Predictive Modeling For New User Bookings

Nick Kessler

Overview

Users of online services supply a wealth of information about themselves to providers of these services. Attributes such as age, gender, race, and location are often supplied by the users themselves, but many other attributes about the user are also collected in the background simply through the user's interaction with the service. Providers can garner all sorts of facts about a user such as what brand of computer or phone a user chooses, what social networking sites they use, or which browser or search engine the user prefers. Beyond this, statistics about how the user interacts with the site, such as how many clicks or searches a user engages in or how long a user's session lasts, can also be harvested through standard logging. All these data points might seem unassuming, but in aggregate they can be used to create a detailed profile of a user's preferences, and they may even hold clues about a user's future choices.

Airbnb is the world's largest online marketplace for peer-to-peer lodging and hospitality. Like other online services, a user provides a wide assortment of information, both actively and passively, about themselves when they create a profile.

In this project, I aim to determine to what degree this user supplied information can be applied to infer where a user will select as the destination for their first booking using Airbnb.

Data Preparation

Data for this project was provided by Airbnb as part of a Kaggle competition.

Data Files

File	Description
train_users.csv	Training set of user data. Supplied for model training. The training set includes user data going back to 2010.
test_users.csv	Test set of user data. Supplied for generating contest submissions. This test set does not include the target variable, *country destination*, as a column. The test set includes only users who signed up after 7/1/2014
sessions.csv	Web sessions log for users. Web session data is only available from 1/1/2014 onward.
countries.csv	Summary statistics about some of the target destinations.
age_gender_bkts.csv	Summary statistics showing number of users by age group, gender, and destination country.
sample_submission.csv	A sample contest submission.

User Data Set

The user data set provides information on the users themselves. This includes user-supplied information such as gender or age, as well as browser supplied information such as operating system, device type, and affiliate providers if applicable. This data is composed primarily of categorical variables.

User Data Set

Column	Type	Description
id	String	User identifier.
date_account_created	Date	Date of account creation.

timestamp_first_active	String	Timestamp of first activity. This can occur before account creation as users can browse the site before signing up. Stored as a string value in the dataset.
date_first_booking	Date	Date of first booking. This value is null if the user never booked a trip.
gender	Categorical	User supplied gender. Male, Female, Other or Unknown.
age	Numeric	User supplied age value.
signup_method	Categorical	User signup method. Indicates whether the user created a direct Airbnb account, or signed up with their Facebook or Google account.
language	Categorical	User language preference
affiliate_channel	Categorical	Type of paid marketing, if applicable.
affiliate_provider	Categorical	Where marketing came from. E.g. Google, craigslist, etc.
first_affiliate_tracked	Categorical	First marketing the user interacted with prior to sign up.
signup_app	Categorical	The app the user used to sign up. E.g. Web, iOS, Android.
first_device_type	Categorical	The device type used to sign up. E.g. Windows PC, Mac, iPhone.
first_browser	Categorical	Web browser used to sign up, if applicable
country_destination	Categorical	The destination chosen by the user for their first booking. This is the target variable in this analysis.

The session data provides information about the user's web sessions, including actions such as clicks and searches, device type, and time elapsed. This session data is only available for logins after 7/1/2014 and thus only is available for roughly 1/3rd of the user IDs in the User dataset. The use of the session data did not prove useful in improving model accuracy.

Session Data

Column	Type	Description
user_id	String	User identifier. Corresponds to id column in User dataset.
action	Categorical	Description of the action
action_type	Categorical	Action category
action_detail	Categorical	Detailed action description
device_type	Categorical	Type of device (e.g. Windows Desktop, Mac, iPhone)
secs_elapsed	Numeric	Number of seconds elapsed in session.

Data Cleaning

Though the dataset was mostly pre-cleaned and formatted, some additional cleaning steps needed to be applied before the data could be used.

- The *gender* column had many missing values, as well as values labeled "-unknown-". To improve consistency, both missing values and "-unknown-" values were replaced with "UNKNOWN", creating four unique possible values for *gender* – MALE, FEMALE, OTHER, and UNKOWN.
- Some of the *age* values contained extreme outliers, such as 1960. This is likely because the user entered their birthyear instead of their age. To correct these errant age values, ages above 1000 were changed to calculate the *age* by subtracting the supplied birthyear from the current year.

- The *timestamp_first_active* values were supplied as strings and needed to be converted to datetimes before I could use this data.

Feature Engineering

While the user data provides a wide assortment of data about the users, to provide a more complete picture of the users, additional features were derived from the existing features provided in the User dataset. The following features were derived from features in the main dataset.

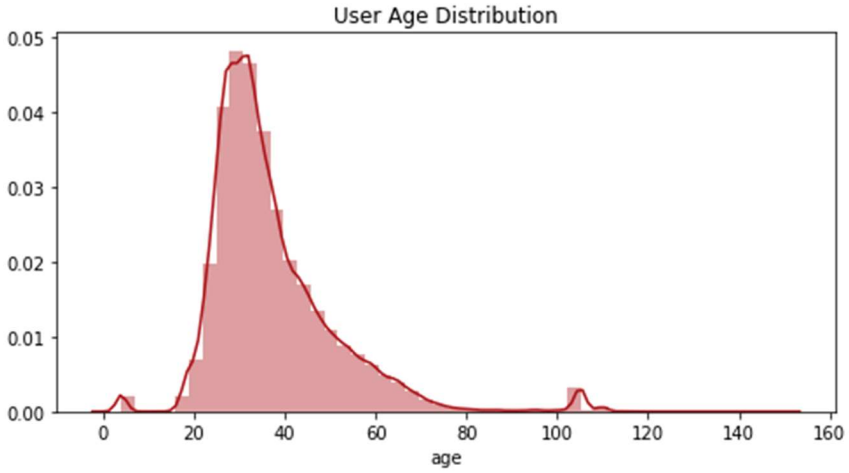
Derived Features – User Data

Column	Type	Description
age_group	Categorical	Age-range group. This feature was designed to alter age from a numerical feature into a set of 6 groups.
days_to_book	Numeric	Because date features cannot be used by Scikit-learn classifiers, a derived feature to determine the time it took a user to book was added. Days to book is the difference in days between the account creation date and the date of first booking. For users who never booked a trip, this value is not available.
device_class	Categorical	The user dataset provides a large assortment of device type categories. To better pinpoint the specific attributes of the user, a derived column for device class was added. Possible values are *PC, phone, tablet,* or *other*.
affiliate_provider_type	Categorical	Many different affiliate providers are available in the user data. This derived feature attempts to group the possible affiliate providers into categories. Possible values are *direct, search engine, social media,* or *other.*

Exploratory Data Analysis

Before running any modeling on the data. I first wished to examine the data visually and see if any apparent patterns were observable.

