. What we aimed to predict was the destination country that a customer is most likely to make their first booking for.

We used a multinomial model, Naive Bayes model, Neural Networks, Random Forest and Boosting in order to classify and predict our dependent variable and we got mixed results. We recorded the lowest accuracy while using Neural Networks where we got an accuracy of approximately 50 %. However, for the rest of the models, we achieved a higher accuracy of approximately around 70 % with Random Forest and Boosting recording the highest. We also attempted to forecast the number of new account creations and used the Holt-Winters method to do so and noticed that the number of new account creations increases in a trend like pattern. Analyzing this dataset was a challenge as we dealt with large records and were attempting multi-class classification and prediction (11 classes).

Data Description

Data source

The dataset that we used for this project contains AirBnb data over a 4- year period from 2010 to 2014 that includes all kinds of customer information such as the dates they first created an account and made a booking as well as the destination country where they booked their AirBnb stay. This dataset was obtained from Kaggle.

Link - https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

Data variables

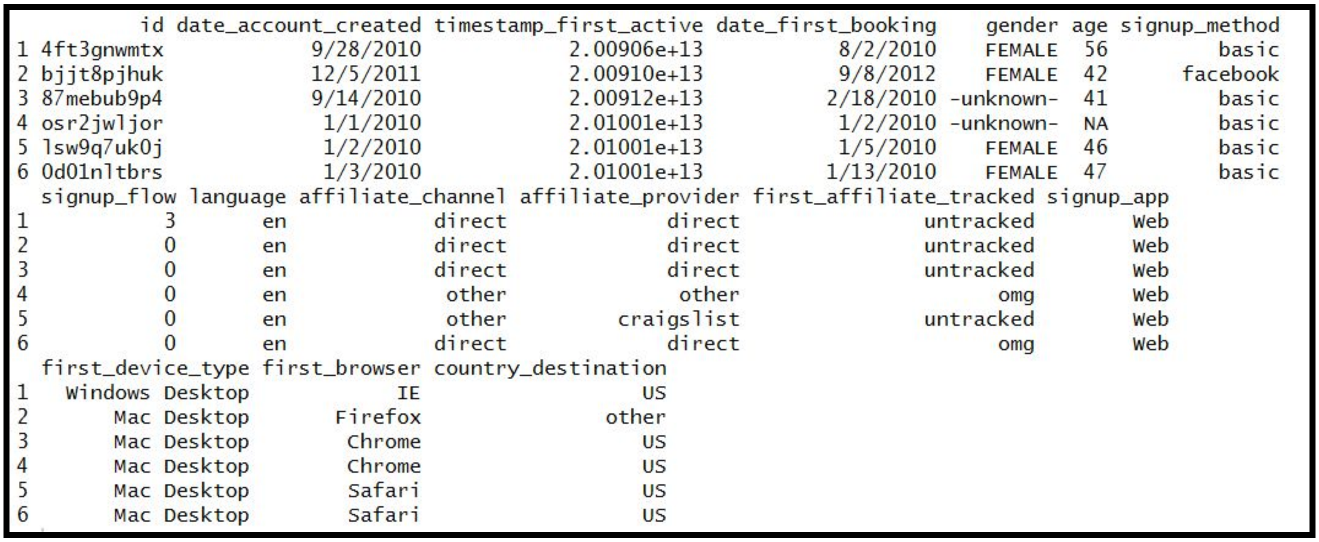
- **id** - records customer id. Factor type which is later omitted from the dataset during our analysis.
- **date_account_created** - records the date that the customer account was first created. Originally factor type converted to date type.
- **timestamp_first_active** - Records the time customer was first active. Numeric type which is omitted from our analysis.
- **date_first_booking** - records the date the customer first made a booking. Originally a factor type which we have converted to a date type.
- **gender** - records the gender of the customer. Treated as categorical.

- **age** - records the age of the customer. Treated as a numerical variable.
- **signup_method** - This indicates through what channels customers have registered for AirBnb. This is taken as a categorical variable.
- **signup_flow** - Numeric data type that indicates the number of channels the customer bounced before registering on AirBnb.
- **language** - Categorical variable that indicates the language the customer accesses the site in.
- **affiliate_channel** - Categorical variable indicating how customers reached AirBnb.
- **affiliate_provider** - Categorical variable that indicates Airbnb partner channel.
- **first_affiliate_tracked** - Categorical variable indicating if affiliate channel is tracked or not.
- **signup_app** - This is a categorical variable that indicates if users signed up on Android, ios or the web.
- **first_device_type** - Categorical variable that indicates the type of device the customer first used to make AirBnb booking.
- **first_browser** - Categorical variable indicating which browser the user first used to make AirBnb booking.
- **country_destination** - Categorical variable indicating destination country of customer's first booking.

