Predictive Modeling Project: OPIM 5604

Project Report

On

Prediction of First Destination country of Airbnb New Customer

[TEAM: 9]

Arjun Sawhney

Anubha Pant

Mihir Gada

Suraj Kumar

Tingting Deng

Contents

SUMMARY

Airbnb is a peer-to-peer online marketplace and homestay network enabling people to list or rent short-term lodging in residential properties, with the cost of such accommodation set by the property owner. The company receives percentage service fees from both guests and hosts in conjunction with every booking. It has over 2,000,000 listings in 34,000 cities and 191 countries.

Project: Airbnb has provided a list of users along with their demographics, web session records, and some summary statistics. The goal of this project is to predict which country a new user's first booking destination will be. All the users in this dataset are from the USA.

There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'. 'NDF' is different from 'other' because 'other' means there was a booking, but is to a country not included in the list, while 'NDF' means there wasn't a booking.

Business Value: Airbnb wants to increase its sales by targeting users with marketing campaigns most suitable to the user. By predicting in advance the most probable destination a user is likely to book Airbnb will present its customer with deals and customized marketing when a user logs in to the Airbnb website.

Data Preprocessing

Since we started with 6 different files with approximately 8 odd variables each, we had to perform significant processing before we began modeling. We performed that in 4 parts:

Correcting erroneous values

For instance, the age field in our database had some values expressed as years (1974,1991 etc.). We took all values above 1900 and subtracted them from 2016 to get the age. Additionally, we dropped values above 100 and below 5. We allowed values between 5 and 15 under the assumption that a child may be booking on behalf of the family.

Imputing missing values

Unlike most modeling problems, we did not impute missing values with mean or median, nor did we drop the value. In fact, in our data, the missing values told a story. Whenever, age or gender data was missing, it was less likely (30:70) chances of booking as against when this information was present (60:40).

Standardization

Most of our data was on the scale of 0-100 with age having the highest range, so there was no need of standardization. Additionally, the models we used – Decision Tree and Logistic regression are not sensitive to range.

Binning

Instead of using age as a continuous variable, we binned it in buckets of 5 years (eg. 0-5, 6-10,11-15, etc.) We noticed that this gave us better results than using it as a continuous variable
Additionally, we split the date fields into – Day, month, year, quarter, time of the day and other such fields. This helped us identify seasonality trends.

Secondary parameters

We had information on the first browser used by a user. However, it had 80 unique entries. We binned different forms of the same browser (chrome, chromium, chrome mobile etc.) and aggregated in to 5 classes.

One hot encoding

In addition to the customer parameters in the training file, we had a file called sessions.csv which contained the browsing behavior of users from 2014. This had a sizeable proportion of both training and test data. However, it had 10 million rows with each click of user recorded with the time spent doing it. We transposed the data such that each activity (click, search, about us, etc.) became columns and the count of clicks and the time spent doing it became the parameters. This helped us compress our data quite significantly and we could finally make sense out of the large and uninterpretable browsing data.

Exploratory Analysis

1. Airbnb data accounts created was disproportionately split across 4 years. This was a result of the whopping growth across the 4 years. It seemed to plateau in the second half of 2014, but that was because that data was used in the test set.

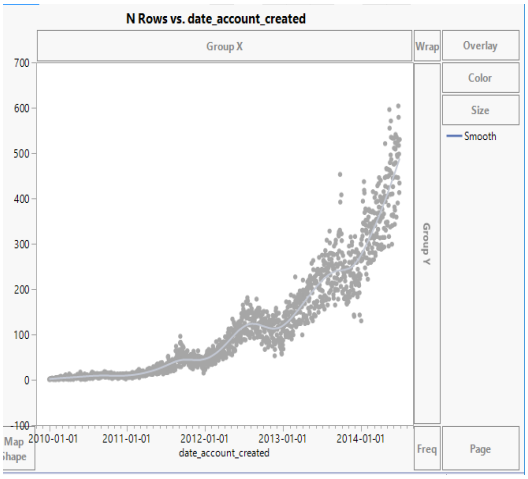


Figure 0.1

2. Also, 90% of the users either booked US or did not book at all which left only 10% data for the other 9 countries

Year	Accounts created	Accounts created (2014 extrapolated)	Y-o-Y growth
2010	2557	2557	
2011	10890	10890	326%
2012	35702	35702	228%
2013	74419	74419	108%
2014	68502	137004	84%