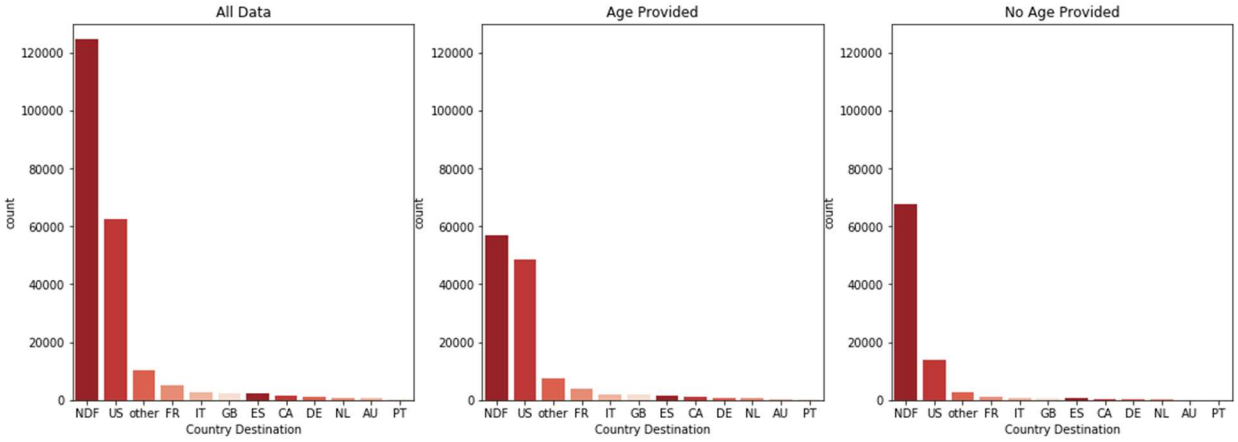
To begin, I looed at the distribution of users by age, after correcting for values where the user entered their birthyear instead of age.



The age distribution plot shows that the bulk of the users are aged between 20 and 40, with a long tale between ages 40 and 80 before dropping off. The plot also reveals a small number of outliers with a handful of observations showing users younger than 10 and older than 100. These values are likely errant entries but are miniscule in quantity. This distribution is in line with expectations, as users younger than 18 are not able to book on Airbnb, and the distribution reflects what I expect for a web-based hospitality booking service.

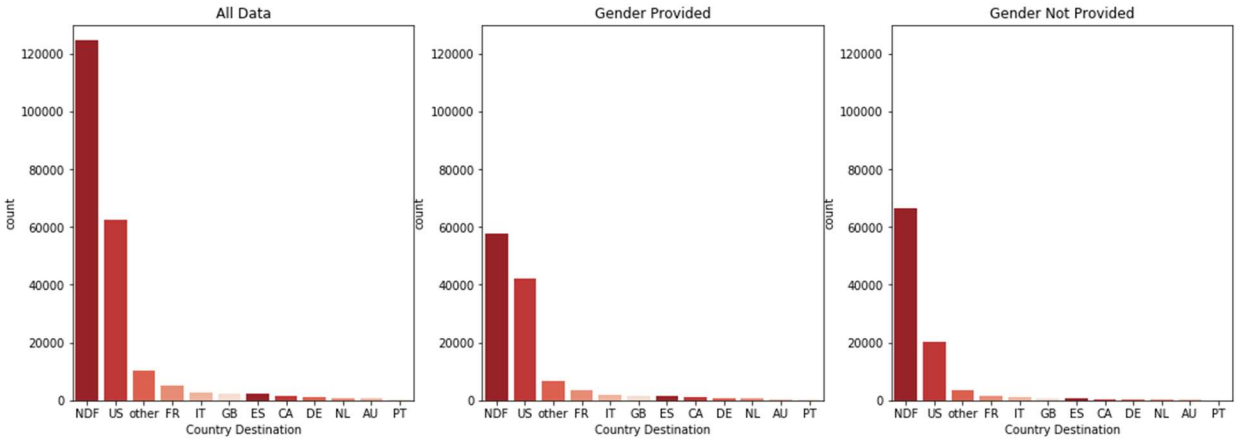
When initially exploring the data, I noticed that there was a large quantity of missing values for attributes like age and gender. I wanted to examine to see if missing data was correlated with the destination chosen (or no destination chosen).

Destination – Age Provided or Not Provided



The resulting plot is revealing. While NDF (No Destination Found, i.e. the user did not book) was the top result for all data, being nearly double the top booked destination which is the United States, when selecting only users where an age is provided, we see that NDF is nearly equal to the top destination and all destinations combined outnumber the NDFs. However, when looking at the users where no age was provided, we can see that NDF is by far the most common result and outnumbers all other destinations combined. From this, we see that users who supply age information are far more likely to have booked a trip than those who did not supply an age.

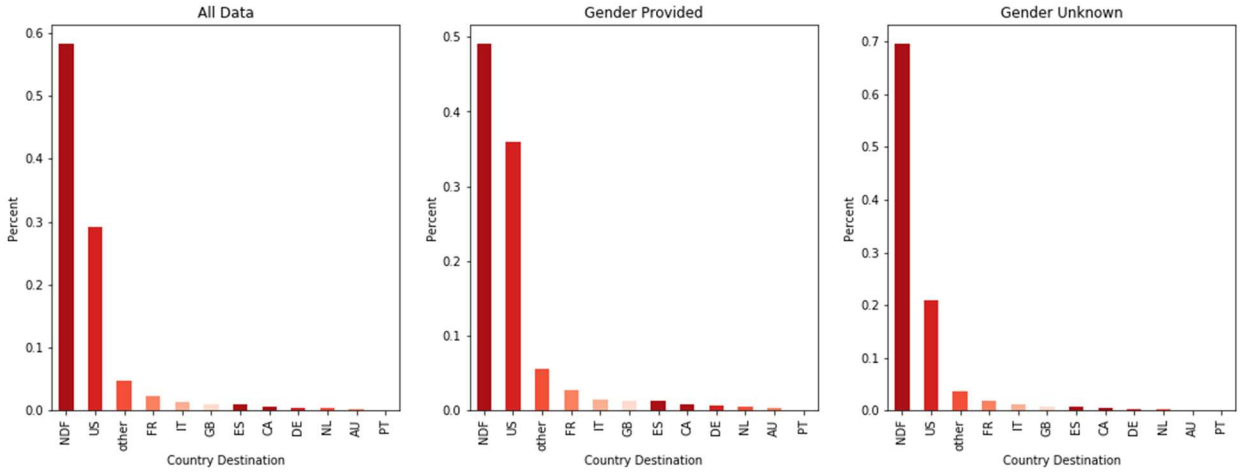
Destination – Gender Provided or Not Provided



Running a similar plot for the gender value shows a similar result. While not quite as dramatic as with the age value, we can see that users who supplied a gender were far more likely to book a trip than those who did not supply a gender.