Sample size and number of variables

Our sample dataset contains 88,908 records and 16 variables.

Sample view of data



```
          id date_account_created timestamp_first_active date_first_booking    gender age signup_method
1 4ft3gnwmtx            9/28/2010            2.00906e+13           8/2/2010    FEMALE  56         basic
2 bjjt8pjhuk            12/5/2011            2.00910e+13           9/8/2012    FEMALE  42      facebook
3 87mebub9p4            9/14/2010            2.00912e+13          2/18/2010 -unknown-  41         basic
4 osr2jwljor             1/1/2010            2.01001e+13           1/2/2010 -unknown-  NA         basic
5 lsw9q7uk0j             1/2/2010            2.01001e+13           1/5/2010    FEMALE  46         basic
6 0d01nltbrs             1/3/2010            2.01001e+13          1/13/2010    FEMALE  47         basic
  signup_flow language affiliate_channel affiliate_provider first_affiliate_tracked signup_app
1           3       en            direct             direct               untracked        Web
2           0       en            direct             direct               untracked        Web
3           0       en            direct             direct               untracked        Web
4           0       en             other              other                     omg        Web
5           0       en             other         craigslist               untracked        Web
6           0       en            direct             direct                     omg        Web
  first_device_type first_browser country_destination
1   Windows Desktop            IE                  US
2       Mac Desktop       Firefox               other
3       Mac Desktop        Chrome                  US
4       Mac Desktop        Chrome                  US
5       Mac Desktop        Safari                  US
6       Mac Desktop        Safari                  US
```

Data interest

We were interested in this particular dataset as it allowed us to employ various models in order to answer our main questions that centered around predicting what country a customer is likely to book their first AirBnb stay given the data.

Research Questions

Key Question

What would be the first country destination that a new user who signs up on airbnb book for?

Questions of interest

If we have capital to promote only one country through the marketing channel, which country will possibly fetch us maximum profit?

Is their any specific age group that should be our target users?

Which all marketing channels get us the most user conversions?

Given a new user which country is he likely to book his destination of travel?

In what seasons would it be most profitable to roll out the discount deals?

If target marketing is to be done to users for more than one country, which should it be in the order of their preference?

Methodology

We used several data mining techniques but because of the complexity of the data and having multiclass predictor many of them couldn’t give us proper results. Also, our dataset has a lot of train cases have destination country US (around 70% of our data) so there is a high bias towards that. In order to deal with that we tried numerous approach and in the end, a few models did show better results.

Models

- **Multinomial**

We used Multinomial Logistic model to classify and predict the classes for multiple country destinations. We ran it twice with multiple classes and a second time with just two classes indicating US and Non-US destinations. We got an accuracy of 70.28% and 70.4% respectively.

- **Time Forecasting**

We attempted to forecast the number of new customer account creations for the next 18 months using the Holt-Winters method and our analysis showed that this number continues to increase in a trend like fashion.

- **Naive Bayes**

Since our dataset predicts multiclass categorical variable. We used this method to predict the classes directly using this method as naive bayes. Using the probabilities of each class Naive bayes is capable of giving a category as an output. Hence, we tried it to get better accuracy with the results.

Performance Measure:
Accuracy achieved: 69.4%
Recall of US (top class): 0.70
Precision of US: 0.99

- **Rpart/Boosting**

During our attempt to perform Boosting to find a good result from weak learners, which we have quite a few, we realized that due to enormous amount of data giving a weight to misclassified points was causing memory to run out and even 8gb ram couldn’t handle it. At that point we realized the Boosting method uses “Control” parameter which actually sets the rpart algorithm use. So, we decided to implement rpart instead and got arguably good results.

Performance Measures:
Accuracy: 70.4%

- **Neural Network**

Since neural networks is suited for datasets that need modelling of complex relationships, we thought of using Neural Networks. The resulting model that we got from neural networks had very poor accuracy of about 50%. Since Neural networks is a black box there was no way to tweak the model or perform variable selection.

Performance Measures:
Accuracy: 50%

- **Market Basket Analysis**

In order to discover the customer booking behavior, we conducted market basket analysis on the dataset with consequent as 'country_destination'. Due to the high bias present in the consequent with maximum countries being US, we conducted the analysis in the two parts:

1. With including US in the consequent list of country destinations.
2. Excluding US from the consequent list of country destinations.

Graphs of the association of the most frequent antecedent sets for the given consequent set for both the parts have been attached in the Appendix.