Figure 0.2

3. Users who input their age and gender details were significantly more likely to book than users who did not.

	Age absent	Age present
No Booking	73%	41%
Booking	27%	59%

	age_present		
NDF?	0	1	Column %
0	56255	46907	53.71%
1	20750	68158	46.29%

Figure 0.3

4. August through October, there was a significant rise in accounts created. This was due to likely more bookings for the upcoming thanksgiving and year-end holiday season

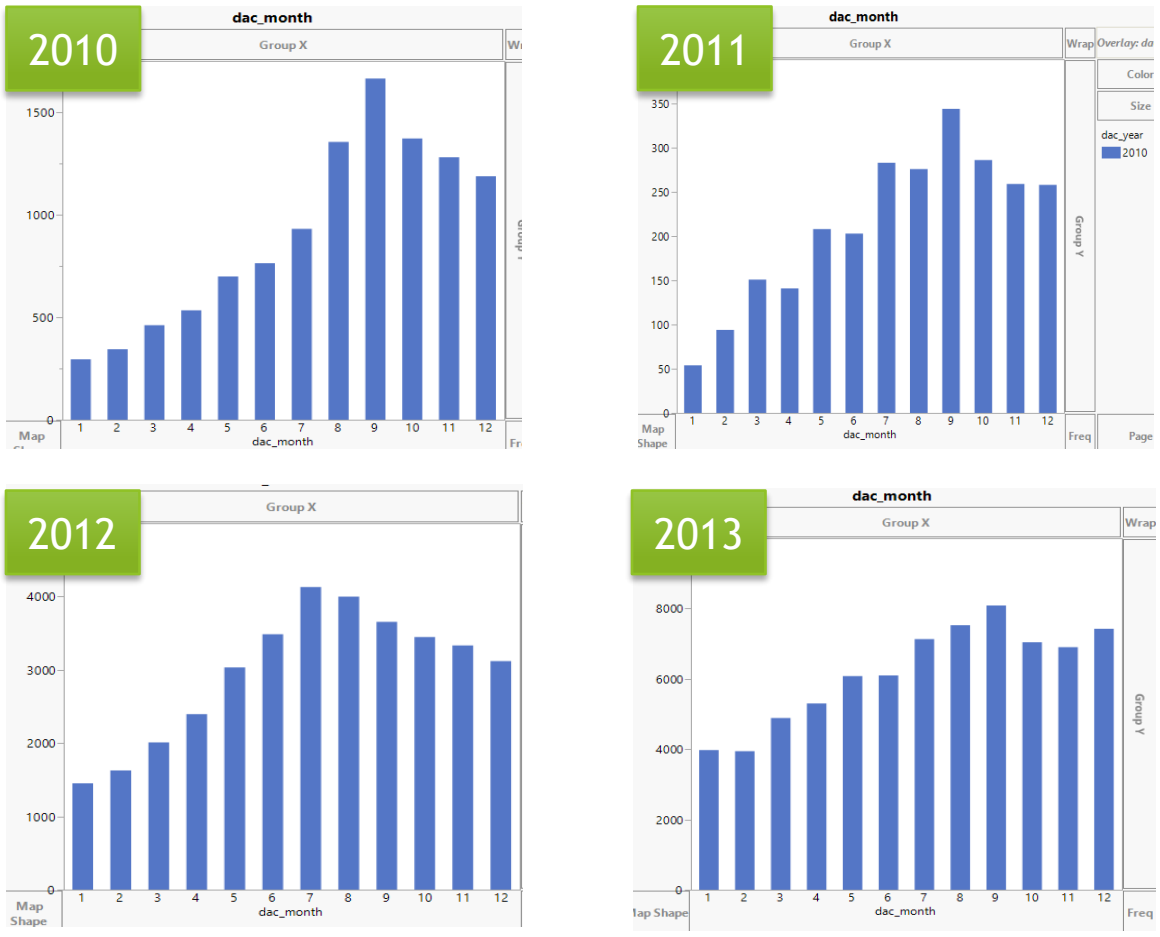


Figure 0.4

5. The device used for booking reflected success rate of booking. Mobiles, tablets and laptops phones were increasingly successful in that order

