University of Illinois Urbana-Champaign
Department of Statistics

Airbnb New Users Booking Prediction

Group Members:
***Lead**: Anushree Vilas Pimpalkar (avp4)*
Shubham Mehta (mehta45)
Umesh Karamachandani (umeshk2)

A project report submitted for the course of
STAT-542 Statistical Learning

Content

Introduction **3**
- Goal 3
- Approach 3
- Conclusion 3

Literature Review **4**

Data Description **4**
- Training Set of Users 4
- Web sessions log for the users 5
- Summary Statistics of destination countries 5
- Summary statistics of users' age group, gender, country of destination 5

Exploratory Data Analysis **5**

Evaluation Metric **7**

Data Preprocessing and Feature Engineering **8**

Modeling **10**
- Naive Bayes 10
- Logistic Regression 10
- Random Forest 10
- Extra Trees 11
- XGBoost 11
- LightGBM 11
- Combined Result 11

Inference **12**

References **13**

1. Introduction

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. 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.

1.1. Goal

In this project, our objective of the project is to predict the top 5 destination countries that a new user on airbnb will likely book. The project also compares and contrasts the performance of different machine learning models for the given multi-class classification problem. These model performances are evaluated with the normalized discounted cumulative gain score.

1.2. Approach

A list of users along with their demographics, web session records, and some summary statistics were given. Our goal is to predict the top 5 destination countries a new user's first booking destination will be. All the users in this dataset are from the USA. First we combined the users and sessions data then we cleaned the data extensively by treating outliers and missing values. We performed extensive data preprocessing steps like tf-idf vectorizing, count vectorizing, ordinal encoding, one hot encoding before applying machine learning models. Results obtained through different modeling techniques using the final data were compared. While understanding and interpreting these results obtained with the models, we extracted some important features that play the most important role in identifying the choice of destination country for a user.

1.3. Conclusion

Boosting models like XGBoost and LightGBM gave the highest accuracy. Finally we interpreted the feature importance from these 2 models. Features like Gender, Age, signup method, presence of certain action types in the user sessions history appear to be significant features in predicting the destination for a new user.

2. Literature Review

This dataset is a part of a kaggle competition hosted a few years ago. There are similar datasets and competitions available on the platform. One similar dataset was mentioned in [1], which proposed an accurate approach for the multi-classification prediction task: Allstate Purchase Prediction Challenge(APPC). This competition, hosted by Kaggle from February, 2014 to May, 2014 is quite similar to our prediction task. A potential customer who shops an insurance policy, receives a number of quotes with different coverage options before purchasing a plan. The goal is to predict the purchased coverage options using a limited subset of the total interaction history. Their work involves an interesting method of prediction, called ”Voting Mechanism” (VM), which proposes that, for a given testing data, there could be good predictions in a bad model, while having a bad prediction in a good model. Therefore, combining several models resulted in better results adopting VM. The author was able to achieve the best score of 0.53266, while the best score of the winner for the competition was 0.53743, suggesting that even a slight improvement in predictions contributed to a huge jump in the score and the rankings.

Resembling closer to our work, [2] performed similar analysis on the same dataset by implementing a two-level classification model. Their first level was a binary classifier combining linear, logistic and polynomial regression to Voting Mechanism (VM). Their next level was a multiclass classifier which was the combination of SVM and multiclass one-against-rest logistic classification. Their work included baseline description, feature engineering and representation, model selection, reasoning and description, and parameter tuning. Among three models that they performed, polynomial regression performed the best. Whereas, linear regression and logistic regression did not perform well. They found that the three models when combined into the Voting Mechanism, outperformed them all. They also found that the vector feature representation of 0,1 performed better than digit feature representation on most features within the data. For example, a better way to represent the “age” feature is to split it into intervals of size 5 and then represent it with a vector of 0,1. Here, 1 means that the age is present in the given interval. This method outperformed as compared to when the feature “age” was used as a number directly as a feature.

