**Prediction for Airbnb New User Bookings**

**Summary**  
  
In the recent past, the tourism industry has been revolutionized by a vacation rental company called Airbnb. This website has successfully leveraged technology to bring people closer, instead of drawing people apart. The innovative internet company is valued at more than $30 billion, has served millions of travelers, and has formatively shaped the sharing economy phenomenon. Airbnb has figuratively and literally changed how people sleep when away from home by providing desirable alternative places to stay while traveling.

In this project, we predict the Airbnb’s new user’s booking destination country based on their features such as age, gender, demographics and session data, language etc. using different machine learning techniques. We begin with a comprehensive analysis on the dataset to explore all the given features. Techniques like Logistic Regression, Random Forest and XGBoost to predict where their users were going to travel based on data from their website. We compare and conclude that Random Forest is the most useful technique for this prediction. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

1. **Introduction/ Motivation**

It is a great, big world out there with billions and billions of people, who each day live their life and have their own unique experiences. In the past few decades, travelling has made it possible to reach the remotest corners of the world. Travel has become one of the most preferred leisure activities leading to people willingly investing in it. Travel has been extremely instrumental in the rising demand for travelling house booking. In such scenarios, companies like AirBnB struck gold by providing a platform that would help travelers book spaces to stay in. Airbnb, Inc. is an American vacation rental online marketplace company based in San Francisco, California, United States. Airbnb offers arrangement for lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking. With listings in 100,000 cities across the world, 6 guests check into an AirBnB listing every second. As of today, over 150 million worldwide users have booked over 800 million stays. The growing popularity of AirBnB over the years inspired us to investigate into different machine learning techniques that will model and predict customer behavior patterns for the company. On comparing the results from techniques like Logistic Regression, Random Forest and XGBoost, we will choose the best one for predicting new users’ first booking destination of Airbnb.

1. **Approach**

In order to solve the above stated problem, we followed an approach consisting of 5 major steps  
  


1. **Major Findings**  
     
   In order to compare the machine learning models, we studied the use of   
   Normalized Discounted Cumulative Gain (NDCG). The metric measures the performance of a recommendation system based on the graded relevance of the recommended entities. Among all the models fitted, Random Forest obtains the highest NDCG score and is hence, the best method for prediction of new users’ first booking destination. The NDCG scores of all the techniques are given below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Logistic Regression** | **Random Forest** | **XGBoost** |
| **Score** | **0.8157** | **0.9511** | **0.9066** |

***Table 1: Model Comparison***

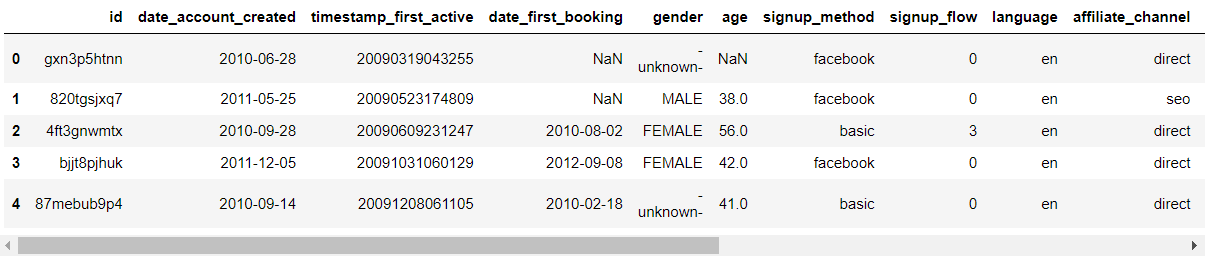
1. **Data Exploration**

For this analysis, we used the Airbnb New user Bookings datasets available in Kaggle as a part of the Kaggle Airbnb Recruiting Competition. The dataset has 275547 rows and 16 columns.  
This data frame contains the following columns:

**Id**: User id  
 **date\_account\_created** the date of account creation   
**timestamp\_first\_active** timestamp of the first activity   
**date\_first\_booking** date of first booking   
**gender** selected gender of the user   
**age** user’s age   
**signup\_method** user’s method for signing up   
**signup\_flow** the web page where the user originated to sign up   
**language** user’s language setting   
**affiliate\_channel** paid market type   
**affiliate\_provider** paid market name   
**first\_affiliate\_tracked** the first previews market of user before signing up   
**signup\_app** the application user used to sign up   
**first\_device\_type** the device user used to login for the first time   
**first\_browser** the browser user used to login for the first time   
**country\_destination** the target variable to predict

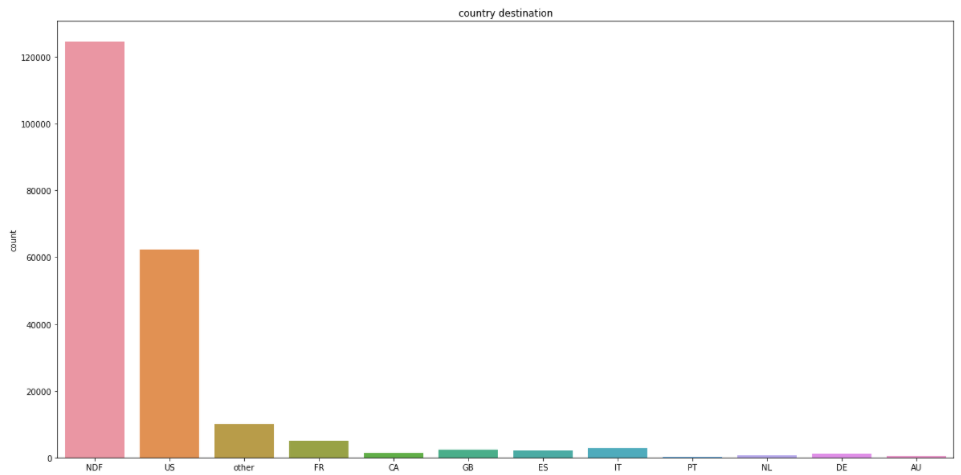
The flat files used for analysis

(1) Train. csv: This dataset has 213451 rows with 16 features



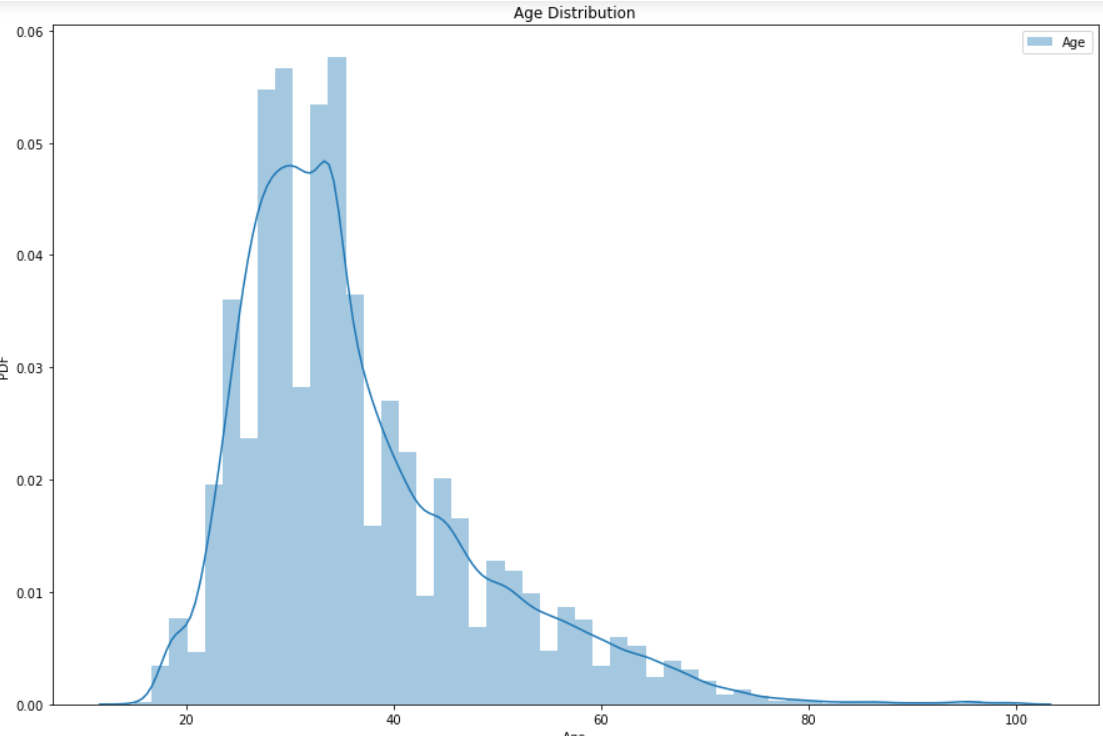
(2) Test.csv: the test dataset has 62096 rows with 16 features