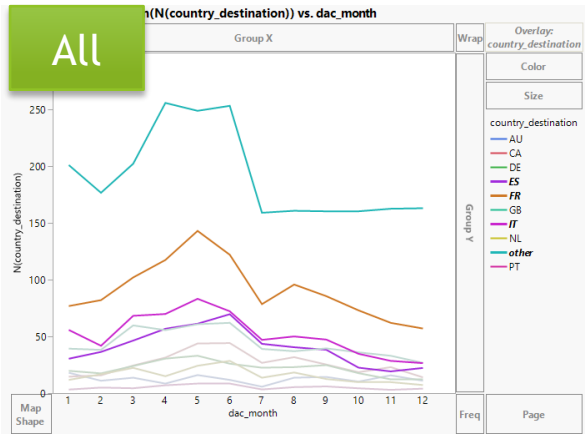
	NDF?		
	0	1	
first_device_type	Row %	Row %	N
Android Phone	70.84%	29.16%	2452
Android Tablet	59.79%	40.21%	1144
Desktop (Other)	49.54%	50.46%	1090
iPad	57.47%	42.53%	12846
iPhone	64.90%	35.10%	18281
Mac Desktop	48.16%	51.84%	81502
Other/Unknown	69.01%	30.99%	9276
SmartPhone (Other)	62.12%	37.88%	66
Windows Desktop	53.91%	46.09%	65413

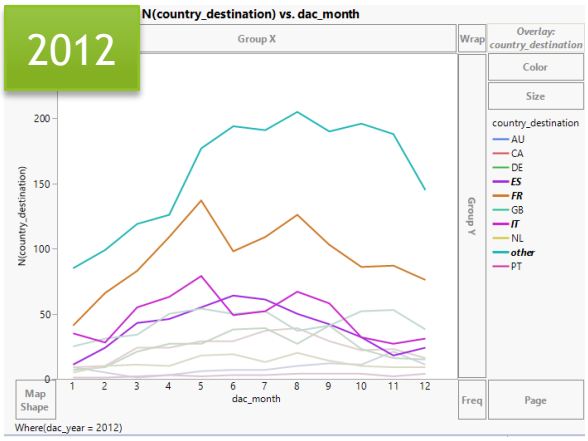
Device	Success
Desktop	55%
Tablets	40%
Mobiles etc.	30%

Figure 0.5

Exploratory analysis for other countries

1. Bookings in Europe were peaking during April through June as against the year-end. July 2012 showed a spike in bookings driven by Europe because the company decided to focus on Europe at the time and opened 6 offices there starting with Germany.