Breaking this down by percent further illustrates this point.

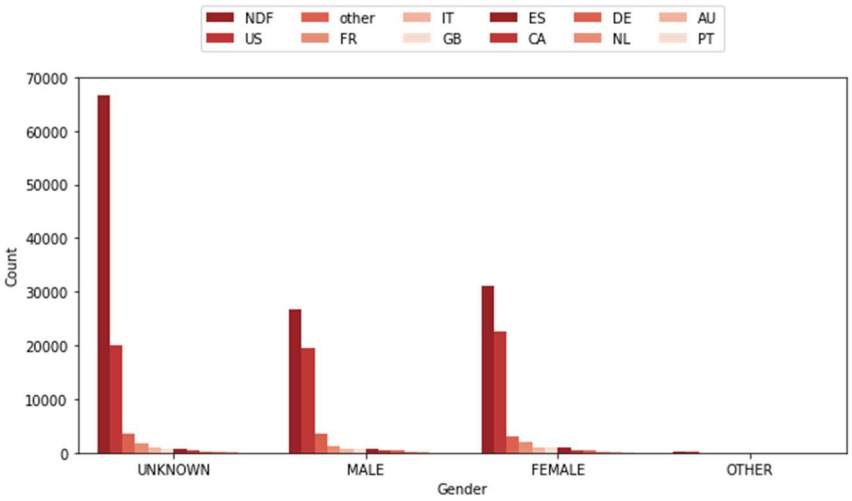
Destination – Gender Provided or Not Provided - Percent



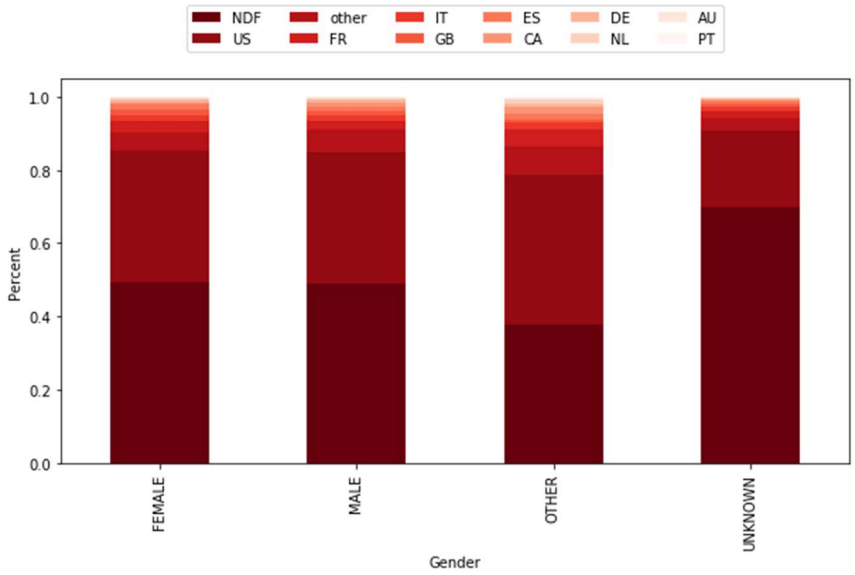
In the above visualizations, we can see that users who supply more information about themselves when signing up, such as gender and age, are much more likely to book a trip after signup than users who do not supply this information. This clue indicates that the more willing a user is to share data about themselves, the more likely it is that they're serious about booking through Airbnb.

I also wanted to see if there was a relationship between the user's age group or gender and the destination country chosen.

Destination by Gender – Count



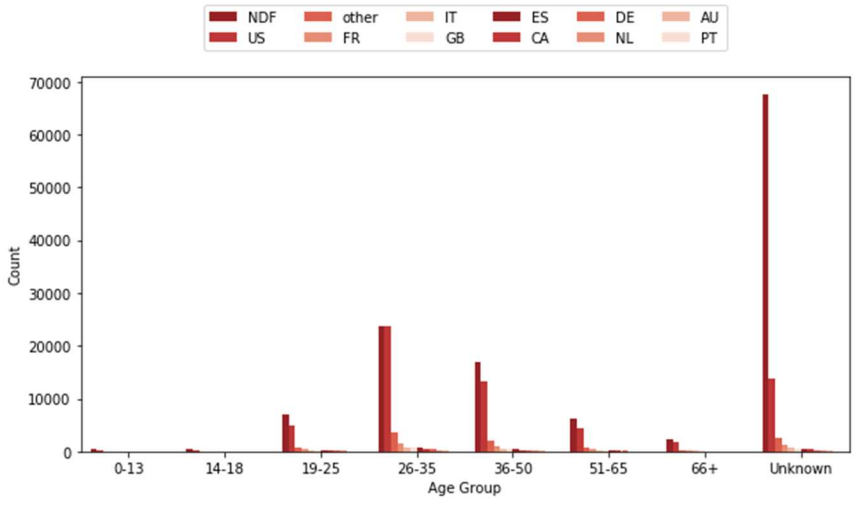
Destination by Gender - Percent



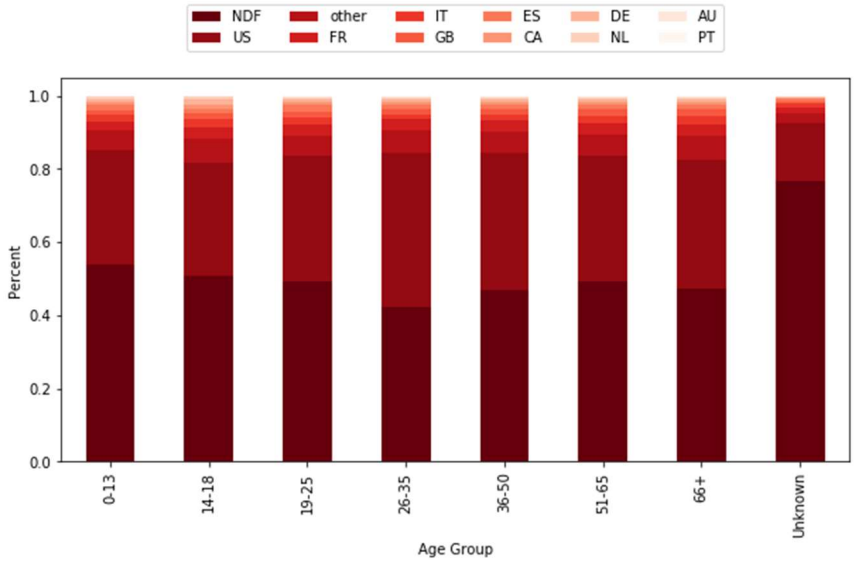
The above plots show that there's very little difference between male and female destination choices. Users who selected Other as their gender do appear to be somewhat more likely to book a trip, however as seen in the count plot above, this number is so miniscule that it's likely not significant enough to infer any conclusions. Again, we see that users who did not supply a gender were far less likely to book.

