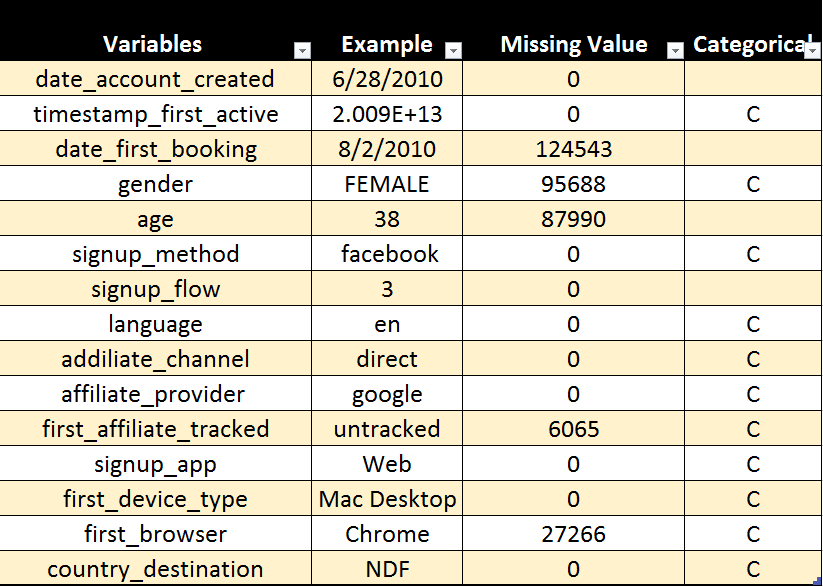
Airbnb New US Users:

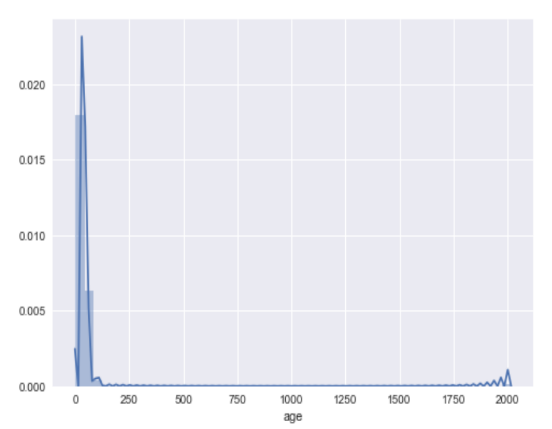
What are their first destinations? How long does it take for them to book?

**Introduction**

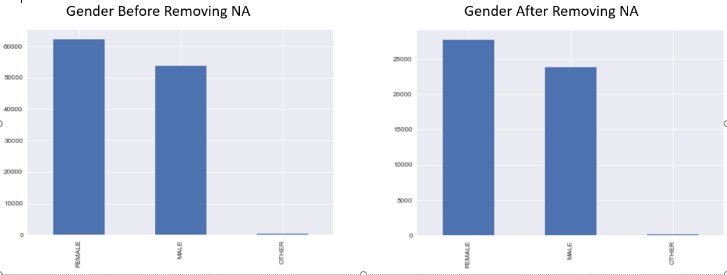
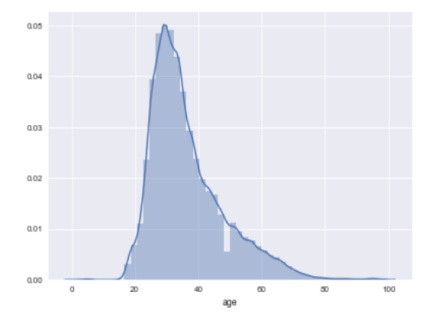
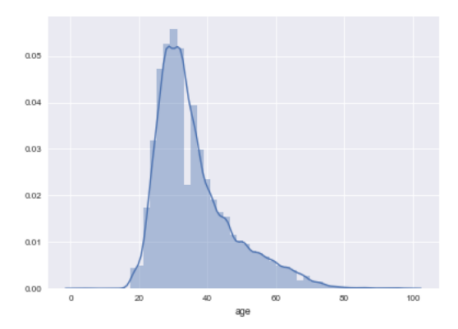
Airbnb is an online housing service that allows individuals to rent out their properties for short periods of time. As a global business, Airbnb has millions of customers that travel both domestically and internationally. Being an American company, with predominately American customers, Airbnb has compiled a data set 275,562 of their U.S. users (213,466 in training, 62096 in testing), containing demographic (age, gender, etc.) and account information (date account was created, device account was created on, etc.).

