Predicting Airbnb user destination using user demographic and session information

Srinivas Avireddy, Sathya Narayanan Ramamirtham, Sridhar Srinivasa Subramanian

***Abstract*—In this report, we develop a model to predict the Airbnb user's booking destination country based on their demographics and session data. This model is very helpful in providing personalized recommendations and targeted marketing to enrich user experience and optimize business conversion. The dataset we used was provided by Airbnb's user information. The given problem is modelled as a classification problem and found random forest classifier pruned with importance of features to be a very good model to predict user preferred destination. The proposed model has an accuracy of 88% which is better than the baseline model and decision tree classification model.**

***Keywords*—Data Analysis, Customer Segmentation, Random Forests.**

I. Introduction

Airbnb is a trusted community marketplace for people to list, discover, and book unique accommodations around the world. Airbnb is increasingly becoming the go-to place for travellers worldwide. With it's presence in 34000+ cities in 190+ countries, the users no longer have to be worried about needing to stay in expensive hotels. Users can use the web application or the Android/iOS application. The user's can browse through listings in several countries and book accommodation with a single click in the app. In order to maximize business conversion, we have the challenge of identifying user's intent and showing them relevant recommendations based on their session behavior and demographics.

Certain behavior/data can be indicative of whether the user has an intent to visit a specific country. For eg., user browsing through certain country's listings, language spoken by the user, country specific seasonality etc., In this project, we try to identify potential features from the dataset and see how they correlate to the countries they make their booking.

In the second section, we describe the dataset and provide the insights got from exploratory analysis of the data set. we dive deep into the data in order to identify distribution in data, patterns, features and biases etc. In the third section, we briefly describe the predictive task. In the fourth section, we analyze the previous work that has been done in this area in data mining. More specifically we analyze the research that has been carried out in this particular problem. Finally adding the impact each of the features had on our take on the model. This follows the evaluation the proposed model with other models.

II. Dataset

Airbnb is a trusted community marketplace for people to list, discover, and book unique accommodations around the world. Users can use the web application or the android/iOS application. We are given a list of users along with their demographics, web session records, and some summary statistics. With this dataset, we should be able predict which country a user's booking destination will be. All the users in this dataset are from the USA. The training dataset contains the following information:

1) user id
2) language
3) age
4) gender
5) information about the users sessions
6) date of creating the account
7) date of first booking
8) signup method - Facebook, Basic
9) first device type - Mac, Windows, iPhone, etc

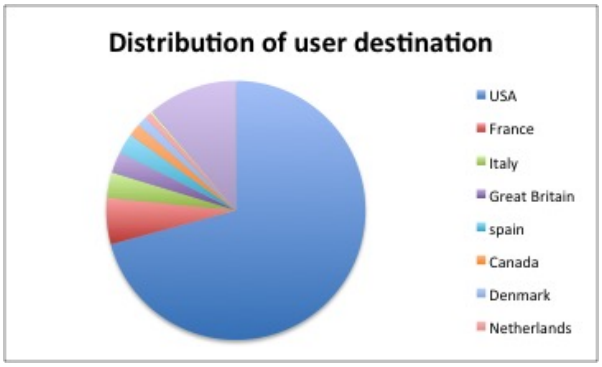

Fig. 1. Pie chart showing the distribution of user preferred destinations

III. Exploratory Analysis

The dataset used for this project was available as part of the Kaggle challenge. The training data had the demographic and session information for 171240 users and the challenge was to predict the next booking destination of 43674 users in the test set. All the users are based in US.

A. User preference distribution in Training set

It was found more than 50% data had destination as NDF, which meant the user hasn't been to any destination yet. This was an interesting observation and is quite true, as most user generally browse destinations but do not actually make a trip. The overall distribution of other countries is shown in Figure 1. It is found that USA tops the user preference.