3. Data Description

3.1. Training Set of Users

	id	date_account_created	timestamp_first_active	date_first_booking	gender	age	signup_method	...	country_destination
0	gxn3p5htnn	2010-06-28	20090319043255	NaN	-unknown-	NaN	facebook	...	NDF
1	820tgsjxq7	2011-05-25	20090523174809	NaN	MALE	38.0	facebook		NDF

Training data captures the users demographics like their age, gender, date of account creation, signup method etc. along with response 'destination country' that the user has booked. In this dataset, with all users from the 'US', we have 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'.

3.2. Web sessions log for the users

	user_id	action	action_type	action_detail	device_type	secs_elapsed
0	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	319.0
1	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	67753.0
2	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	301.0

Web sessions capture how the various users have interacted with Airbnb platform. It tracks important information like time spent by a user in a particular session and what action & its details a particular user has taken while interacting with the platform. In addition to that it also captures the device that a particular user has used for that session.

3.3. Summary Statistics of destination countries

	country_destination	lat_destination	lng_destination	distance_km	destination_km2	destination_language	language_levenshtein_distance
0	AU	-26.853388	133.275160	15297.7440	7741220.0	eng	0.00
1	CA	62.393303	-96.818146	2828.1333	9984670.0	eng	0.00

This data captures the latitude, longitude, distance from the US, area and language of the destination.

3.4. Summary statistics of users' age group, gender, country of destination

	ucket	country_destination	gender	population_in_thousands	year
	100+	AU	male	1.0	2015.0
1	95-99	AU	male	9.0	2015.0
2	90-94	AU	male	47.0	2015.0

4. Exploratory Data Analysis

In order to understand the data better and to check which factors could play an important role in deciding the choice of destination, some basic data analysis was done. Majority of the users have not booked any destination. United States was the most preferred destination among the users. Since the data has users only from the United

States, this was a bit expected as a person is more likely to travel within the country as compared to a foreign country.

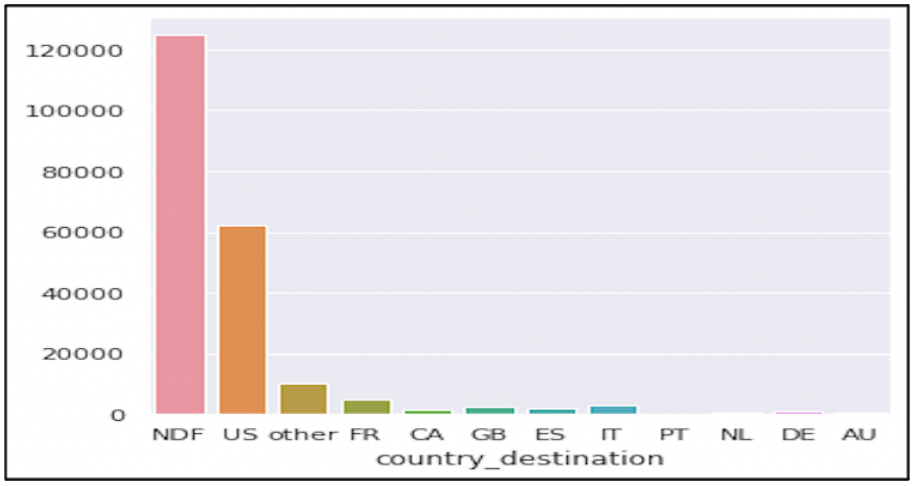

Fig 1 : Users distribution across various destination countries

The effect of gender and age of the users on the choice of destination was explored. It was observed that users who have not given gender information (‘Unknown’ Gender) did not book any destination. This could mean that such users are not serious about booking on the platform. The median age of the users did not vary much across various destination countries. Great Britain has the highest median age and Spain is more popular amongst younger travelers