Using the data Airbnb has provided, we created a model that predicts what the first country that new U.S. users chose to travel to. This model would be especially helpful for helping focus Airbnb’s marketing strategies. Being able to accurately predict where a new user will first travel to allows their marketing algorithms to suggest rental locations more accurately to users, which in turn will decrease booking times and help increase the overall number of bookings.

**Data Cleaning**



To get the data ready for modeling, we preformed multiple data cleaning steps. The first issue that we noticed was that there were large outliers in the age data set. By observing the histogram of users’ ages, we noticed that many data had ages ranging over above 100, going passed 2000. Since the ages are self-reported, ages can be very noisy. I simply remove all the data whose ages are over 100. Second, I formatted time from UNIX timestamp, and extracted detailed time information, like month and day. Similar to the date\_account\_created and date\_first\_booking, I extracted its month and day.

The data set has a lot of NAs. For example, gender has 94982 NAs, and age has 87990 NAs. To remove NAs, we need to satisfy two conditions. First, after removing NAs, the number of observation should still be sufficient for training. Second, removing NAs does not result in any significant change in distributions in any variable. We examined fields with significant numbers of NA entries: Age and Gender. The histogram shows that after removing NAs, there is no significant change in the distribution of these two fields.

Age before removing NA Age after removing NA:

To create a model to predict the time it takes for a particular user to book, we had to create a variable that would store this information. The data Airbnb provided us contained two pieces of information for this, the date an account was created and the date that the booking was made. Using these two pieces of information, we stored the information on the month the booking was made into a variable called First\_booking\_month, and we stored the number of days between the two dates into a variable labeled time\_delta\_bc. To ensure that the numerical representation of the month was used ordinally, we converted First\_booking\_month into categorical data. Since the first active variable contained a great deal of information, we decided that it would be best to use the time of day as a reference instead of a specific time, and so we replaced the variable with only the hour that first activity was made (e.g., 2:23 pm became 14).

This data set has too many categorical variables. Each variable, like Signup\_method, has over ten different categories. If I use one hot encoding, I will get over 200 variables, which is tough to interpret. It is also hard to choose variables when coding. Moreover, I haven’t learned PCA and dimensional reduction. Hence, I use ordinal encoding for my categorical variables. Last but not least, I use z-scaling to have all the features standardized.

**Question I**

To further study ways to increase Airbnb’s profitability and marketing effectiveness, we also created a model that will predict the amount of time that it will take for a new US user to make their first booking. Having a prediction for how long it will take an individual to book will provide Airbnb with a basis to focus their marketing efforts. For example, using this model Airbnb can identify cases that are predicted to have a long gap between their account creation and first booking. After identifying these individuals, Airbnb can implement marketing strategies to attempt to narrow that time gap. This model can also be used to perform experiments on these strategies as well, to confirm whether their efforts are having any effect. This will help to save Airbnb money if they are using a technique that they believe is lowering time gaps, but is having no statistically significant effect.