Figure 2 - Plots Top 5 associations for Part 1 and Figure 3 -Plots Top 5 associations for Part 2.

As seen in the Figure 2: 5 % of the customers that have a 'language' preference as English with 'sign-up method' as basic and craigslist as 'affiliate provider' book their destination country as US. This rule has a healthy support in the dataset as seen by the size of the bubble and reasonable confidence as denoted by the shade of the bubble and so on.

- **Random Forest**

The Random Forest was the first ensemble method we used in our prediction. The aim of this particular algorithm was to avoid overfitting and reduce the variance in order to average out the results obtained.

Accuracy: 70.42%

- **XGBoost**

As part of our analysis, we wanted a way where we can rank the multi-class dependent variable w.r.t. its probability for each user and at the same time, be able to reduce the computation time. The answer to this particular approach was the Extreme Gradient Boosting (XGBoost) algorithm. So for each of the user, we create its preference list.

For example, for user with User ID: 87mebub9p4, his/her country preference is as seen in appendix Fig 1 .

Thus, his/her first country preference is USA. Also, it is to be noted that he/she is least likely to go to Portugal. Similarly, we have found the preferences for all the users.

- **Ensemble with a Twist**

We were able to develop some good models and identify algorithms that fit best with our dataset. To go one step further, we decided to pick the best models and combine the results using a majority vote but the results of biased majority were biased too. Hence, it doesn't work in this scenario.

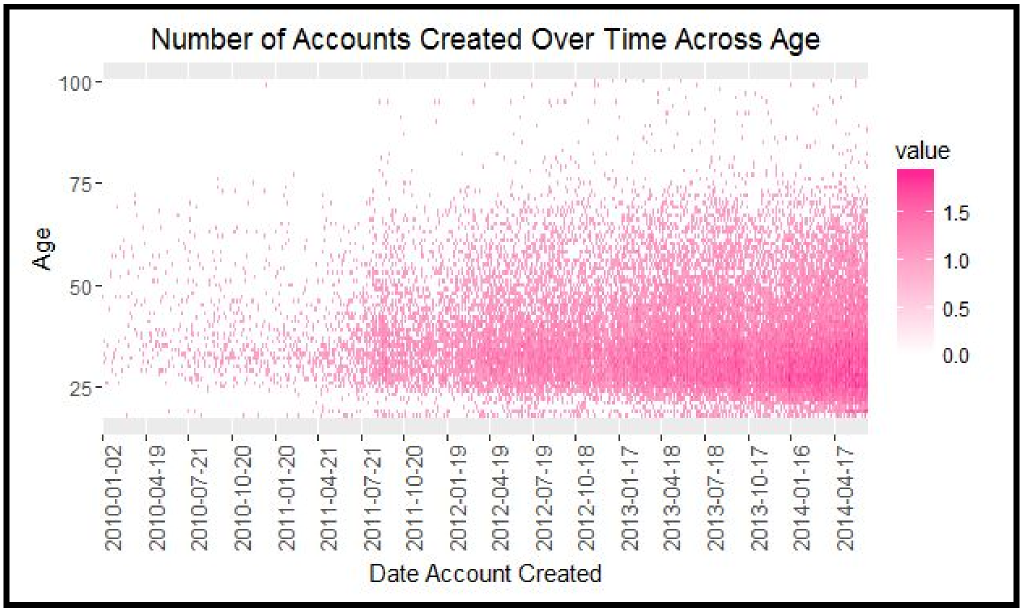
Results and Findings

Descriptive Analytics

We decided to explore the dataset to understand the significance of various parameters and the associations between them. Some of the discoveries that we made were:

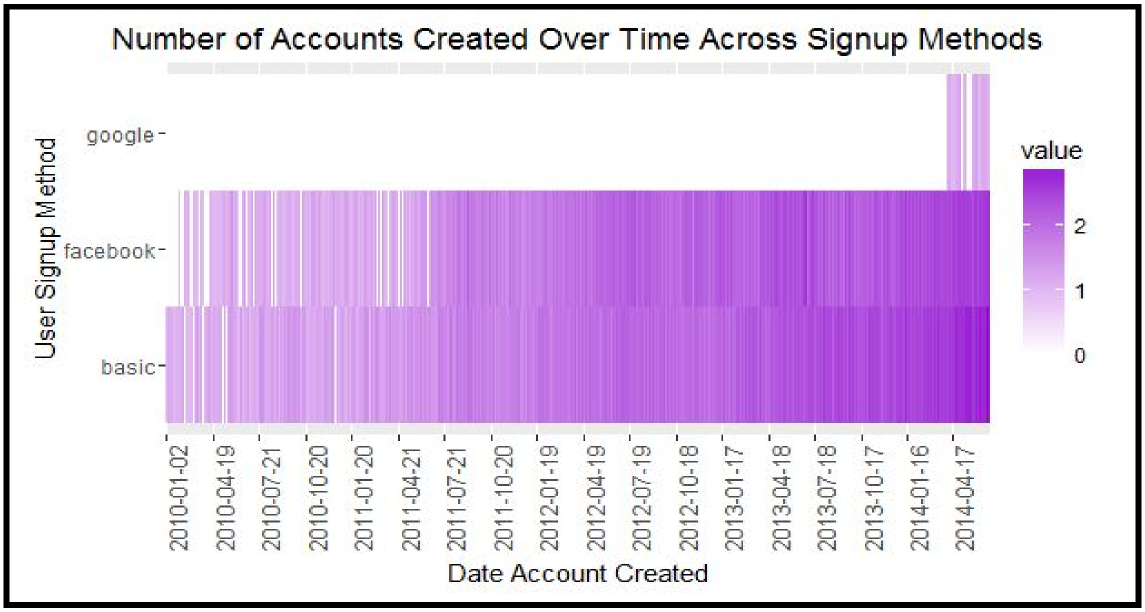
Is there any specific age group that should be our market target?

Initially the age group of users on AirBnb was concentrated amongst the 20-35 years bracket, but over the years it is much more spread out.



Which all marketing channels get us the most user conversions?

Certain channels such as facebook play a significant role in the marketing of AirBnb. The number of users signing up on AirBnb directly and through Facebook is just about the same. Even though signup through Google was not present until very recently, its contribution now is considerable.



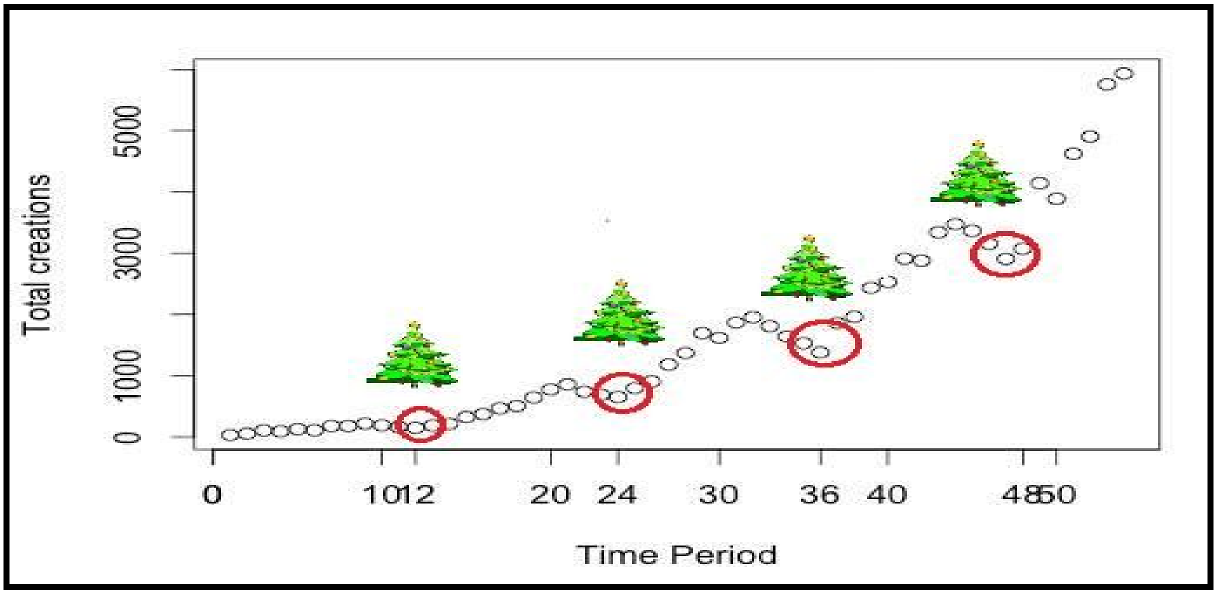
Other interesting comparisons like the number of users active through phones v/s computers was made where phone users are coming close to surpassing laptop users. This comparison shows that amongst all phones; iphone users have consistently ranked highest.

If target marketing is to be done to users for more than one country, which should be it in the order of their preference?

