

**FINAL PROJECT**

**OPIM 5604 – Fall, 2016**



**TEAM #2**

**CLASS SECTION – AFTERNOON**

**TEAM MEMBERS:**

Mohammad Yasar Arshad  
Rikdev Bhattacharya  
Yunong Liu  
Suriyaa Sugumar Valantina  
Wenyi Xu

*"The work contained and presented here is the team's work and the team's work alone."*

## Contents

1. Executive Summary.....	4
2. Data Sampling .....	5
3. Data Exploration .....	6
3.1 Variable Classification .....	6
3.2 Univariate Analysis.....	7
3.2.1 Nominal Variables .....	7
3.2.2 Continuous Variables .....	11
3.3 Bivariate Analysis .....	14
3.3.1 Between continuous variables.....	14
3.3.2 Between Nominal Variables.....	16
3.4 D-Variable Analysis .....	17
3.5 Identifying the Target Variable .....	18
4. Data Modification .....	19
4.1 Variable Creation .....	19
4.2 Removing Inconsistencies.....	20
4.3 Missing Value Treatment .....	23
4.4 Outlier Treatment .....	24
4.5 Standardization .....	26
4.6 Dimensionality Reduction .....	26
4.6.1 Removing Non – Explanatory variables .....	26
4.6.2 Two by Two Factor Correlation.....	27
4.6.3 Principal Component Analysis.....	27
4.7 Dummy Variable Creation.....	29
4.8 Model Sampling .....	29
5. Data Modelling and Assessment.....	32
5.1 Logistic Regression .....	33
5.2 Discriminant Analysis .....	35
5.3 Neural Network.....	36
5.4 Decision Trees .....	38
5.5 Bootstrap Forest .....	40
5.6 Ensemble Model .....	42
5.7 Final Model Comparison .....	45
6. Analysis and Communication to Business.....	47

## Team 2: Expedia Hotel Bookings: Project Report

6.1	Key Findings .....	47
6.2	Insights .....	49
6.3	Business Recommendations .....	50
7.	Appendix .....	51

### List of Figures:

Figure 1: Data Summary .....	6
Figure 2: Bi-Variate Analysis between continuous variables .....	15
Figure 3: Expedia Website Booking Date .....	20
Figure 4: Expedia Website Rooms, Adults and Children Selection .....	21
Figure 5: Summary Table for Inconsistencies .....	21
Figure 6: Missing Data Pattern .....	24
Figure 7: Outlier Analysis .....	25
Figure 8 Correlation between Destination Variables .....	27
Figure 9: Principal Component Analysis of Destination Variables .....	28
Figure 10: Principal Components of the Destination Variables showing column contribution .....	28
Figure 11: Stratified Sampling .....	30
Figure 12: Training, Validation, and Testing after Stratified Sampling .....	30
Figure 13: Snapshot of dataset after sampling .....	31
Figure 14: Logistic Regression Model Summary Stats .....	33
Figure 15: Confusion Matrix for Logistic Regression .....	34
Figure 16: ROC Curve for Logistic Regression .....	34
Figure 17: Column Contributions for Logistic Regression .....	34
Figure 18: Overall Statistics – Discriminant Analysis .....	36
Figure 19: Neural Network Summary Stats .....	37
Figure 20: Neural Network Model .....	38
Figure 21: Overall Statistics – Decision Tree .....	39
Figure 22: Small Leaf Report .....	39
Figure 23: ROC Curve – Decision Tree .....	40
Figure 24: Overall Statistics – Bootstrap Forest .....	41
Figure 25: ROC Curve – Bootstrap Forest .....	41
Figure 26: Overall Statistics – Ensemble .....	43
Figure 27: Small Tree Report -Ensemble Decision Tree .....	43
Figure 28: ROC Curve - Ensemble .....	44
Figure 29: Seasonality Trend Actual vs Predicted .....	45

## ***Team 2: Expedia Hotel Bookings: Project Report***

### List of Tables:

Table 1:Variable Classification .....	7
Table 2: Nominal Variable Analysis Summary .....	8
Table 3: Continuous Variable – Analysis Summary.....	12
Table 4: Bi-Variate Analysis between Nominal Variables .....	16
Table 5: Ranking of Destination variables based on their probabilities .....	18
Table 6: Variable Creation .....	19
Table 7: Inconsistency Removal using sense check and Expedia website .....	22
Table 8: Missing Data Treatment.....	24
Table 9: Outlier Treatment .....	25
Table 10: Variable reduction summary .....	26
Table 11: Training and Validation – Sensitivity and Specificity – Logistic Regression .....	33
Table 12: Training and Validation comparison for overfitting .....	35
Table 13: Training and Validation – Sensitivity and Specificity – Discriminant Analysis.....	35
Table 14: Training and Validation comparison for Overfitting- Discriminant Analysis .....	36
Table 15: Training and Validation – Sensitivity and Specificity – Neural Network .....	37
Table 16: Training and Validation Comparison for Overfitting – Neural Network .....	38
Table 17: Training and Validation – Sensitivity and Specificity – Decision Tree .....	39
Table 18: Training and Validation Comparison for Overfitting – Decision Tree .....	40
Table 19: Training and Validation – Sensitivity and Specificity - Bootstrap Forest.....	41
Table 20: Training and Validation Comparison for Overfitting – Bootstrap Forest .....	42
Table 21: Training and Validation – Sensitivity and Specificity – Ensemble Model.....	42
Table 22: Training and Validation Comparison for Overfitting .....	44
Table 23: Model Assessment – All model comparison .....	46
Table 24: Data Dictionary .....	51
Table 25: Principal Component Analysis for Destination Variables.....	52
Table 26 : Model Prediction Formulas.....	54

## **1. Executive Summary**

### **Problem Statement:**

With over a thousand visitors every day to the Expedia website and hundreds of hotel options to choose at every destination, it is difficult to understand the user behaviour and to improve the existing conversion rates. Our objective is to develop a thorough understanding of the dataset containing information about each user session and build the best performing model to predict whether an Expedia customer will book a hotel in a specific user session or not. These predictions would be used to provide customized incentives to users to influence their booking behaviour.

### **Approach:**

We followed the SEMMA approach for data analysis. First, we subset the initial dataset for 4 hotel clusters. Second, we performed data exploration via Univariate and Bivariate analysis to understand the basic patterns stored in the data and then identified the target variable for modelling which formed the basis for further data analysis. Third, we performed a few data modification techniques with an objective to identify and fix a few preliminary errors and discrepancies in the dataset. Decisions taken from the data exploration step were used to treat missing values, handle outliers, create derived variables and remove variables with redundant information in the later phases. Fourth, we built predictive models on the model-ready dataset to predict the booking outcome in a user session. Finally, we evaluated the different models based on sensitivity, specificity, AUC, and R-Square values and chose the best model.