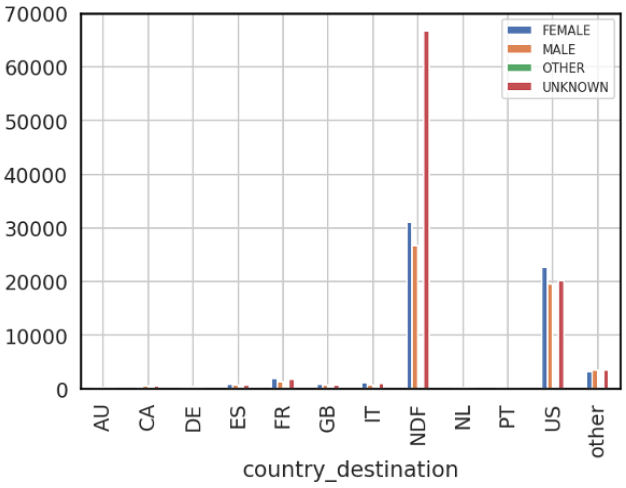

Fig 2: Gender vs. Destination Country

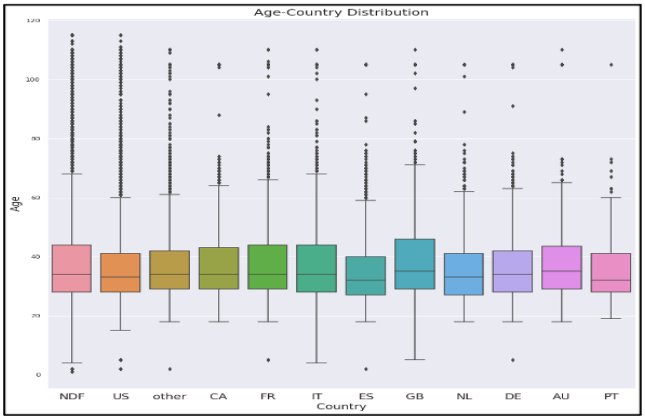

Fig 3: Age vs. Destination Country

The effect of users’ web activity on the choice of destination country was also explored as it could play an important role in decking the destination country. Fig 4, shows how time per session of users vary across various destination countries. Average time per session is lowest for users who did not book any destination. Average time per session is highest for users who booked the United States, Great Britain as destinations. Thus, this metric shows good variation across various destinations users and could be very helpful in predicting the destination country at the modeling stage.

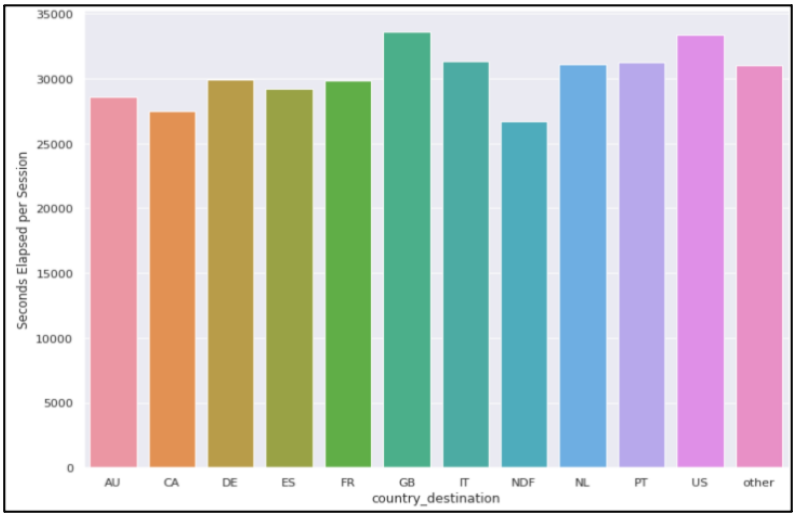

Fig 4: Average time per sessions vs. destination country booked

Similarly, the variation in the number of users session across the destination country was also explored. Average number of sessions is highest for users who booked France, Italy as destinations. Average number of sessions is lowest for users who did not book any destination.