Examining destination by age group yielded a similar result.

Destination by Age Group – Count



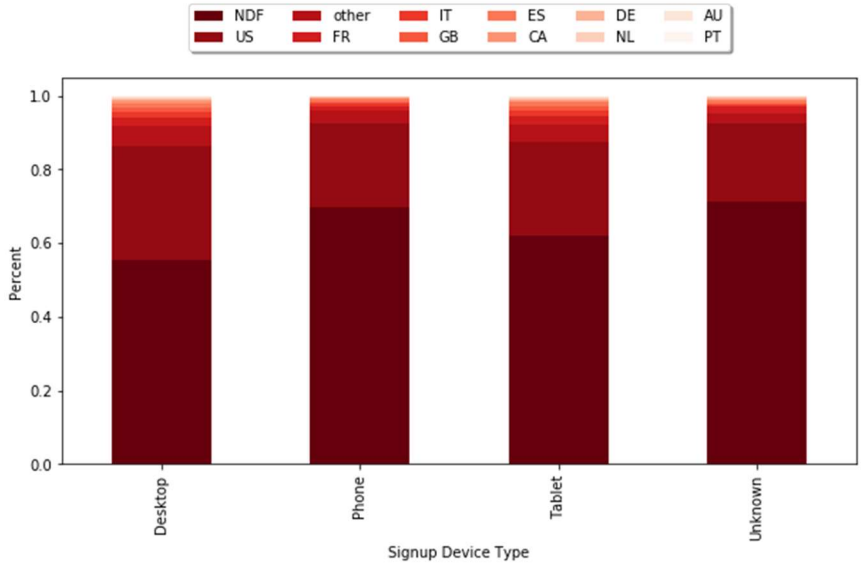
Destination by Age Group - Percent



Once again, it is clear that users who did not provide an age are far less likely to book a trip than those who do. We can also see that users in the 26-35 age group were most likely to book a trip, showing that there is some correlation between age group and destination. However, the plot reveals that most age groups have similar booking outcomes.

I also wanted to look for relationships between some of the other categorical values and destination chosen.

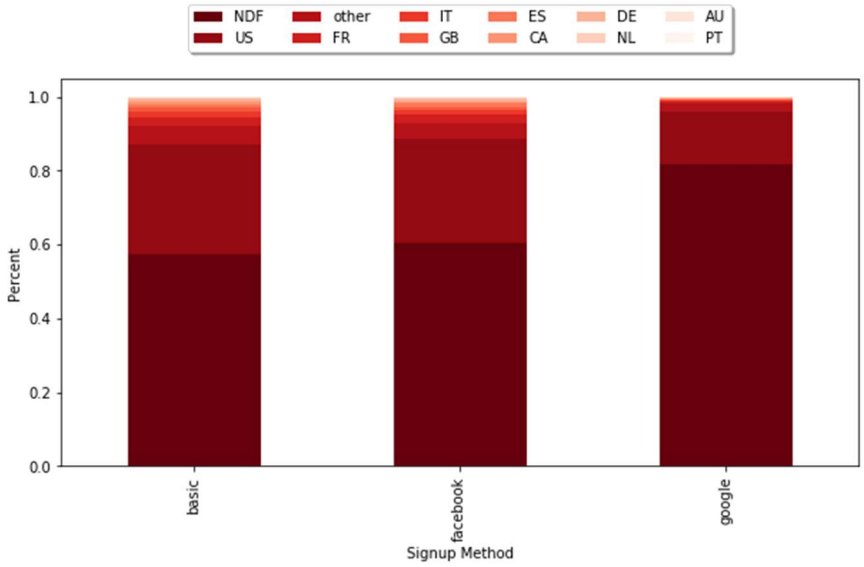
Destination by Signup Device Type - Percent



In the above plot, we see that users who signed up on their desktop or tablet were somewhat more likely to book a trip than those who signed up on their phone. This could indicate that users who sign up on

their phones are only looking, while those that sign up on larger screen devices are more serious about booking.

Destination by Signup Method - Percent



I also looked for a correlation between signup method and destination. We can see that users who signed up with a direct Airbnb account (labeled 'basic') or with Facebook were far more likely to book than those who signed up with their Google account. This is an interesting result though it's not clear from the data why that might be the case.

The above exploratory data analysis indicates that there are clear indicators in the data that reveal whether a user decided to book a trip. Missing information is an apparent sign that the user would not ultimately book a destination. Phone users also seem less inclined to book than users of larger devices. And interestingly, Google users are far less inclined to book than those who signed up using other methods. While the data does make clear which users are likely to book, the visualizations do not reveal any clear patterns between user attributes and the destination chosen. The plots reveal that the distribution of destinations is roughly even between men and women, various age groups, and even different device types. In the next step, I see if our predictive model can better assess the chosen destination based on the data available.

Classification Analysis

Our target variable in this analysis is the destination country selected (or *NDF* for no destination).

Before applying the classifiers to our data, missing values first had to be handled. Most of my data is categorical, so missing values only needed to be considered in the numeric columns for *age* and *days to book.* Observations where the age was missing were filled in with 0s, while missing values for days to book were filled in with 100,000. These extreme values were chosen so that the classifier could easily distinguish between valid values and those that were filled in missing values. Missing gender values were substituted with *UNKNOWN*.

Next, because the data was primarily categorical, one-hot encoding needed to be applied before modeling.

Twelve Category Classification

There are twelve possible values for destination country, so to begin I ran a 12-category classification using various models available in the scikit-learn library. The below table summarizes the results of the various models using the default parameters and a test set size of 10%

Twelve Category Classification Model Results Summary

Classifier	Score	Avg. Precision	Avg. Recall
Decision Trees	0.8116	0.81	0.81
SVM	0.8808	0.80	0.83
Random Forest	0.8494	0.81	0.83
Gradient Booster	0.8808	0.80	0.83
Extra Trees	0.8345	0.81	0.82
Gaussian Naïve Bayes	0.8107	0.81	0.81
MLP Classifier	0.8808	0.80	0.88

All the models achieved a reasonably high degree of accuracy. However, because 87.5% of the data is either NDF or US, this result was largely achieved by the models selecting these values as the target to the exclusion of all others.