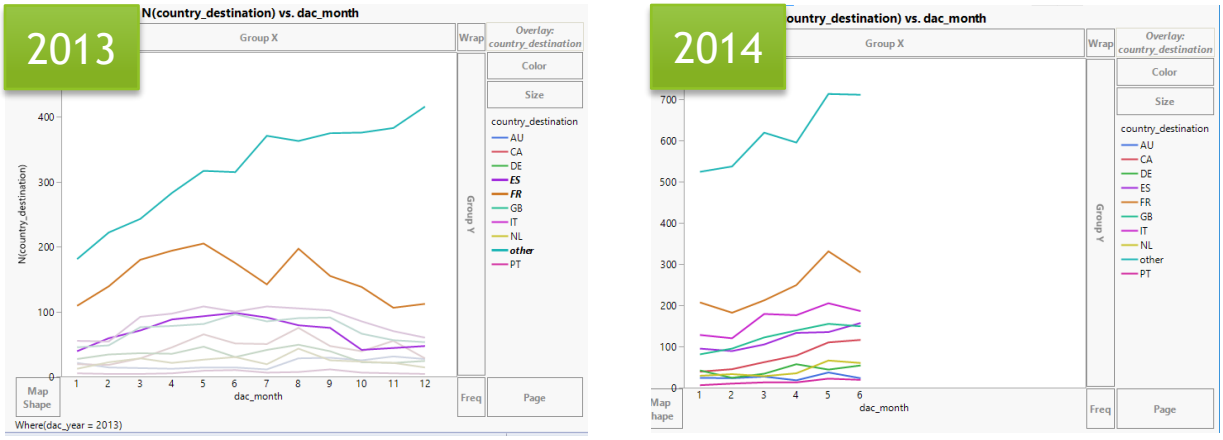


Figure 0.6

2. Women preferred France, Great Britain and Italy more times than men. We want to further test out if these bookings were initiated by women for families, friends or solo trips. We wanted the number of travelers booked for with their gender and the credit card holder information to verify these claims.

	Row %			
	gender			
country_destination	FEMALE	MALE	OTHER	-unknown-
AU	38.40%	34.88%	0.19%	26.53%
CA	31.86%	33.40%	0.35%	34.38%
DE	33.74%	39.21%	0.28%	26.77%
ES	37.93%	30.10%	0.18%	31.79%
FR	39.06%	26.58%	0.26%	34.10%
GB	37.91%	29.35%	0.13%	32.62%
IT	38.48%	24.66%	0.18%	36.68%
NL	33.33%	36.48%	0.39%	29.79%
PT	35.94%	31.80%	0.46%	31.80%

Figure 0.7

3. Language of browsing and hence possibly nationality drove sales to the home country.

Not (language = en)

	language													
	cs	da	de	el	es	fi	fr	it	ja	ko	nl	no	pl	pt
country_destination	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %
AU	0.00%	0.00%	37.50%	0.00%	0.00%	0.00%	25.00%	0.00%	0.00%	12.50%	0.00%	0.00%	0.00%	0.00%
CA	0.00%	0.00%	6.25%	0.00%	0.00%	0.00%	43.75%	6.25%	12.50%	12.50%	0.00%	0.00%	0.00%	0.00%
DE	0.00%	0.00%	53.33%	0.00%	13.33%	0.00%	17.78%	4.44%	0.00%	0.00%	4.44%	0.00%	0.00%	0.00%
ES	1.52%	3.03%	10.61%	3.03%	30.30%	0.00%	16.67%	9.09%	1.52%	3.03%	0.00%	0.00%	3.03%	1.52%
FR	0.00%	0.00%	7.09%	0.71%	12.77%	0.71%	48.94%	2.13%	2.84%	8.51%	2.13%	0.00%	0.71%	0.00%
GB	0.00%	0.00%	12.50%	0.00%	15.00%	0.00%	30.00%	10.00%	5.00%	7.50%	0.00%	5.00%	0.00%	0.00%
IT	0.00%	1.39%	8.33%	0.00%	12.50%	0.00%	15.28%	26.39%	2.78%	12.50%	1.39%	0.00%	1.39%	0.00%
NL	0.00%	0.00%	11.11%	0.00%	5.56%	0.00%	22.22%	11.11%	0.00%	11.11%	22.22%	0.00%	0.00%	5.56%
PT	0.00%	0.00%	14.29%	0.00%	28.57%	0.00%	28.57%	0.00%	0.00%	0.00%	0.00%	0.00%	14.29%	14.29%

16025 rows have been excluded.

Figure 0.8

Modeling

As stated in the data exploration part, there are 3 challenges in this data sets:

1. The objective of this case is to classify 12 levels of outcomes.
2. This data set is of extremely uneven population, 3 outcomes account for 92.3% of observations, 9 outcomes of concern accounts for only 7.7%. However, for marketing purpose, classification of the 9 outcomes is of more business value.
3. When merging these 2 datasets, train_users and sessions, more than 500 variables are created.