### **Results:**

- Identified “is\_booking” as the target variable which denotes the booking outcome in a user session. This variable will be of interest to the company to decide promotions and discounts to increase the likelihood of booking.
- In the initial dataset, we excluded 1,298 observations as inconsistencies based on secondary research/ business checks
- In missing value treatment, we removed 216 observations (from 2 variables) and 1 variable
- In outlier treatment, we removed 28 observations (from 1 variable) as extreme outliers
- We extracted 12 out of 150 features as the dominating ones in the user reviews through frequency analysis, correlation analysis and principal component analysis
- Selected Ensemble decision tree model with an overall accuracy of 62%

### **Recommendations:**

- To improve the overall accuracy of the model to predict the booking outcome, Expedia should consider capturing user web activity metrics such as session duration, bounces,

time spent on booking page, hit rate, traffic source, number of unique searches, search depth.

- To gather demographic details of users which would help in improving the model accuracy, Expedia should provide promotional offers to make first time users register on their portal.
- Provide exact location details to perform intuitive analysis
- Implement exception handling checks in forms and rectify bugs in data storage when capturing certain variables like number of adults, children, days of booking
- Devise differential strategies such as customized time bound discount deals to entice customers to make a booking in the current user sessions. Provide value add packages and incentivize customers on their next booking by making them join the Expedia loyalty program.

### **Next Steps:**

We intend to leverage information from the available user session data and thereby enable the business to devise differential strategies to influence user behaviour to improve current conversion rates. This will increase hotel bookings via the Expedia portals and ultimately increase revenue for the company.

## **2. Data Sampling**

Kaggle provided 2 raw datasets – One dataset pertaining to the user search on the Expedia website with 37.6 Million observations and 24 variables; Second – Destination dataset that contained 62,106 destinations IDs and 150 features extracted from user reviews about each destination.

The first challenge was to combine these 2 datasets. We identified the primary key between the 2 datasets and combined them on the variable “srch\_destination\_id”.

Second challenge was to extract data that was within the computational capability of JMP. We tackled this by sampling 4 hotel clusters out of the 100 present. The clusters were chosen based on the frequency of booking – non - booking events. We ensured that there were at least 10% booking events in each cluster replicating the population trends. After this analysis, clusters 6, 7, 13 and 15 were filtered. We used R to sample the dataset.

### **Data summary:**

- **Total number of observations:** 293,125
- **Total number of attributes:** 174
- **Time:** 2 years (Jan 2013 - Dec 2014)
- **Data Source:** <https://www.kaggle.com/c/expedia-hotel-recommendations/data>

## Team 2: Expedia Hotel Bookings: Project Report

File Edit Tables Rows Cols Dock Analyze Graph Tools Add-Ins View Window Help

Original Set

Source

Columns (23/21)

is\_booking

srch\_destination\_id

date\_time

site\_name

posx

continent

user\_location\_country

user\_location\_region

user\_location\_city

orig\_destination\_distance

user\_id

is\_mobile

is\_package

channel

with\_co

srch\_adults\_cnt

with\_hotels\_cnt

with\_rm\_cnt

srch\_destination\_type\_id

All rows 293,125

Selected 0

Excluded 0

Hidden 0

Labelled 0

	is_booking	srch_destination_id	date_time	site_name	posx	continent	user_location_country	user_location_region	user_location_city	orig_destination_distance	user_id	is_mobile	is_package	channel	srch_id	srch_dt
1	0	8	06/23/2014	2	3	66	442	40518	777839	434327	0	0	0	0	9	06/27/2014
2	1	8	06/04/2013	2	3	66	442	18005	144483	321996	0	0	0	0	1	06/29/2013
3	0	8	02/09/2013	2	3	66	442	51194	842105	398750	0	0	0	0	4	03/22/2013
4	0	8	06/12/2014	2	3	66	442	25538	1583924	43866	0	0	0	0	2	06/14/2014
5	0	8	12/01/2013	2	3	66	442	5875	2803007	411660	0	0	0	0	9	12/01/2013
6	0	8	11/03/2014	2	3	66	442	35390	3148888	629525	0	0	0	0	9	11/03/2014
7	0	8	10/11/2014	34	3	205	354	25315	13438226	988335	0	0	0	0	9	12/03/2014
8	0	8	06/11/2014	2	3	66	442	54240	3783763	175918	0	0	0	0	3	06/21/2014
9	0	8	06/11/2014	2	3	66	442	25538	1558396	43866	0	0	0	0	2	06/14/2014
10	0	8	07/06/2014	26	1	68	447	36752	476990	0	0	0	0	9	08/21/2014	
11	0	8	07/02/2014	2	3	66	174	24100	10806857	169973	1	0	0	0	1	07/02/2014
12	1	8	07/20/2014	2	3	66	174	24100	10806857	169973	1	0	0	0	1	07/20/2014
13	0	8	06/28/2013	2	3	66	442	19334	2972574	421429	1	0	0	0	9	06/28/2013
14	0	8	12/04/2014	2	3	66	442	19349	2923775	11314...	0	0	0	0	1	12/12/2014
15	0	8	12/03/2013	2	3	66	442	5875	2803007	411660	0	0	0	0	9	12/03/2013
16	0	8	06/04/2013	2	3	66	442	18005	144483	321996	0	0	0	0	1	06/29/2013
17	1	8	08/13/2014	2	3	66	174	30055	11794473	526269	0	1	0	0	9	10/12/2014
18	0	8	03/25/2014	2	3	66	442	76	1694218	520566	0	0	0	0	2	03/31/2014
19	0	8	08/03/2014	2	3	66	442	31964	243731	0	0	0	0	9	08/17/2014	
20	0	8	12/08/2014	2	3	66	442	76	1749467	775146	1	0	0	0	1	12/08/2014
21	1	8	11/06/2014	2	3	66	442	20850	2256994	408927	0	0	0	0	0	11/01/2014
22	1	8	06/11/2013	2	3	66	174	38972	38972	38972	0	0	0	0	9	06/21/2013
23	0	8	07/16/2014	2	3	66	442	39840	2134252	861028	0	0	0	0	9	07/16/2014
24	0	8	11/04/2013	2	3	66	442	41566	1797757	11244...	1	0	0	0	9	11/29/2013
25	1	8	03/25/2014	2	3	66	442	76	1694218	520566	0	0	0	0	2	03/31/2014
26	1	8	08/03/2014	2	3	66	442	31964	243731	0	0	0	0	9	08/17/2014	
27	0	8	07/01/2013	2	3	66	442	1010	34973	217868	0	0	0	0	4	07/12/2013
28	0	8	09/22/2013	2	3	66	340	39848	4184927	11244...	0	0	0	0	9	09/23/2013
29	1	8	06/23/2014	2	3	66	442	40518	777839	434327	0	0	0	0	9	06/27/2014
30	0	8	12/09/2014	2	3	66	442	47604	3205154	375146	1	0	0	0	0	12/09/2014
31	0	8	10/30/2014	2	3	66	340	43366	4251322	31217	1	0	0	0	9	10/31/2014
32	0	8	12/04/2014	2	3	66	442	19349	2923775	11314...	0	0	0	0	1	12/12/2014
33	0	8	07/04/2014	2	3	66	442	31132	660686	648810	0	0	0	0	1	07/04/2014
34	0	8	12/11/2014	2	3	66	226	42300	8962011	663071	1	0	0	0	1	12/31/2014
35	0	8	08/16/2014	2	3	66	294	52997	4463477	804043	0	0	0	0	9	08/25/2014
36	0	8	03/16/2014	2	3	66	442	36896	3370182	547818	0	0	0	0	4	03/16/2014
37	0	8	07/21/2013	2	3	66	174	54108	108335	169973	0	0	0	0	9	07/21/2013
38	0	8	08/20/2013	2	3	66	442	16678	1457433	654832	0	0	0	0	9	08/30/2013