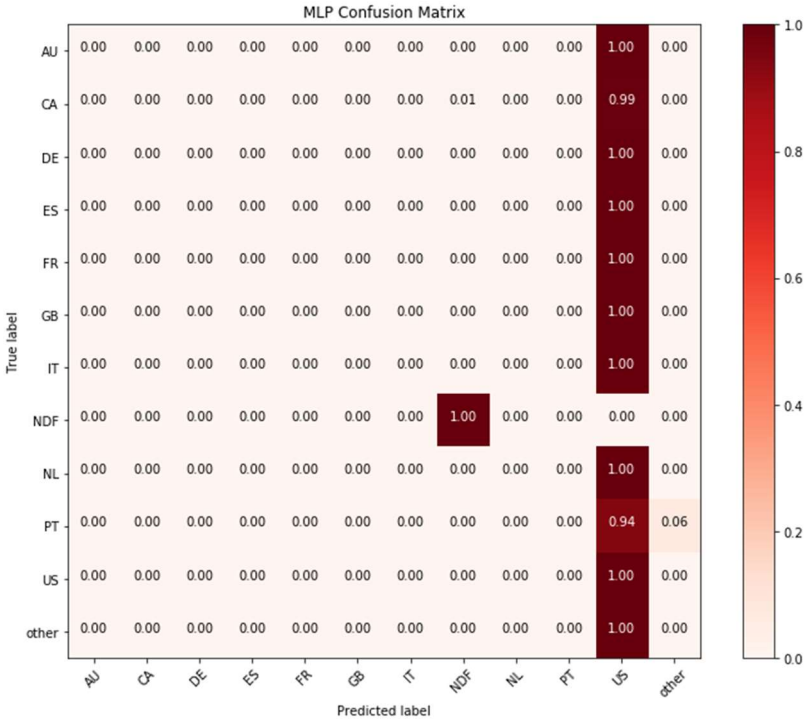
MLP Classifier

I attempted to improve the results, using the MLP classifier which provided the best results with the default parameter values. Using Random Search Cross Validation to refine hyperparameters, I ran the data through the MLP classifier. The below table shows the classification report.

MLP Model Classification Report

Rating	Precision	Recall	F1 Score	Support
AU	0.00	0.00	0.00	59
CA	0.00	0.00	0.00	134
DE	0.00	0.00	0.00	114
ES	0.00	0.00	0.00	204
FR	0.00	0.00	0.00	472
GB	0.00	0.00	0.00	251
IT	0.00	0.00	0.00	286
NDF	1.00	1.00	1.00	12547
NL	0.00	0.00	0.00	57
PT	0.00	0.00	0.00	16
US	0.71	1.00	0.83	6256
other	1.00	0.00	0.00	950
Avg./total	0.84	0.88	0.83	21346

Confusion Matrix – MLP Classifier – 12 Category Classification



As expected, the model was able to perfectly identify which users did not book a trip, as these observations are defined by clear markers. Using the default MLP settings also achieved 100% precision for destination of 'US'. However, for the remaining destinations the model was less accurate. While the overall score was high, this is because the models chose only *US* and *other* as the other possible destination. Nearly all rows that were not identified as NDF were assigned US, thus producing high accuracy at the cost of excluding non-US destinations. As previously stated, these three categories comprise 92.5% of the total data. The model was very accurate at predicting *NDFs* and *US*, but it appears that *US* destinations can't be as easily distinguished from other destinations using the available data.

Other classifiers produced a similar result. The users who did not book a trip were able to be identified with perfect accuracy, and achieved high score, but only by means of predicting US as the destination for all non-NDF destinations.

Session Data

In addition to the features available in the user dataset, session data which logs web session statistics was also available as part of this dataset. This data shows extract from weblogs revealing which actions the user took during their engagement with the website. Each row represents a distinct action taken by the user. With this data, I wanted to see if information about how the user interreacted with the website (e.g. number of clicks, number of searches, etc.) could provide clues that would improve my model's accuracy.

To prepare this data for including in my analysis, I calculated counts for various actions and used these as features. For example, for each possible value of 'action_type' for each user (e.g. *search, click*), I calculated the total count for each action, then returned columns for each action with the count for each user, so that for each user there was a single row. I then joined this to the user dataset.

Because the period covered by the session data is limited compared to the user data, only about 1/3rd of the user IDs in the *user* dataset matched with an ID in the session dataset.

Fitting the joined user and session data to our model did not produce the results I would have liked. Using the MLP classifier, the fitted model proved to be less accurate than when I had fitted using only the *user* data. This may be a result of having fewer observations with which to train the model, or it may be that the features engineered from the session data did a poor job of predicting the target destination.

The Random Forest classifier, though, likewise yielded disappointing results and provided less precision than training with just the user data alone.

Twelve Category Classification Model Results Summary – Incl. Session Data

Classifier	Score	Avg. Precision	Avg. Recall
MLP	0.8096	0.81	0.81
Random Forest	0.8845	0.80	0.88

The inclusion of the session data did not provide the improvements I was looking for. This is likely due to the fact that the session data is incomplete and by using it I had to provide fewer observations to my models. It's also possible that the features I engineered were not useful in predicting the destination country. With more complete session data, it might be possible to get improved results.

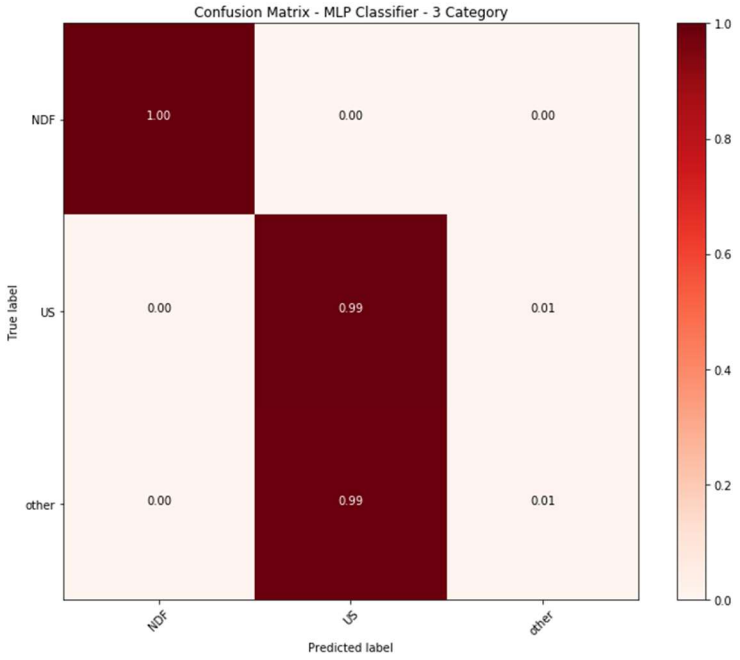
Three Category Classification

Given that the MLP classifier only seemed to be able to correctly identify targets of *NDF, US,* and *other*, I attempted to reformat the data into a 3-category classification problem to see if this resulted in an improvement of model accuracy. I achieved this by putting all destinations other than *NDF* and *US* into the *other* category. The result when using the MLP classifier was very similar to the 12-category classification. Users who did not book a destination (target of NDF) were identified with perfect accuracy, but nearly all other observations were classified as US.