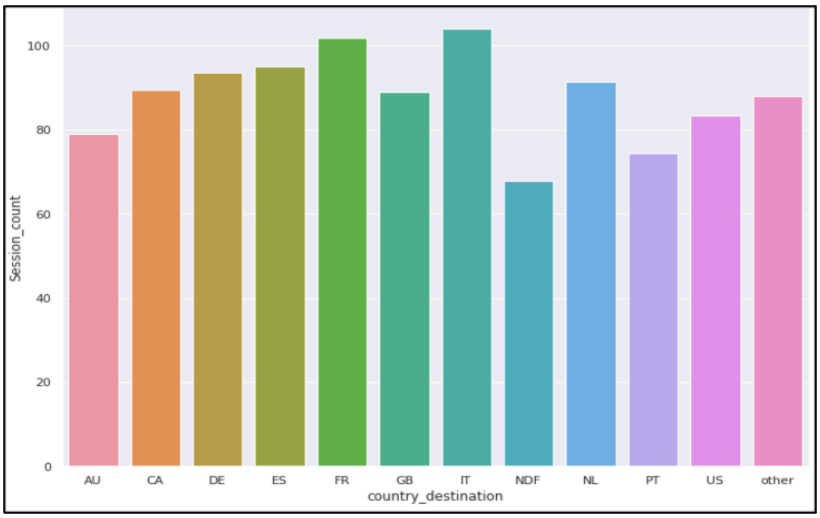

Fig 5: Average session count vs. destination country booked

5. Evaluation Metric

Since the goal is to predict top 5 destinations a new user is likely to book; the rank of the relevant destination is an important consideration. Thus we can't simply use accuracy as the classification metric. Instead we used Normalized Discounted Cumulative Gain (NDCG) as the evaluation metric which takes care of the ranking of the ground truth into context. Further details on how the metric is evaluated is described below:

$$\mathrm{DCG} = \sum_i \frac{2^{\mathrm{rel}_i} - 1}{\log_2(i+1)}$$

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}}$$

where, IDCG is the maximum possible (ideal) DCG. The ground truth country is marked with relevance of 1, while the rest have relevance of 0

Predictions (For Ground Label : FR)	Score
[FR,US,GB,IT,NZ]	1
[US,FR,GB,IT,NZ]	0.63
[US,GB,FR,IT,NZ]	0.5
[US,GB,IT,FR,NZ]	0.43
[US,GB, IT,NZ,FR]	0.39

Table 1: Sample predictions vs. NDCG Score

The above table shows how the position of the ground truth affects the NDCG score. As the ground truth “FR” position moves away from its actual position the NDCG score decreases too.

6. Data Preprocessing and Feature Engineering

Before modeling, we performed some data preprocessing and Feature Engineering. The training data has an “age” column. The values in this column have a huge range, with the maximum value as 2014. Clearly, the data in the “age” column is not right. Therefore, instead of removing the rows with these outliers in the “age” column, the values above 120 were changed to 120. The figure below provides an intuition to select this particular value.

The web sessions data contains information about the customer interaction with the airbnb platform. To capture the interaction with the website/app, for every user, summary statistics like mean time elapsed, number of unique actions and number of sessions were created from seconds elapsed column. Additionally, unique values for all the interaction types like actions, actions type, action details were extracted. To capture the importance of these Actions performed w.r.t booking, the text data is vectorized through Term Frequency – Inverse Document Frequency (TFIDF), which works by proportionally increasing the number of times a word appears in the document. An important thing to note here is that TFIDF penalizes the rank by the increase in the number of documents in which that particular word is present. Therefore, words like ‘the, ‘is’ etc., which are very common in any document will not be given a very high rank. Moreover, a word which is present in the document too many times will be given a higher rank, because such words can indicate the context of the document.