(3) countries.csv - This dataset contains details like latitude and longitude of ten different destination countries. The graph below shows the count of visitors for each of the 10 listed destination countries



***Fig 1: User specified country destination***

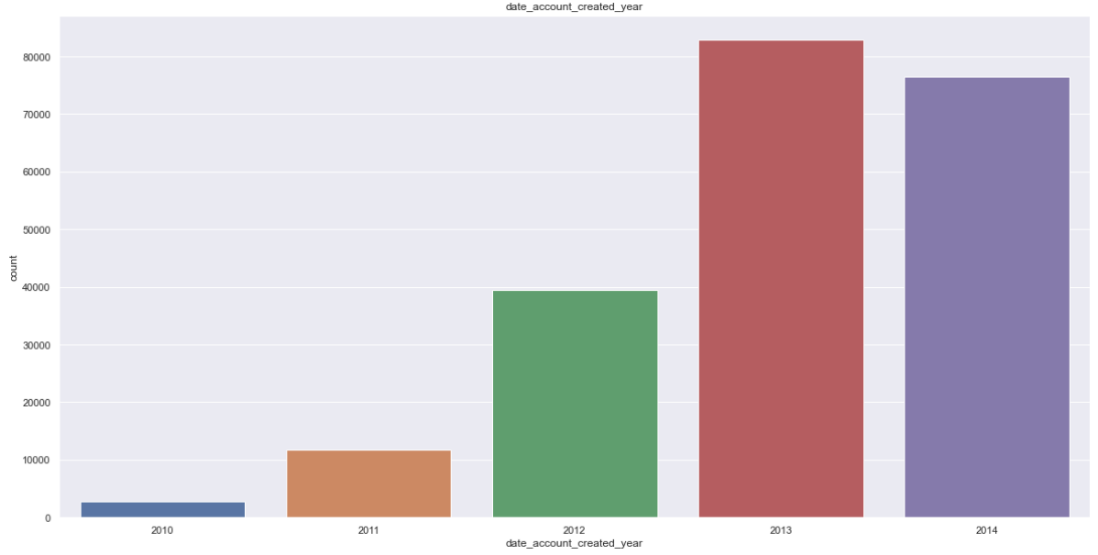
The above bar plot shows most users proceeded without specifying the destination country  
  
(4) age\_gender\_bkts.csv – Contains demographic details for different age brackets. The graph below shows the probability distribution function for different age groups



***Fig 2: User’s age distribution***

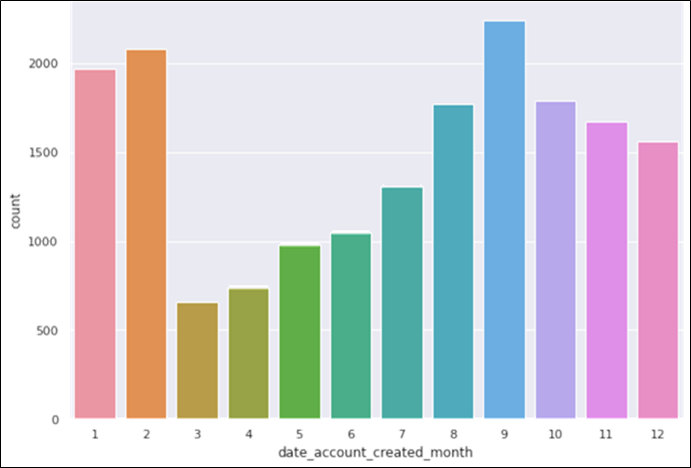
The largest share of Airbnb users lies in middle-aged (30-59), followed by the youth (under 30) and lastly the senior (above 60).