MLP 3-Cateogry Model Classification Report

Rating	Precision	Recall	F1 Score	Support
NDF	1.00	1.00	1.00	12547
US	0.71	0.99	0.83	6256
other	0.32	0.01	0.02	2543
Avg./total	0.83	0.88	0.83	21346

Confusion Matrix for MLP Classifier – 3-category



Applying the selection of models above, including SVM, Random Forest, etc., on the 3-category data yielded nearly identical results as MLP, where *NDF* targets were identified with 100% accuracy, and other observations predicted *US* as the destination. While this does achieve a high score by virtue of the data being heavily skewed towards *NDF* and *US* destinations, this result is only achieved by excluding all other classes from the models' predictions and is thus a suboptimal result.

Balanced Data

One of the issues with this data is that it is very uneven in its distribution of the target variables. Users who did not choose a destination (*NDF*) comprise 58.3% of the data. Trips booked in the United States, the top destination, compose 29.2% of the data. And destinations other than the top countries listed are another 4.7%. These top three categories comprise 92% of the data. Less popular destinations, such as the Netherlands, make up only 0.35% of the available data. This small number of observations may be reducing the model's capacity to accurately classify the data.

Because this is the only data available, to be able to provide additional observations for less popular destinations, I attempted to create and evenly distributed dataset with an equal number of observations for each target variable. I achieved this by randomly sampling (with replacement) observations from each target variable, so that each destination was represented 124,543 times (i.e. each destination has a number of observations equal to the total number of NDF observations). The resulting dataset had 1,494,516 total observations.

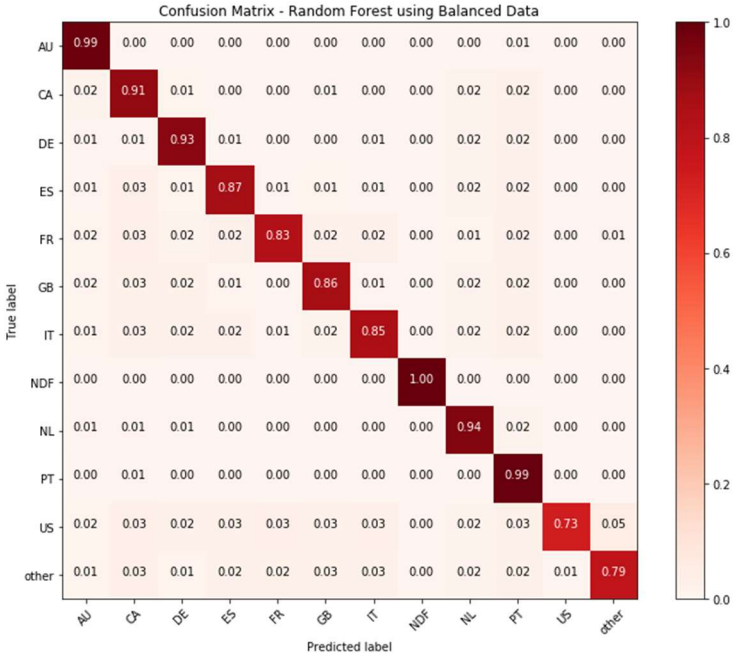
Using this balanced dataset, I then applied the classification models used before. For this data, I applied MLP and Random Forest Classifiers.

Twelve Category Classification Summary – Balanced Data

Classifier	Score	Avg. Precision	Avg. Recall
MLP	0.3696	0.34	0.34
Random Forest	0.8904	0.90	0.89

With the bootstrapped data, we can see that the MLP classifier resulted in a large decrease in model accuracy, indicating this approach is not well-matched with the MLP algorithm. However, Random Forest saw a marked improvement in accuracy both in total score and in target precision for each category.

Confusion Matrix – Random Forest Classifier – 12 Category Classification on Balanced Data



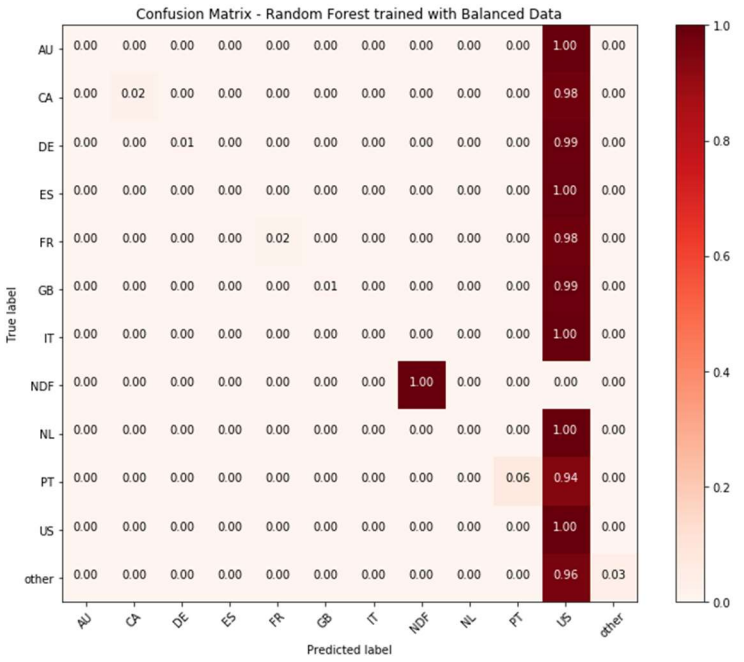
However, this result is based on scoring the balanced data. When I attempted to score the original data using the Random Forest model fitted with the balanced data, the improvement was less dramatic.