The term frequency is calculated as:

$tf_{i,j}$ = $\frac{n_{i,j}}{\sum n_{i,j}}$, where $n_{i,j}$ is the number of times a word (i) appears in a document (j) and the denominator is the total number of words in the document.

The inverse document frequency is calculated as:

$idf(d) = log\frac{N}{d_n}$, where N is the total number of documents and d_n is the number of documents that contain the word.

TFIDF is then calculated by multiplying the term frequency with the inverse document frequency.

$$TFIDF = tf_{i,j} * idf(d)$$

Bag of Words was used to process the device type column. These pre-processing techniques resulted in the final web session data with summary statistics and vectorized column data. This was then merged with the training data to have this information for every user. The account created date in the training data was transformed to year, month and day, to capture the information or influence of these in the choice of destination for the users. The final data thus had 413 columns.

Encoding:

Since most machine learning models accept only the numeric variables, preprocessing the categorical variables is very important, so that the model is able to understand and extract valuable information from them. We have multiple categorical variables in our data. Therefore we performed the following encoding methods before moving on to the modeling.

1. Ordinal Encoding
2. One-hot encoding

Ordinal Encoding: This type of encoding is used when the categorical variables in the data are ordinal. Ordinal encoding converts each label of a particular categorical variable into integer values and thus the encoded data represents a sequence of labels. Here, ordinal encoding was implemented in the Tree-based algorithms. Using ordinal encoding in models like logistic regression, naive bayes etc can result in the algorithm assuming some natural order within the sequence of labels for any particular categorical variable.

One-hot Encoding: This type of encoding technique is used when the features in the data do not have any order. In one hot encoding, for each level of a categorical feature, a new variable is created, and each category is mapped to a binary variable containing

either 0 (representing the absence of that category) or 1 (representing the presence of that category). These newly created binary features are known as Dummy variables and the number of dummy variables depends on the levels present in the categorical variable.

7. Modeling

7.1. Naive Bayes

This is a classification algorithm based on the bayes theorem. Naïve bayes classifier assumes that there is conditional independence between every pair of features given the value of the class variable.

7.2. Logistic Regression

Logistic Regression is a Supervised machine learning algorithm that can be used to model the probability of a certain class. It can be used for both binary and mutli-class classification problems. One of the disadvantages of logistic regression is that it can only separate the classes linearny and can not make non-linear decision boundaries.

7.3. Random Forest

A random forest algorithm trains many decision trees on various sub-samples of the data (creating a forest) and uses them as a classifier. It uses averaging to improve the predictive accuracy and control over-fitting. Important hyper-parameters like node size and number of features were tuned by using k-fold cross validation.

7.4. Extra Trees

Extremely Randomized Trees Classifier(Extra Trees Classifier) is an ensemble learning method which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output its classification result. It is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest. This algorithm trades more bias for a lower variance. It also makes ExtraTrees much faster to train than regular random Forests, because finding the best possible threshold for each feature at every node is one of the most time-consuming tasks of growing a tree.

7.5. XGBoost

XGBoost (Extreme Gradient Boosting) belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core. It combines several weak learners (mostly decision trees) into a strong learner. It tries to train predictors sequentially, correcting the mistakes of its predecessor.

7.6. LightGBM

LightGBM (Light Gradient Boosting machine) belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core. It combines several weak learners (mostly decision trees) into a strong learner. It tries to train predictors sequentially, correcting the mistakes of its predecessor. The difference between LightGBM and XGBoost is in the way the decision tree is constructed.In XGBoost, trees grow depth-wise while in LightGBM, trees grow leaf-wise. This makes LightGBM much faster than XGBoost

7.7. Combined Result

Model	Training: NDCG Score	Testing: NDCG Score
Naive Bayes	0.447	0.444
Logistic Regression	0.686	0.806
Random Forest	0.741	0.734
XGBoost	0.867	**0.858**
LightGBM	0.883	**0.858**
ExtraTrees	0.894	0.847

Table 2: NDCG Score comparison across 6 different models

- Naïve Bayes gives the lowest NDCG score on training and testing data. This is because it assumes that features are independent, which is not true in most cases.