Figure 1: Data Summary

### 3. Data Exploration

The first step was to explore and understand every variable in the dataset. The key things that we focused on were

- Variable type misclassification
- The distribution of each variable
- Understanding the explanatory power of each variable
- Business and sense checks to identify abnormalities, inconsistencies, and potential outliers
- Relationship between variables
- Identification of data redundancy

#### 3.1 Variable Classification

The variable type was retained or changed depending on their nature. By default, all variables were categorized as continuous. The following table lists the variables and their modified variable type.

*Table 1: Variable Classification*

Variable	Variable Type
date_time	Nominal
site_name	Nominal
posa_continent	Nominal
user_location_country	Nominal
user_location_region	Nominal
user_location_city	Nominal
orig_destination_distance	Continuous
user_id	Nominal
is_mobile	Nominal
is_package	Nominal
Channel	Nominal
srch_ci	Nominal
srch_co	Nominal
srch_adults_cnt	Continuous
srch_children_cnt	Continuous
srch_rm_cnt	Continuous
srch_destination_id	Nominal
srch_destination_type_id	Nominal
hotel_continent	Nominal
hotel_country	Nominal
hotel_market	Nominal
is_booking	Nominal
Cnt	Continuous
hotel_cluster	Nominal
d1-d149	Continuous

## **3.2 Univariate Analysis**

Each variable was carefully explored to understand the inconsistencies, potential outliers, and their explanatory power. The nominal and continuous variables were analysed in different ways.

### **3.2.1 Nominal Variables**

The nominal variables were checked for inconsistencies and missing values.

#### **Approach:**

Cols → Utilities → Recode



## Team 2: Expedia Hotel Bookings: Project Report

Table 2: Nominal Variable Analysis Summary

Variable	Explanation	Recommendation	Decision
date_time	This variable has 724 levels and the user search dates range from 7th January 2013 to 31st December 2014. We observed an increased user traffic in the Expedia website over time. Does not have direct explanatory power.	Recommend business to capture dates in the same format	Create useful variables that would impact the booking event. Refer Section 4.1
site_name	This variable has 42 levels and we noticed that 64% of the observations are from point of sale Site 2. No missing values were observed.	Recommend business to focus on point of sale Site 2 since any minor lapse would have major impact on the traffic.	Remove the variable since it has no explanatory power. Refer Section 4.6.1
posa_continent	This variable has 5 different levels. 78% of the traffic is from continent 3. No missing values observed.	<ul style="list-style-type: none"><li>- Recommend business to provide exact continent names instead of numeric levels</li><li>- Recommend business to devise specific strategies for continent 3 since it generates the maximum traffic</li></ul>	Retain the variable and create dummy variables to treat as continuous for specific models. Refer Section 4.7
user_location_country	This variable has 197 different levels. 58% of the traffic is from country 66. No missing values observed	<ul style="list-style-type: none"><li>- Recommend business to provide exact country names instead of numeric levels</li><li>- Recommend business to devise specific strategies for country 66 since it generates the maximum traffic</li></ul>	Retain the variable
user_location_region	This variable has 770 different levels. 10% of the traffic is from region 174.	- Recommend business to provide exact region names	This variable is flagged as non-explanatory and

**Team 2: Expedia Hotel Bookings: Project Report**

	No missing values observed	instead of numeric levels - Recommend business to devise specific strategies for region 174 since it generates significant traffic	removed. Refer Section 4.6.1
user_location_city	This variable has 13,152 different levels. There is no significantly high traffic from any city. No missing values observed	- Recommend business to provide exact city names instead of numeric levels	This variable is flagged as non-explanatory and removed. Refer Section 4.6.1
user_id	This variable has 92,659 different levels. ID variable and has no explanatory power. No missing values observed		This variable is flagged as non-explanatory and removed. Refer Section 4.6.1
Channel	This variable has 11 levels. 55% of the traffic is generated from channel 9. No missing values observed.	- Recommend business to provide exact channel names instead of numeric levels and capture more information on the frequency and cost of touch-points - Recommend business to invest more on channel 9 since it generates highest user traffic	Retain the variable and create dummy variables to treat as continuous for specific models. Refer Section 4.7
srch_ci	This variable has 1,078 levels and the user check-in dates range from 7th January 2013 to 18th February 2016. We observed an increased user check-in searches in the Expedia website during the last 2 weeks of December and first 2 weeks of January. Does not have direct explanatory power. 306 missing values observed.	Recommend business to capture dates in the same format	Create useful variables that would impact the booking event. Refer Section 4.1

**Team 2: Expedia Hotel Bookings: Project Report**

srch_co	This variable has 1,082 levels and the user check-out dates range from 8th January 2013 to 19th February 2016. We observed an increased user check-out searches in the Expedia website during the last 2 weeks of December and first 2 weeks of January. Does not have direct explanatory power.	Recommend business to capture dates in the same format	Create useful variables that would impact the booking event. Refer Section 4.1
srch_destination_id	This variable has 7,024 levels and does not have explanatory power. No missing values observed.		This variable is flagged as non-explanatory and removed. Refer Section 4.6.1
srch_destination_type_id	This variable has 8 levels. 58% of the searches are for destination type 1. No missing values observed.	<ul style="list-style-type: none"> <li>- Recommend business to provide exact destination type instead of numeric levels</li> <li>- Recommend business to invest more on collaborating with hotels in destination type 1</li> </ul>	Retain the variable and create dummy variables to treat as continuous for specific models. Refer Section 4.7
hotel_continent	This variable has 7 levels. 68% of the searches are for hotels in continent 2. No missing values observed.	<ul style="list-style-type: none"> <li>- Recommend business to provide exact continent names instead of numeric levels</li> <li>- Recommend business to invest more on collaborating with hotels in continent 2</li> </ul>	Retain the variable and create dummy variables to treat as continuous for specific models. Refer Section 4.7
hotel_country	This variable has 84 levels. 60% of the searches are for hotels in country 50. No missing values observed.	<ul style="list-style-type: none"> <li>- Recommend business to provide exact country names instead of numeric levels</li> <li>- Recommend business to invest</li> </ul>	Retain the variable