We used XgBoost algorithm to find the second preference due to high bias in the data.
To probe the question of identifying user preferences in further detail we analyzed the data to understand not their first destination of travel but an ordered preference list of all the available destinations (11 classes). For eg. if a user is most likely to make a booking in the US, which is the 2nd highest most likely destination. The XgBoost algorithm gave us a comprehensible list of each users preference and ranked them in order.

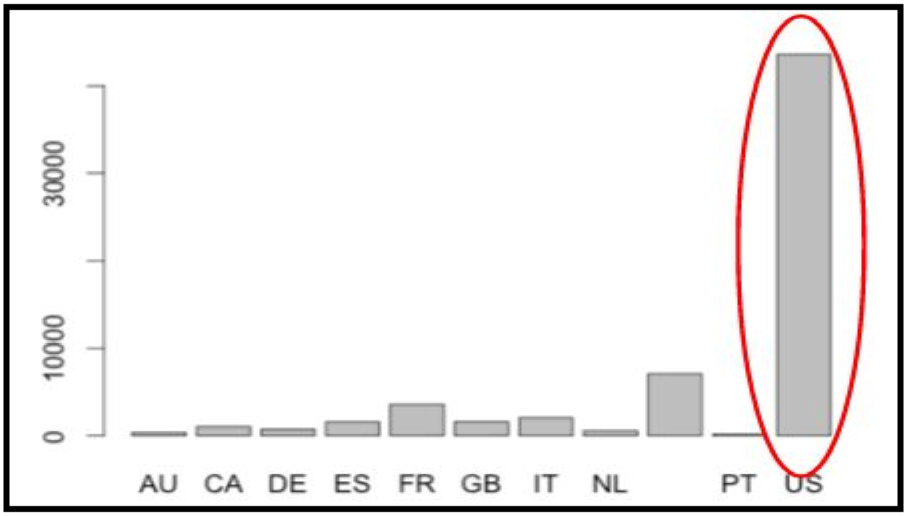
In what seasons would it be most profitable to roll out the discount deals?

While studying the time trends we observed a dip in the booking at end of each year which also overlaps with the thanksgiving and christmas. As is observed in the market trend, people tend to stay home around these festive seasons. Even if people travel they visit their family and so the dip during end season is pretty prominent. This we were able to find through time series forecasting when we tried to study the data time-wise. So, there should be more lucrative deals that should be rolled out during these seasons to attract more customers and increase the number of bookings.



If we have capital to promote only one country through the marketing channel, which country will possibly fetch us maximum profit?

As all our models show a high bias towards the US, we would concentrate all our marketing efforts on the US.



Main Question: What would be the first country destination that a new user who signs up on airbnb book to?

To answer this question we implemented multiple classification algorithms and predicted values for the country of destination. Each model gave us some insight into the data. Based on all the models some of our findings were

- The dependent variable (country_destination) is highly skewed in the sense that majority of the records have the values ‘US’. Hence we realised that in case of some algorithms the accuracy value was misleading. It is like Naive rule; if all records are predicted as

‘US’, accuracy may still be a high number as the actual number of records with values ‘US’ are high. This led us to use other measures such as Precision and Recall for multiple classes of interest.

- Just like the dependent variable, some of the predictor variables are also biased. In some cases, we had to extract month or year of data to make sense of the analysis.
- Variable selection was important to identify important variables so that the most efficient models are created.

Conclusion

Rigorous exploration and analysis of the data set lead us to some interesting correlation between variables.

Using this, we think there can be an increase in business if certain trends are paid attention to and if some marketing channels are implemented to help enhance the opportunities that these trends create. Some of such suggestions to airbnb would be:

- Advertise more on facebook and less on google as user conversion is way higher on facebook.
- Come up with cheap lucrative deals around thanksgiving and christmas to increase the booking around that time as bookings tend to drop around that time.
- Most of the users who book lie between the age group of 25-45, so focus more on them.

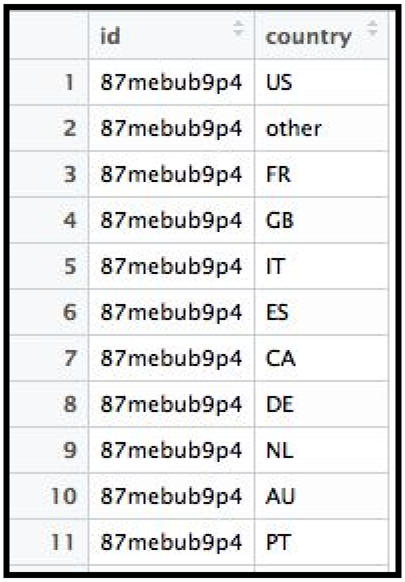
About data set and data mining observations:

- Since individually a lot of variables are weak learners but combining their data can bring in interesting observation   [For example. Affiliate channel along with Affiliate Provider gives us more fruitful results]
- Hence, boosting with more meaningful data should produce higher accuracy and better predictions.