- Logistic regression produces linear decision boundaries resulting in lower accuracy as compared to LightGBM / XGBoost / ExtraTrees

- ExtraTrees, XGBoost and LightGBM produces better accuracy among these 6 models.

8. Inference

It is essential to determine which are the factors that are most relevant to predict the next booking location of the user. The importance of the features has been extracted through the feature_importance_ attribute of the tree based models. In this study LightGBM and XGBoost have been used for this purpose. The following two plots depict the order of features(y-axis) based on decreasing order of importance.

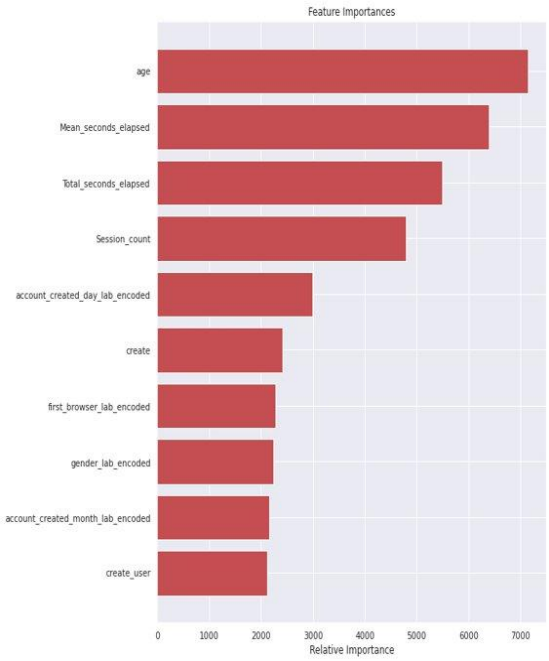

Fig 6: LGBM Feature Importances

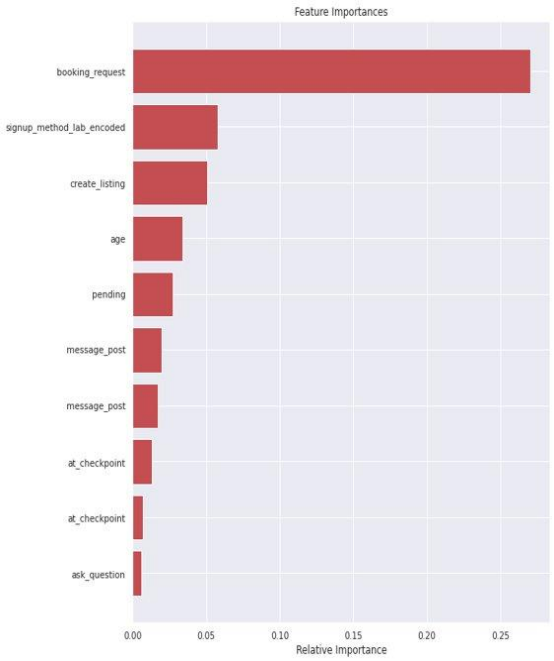

Fig 7: XGBoost Feature Importance

Number of interpretations have been made on the basis of the above plots. **Age** Comes out to be among the important features since the absence of it may indicate that the user is not serious about the booking and the category “NDF” may be preferred for the customer. It can also be observed that metrics like average time per session, number of sessions play a significant role in deciding the destination since a serious user would spend a higher amount of time looking at the details of the site, the customer may want to book. It can also be interpreted that actions like booking requests when performed by the customer may indicate that he is closer to completing the process of booking and thus the category “NDF” would be less preferred for that customer. Metrics like signup method and gender also appear to be deciding factors.

9. References

[1]https://aicsjournal.files.wordpress.com/2014/01/a049-predicting-purchased-policy-saba-arslan.pdf

[2]https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/038.pdf