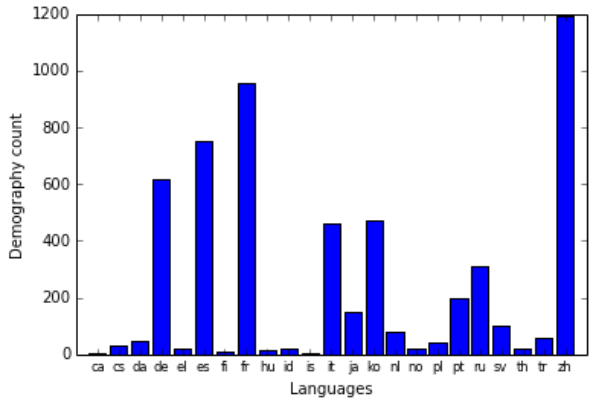

Fig. 2. Histogram showing the distribution of demographics in the training set with English language removed.

B. Language Demographics and Country Preferences

We initially plotted a histogram of users visiting each country as per their language. Fig. 1 shows the demographics of the population in the training set.

As the users were primarily based in US, the language distribution was heavily skewed towards English speaking people, we removed English speaking people from this particular plot. We then normalized the count as per the number of people in the particular demographic. We were able to obtain relative preferences to countries as per the language. We will look at some interesting plots and our corresponding observations.

Figs.3, 4 and 5 show normalized count as per language of people with booking destination as United States, Great Britain and Spain. It can be seen that higher proportion of certain demographics like 'catalan' and 'norwegians' book primarily in US locations. Also, 'norwegians' in the data set have gone only to United States and Great Britain. It could be related to scandinavian migration to United States and Great Britain (probably family members and friends living around). Also, the effect of distance and cultures seem to be evident. People from Asian countries like China and Japan are less likely to go to European countries whereas they are more likely to go to United States. Hence, these culture specific biases could be used for determining the likelihood for going to a specific country.

C. Seasonality

Next, we plotted the number of visitors in each month for different countries. The intuition behind this is that, given a month, some countries are more likely to be visited owing to seasonality in terms of weather/festivals etc., Figs. 5 and 6 shows a sample plot of distribution of number of visitors in each month between two countries with contrasting weather conditions Australia and Canada. Australia with a relatively hotter weather has more visitors during winter whereas the influx seems to be more during summers in Canada.

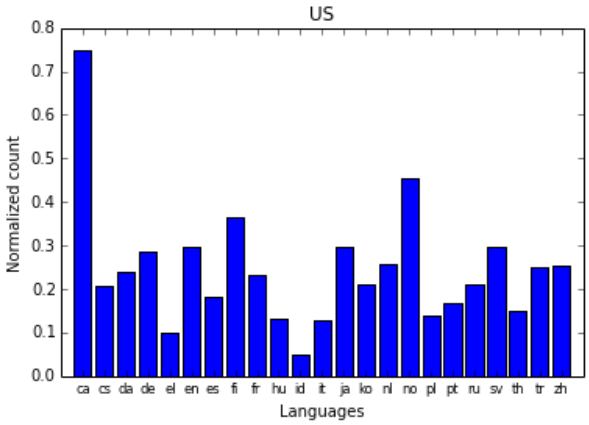

Fig. 3. Demographics of people whose booking destination is US as a function of normalized count

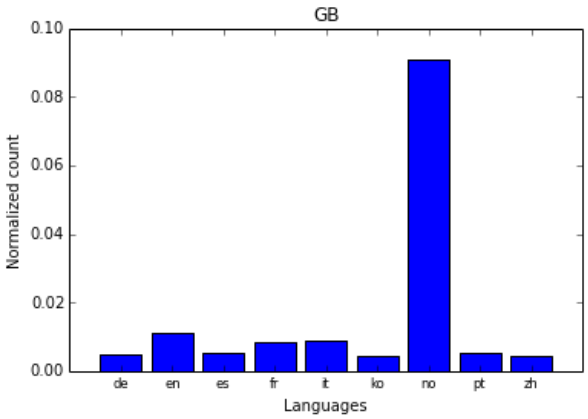

Fig. 4. Demographics of people whose booking destination is Great Britain as a function of normalized count