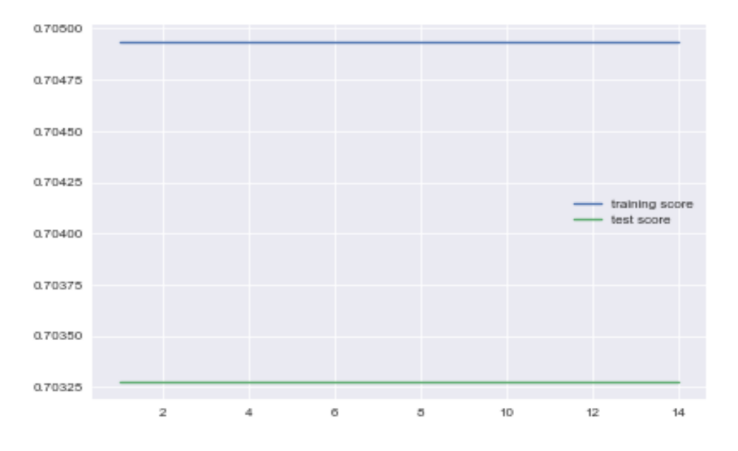
After completing the data cleaning, we began creating our models. To build the first model, predicting the first country a new user would visit, we used decision trees, multinomial logistic regression, support vector machining, and ensembles.

The resulting models in each grouping showed fairly similar results. Our decision tree model came back with an accuracy score of 74.553% on the training data set, and 62.165% on the testing data set. The multinomial logistic regression returned scores of 70.494% and 70.327% on the training and testing sets respectively. Interestingly, these results remaining constant across all the models that our code produced, as can be seen in the graphic below. The original model consisted of gender alone, and would imply that gender is the most statistically significant predictor that we have. The same was also seen when we used SVM techniques, which returned a model of only gender, with consistent scores across all levels. Interestingly enough, these models returned identical scores. We conclude this interesting observation in SVM is the result of the fact that SVM only uses support vectors as the boundary. Other data points are irrelevant.

SVM

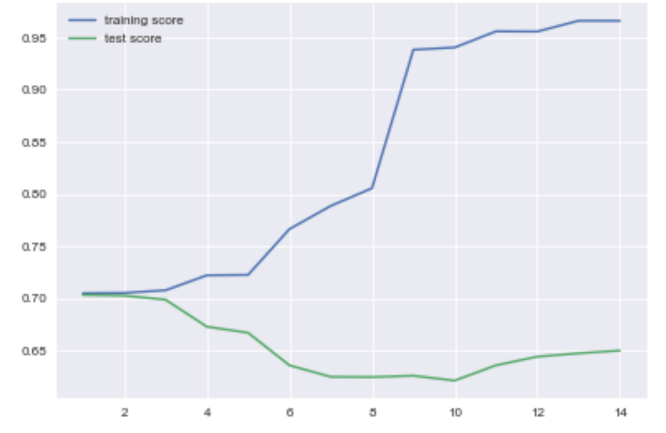


Multinomial Logistic Regression



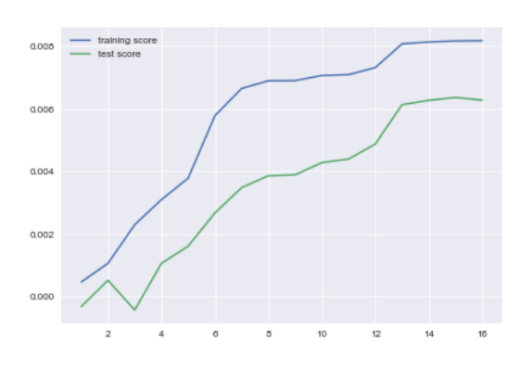
In direct contrast to the others, our Random Forest model returned the first model that decreased in accuracy in the testing as variables were added. As seen in the graphic, the training set consistently overfit the data, approaching 100%, as the number of variables increases but does the opposite for most of the testing set. Hence, the model we would select in this case is gender alone, with 70.495% accuracy on the training set, and 70.327% on the testing set.