(5) sessions.csv - web sessions log for users: 1048575 rows with 6 features. The graph below shows that the greatest number of accounts were created during the year 2013 followed by 2014.



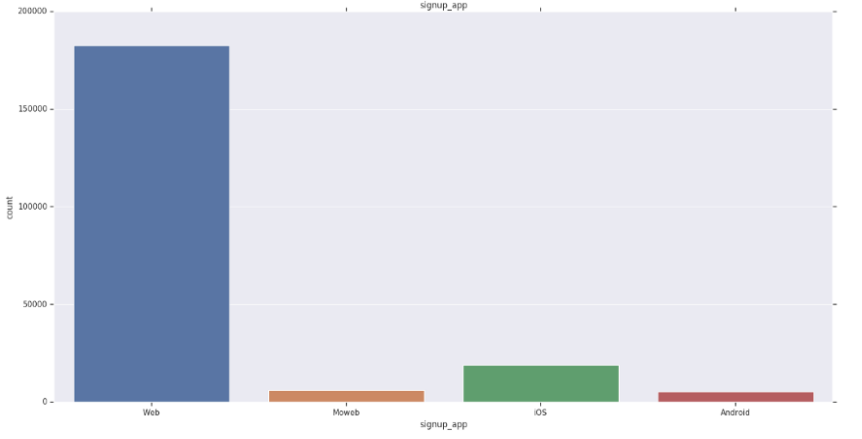
***Fig 3: Year of account creation***

The graph also shows that Airbnb’s popularity has increased exponentially from 2010 to 2014. The dip in 2014 indicates test data has been taken from 2014.



***Fig 4: Date of the month for account creation***

As can be seen from the above plot, most accounts created in May and June. Also, the least number of accounts are created in March and April.



***Fig 5: Signup method***

The above graph shows that majority of users’ signup app is ‘Web’ while Moweb and Android have the lowest share

1. **Data wrangling and Feature Engineering**

Before we proceed towards prediction of the response variable using the mentioned algorithms, we perform feature engineering which is one of the most important steps in any analysis that uses machine learning techniques is feature engineering. This step involves using domain knowledge of the data to create features that make machine learning algorithms work which is fundamental to the application of machine learning.   
  
As a part of data wrangling and feature engineering, we perform the following steps:

* Check null values in the train dataset. Check and drop null values in the sessions dataset keeping user\_id as the primary key
* Concatenate the sessions by user\_id field with fields action,action\_type, action\_detail, device\_type, secs\_elapsed as lists with details of every transaction for one single user\_id
* Convert the lists in the concatenated sessions data into strings
* Merging sessions and training dataset to include KPIs present in the dataset. 35% of users present in the train dataset have details in the session dataset. The merged dataset has 73815 rows and 22 columns
* Merging sessions and test dataset. 99% users in the test dataset have information present in the session dataset. The merged dataset has 62096 rows and 22 columns
* Headers like Day, month, year of date\_account\_created added to new train and test dataset
* Timestamp\_first\_active added to new train and test dataset
* Outlier treatment performed on the age column. Ages above 100 and below 15 replaced by median age 34
* Age buckets of width 5 created basis age column
* To remove untracked entries, if column first\_affiliate\_tracked = untracked, the entries have been replaced by NA
* Column “Country\_Destination” is our response variable and has been renamed as y
* Redundant columns removed from the train and test dataset
* One hot encoding performed on categorical columns like: gender, signup method, language, affiliate\_channel, affiliate\_provider, first\_affiliate\_tracked, signup\_app, first\_device type and first\_browser. As a result, new train and test dataset have 139 columns each.
* Tokenization of strings performed resulting in comma separated token. TF-IDF vectorization performed on tokens in train/ test dataset
* Train dataset and test dataset finalized by removing redundant columns and stacking the observations generated from TF-IDF. We use these 2 datasets for modelling