Random Forest Model Classification Report – Trained on Balanced Data

Rating	Precision	Recall	F1 Score	Support
AU	0.00	0.00	0.00	59
CA	1.00	0.02	0.04	134
DE	0.50	0.01	0.00	114
ES	1.00	0.00	0.01	204
FR	0.89	0.02	0.02	472
GB	1.00	0.00	0.01	251
IT	0.00	0.00	0.00	286
NDF	1.00	1.00	1.00	12547
NL	0.00	0.00	0.00	57
PT	0.50	0.06	0.11	16

US	0.71	1.00	0.83	6256
other	0.77	0.03	0.05	950
Avg./total	0.89	0.88	0.83	21346

Confusion Matrix – Random Forest Trained on Balanced – 12 Category Classification on Original Data



The classification report shows that when predicting the destination using the original data, but fitting the model with the balanced data, the accuracy of less popular destinations is still very low. However, we see that the Random Forest Classifier can now distinguish between *US* and *other* destinations at least sometimes. Based on this result, it appears that by balancing the data and providing additional observations for each destination to the model, I can improve the accuracy for destinations with a small number of observations, though only to a slight degree. Additional tuning of the model's hyperparameters or adjusting the observation size for each target may results in even better precision.

Three Category Classification with Balanced Data

Lastly, I returned to my three-category classification model using the balanced data to see if this balanced data approach could also result in improved accuracy.

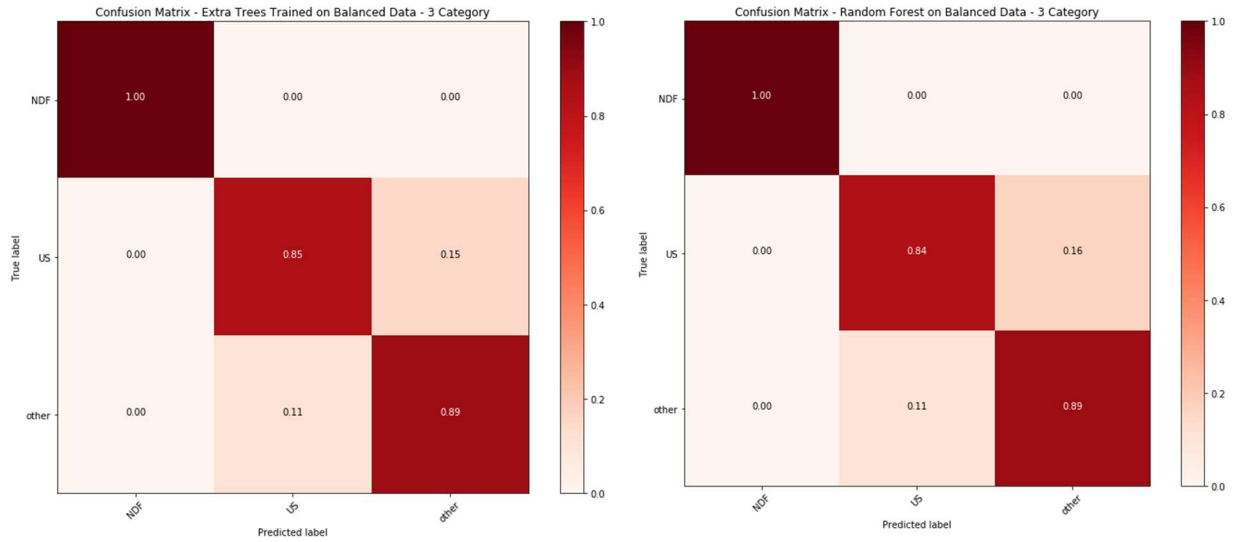
Indeed, when I applied the bootstrapping method to even the distribution of target variable observations, the three-category model showed improved accuracy. The table below shows the classification report fitted using the bootstrapped data and applied to the original data using the Random Forest classifier. The *other* category showed significant improvement in precision with this approach. With this approach, the model can now distinguish *other* destinations from *US* with a reasonably high degree of accuracy.

Random Forest 3-Cateogry Model Classification Report – Bootstrapped Data

Rating	Precision	Recall	F1 Score	Support
NDF	1.00	1.00	1.00	12547
US	0.95	0.84	0.82	6256
other	0.69	0.89	0.27	2543
Avg./total	0.95	0.94	0.83	21346

The Extra Trees Classifier also returned a similar result, with high accuracy in distinguishing *US* and *other* destinations when trained using the balanced dataset.

Confusion Matrices for Extra Trees and Random Forest Classifiers – Trained on Balanced Data

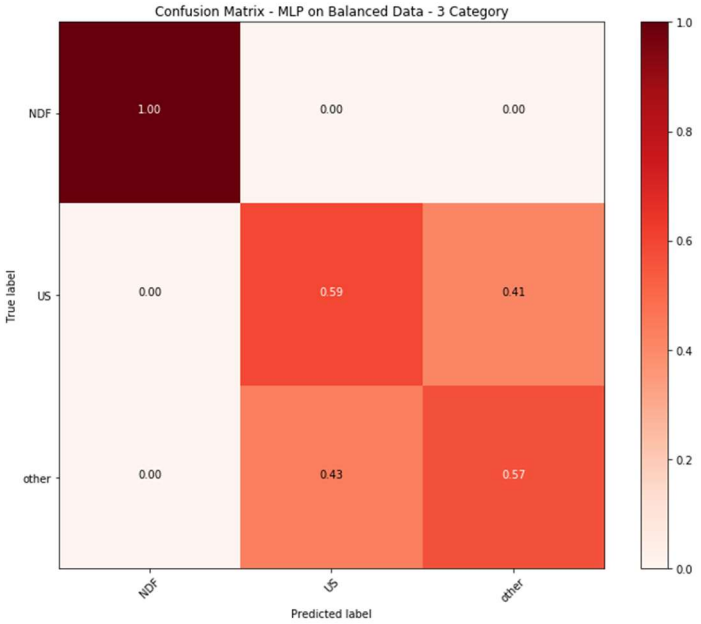


Additionally, I attempted fitting the balanced data using the MLP classifier. Here we see almost no difference between this result and the model fitted with the original data. This once again indicates that MLP does receive the same benefit from the bootstrapping method that ensemble methods like Random Forest and Extra Trees do.