### Team 2: Expedia Hotel Bookings: Project Report

		more on collaborating with hotels in country 50	
hotel_market	This variable has 1,189 levels and does not have explanatory power. No missing values observed.		This variable is flagged as non-explanatory and removed. Refer Section 4.6.1
is_booking	This variable denotes whether a booking occurred or not.		Use this variable as the target. Refer Section 3.5
hotel_cluster	This variable has 4 levels chosen based on the frequency of booking events (at least 10% booking)		Retain the variable and create dummy variables to treat as continuous for specific models. Refer Section 4.7
is_mobile	This variable has 2 levels and denotes whether the search is made from a mobile or not		Retain the variable and treat as continuous for specific models
is_package	This variable has 2 levels and denotes whether the search is through a package		Retain the variable and treat as continuous for specific models

#### 3.2.2 Continuous Variables

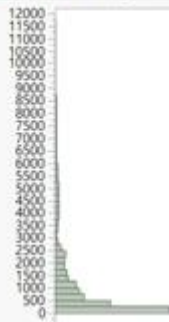
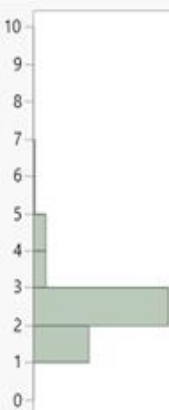
The continuous variables were explored mathematically by understanding their measures of location and variability. Statistics like min, max, mean, median, mode, percentile, standard deviation, variance etc. were studied to identify inconsistencies, outliers, and missing values.

##### Approach:

Analyse → Distribution

## Team 2: Expedia Hotel Bookings: Project Report

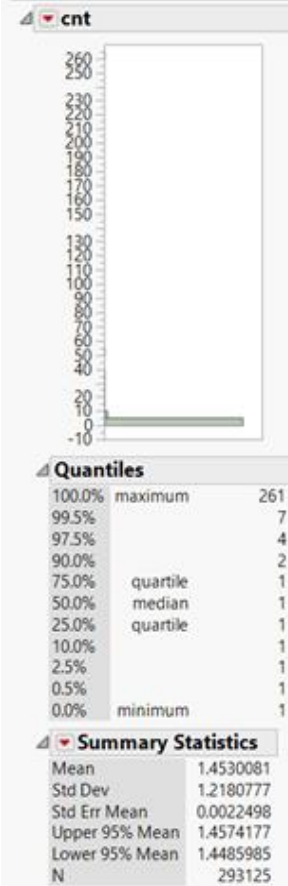

Table 3: Continuous Variable – Analysis Summary

Variable	Distribution Statistics	Explanation	Decision																																													
orig_destination_distance	<div><div>Distributions</div><div>orig_destination_distance</div><div>Quantiles</div><table><tr><td>100.0%</td><td>maximum</td><td>11734.6474</td></tr><tr><td>99.5%</td><td></td><td>9596.1654</td></tr><tr><td>97.5%</td><td></td><td>6904.5757</td></tr><tr><td>90.0%</td><td></td><td>4227.56854</td></tr><tr><td>75.0%</td><td>quartile</td><td>1668.12385</td></tr><tr><td>50.0%</td><td>median</td><td>547.94175</td></tr><tr><td>25.0%</td><td>quartile</td><td>179.6338</td></tr><tr><td>10.0%</td><td></td><td>73.8876</td></tr><tr><td>2.5%</td><td></td><td>8.2232</td></tr><tr><td>0.5%</td><td></td><td>0.6309</td></tr><tr><td>0.0%</td><td>minimum</td><td>0.0056</td></tr></table><div>Summary Statistics</div><table><tr><td>Mean</td><td>1351.1793</td></tr><tr><td>Std Dev</td><td>1882.1204</td></tr><tr><td>Std Err Mean</td><td>4.2197464</td></tr><tr><td>Upper 95% Mean</td><td>1359.4499</td></tr><tr><td>Lower 95% Mean</td><td>1342.9087</td></tr><tr><td>N</td><td>198940</td></tr></table></div>	100.0%	maximum	11734.6474	99.5%		9596.1654	97.5%		6904.5757	90.0%		4227.56854	75.0%	quartile	1668.12385	50.0%	median	547.94175	25.0%	quartile	179.6338	10.0%		73.8876	2.5%		8.2232	0.5%		0.6309	0.0%	minimum	0.0056	Mean	1351.1793	Std Dev	1882.1204	Std Err Mean	4.2197464	Upper 95% Mean	1359.4499	Lower 95% Mean	1342.9087	N	198940	This variable is right skewed and has 33% missing values	Perform missing value treatment. Refer Section 4.3
100.0%	maximum	11734.6474																																														
99.5%		9596.1654																																														
97.5%		6904.5757																																														
90.0%		4227.56854																																														
75.0%	quartile	1668.12385																																														
50.0%	median	547.94175																																														
25.0%	quartile	179.6338																																														
10.0%		73.8876																																														
2.5%		8.2232																																														
0.5%		0.6309																																														
0.0%	minimum	0.0056																																														
Mean	1351.1793																																															
Std Dev	1882.1204																																															
Std Err Mean	4.2197464																																															
Upper 95% Mean	1359.4499																																															
Lower 95% Mean	1342.9087																																															
N	198940																																															
srch_adults_cnt	<div><div>srch_adults_cnt</div><div>Quantiles</div><table><tr><td>100.0%</td><td>maximum</td><td>9</td></tr><tr><td>99.5%</td><td></td><td>6</td></tr><tr><td>97.5%</td><td></td><td>4</td></tr><tr><td>90.0%</td><td></td><td>3</td></tr><tr><td>75.0%</td><td>quartile</td><td>2</td></tr><tr><td>50.0%</td><td>median</td><td>2</td></tr><tr><td>25.0%</td><td>quartile</td><td>1</td></tr><tr><td>10.0%</td><td></td><td>1</td></tr><tr><td>2.5%</td><td></td><td>1</td></tr><tr><td>0.5%</td><td></td><td>1</td></tr><tr><td>0.0%</td><td>minimum</td><td>0</td></tr></table><div>Summary Statistics</div><table><tr><td>Mean</td><td>1.9859412</td></tr><tr><td>Std Dev</td><td>0.9385012</td></tr><tr><td>Std Err Mean</td><td>0.0017334</td></tr><tr><td>Upper 95% Mean</td><td>1.9893386</td></tr><tr><td>Lower 95% Mean</td><td>1.9825437</td></tr><tr><td>N</td><td>293125</td></tr></table></div>	100.0%	maximum	9	99.5%		6	97.5%		4	90.0%		3	75.0%	quartile	2	50.0%	median	2	25.0%	quartile	1	10.0%		1	2.5%		1	0.5%		1	0.0%	minimum	0	Mean	1.9859412	Std Dev	0.9385012	Std Err Mean	0.0017334	Upper 95% Mean	1.9893386	Lower 95% Mean	1.9825437	N	293125	The user searches have been done for number of adults ranging between 0 -10. We observed that on an average 2 adults were involved in a search event. No missing values observed.	Perform business/ sense checks to see if a search event can happen with 0 adults. Refer Section 4.2
100.0%	maximum	9																																														
99.5%		6																																														
97.5%		4																																														
90.0%		3																																														
75.0%	quartile	2																																														
50.0%	median	2																																														
25.0%	quartile	1																																														
10.0%		1																																														
2.5%		1																																														
0.5%		1																																														
0.0%	minimum	0																																														
Mean	1.9859412																																															
Std Dev	0.9385012																																															
Std Err Mean	0.0017334																																															
Upper 95% Mean	1.9893386																																															
Lower 95% Mean	1.9825437																																															
N	293125																																															