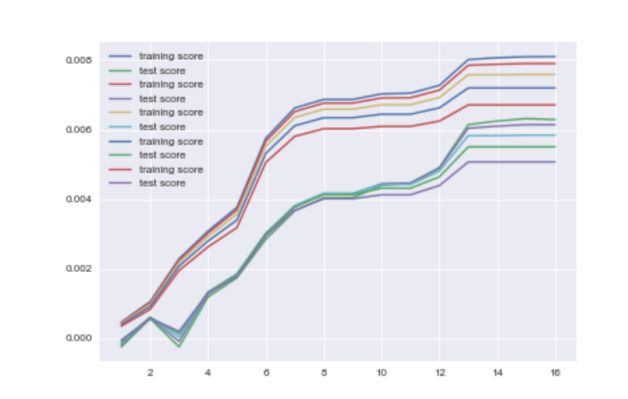
Random Forest



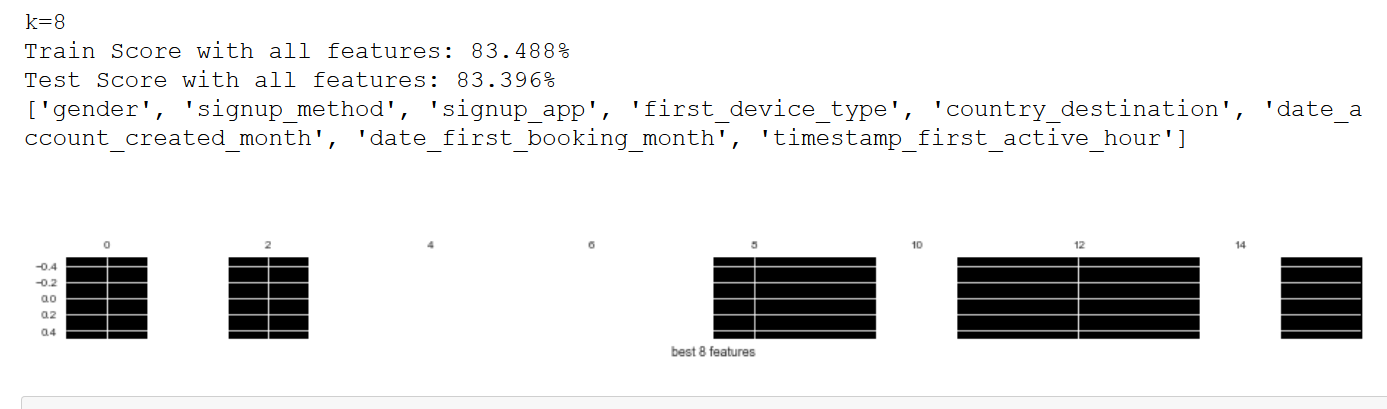
**Question II**

For our second question, we try to find a regression model to predict the time it takes for an individual to book, we decided to try Linear Regression, Lasso Regression, and Decision Tree Regressor. We made the initial mistake of leaving in the month of first booking, and were pretty naively excited by the good results (over 83% for both training and test R square). As we soon realized, the reason these models do such a good job predicting our results is because our models contain the information we are trying to predict. If we knew the month that they would book, then we would also already know how long it took them to book. This inherently defeats the purpose of the model, since you don’t model something that you already know. After adjusting dropping the columns relating to booking dates from the data set, we came up with a much different result. The linear regression came back with a R squared of only 0.812% and 0.622% between the training and testing groups, and the Lasso came back similarly different, although in a positive way. While previously we had scores of 0.671% and 0.507%, we saw an increase to 0.795% and 0.608%, as can be seen in the graphic. Unfortunately, this did not extend to the decision tree model, as we came up with a model that was vastly different and less effective than previously. Many of the models even managed to be consistently opposite from our intended results, as we can see in the graphic.

Linear Regression: 

Lasso Regression:

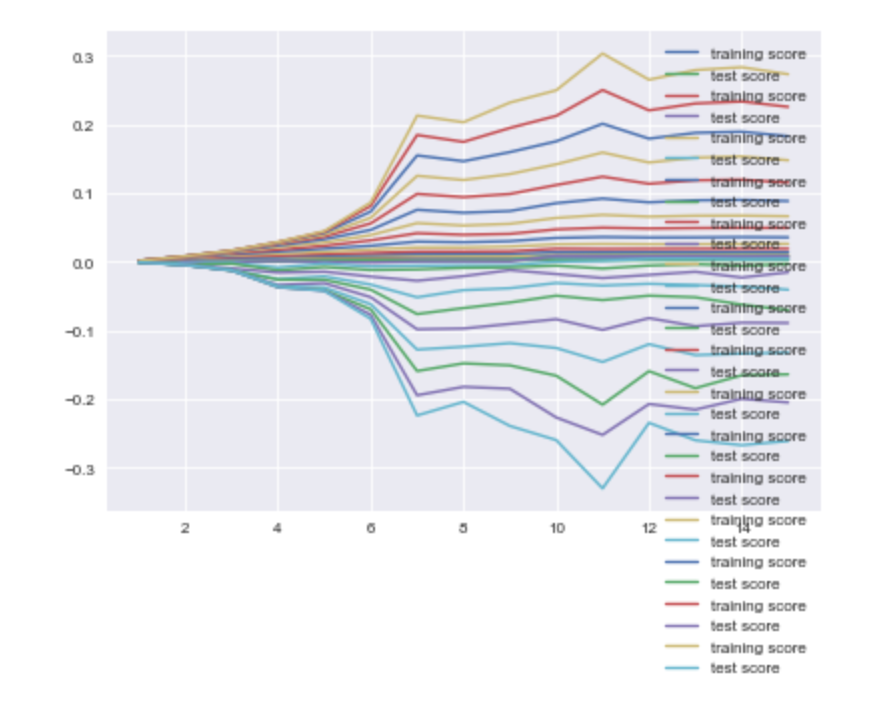
Decision Tree:

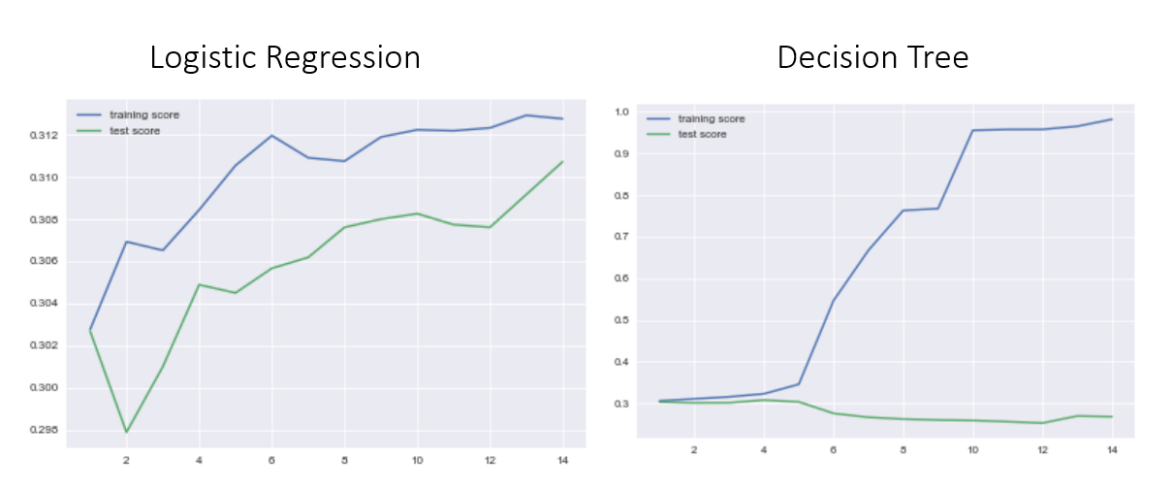


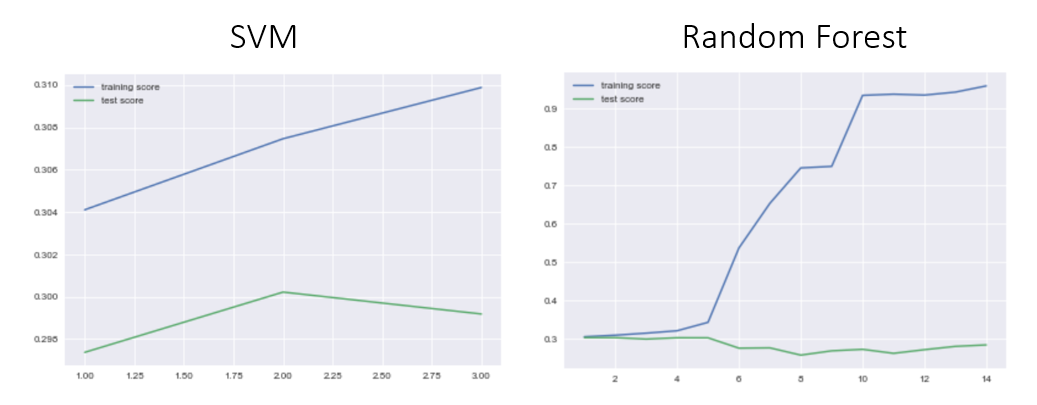
Lasso (without booking date)