MLP 3-Cateogry Model Classification Report – Balanced Data

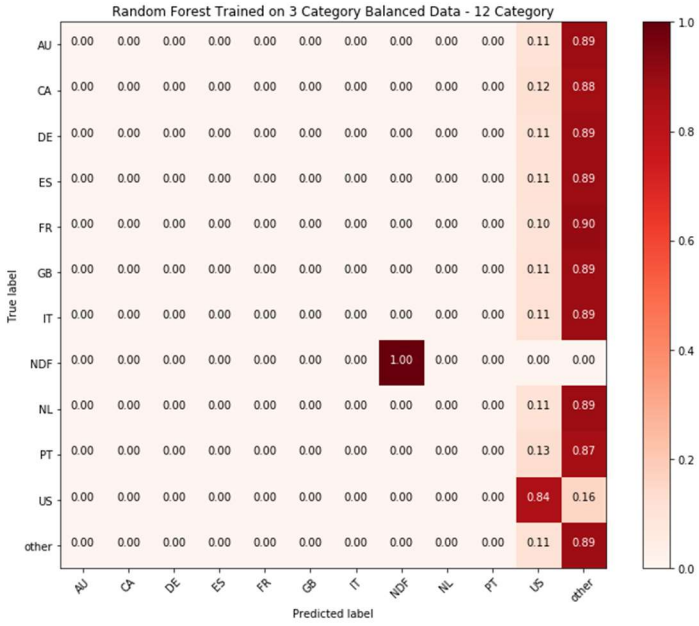
Rating	Precision	Recall	F1 Score	Support
NDF	1.00	1.00	1.00	12547
US	0.58	0.59	0.55	12533
other	0.58	0.57	0.40	12448
Avg./total	0.84	0.79	0.80	21346

Confusion Matrix for MLP Classifier – 3-category – Trained on Balanced Data



Lastly, I attempted to predict destination with the 12-category target variable using the model trained on the 3-category data. This model could accurately predict all the three categories it was trained on, but, unsurprisingly, did not predict any of the other nine categories. Though this model achieved a very slightly lower overall score than the earlier 12-category model, it is the best at being able to distinguish US and non-US destinations.

Confusion Matrix – Random Forest Trained on 3-Category Balanced Data – 12 Category Classification

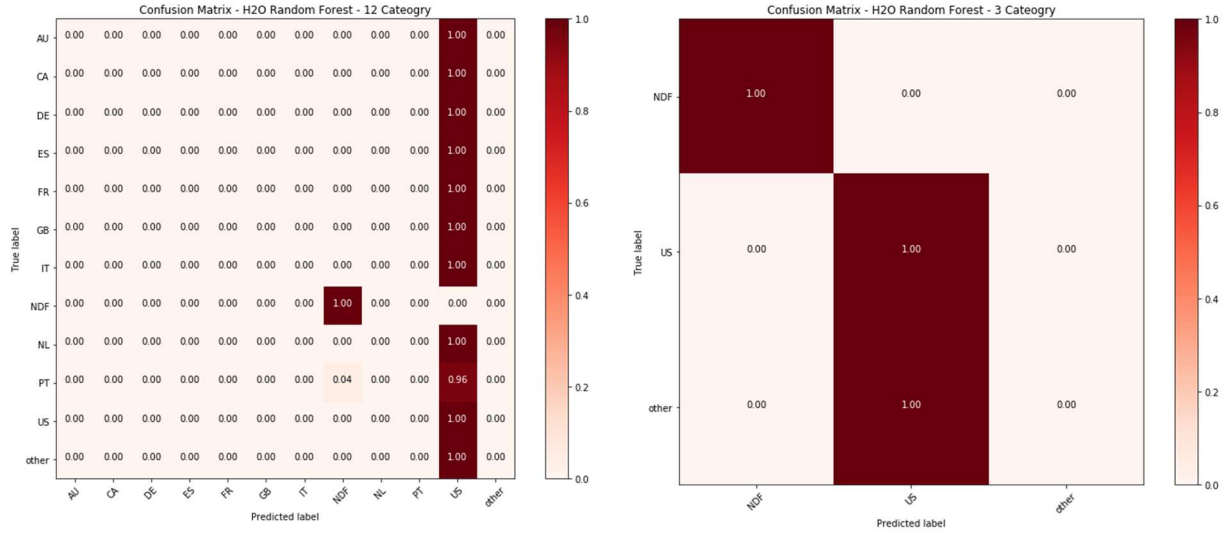


H2O Estimators

In addition to *scikit-learn* classifiers, I also attempted using H2O to see if I could obtain a better result by using H2O's machine learning libraries. H2O has features not available in scikit-learn that I felt might benefit my analysis. For instance, H2O does not need categorical data to be one-hot encoded, considers missing values as separate categories, and it can also balance classes in the data automatically instead of having to manually create a manually balanced dataset as I had to do with the scikit-learn models.

However, despite these useful capabilities, the results produced by H2O were very similar to what scikit-learn was able to produce. Just as before, the models produced by H2O's estimators achieved high scores, but this was done by labeling the easily identifiable NDF values and then choosing US for all other observations. This result was seen in both the 12-category and 3-category models. Using the balanced classes argument did not improve the results the way it did when balanced data was used to train the scikit-learn models.

Confusion Matrices for H2O Random Forest Classifier – 12- and 3-category



Conclusion

The results of my analysis show that we can predict a user's destination using the available data with a relatively high degree of accuracy. However, that accuracy largely comes about because most of the data is clustered into the top 2 categories. Users who did not book a destination can be identified with 100% accuracy, but that's due to features such as "days to book" that clearly distinguish *NDF* observations from the others.

When attempting to pinpoint specific destinations for users who did book a trip, the model was less reliable. This is likely due to the low number of observations available for these destinations. When I attempted to supply additional observations for less frequently booked destinations using the bootstrapping method, the Random Forest classifier showed significant improvement in its predictive ability. If I had

access to a larger dataset, I feel confident that we could train the model to achieve a significantly higher degree of accuracy.

Visual exploration of the data also revealed clear patterns that were helpful in identifying associations between various features in the data and the destination. We can see that users who provided more information during signup were much more likely to book, again allowing the model to easily separate out the *NDF* values from others. We can also see some slight variation in destination choices between different categories of users, such as the fact males are slightly more likely to book trips to *other* destinations than females.

Engineering additional features, such as binning features into narrow categories, also appears to have improved accuracy of the modeling. Accuracy would likely be further improved with access to additional features that were not available in this dataset. However, this is not always the case, as seen by the reduced accuracy when attempting to include features derived from the session data. It appears that demographic information such as age, language, and even signup device information such as OS and device type, provide better clues about what destination the user is likely to choose. If the quantity of this type of information was increased, accuracy could likely see a sizable boost. Additional session data describing user interaction with the Airbnb website may also improve model accuracy, though for this dataset session data was only available for a fraction of the users, thus limiting its capacity for improving my predictions.

Extra observations for less frequently chosen destinations would also likely improve the result. When I trained the models using a balanced dataset created by duplicating an equal number of observations in each category, certain models were able to see a significant improvement in their ability to differentiate the destinations. If additional observations for each destination were available, I believe the model accuracy could be greatly improved.