Running Models

We introduced 3 levels models to overcome above challenges by running the following:

1. **Booked/not-booked** - Decision Tree Model
2. **Booked US/booked other country** - Regression Model (with Non-booking observation excluded).
3. **Classify the remaining 10 countries** - Decision Tree Model (with US observations excluded).

First Stage Model

Refer to Figure 1.1 and Figure 1.2 for the result of this model, a decision tree from 6 splits with 4 input variables.

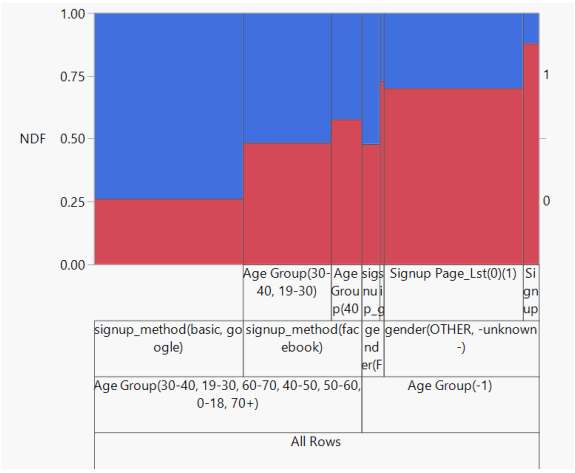


Figure 1.1

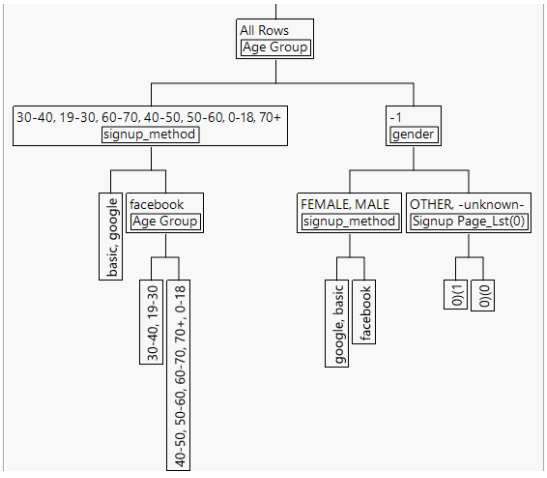


Figure 1.2

Model Performance

Refer to Figure 1.3 for the confusion matrix of this model.

Confusion Matrix

Training

Actual	Predicted	
NDF	0	1
0	31736	21608
1	13825	39519

Validation

Actual	Predicted	
NDF	0	1
0	14594	10315
1	4628	13154

Test

Actual	Predicted	
NDF	0	1
0	14682	10227
1	4631	13151

Figure 1.3

In the aspect of confusion matrix, the result is acceptable, since the misclassification rate in validation data set is 34.7%, 34.8% in test data set; 58.6% accuracy rate of Non-bookings (0,0) in validation data set, 59.0% in test data set, and 74% accuracy rate of Bookings (1,1) both in validation and test data sets.
In the aspect of accuracy rate of Non-bookings and accuracy rate of Bookings, it's a good model. Because the goal of first stage is filtering non-bookings and correctly classify bookings so that in the implementing stage, there will be higher accuracy in second stage classification.

Sampling Method of First Stage

Proportions of traning data set, validation data set and test data set are 50%, 25% and 25% perspectively.Make the proportions of Bookings and Non-bookings even in training data set, not even in validation and test.

Selected Variables

Refer to Figure 1.7 for the 4 input variables selcted after other combinations. All of the 4 variables are categorical.

Term	Number of Splits	G^2	Portion
Age Group	2	11175.8976	0.6604
signup_method	2	4347.82385	0.2569
gender	1	768.120549	0.0454
Signup Page_Lst(0)	1	630.105796	0.0372

Figure 1.7

There are 8 levels of age groups. -1 represents the outliers and missing value in the origin data set. Roughly, the remain age data range from 0 to 90. To make the decision tree split less times and not lose value of information, we think group the observations generally every 10 years would be adequate.