## Team 2: Expedia Hotel Bookings: Project Report

srch_children_cnt	<div><div><div><div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div>&lt;/</div></div></div></div>
-------------------	---

## Team 2: Expedia Hotel Bookings: Project Report

Cnt	 <p><b>Quantiles</b></p> <table><tr><td>100.0%</td><td>maximum</td><td>261</td></tr><tr><td>99.5%</td><td></td><td>7</td></tr><tr><td>97.5%</td><td></td><td>4</td></tr><tr><td>90.0%</td><td></td><td>2</td></tr><tr><td>75.0%</td><td>quartile</td><td>1</td></tr><tr><td>50.0%</td><td>median</td><td>1</td></tr><tr><td>25.0%</td><td>quartile</td><td>1</td></tr><tr><td>10.0%</td><td></td><td>1</td></tr><tr><td>2.5%</td><td></td><td>1</td></tr><tr><td>0.5%</td><td></td><td>1</td></tr><tr><td>0.0%</td><td>minimum</td><td>1</td></tr></table> <p><b>Summary Statistics</b></p> <table><tr><td>Mean</td><td>1.4530081</td></tr><tr><td>Std Dev</td><td>1.2180777</td></tr><tr><td>Std Err Mean</td><td>0.0022498</td></tr><tr><td>Upper 95% Mean</td><td>1.4574177</td></tr><tr><td>Lower 95% Mean</td><td>1.4485985</td></tr><tr><td>N</td><td>293125</td></tr></table>	100.0%	maximum	261	99.5%		7	97.5%		4	90.0%		2	75.0%	quartile	1	50.0%	median	1	25.0%	quartile	1	10.0%		1	2.5%		1	0.5%		1	0.0%	minimum	1	Mean	1.4530081	Std Dev	1.2180777	Std Err Mean	0.0022498	Upper 95% Mean	1.4574177	Lower 95% Mean	1.4485985	N	293125	This variable is right skewed with few extreme outliers. No missing values observed	Perform outlier treatment. Refer Section 4.4
100.0%	maximum	261																																														
99.5%		7																																														
97.5%		4																																														
90.0%		2																																														
75.0%	quartile	1																																														
50.0%	median	1																																														
25.0%	quartile	1																																														
10.0%		1																																														
2.5%		1																																														
0.5%		1																																														
0.0%	minimum	1																																														
Mean	1.4530081																																															
Std Dev	1.2180777																																															
Std Err Mean	0.0022498																																															
Upper 95% Mean	1.4574177																																															
Lower 95% Mean	1.4485985																																															
N	293125																																															
d1-d149		This variable is log transformed probability values of the features occurring in each user review about the hotel	Perform dimensionality reduction on these 149 variables. Refer Section 3.4, 4.6.2 and 4.6.3																																													

### 3.3 Bivariate Analysis

The interaction and relationship between variables were studied using Scatter Plots. This step was essential to identify the redundant information captured.

#### 3.3.1 Between continuous variables

##### Approach:

Analyze → Multivariate Methods → Multivariate

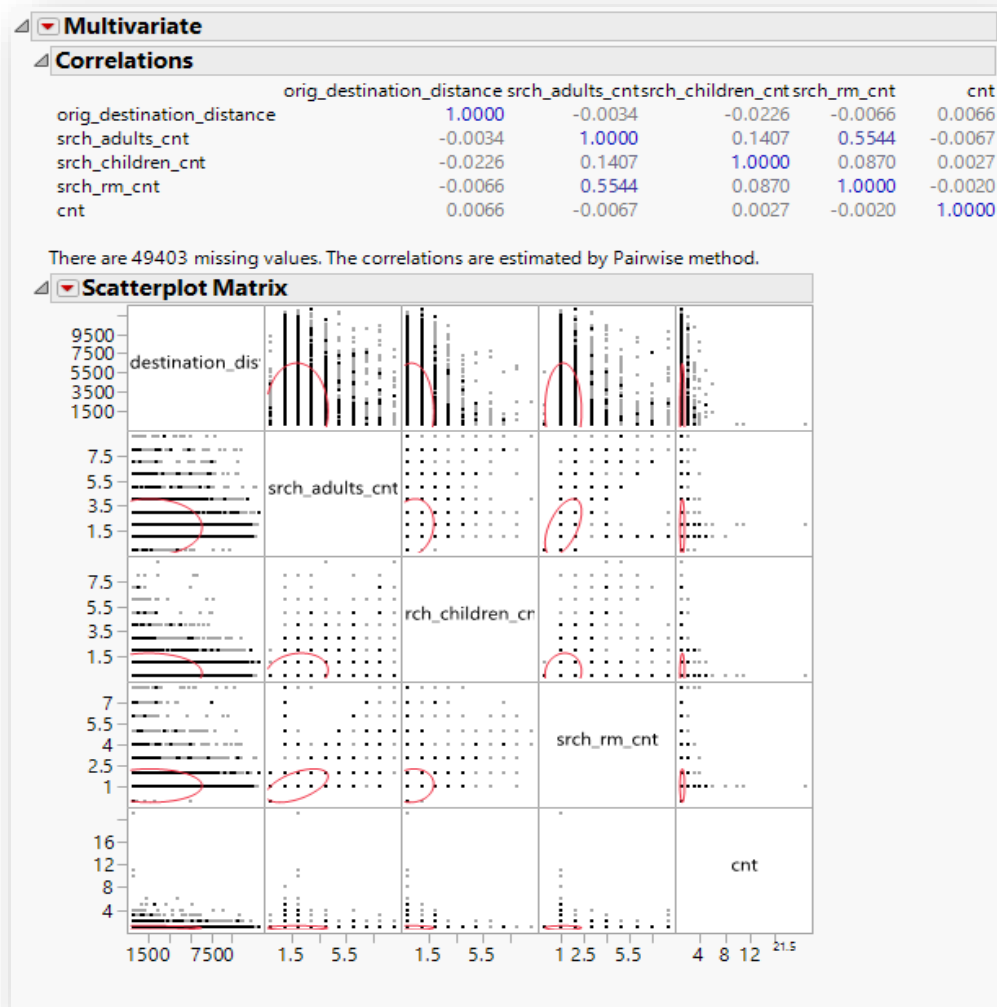


Figure 2: Bi-Variate Analysis between continuous variables

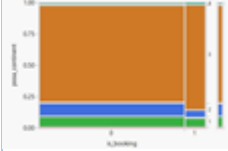
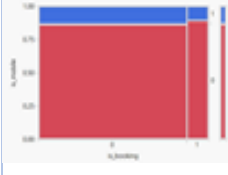
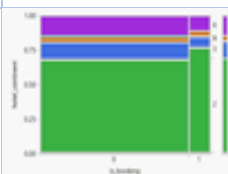
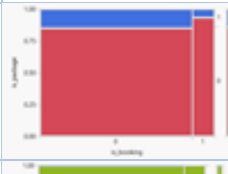
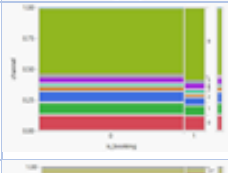