We also think that had we been given the number of sites that can be visited in every country that we predict, we might have been able to see a good correlation between the number and probability of the country being predicted. Combining such interesting facts about the predictor column can help us discover strong results and with higher confidence.

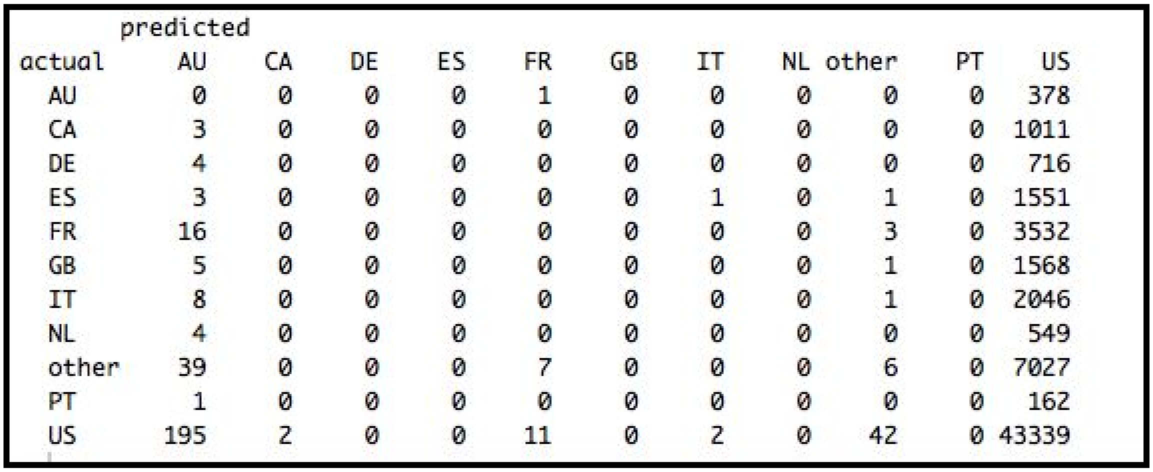
Appendix

Fig 1. - XGBoost



	id	country
1	87mebub9p4	US
2	87mebub9p4	other
3	87mebub9p4	FR
4	87mebub9p4	GB
5	87mebub9p4	IT
6	87mebub9p4	ES
7	87mebub9p4	CA
8	87mebub9p4	DE
9	87mebub9p4	NL
10	87mebub9p4	AU
11	87mebub9p4	PT

Fig 2. - Naive Bayes



```
       predicted
actual     AU    CA    DE    ES    FR    GB    IT    NL other    PT    US
  AU        0     0     0     0     1     0     0     0     0     0   378
  CA        3     0     0     0     0     0     0     0     0     0  1011
  DE        4     0     0     0     0     0     0     0     0     0   716
  ES        3     0     0     0     0     0     1     0     1     0  1551
  FR       16     0     0     0     0     0     0     0     3     0  3532
  GB        5     0     0     0     0     0     0     0     1     0  1568
  IT        8     0     0     0     0     0     0     0     1     0  2046
  NL        4     0     0     0     0     0     0     0     0     0   549
  other    39     0     0     0     7     0     0     0     6     0  7027
  PT        1     0     0     0     0     0     0     0     0     0   162
  US      195     2     0     0    11     0     2     0    42     0 43339
```