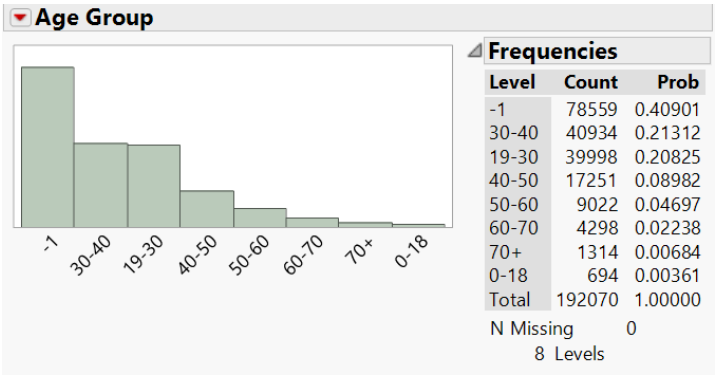


Figure 1.8

Also, we use new-created dummy variable, Signup Page_Lst, which indicates whether the user login in the last page or not. Refer to Figure 1.9 for the distribution for this dummy variable.

Figure 1.9

Second Stage

Refer to Figure 2.1 for the result of this model, the misclassification rate is 44% both in validation and test data set; 59.4% accuracy rate of US (0,0) both in validation and test data set, 47.6% accuracy rate of Non-US (1,1) in validation data set, 47.5% in test data set.

Confusion Matrix

Training

Actual US?	Predicted 0	Predicted 1
0	6965	4752
1	6111	5606

Validation

Actual US?	Predicted 0	Predicted 1
0	5669	3868
1	2046	1860

Test

Actual US?	Predicted 0	Predicted 1
0	5667	3870
1	2049	1857

Figure 2.1

Considering this high misclassification rate of 44%, which is almost as much as the probability of picking up US randomly without a model, the model does not work well.

Although 59.4% accuracy rate of US is not very low, 47.6% accuracy rate of Non-US is below the base line (50%), so the result is unacceptable. Because correctly classifying US and Non-US is of equal importance.

However, we've tried combinations of different variables, the below variables (Figure 2.2) selected yields the best result, though the result is unacceptable. And we've tried decision tree

and K-nearest neighbors. The decision tree has much higher misclassification rate, and K-nearest neighbors consumes lots of time and does not work at all.

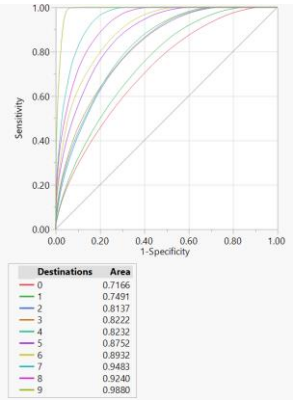
Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
gender	3	3	8.55653558	0.0358*
Age Group	6	6	27.2896261	0.0001*
signup_app	3	3	20.6957151	0.0001*
First_device_type	3	3	10.3873013	0.0155*
ajax_refresh_subtotal	1	1	26.3037609	<.0001*
show	1	1	10.6112795	0.0011*
message_post.x	1	1	6.38583798	0.0115*
message_thread	1	1	13.0971879	0.0003*
update_listing	1	1	5.22374696	0.0223*
user_social_connections	1	1	12.0762577	0.0005*
Quarter Account	3	3	50.7028095	<.0001*

Figure 2.2

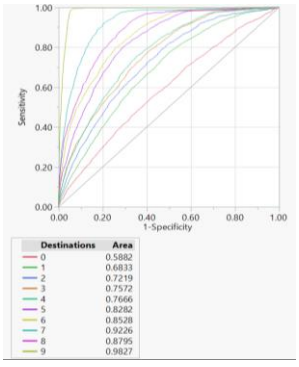
Result of Third Stage

Refer to Figure 3.1 for ROC curve and confusion matrix of training data set in model for 3rd stage; Figure 3.2 for validation data set; Figure 3.3 for test data set. The misclassifications of training, validation and test data sets are 57.7%, 64.4% and 64.4% respectively.