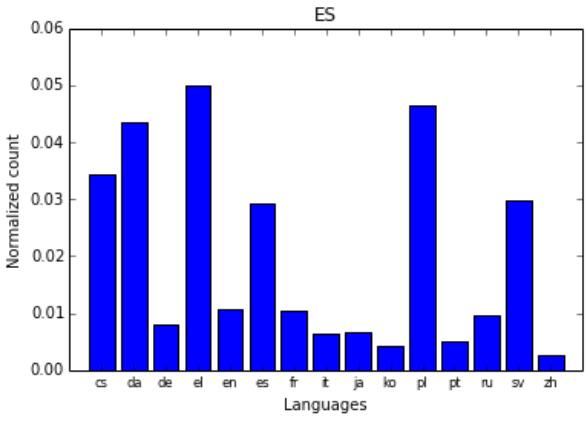

Fig. 5. Demographics of people whose booking destination is Spain as a function of normalized count

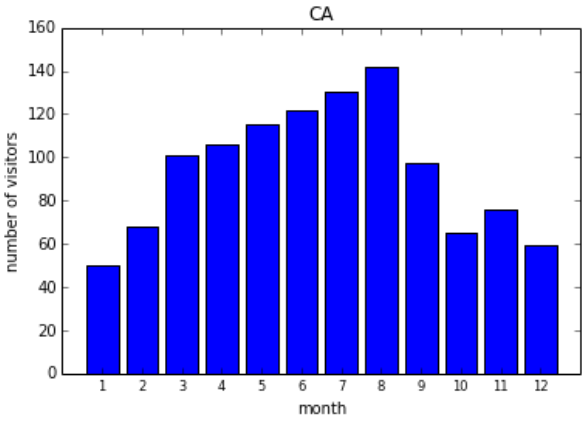

Fig. 6. Number of visitors each month: Canada

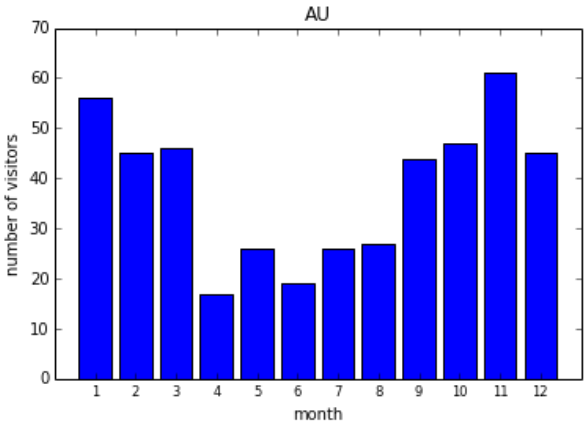

Fig. 7. Number of visitors each month: Australia

D. Gender Distribution

Whilst the overall gender distribution stands at a male to female ratio of 48:52, certain countries like France and Italy see a higher influx of female visitors. It could be because these places are a good vacation spots and fashion destinations. The gender distribution in France for eg., is shown in the Fig. 7.

IV. Predictive Task

In this project, we predict the booking destination of a user, given many useful features about the user. The input to our model is a set of features like the age, gender, language of the user. The date of the user's first booking and session activity play a vital role in the predictive task.

A. Classification problem

The prediction task is a multiclass classification problem since the output is one of eleven countries (AU, CA, DE, ES, FR, GB, IT, NL, PT, US) or NDF, which means no destination found. If we predict our outcome to be a country other than these, it is classified as other. In this problem the following

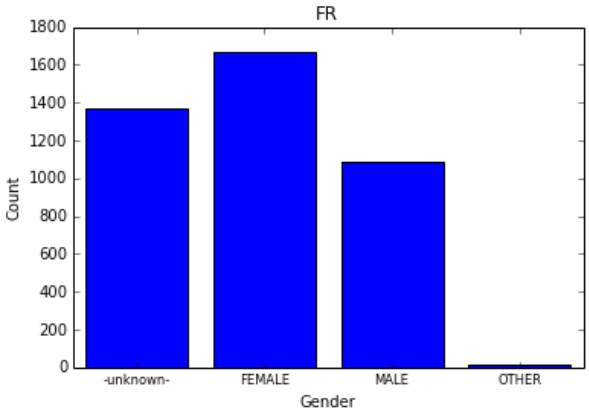

Fig. 8. Gender distribution of visitors: France

are the list of models we considered to predict the user's next booking destination.