Fig 3. - Random Forest

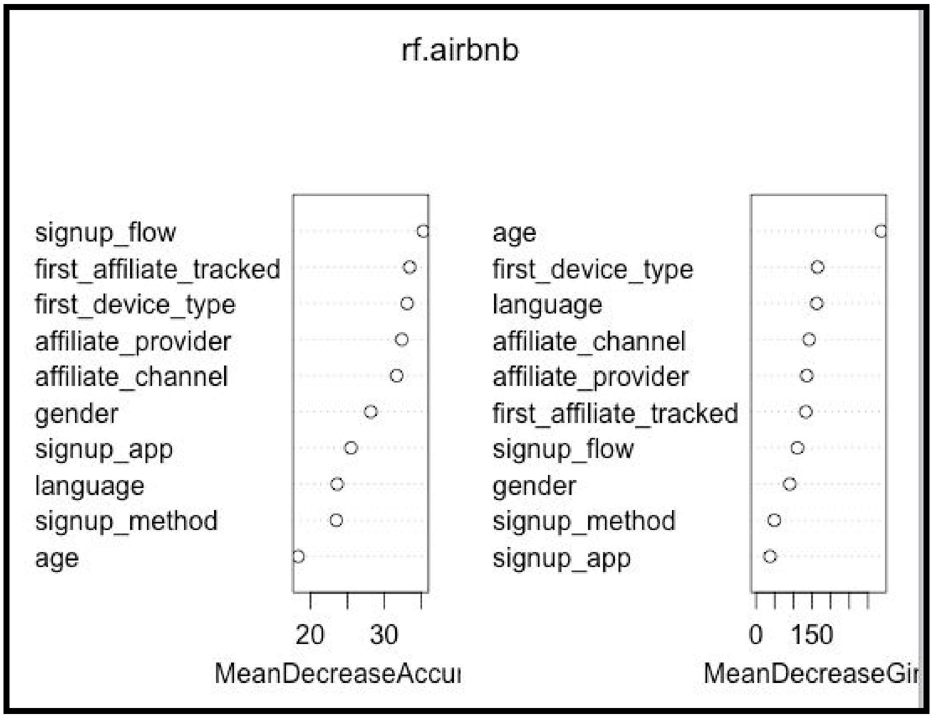
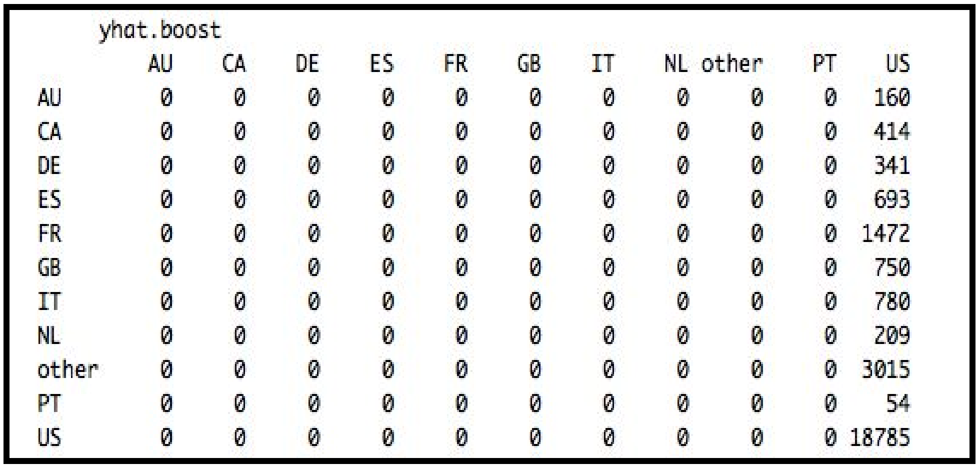


Fig 4. - Boosting



```
       yhat.boost
         AU   CA   DE   ES   FR   GB   IT   NL other   PT    US
  AU      0    0    0    0    0    0    0    0     0    0   160
  CA      0    0    0    0    0    0    0    0     0    0   414
  DE      0    0    0    0    0    0    0    0     0    0   341
  ES      0    0    0    0    0    0    0    0     0    0   693
  FR      0    0    0    0    0    0    0    0     0    0  1472
  GB      0    0    0    0    0    0    0    0     0    0   750
  IT      0    0    0    0    0    0    0    0     0    0   780
  NL      0    0    0    0    0    0    0    0     0    0   209
  other   0    0    0    0    0    0    0    0     0    0  3015
  PT      0    0    0    0    0    0    0    0     0    0    54
  US      0    0    0    0    0    0    0    0     0    0 18785
```