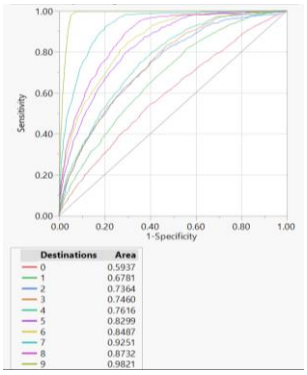
Actual	Predicted									
Destinations	0	1	2	3	4	5	6	7	8	9
0	543	498	747	514	397	501	415	397	294	254
1	250	1193	806	473	420	525	377	287	339	215
2	167	422	2015	397	365	365	297	210	276	143
3	136	391	538	1615	299	423	258	218	249	212
4	163	392	681	342	1319	461	291	260	255	138
5	175	304	419	314	178	1963	195	210	134	122
6	133	215	361	186	194	340	1868	164	138	108
7	39	184	204	144	74	112	134	1728	59	52
8	88	125	279	178	142	192	158	140	1481	126
9	0	31	39	28	39	4	10	2	14	2824

Figure 3.1



Actual	Predicted									
Destinations	0	1	2	3	4	5	6	7	8	9
0	340	500	860	542	447	502	423	414	336	281
1	241	943	822	534	432	520	447	318	333	215
2	183	494	1428	414	477	412	290	301	298	212
3	173	506	618	1318	310	441	353	279	282	198
4	185	472	719	377	1127	473	295	236	236	186
5	200	302	504	349	218	1599	190	273	196	148
6	172	265	387	212	214	409	1659	225	160	96
7	58	191	252	157	122	133	161	1628	57	69
8	79	226	298	206	169	249	201	152	1224	153
9	0	44	58	27	33	4	8	4	13	2647

Figure 3.2



Actual	Predicted									
Destinations	0	1	2	3	4	5	6	7	8	9
0	179	275	352	272	238	304	205	200	166	120
1	128	508	418	265	206	259	202	157	173	110
2	104	231	773	223	218	207	149	118	191	88
3	82	238	298	621	183	205	147	121	117	106
4	95	239	320	168	525	245	163	117	154	92
5	92	152	211	202	109	805	126	129	111	73
6	79	124	213	123	118	245	807	96	69	48
7	28	135	132	87	40	63	66	775	39	34
8	38	94	166	111	93	122	100	89	601	76
9	0	11	39	24	24	3	10	4	6	1354

Figure 3.3

To further analyze the misclassification rate and accuracy rate of this model, refer to Figure 3.4, which represents the confusion matrix of test data set, and which indicates the base line of misclassification rate (90%) and accuracy rate(10%).

Destinations		Predict														
		Others	FR	IT	GB	ES	CA	DE	AU	NL	PT	Total	Acc%	BL for Acc%	Mis%	BL for Mis%
Actual	Others	179										2311	7.7%	10.0%	92.3%	90.0%
	FR		508									2426	20.9%	10.0%	79.1%	90.0%
	IT			773								2302	33.6%	10.0%	66.4%	90.0%
	GB				621							2118	29.3%	10.0%	70.7%	90.0%
	ES					525						2118	24.8%	10.0%	75.2%	90.0%
	CA						805					2010	40.0%	10.0%	60.0%	90.0%
	DE							807				1922	42.0%	10.0%	58.0%	90.0%
	AU								775			1399	55.4%	10.0%	44.6%	90.0%
	NL									601		1490	40.3%	10.0%	59.7%	90.0%
	PT										1354	1475	91.8%	10.0%	8.2%	90.0%

Figure 3.4

Sampling Method

Bootstrap Augmentation is the method used to resample the data set. See Figure 3.5, the graph left side shows the distributions of remaining 10 outcomes, 10 different countries in the training, validation and test data sets which are randomly sampled; the right side one shows the distributions of the 10 outcomes in the data sets with adjusted proportions which are created by Bootstrap Augmentation. And the size of sample increased from 19,529 observations to 97,809 observations. Refer to Figure 3.6 for ROC curves yields by different sampling, the left one responses to the random sample, the right one responses to the sample with adjusted proportions.