- Predicting all destinations to be NDF.
- Decision tree model.
- Random forests model.
- Random Forest model, with NDF binning.

B. Features selection

We started out by identifying relevant features. Based on our exploratory analysis, these were the set of features we considered.

- Age - We see that users over 40 travelled more compared to users under 40.
- Gender - The dataset contains more number of female users in comparison to male users. We also observed that a greater proportion of female users have travelled to countries than men i.e. smaller proportion of NDF for female users.
- Date first booked - The presence of this field indicates that the user has travelled to some destination. And the absence means the destination is NDF. Using this feature, we also pruned the dataset given for training, and testing. In our prediction, we eliminated the entries without this field, and predicted the output, to be NDF.
- Seconds elapsed (session) - The more time a user spends on the website, the more probability there is for them to book a destination i.e a factor in identifying a 'no booking' and a booking
- Language - As we discussed in the exploratory analysis, language played a major role in deciding the choice of users destination.
- API calls - The API calls in the user session to Google Translate/keywords like country names could play a major role in predicting the destination.

After collecting individual features, in order to prune unnecessary or redundant features, we plotted a heatmap of various attributes based on their linear correlation. Fig. 9 shows the same.

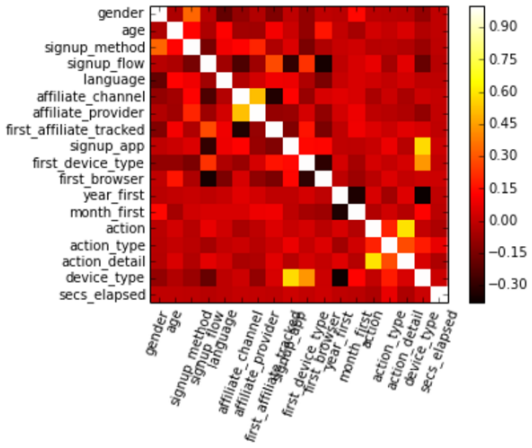

Fig. 9. Linear correlation of attributes

TABLE I.
TOP 5 FEATURE IMPORTANCE OF RANDOM FOREST CLASSIFIER

Feature	Importance Value
Age	0.1707
Language	0.1228
Month of first booking	0.0901
browser used	0.0843
Gender	0.0613

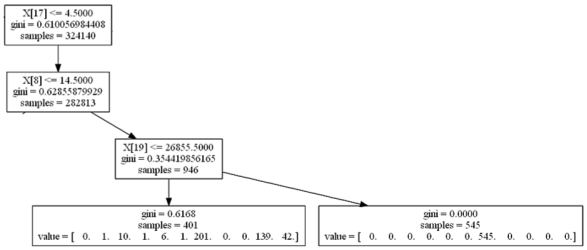

Fig. 10. Decision tree sample for the feature; The difference in the predictions for true/false condition on an attribute can be seen clearly

V. MODEL DESCRIPTION

We started out with Decision Tree and Random Forest models because the decision boundary is not linear. Decision tree follows this approach to form a predictor f(x) = y. It forms a tree whose nodes are features, then it decides which features to consider first in predicting y from x. It then uses recursion to form sub-trees based on attributes to form a decision. The Decision Tree Classifier is easy to use, and easy to interpret. The Random Forest Classifier works in a similar way as the decision tree classifier, but it grows multiple classification trees. If the original feature vector has d features, each tree uses a random selection of features. All the associated feature space is different, but fixed for each tree.

A. Data pruning

To prune the initial data, we used imputation to fill in the missing values like age, first-affiliated, etc. The imputation method chosen was 'mean' i.e. if data is not available in a particular field, mean of that column (feature) will be used. We also wanted to check how our input features correlated with each other, so that redundant features can be removed before training. Fig. 9, shows the correlation between features. The following features which were strongly positively correlated or strongly negatively correlated were removed - timestamp-first-active, signup-app, signup-flow, and affiliate-provider.

B. Baseline Model - Predicting the destination to be NDF

We categorically predict that the user will not be traveling anywhere. This gave 62% accuracy in the test set.