Fig 5- Market Basket Analysis with US included in the consequent list

Top 5 Association Rules

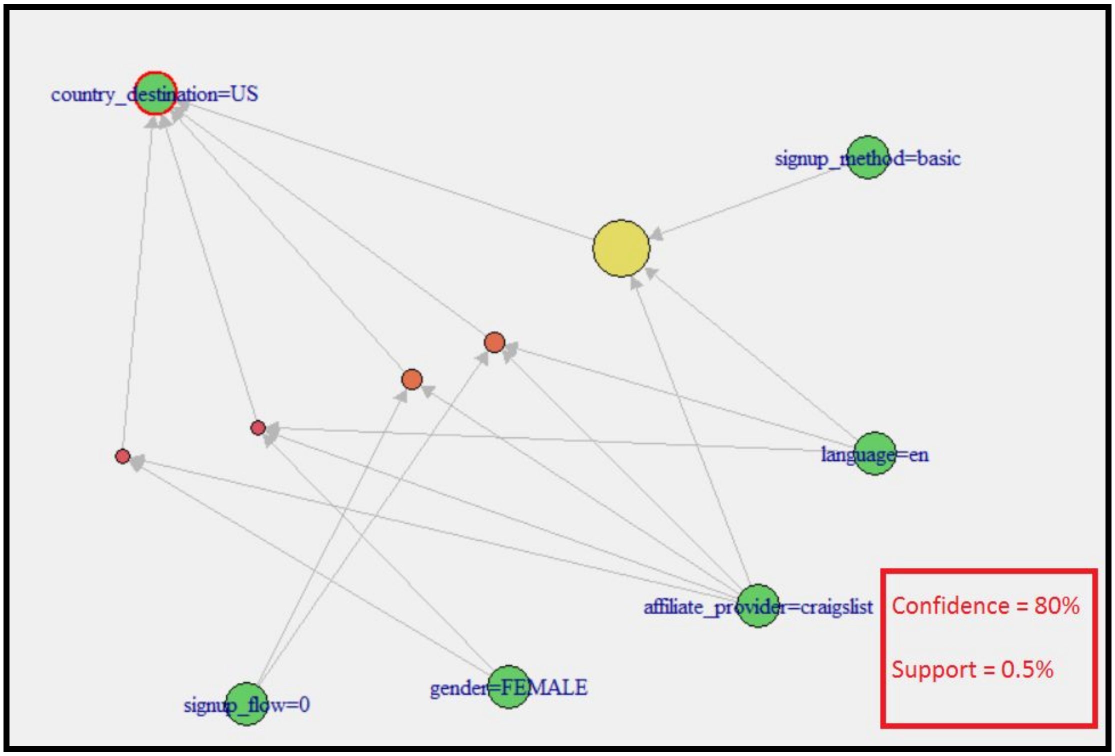
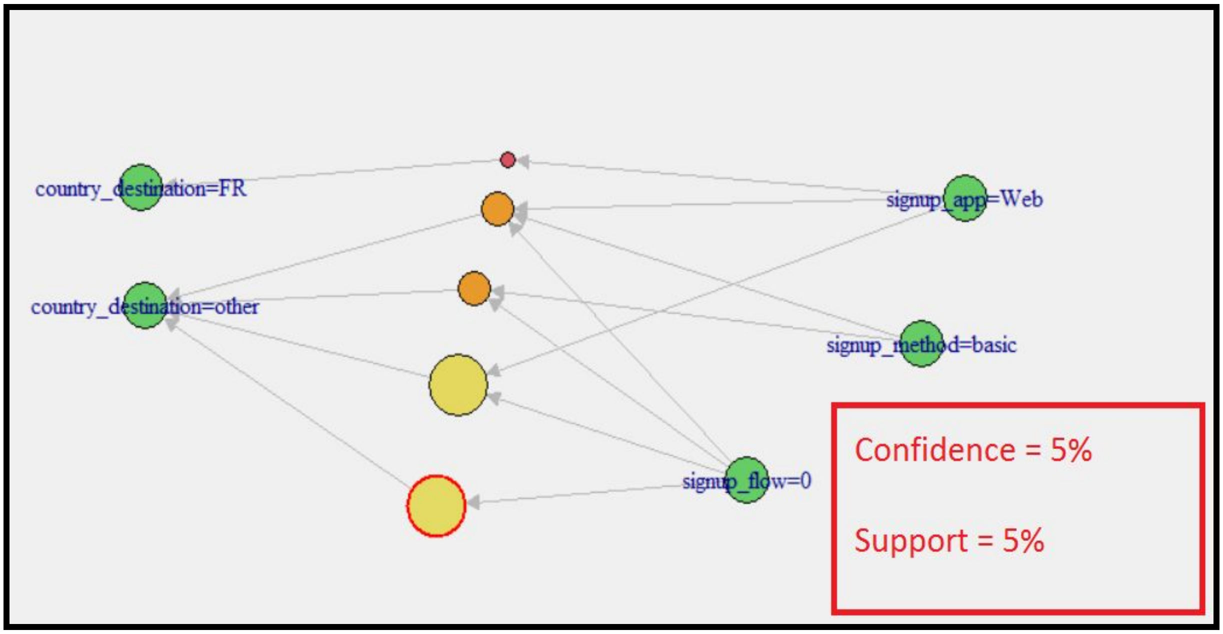


Fig 6. - Market Basket Analysis excluding US from the consequent list.

Top 5 Association Rules


country_destination=FR
signup_app=Web
country_destination=other
signup_method=basic
signup_flow=0
Confidence = 5%
Support = 5%