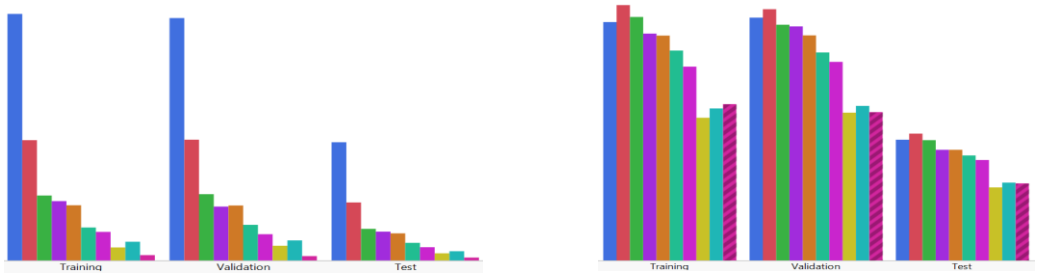


Figure 3.5

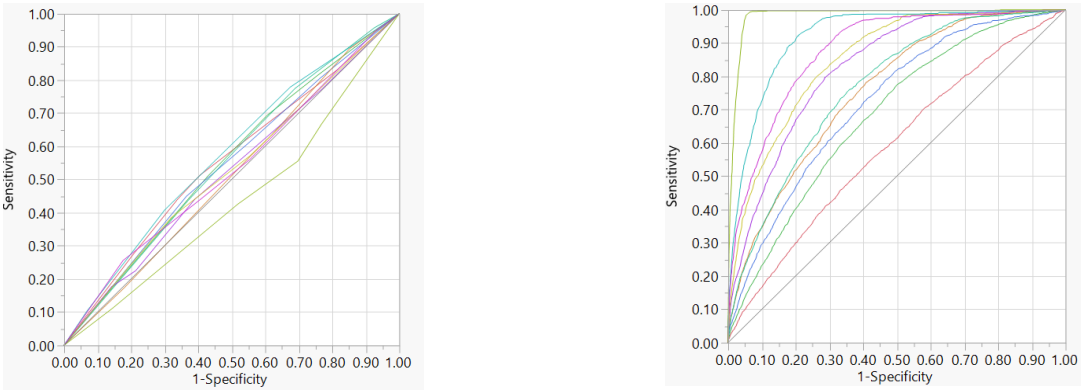


Figure 3.6

Conclusion

Airbnb wanted us to predict the first visit of a new user based on the customer demographics and their browsing history. Airbnb wanted to use the solution to identify potential countries to channelize its efforts in the new markets. In efforts to solving this, we tried to formulate traits of the customers choosing other countries as their first holiday destination. This would allow Airbnb to do a targeted marketing campaign for them. Following were our top suggestions:

1. On monitoring the clicks and time spent on each page, we realized that **changing language of browsing had a high influence on users choosing its corresponding indigenous country**. So, if a customer changed browsing language, Airbnb could increase advertisements corresponding to that parent country or even offer discounts to convert that lead.
2. Even for the US demographic, **Europe was a preferred destination in the summers – May through July** and not during the holiday season at the end of the year. This could primarily be due to the destination country weather.
3. **Women were significantly more likely to book Great Britain, France or Italy** than men. This jump should further be analyzed to learn the nature of this trip to do a targeted marketing. This can be checked using the number of people visiting during these trips and the owner of the card used to make this booking, among other things.
4. **People who input their personal details** like age, gender etc. **are significantly more likely to proceed with bookings**.
5. **Gender, age group, month of the year and first device used were the most determining parameters** to decide the destination country and further analysis on the nature of these parameters might reveal more information.