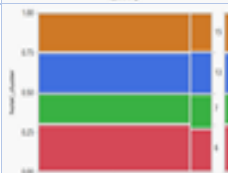
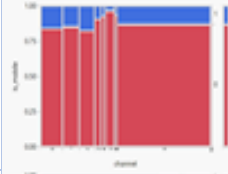
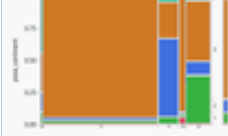
We observed that there was no strong correlation between any of the continuous variables except between 'srch\_rm\_cnt' and 'srch\_adults\_cnt'. The correlation value was 0.5544 which implies that for every room searched there was an adult. Since there is no significant correlation, we retained both variables.


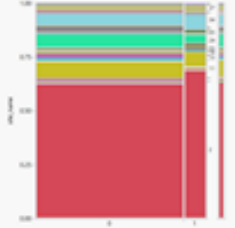
The remaining 150 destination feature variables were treated separately to understand their significance and reduce dimensions. Refer Section 3.4



### 3.3.2 Between Nominal Variables

Table 4: Bi-Variate Analysis between Nominal Variables

Variable 1 Y-Axis	Variable 2 X-Axis	Visualization	Explanation
POSA_continent	Is_booking		Data distributed sparsely between continents. Majority of the bookings are done from Continent 3.
Is_mobile	Is_booking		Considerable traffic through mobile devices, but that does not necessarily convert into actual bookings. No significant difference in booking event based on the traffic type.
Hotel_continent	Is_booking		Data distributed sparsely among the different continents. Considerably higher interest in destinations of continent 2 as compared to the others.
Is_package	Is_booking		Customers are more inclined to book non-package hotels.
Channel	Is_booking		The proportion of traffic from different channels vary considerably. Channel 9 has significant impact on bookings.
Site_name	Is_booking		The proportion of traffic from different sites vary considerably. Site 2 has the most significant impact on bookings.
Hotel_cluster	Is_booking		The proportion of interests in each hotel cluster seems to be constant along with the corresponding booking outcomes in each.
Is_mobile	Channel		Mobile traffic across each of the channel seems to remain constant
Posa_continent	Hotel_continent		Significantly high interest in users from continent 3 to travel to continent 2.

Hotel_cluster	Hotel_continent		The hotel clusters have equal distribution across the different continents
Site_name	Is_mobile		Site 2 receives significantly higher mobile traffic compared to the other sites.

### 3.4 D-Variable Analysis

#### Destination latent vector analysis:

D0-D149 listed in **destination.csv** are latent description of hotel reviews that are related to a given search destination. These columns correspond to different facets (e.g. swimming pool, beach, ski, bar etc.) and values are (log) probabilities that a customer would endorse a hotel in the destination for a specific facet. Each search destination has 150 features and each feature has been assigned a probability based user reviews. The feature with highest probability is the dominant feature of the hotel and can play significant role in the booking.

**Objective:** To extract the dominating endorsed features that would influence the user booking event from the 150 features available for each search destination

#### Approach:

1. We identified the top 4 dominating features in each destination review because they contributed to approximately 50% probability of occurrence and each of the remaining 146 features had negligible probabilities
2. We recorded them in 4 groups; G1 listed the most dominating feature, G2: Second dominating feature, G3: Third dominating feature, G4: Fourth dominating feature
3. We calculated the frequency of each of the features in the 4 groups; where we shortlisted 15 features within each group which contributed to 96% of the total occurrences. (Refer Table 5)
4. We observed that only 24 of them were unique among the 4 groups
5. These 24 features were further reduced using Dimensionality reduction techniques. Refer Section 4.6.2 and 4.6.3

## Team 2: Expedia Hotel Bookings: Project Report

Table 5: Ranking of Destination variables based on their probabilities

G1	G2	G3	G4
Most Important Feature	2 <sup>nd</sup> Most Important Feature	3 <sup>rd</sup> Most Important Feature	4 <sup>th</sup> Most Important Feature
d37	d93	d93	d49
d9	d37	d49	d139
d108	d43	d43	d43
d43	d108	d108	d108
d131	d49	d139	d93
d58	d103	d6	d6
d49	d58	d37	d16
d71	d9	d58	d58
d20	d79	d9	d132
d6	d139	d16	d80
d139	d94	d122	d37
d110	d6	d71	d9
d141	d71	d29	d103
d88	d88	d94	d122
d79	d146	d103	d29

### Result:

Based on the above results we retained the following variables:

d6, d9, d16, d20, d29, d37, d43, d49, d58, d71, d79, d80, d88, d93, d94, d103, d108, d110, d122, d131, d132, d139, d141, d146.

### 3.5 Identifying the Target Variable

Expedia wants to improve the current conversion rate of the users visiting their portals. The variable 'is\_booking' explains whether the user would make a booking in the existing user session or not. This is identified as our target variable. Is\_booking is a nominal variable with two levels 0 and 1; 0 represents not booking and 1 represents booking.

The focal group in this variable will be "0" which represents users who would possibly not book in a user session. This would help Expedia to devise differential targeting strategies that would help influence these customers to make a booking.

## 4. Data Modification

### 4.1 Variable Creation

Since the dataset contained nominal variables (date variables) that did not have direct explanatory power, we decided to create variables that would provide useful information and would impact the booking event.

Table 6: Variable Creation

Variable	Variable Type	Derivation *	Reason
no_chkin_days	Continuous	<b>no_chkin_days = srch_co - srch_ci</b> This variable denotes the number of days the user wants to book a hotel for	Since the date variables do not have direct explanatory power, this variable was created to see the effect of number of check-in days on the booking event
no_days_to_chkin	Continuous	<b>no_days_to_chkin = srch_ci - date_time</b> This variable denotes the number of days left for desired check-in date	Since the date variables do not have direct explanatory power, this variable was created to see the effect of number of days left to check - in on the booking event
search_month	Continuous	<b>search_month = month(date_time)</b> This variable denotes the user search month	Since the date variables do not have direct explanatory power, this variable was created to see the effect of the search month on the booking event
ci_month	Continuous	<b>ci_month = month(srch_ci)</b> This variable denotes the user check-in month	Since the date variables do not have direct explanatory power, this variable was created to see the booking patterns

## Team 2: Expedia Hotel Bookings: Project Report

			based on the check-in months
co_month	Continuous	<b>co_month = month(srch_co)</b> This variable denotes the user check-out month	Since the date variables do not have direct explanatory power, this variable was created to see the booking patterns based on the check-out months

\* Since the date fields were in different formats, we used R for manipulation and calculating the above variables

### 4.2 Removing Inconsistencies

We performed business/ sense checks by exploring Expedia website to get the permissible values for certain user search variables and found the following abnormalities in our current dataset.