Decision Tree:



Due to this change, we decided to try to create a classification model instead. We grouped the data into their respective quartiles; under one day, between one and four days, between four and forty-six days, and between forty-six and three-hundred and sixty-five days. We encoded these levels as 0, 1, 2, and 3. From here we implemented logistic regression, re-ran a decision tree model, random forest, and implemented boosting. We got consistently better results than we saw in the previous models. All the models returned a score of approximately 30%, with the logistic regression coming in slightly higher at 31%. Although the logistic regression carried the most variables in its calculation, while others only held one or a few at similar scoring levels. 



Our two problems came with very different results. While we were successfully able to create models that would predict an individual’s target destination 70% of the time, this is not as helpful as it may initially appear. It is true that since we have twelve potential destinations, we should expect to randomly guess the travel destination about 8.33% of the time. In contrast to 8.33%, our accuracy rating of 70% seems extremely good. However, upon examination of the data, we noticed that the distribution of destinations picked is very skewed. About 58% of our data has a designation of “NDF” or no designation found, meaning that these individuals never booked. Following that there is the traveling domestically inside the US at 29%, and all others come in at under 5%. That means that about 70% of the cases that individuals do engage in travel, travel domestically in the United States. Our model does a good job predicting the domestic traveling, but not other country destinations. Given this information, we can say that the model is biased and will likely not provide us with any useful information. The model will likely perform well in the sense that it will predict whether or not a person travels inside the United States or does not travel at all. This does not, however, help us with our initial question, which was to predict which country they are going to go to given that they are traveling.

**Conclusion**

We applied a variety of machine learning methods to predict the time between creating the account and first booking and predict the first booking country destination. We use both regression and classification methods, such as Ridge Regression, Lasso Regression, Decision Tree Regressor, Decision Tree (classification), Support Vector Machine, Random Forest, Boosting, etc. Among these models, Decision Tree seems to work the best in answering both questions, although all the models return a similar result.

Given the unhelpful nature of our model, if we were to continue our investigation we would like to create two separate models instead. The first model would predict whether or not an individual is likely to book a trip at all. This would help the marketing team identify ‘high risk” individuals that may be an untapped source of income. The second model would be, assuming that the individual is going to travel, where will they go. Having both of these models would maintain the useful predictions on if they are at risk of not traveling, and then given either outcome where they will go. By doing this Airbnb can target high-risk and low-risk individuals with personalized ads/specials for locations that the model predicts they will enjoy. These combined methods should see an increase in sales, and maintain both pieces of predictive power. In both models, we would have also liked to try different methods for accounting for the outliers and missing information in age, gender, and other pieces of the data set. Another possible approach is to use Clustering (Unsupervised Learning) to divide the customers into different groups, which can be understood as different target markets. Repeat the supervised learning methods on each cluster and see if we can get any improvements.

Our second model was also not particularly helpful. With scores ranging in the 30%, we do gain some information. Although we would have been right 25% of the time had we randomly guessed which classification an individual would belong to. Thus, this model is also a failure in the practical business sense. If we had more time, we would have liked to see if dividing the cases into fewer, but larger groups, would have made resulted in more robust predictive models. However, given the number of individuals that never traveled at all the focus would have been on completing the changes suggested to model one. Those changes would have covered similarly important data and would be targeting a much larger part of Airbnb’s lost potential revenue than by speeding up the process for those who are already going to book.