C. Model 2: Decision tree classifier

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Tree models where the target variable can take a finite set of values are called classification trees

D. Model 3: Random Forest Classifier

Random Forest classifier grows many classification trees. To classify a new object from an input vector, it puts the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

E. Model 4: Random Forest classifier with NDF binning

On exploring the feature importance in the decision tree model, we observed that the date-first-booking feature had a very high weight in the prediction. This was because whenever the date-first-booking field was absent in an entry, the outcome was always NDF. So, we removed the entries where the first booked element was absent, and predicted that the output of those entries are NDF.

F. Feature importance

After training the data, these were the top features in the decreasing order of importance: age, language, month of first booking, browser used, gender.

TABLE II.
RELATIVE PERFORMANCE OF MODELS

Model	Validation set Accuracy (%)	Test set Accuracy (%)
Primitive	63.7	62.3
Decision Tree Classifier	80.8	79.5
Random Forest Classifier	89.3	87.5
Random Forest Classifier with manual NDF prediction	89.2	87.3

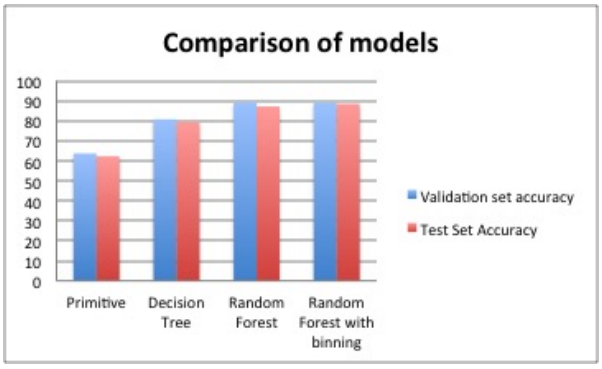

Fig. 11. Comparison of classification models used

VI. Evaluation of Models and Results

We had 171240 user data points in the training set and we had to predict for about 43674 users in the test set. To evaluate the various models, we used a 70-30 split of the given dataset for training and validation.

A. Comparison metrics

The comparison metrics that we used was the error rate of each of these models under discussion. Table 2 shows the comparison of models based on accuracy on validation and test data It was found that the primitive model had the least accuracy and random forest classifier had the best prediction accuracy of 87.5%.

VII. Literature

- Prediction of short-term human behavior is a rapidly growing area of research. There are a lot of models for predicting user intent using Hidden Markov models and Kalman Filters.
- The uber blog on analysing user destination intent was also a very interesting read and relevant to our predictive task.
- he dataset being used is provided by Airbnb. The dataset is available online. We were inspired to work with multi class classification tasks after going through a few interesting competitions on Kaggle.
- To deal with multi class classification tasks, we went through different materials on the same, the most useful being the scikit.learn python library documentation.
- The Airbnb blog on filling in missing values for Random Forests was helpful for us as a large percentage of the training data was unfilled and we used imputation using 'mean' for filling in the values.

VIII. Conclusion

From the above results, we can conclude that the Random Forest classifier algorithm is superior to that of the the baseline model and the Decision Tree model. The Random Forest Classifier model had an 87% accuracy compared to 62% for the baseline model and 79% for the decision tree model.

Originally, we had expected that pruning the dataset given to the Random Forest Classifier will improve the results. But, from the results, we can see that removing features in the random forest classifier model gives no additional benefits. This proves that the classifier will not over-fit the data. Because it generates multiple decision trees and chooses the best one, even if we had a redundant features, there won't be any degradation in the prediction accuracy.

This problem of identifying the next user booking destination was very challenging, open-ended and rewarding. We wish the session information had some data about countries the user searched for, it would have been quite helpful.

References

[1] D. D. Salvucci, Inferring driver intent: A case study in lane-change detection, in Human Factors and Ergonomics Society 48th Annual Meeting, 2004

[2] A. Pentland and A. Liu, Modeling and prediction of human behavior, Neural Computation, vol. 11, pp. 229242, 1999.

[3] Alok Gupta - Overcoming missing values in Random Forest Classifiers

[4] Riley Newman - How we determined most hospitable cities

[5] Random Forests Berkeley - Berkeley Breiman Notes

[6] Uber user destination prediction - Uber user destination intent