#### Approach:

Tables → Summary

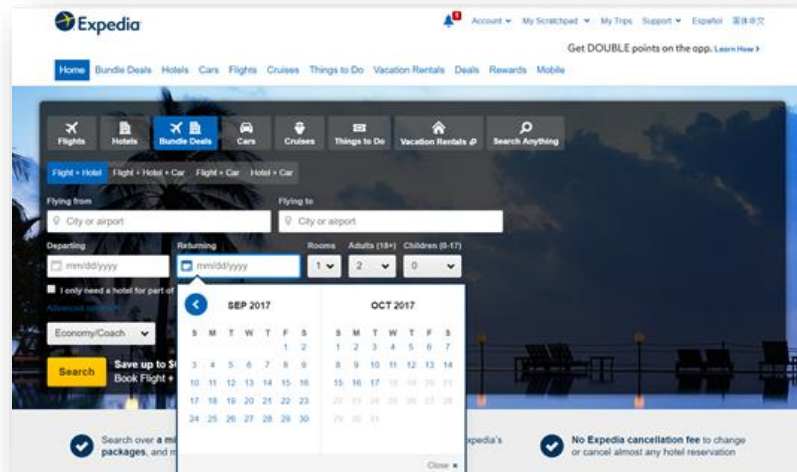


Figure 3: Expedia Website Booking Date

## Team 2: Expedia Hotel Bookings: Project Report

Figure 4: Expedia Website Rooms, Adults and Children Selection

Original Set By (srch\_adults\_cnt, srch\_children\_cnt, srch\_rm\_cnt) - JMP Pro

	srch_adults_cnt	srch_children_cnt	srch_rm_cnt	N Rows
1	0	0	0	7
2	0	0	1	332
3	0	0	2	15
4	0	1	1	5
5	0	2	1	2
6	0	3	1	1
7	1	0	1	65138
8	1	0	2	33
9	1	0	3	61
10	1	0	4	51
11	1	0	5	203
12	1	0	6	152
13	1	0	7	80
14	1	0	8	207
15	1	1	1	6286
16	1	1	2	44
17	1	1	3	5
18	1	1	4	1
19	1	1	5	37
20	1	1	6	20
21	1	1	7	5
22	1	1	8	16
23	1	2	1	1782
24	1	2	2	3
25	1	2	3	5
26	1	2	4	2
27	1	2	5	3
28	1	2	6	8
29	1	2	8	3
30	1	3	1	400
31	1	3	4	2
32	1	3	5	1
33	1	3	6	2
34	1	3	7	5
35	1	3	8	2
36	1	4	1	44
37	1	4	3	1
38	1	4	4	4
39	1	4	5	4

Columns (4/1)

- srch\_adults\_cnt
- srch\_children\_cnt
- srch\_rm\_cnt
- N Rows

Rows

All rows	274
Selected	1
Excluded	0
Hidden	0
Labelled	0

Figure 5: Summary Table for Inconsistencies

## Team 2: Expedia Hotel Bookings: Project Report

Table 7: Inconsistency Removal using sense check and Expedia website

Variable	Business/Sense Check	Decision	Recommendation
no_chkin_days	Negative values - The check-out date cannot be prior to the check-in date. There were 55 instances with negative values	Remove the observations with negative values in no_chkin_days. <b>55 observations were removed.</b>	Recommend business to take exception handling measures or form control while capturing such variables
no_chkin_days	Maximum value - The maximum duration of stay that can be planned per Expedia website is 27 days. There were 259 instances with no_chkin_days more than 27	Remove the observations with no_chkin_days > 27. <b>259 observations were removed.</b>	Recommend business to take exception handling measures or form control while capturing such variables
no_days_to_chkin	Negative values - The check-in date searched should be at least on the day of search and not prior to it. There were 30 instances with negative values.	Remove the observations with negative values in no_days_to_chkin. <b>30 observations were removed.</b>	Recommend business to take exception handling measures or form control while capturing such variables
no_days_to_chkin	Maximum value - The maximum number of days within which a user can search hotel per Expedia website is 334 days. There were 554 instances with no_days_to_chkin values more than 334 days.	Remove the observations with no_days_to_chkin > 334. <b>554 observations were removed.</b>	Recommend business to take exception handling measures or form control while capturing such variables
srch_children_cnt	The number of children can be 1 or more provided there is at least 1 adult in the search per the Expedia website. There were 8 observations with number of adults searched as 0	Remove the observations with srch_adults_cnt = 0 and srch_children_cnt > 0. <b>8 observations were removed.</b>	Recommend business to take exception handling measures or form control while capturing such variables

### Team 2: Expedia Hotel Bookings: Project Report

srch_children_cnt	Maximum value - The number of children can be at most 6 provided 1 room is searched per the Expedia website. There were 6 observations with number of children greater than 6.	Remove the observations with srch_rm_cnt = 1 and srch_children_cnt > 6. <b>6 observations were removed.</b>	Recommend business to take exception handling measures or form control while capturing such variables
srch_adults_cnt	Minimum value - The minimum number of adults required to perform a search is 1 per the Expedia website. There were 353 observations with number of adults searched as 0	Remove the observations with srch_adults_cnt = 0. <b>354 observations were removed.</b>	Recommend business to take exception handling measures or form control while capturing such variables
srch_adults_cnt	Maximum value - The number of adults can be at most 6 provided 1 room is searched per the Expedia website. There were 32 observations with number of adults greater than 6.	Remove the observations with srch_rm_cnt = 1 and srch_adults_cnt > 6. <b>32 observations were removed.</b>	Recommend business to take exception handling measures or form control while capturing such variables
srch_rm_cnt	Minimum value - The minimum number of rooms required to perform a search is 1 per the Expedia website. There were no observations with number of rooms searched as 0	Remove the observations with srch_rm_cnt = 0. <b>No observations were removed.</b>	Recommend business to take exception handling measures or form control while capturing such variables

### 4.3 Missing Value Treatment

The next step was to understand the missing data pattern

#### Approach:

Table → Missing Data Pattern



## Result:

Count	Number of columns missing	Patterns	is_booking	srch_destination_id	date_time	site_name	posa_continent	user_location_country	user_location_region	user_location_city
198729	0	00000000000000000000000000000000	0	0	0	0	0	0	0	0
211	2	0000000000000011000000000	0	0	0	0	0	0	0	0
94090	1	0000000010000000000000000	0	0	0	0	0	0	0	0
95	3	0000000010000110000000000	0	0	0	0	0	0	0	0

Figure 6: Missing Data Pattern

Table 8: Missing Data Treatment

Variable	# Missing Values	Decision
orig_destination_distance	94,090 (33%)	Removed the variable because: - Imputation was not possible because there was no strong correlation for orig_destination_distance with any of the other continuous variables - Removing the missing observations was not possible because we would lose one-third of the entire dataset
srch_ci	216 (0.001%)	Insignificant number of missing values. Hence, we removed 216 observations
srch_co		

## 4.4 Outlier Treatment

The next step was to analyze the distribution of continuous variables, identify and treat outliers.