1. **Analysis – Methodologies and Results**  
     
   The objective is to identify the best machine learning algorithms used to predict which country a new user will make the first booking on Airbnb. We have used NDCG (Normalized Discounted Cumulative Gain) metric for evaluation of the final prediction model. NDCG measures the performance of a recommendation system based on the relevance of the recommended entries. It varies from 0.0 to 1.0, with 1.0 representing the ideal ranking of the entities. This metric is commonly used in information retrieval and to evaluate the performance of web search engines. Performances of all the models were measured using NDCG. The model technique with the highest NDCG score will be our most effective choice in this scenario. To validate our models, we used 5-fold cross validation. In this way, each user record can be used for both training and validating the model.

***6.1 Logistic Regression***   
  
Logistic regression (LR) is a statistical method like linear regression since LR finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X. However, unlike linear regression the response variables can be categorical or continuous, as the model does not strictly require continuous data. To predict group membership, LR uses the log odds ratio rather than probabilities and an iterative maximum likelihood method rather than a least squares to fit the final model. This means the researcher has more freedom when using LR and the method may be more appropriate for non-normally distributed data or when the samples have unequal covariance matrices. Logistic regression assumes independence among variables, which is not always met. However, as is often the case, the applicability of the method (and how well it works, e.g., the classification error) often trumps statistical assumptions.   
  
On using Logistic Regression on the train and test dataset, the Normalized discounted cumulative gain (NDCG) at a 5-fold cross validation comes out to be **0.8158.**

***6.2 Random Forest***

A random forest is a machine learning technique that is used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. This algorithm consists of many decision trees. The ‘forest’ generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

On using random forest on the train and test dataset, the Normalized discounted cumulative gain (NDCG) at a 5-fold cross validation comes out to be **0.9511**.

***6.3 XGBoost***

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now.

On using XGBoost on the train and test dataset, the Normalized discounted cumulative gain (NDCG) at a 5-fold cross validation comes out to be **0.9066**.

1. **Model Comparison**

Normalized discounted cumulative gain is calculated for each of the methods used. The NDCG score is simply the ratio of the participant's Discounted cumulative gain (DCG) at rank K over the ideal ranking's DCG score. Thus, the NDCG metric can be interpreted as the extent to which a user submitted ranking agrees with the ideal ranking, considering the relevance of each element in that list of things to rank. . Among Logistic Regression, Random Forest and XGboost, the best NDCG score is obtained on Random Forest.

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Logistic Regression** | **Random Forest** | **XGBoost** |
| **Score** | **0.8157** | **0.9511** | **0.9066** |

***Table 2: Model Comparison***

1. **Conclusion and Recommendations**  
     
   From the exploratory data analysis and model development/ evaluation performed, we have the following key takeaways and recommendations:

* Random Forest has been identified as the best technique for predicting new users’ first booking destination of Airbnb.
* The platform can use predictions obtained from this technique to personalize the recommendation information to different users to improve the booking rate.
* Highest number of Airbnb users lies in middle-aged (30-59), followed by the youth (under 30) and lastly the senior (above 60). The marketing team can use these statistics as a good reference for distributing marketing resources.
* Since the dataset shows destination countries that are more popular with the users, Airbnb can implement targeted marketing only specific to these popular destinations. This means, focusing marketing strategies for these specific countries to the users identified in the above exercise.
* Most users on mobile devices belong to the NDF outcome (not booking any accommodation) than web users, which means Web users more end up booking an accommodation. This could indicate the need for improving user experience on mobile devices, especially for Android and Moweb users.
* Airbnb can perform clustering based on the booking patterns of the user, i.e., different destinations can be bucketed basis user bookings
* Users at the explosion stage do not fill out a lot of personal details. New users like these should be given great offers to ensure their conversion