### Approach:

Analyze → Distributions

Result:

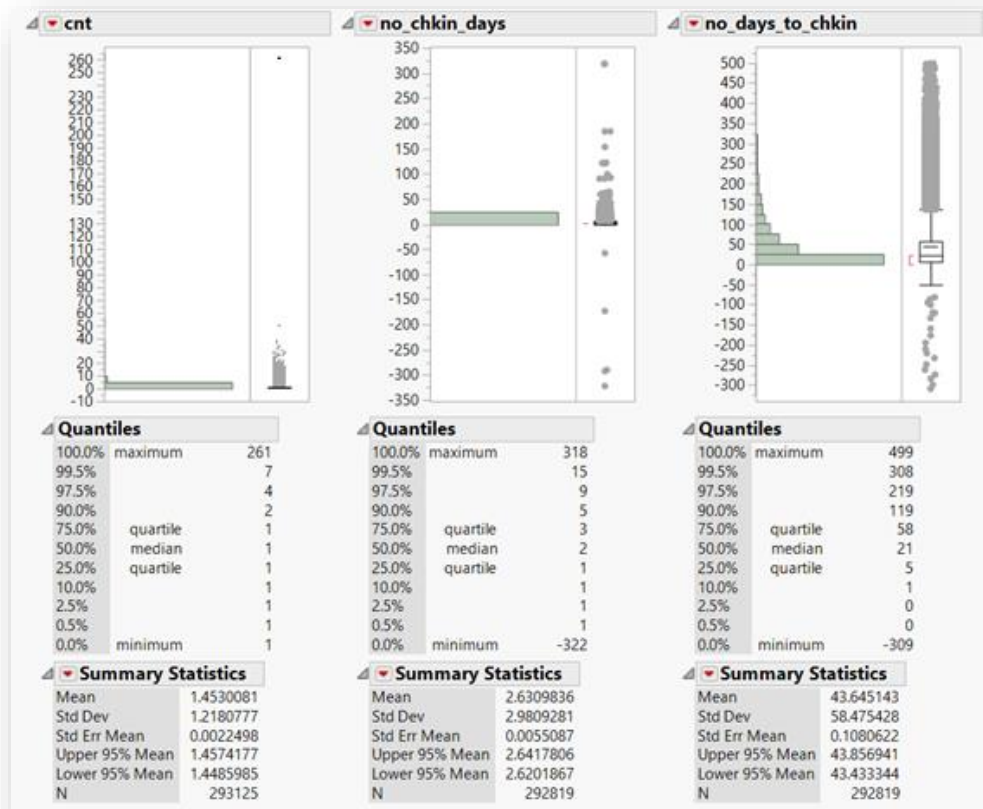


Figure 7: Outlier Analysis

Table 9: Outlier Treatment

Variable	Frequency above 1.5 IQR	Decision
Cnt	72,518	Limited the number of similar events in the context of same user session to at most 20. 28 observations with cnt values > 20 were considered insignificant and removed
no_chkin_days	18,483	These values are permissible per the business check from the Expedia website. Refer Section 4.2
no_days_to_chkin	22,575	

## 4.5 Standardization

Continuous variables have different scales due to different units of measurement. All variables were transformed to a scale of 0 to 1 so that one variable doesn't dominate other variables while building the model and are comparable.

### Formula:

$$\text{Standardized value} = \frac{Xi - \text{col min}(X)}{\text{col max}(X) - \text{col min}(X)}$$

### Approach:

Right click on the column → New Formula Column → Distributional → Range 0 to 1

### Result:

The following variables were standardized:

- srch\_adults\_cnt
- srch\_children\_cnt
- srch\_rm\_cnt
- No\_chkn\_days
- No\_days\_to\_chkin

## 4.6 Dimensionality Reduction

### 4.6.1 Removing Non – Explanatory variables

Based on univariate analysis, the following variables have been removed.

Table 10: Variable reduction summary

Variable	Reason
date_time	This variable does not have direct explanatory power and was used to create derived variables. Hence this variable was removed to avoid redundancy
site_name	This variable was too granular and hence was removed
user_location_region	This variable was too granular and hence was removed
user_location_city	This variable was too granular and hence was removed
user_id	This variable is the ID which is unique for every user session and hence was removed
srch_ci	This variable does not have direct explanatory power and was used to create derived variables. Hence this variable was removed to avoid redundancy

srch_co	This variable does not have direct explanatory power and was used to create derived variables. Hence this variable was removed to avoid redundancy
hotel_market	This variable was too granular and hence was removed
srch_destination_id	This variable is the ID which is unique for every destination and hence was removed
hotel_region	This variable was too granular and hence was removed
hotel_city	This variable was too granular and hence was removed

#### 4.6.2 Two by Two Factor Correlation

Based on multivariate analysis results, we extracted top 24 destination features based on the user reviews. We checked the correlation between each of the 24 variables.

	d6	d9	d16	d20	d29	d37	d43	d49	d58	d71	d79	d80	d88	d93	d94	d103	d108	d110	d122
d6	1.0000	0.1411	0.3152	-0.1912	0.2882	0.6250	0.2292	0.5504	0.1535	-0.2121	-0.3475	0.1414	0.0662	0.6247	-0.0973	0.0562	0.2891	-0.3846	0.2900
d9	0.1411	1.0000	-0.3077	-0.0968	-0.1177	0.0472	0.1494	0.0866	-0.1632	-0.2093	-0.1917	-0.2886	0.2409	0.0351	-0.3137	0.6273	-0.0498	-0.2459	-0.1614
d16	0.3152	-0.3077	1.0000	-0.0454	0.3614	0.3964	0.1759	0.3406	0.0033	-0.1183	-0.2570	0.1396	-0.0314	0.4029	-0.0892	-0.3033	0.2782	-0.2022	0.6165
d20	-0.1912	-0.0968	-0.0454	1.0000	0.0100	-0.1473	-0.1016	-0.0582	-0.1636	0.0422	0.0342	-0.1738	-0.0907	-0.1547	0.0986	-0.1159	-0.2044	0.0748	0.0459
d29	0.2882	-0.1177	0.3614	0.0100	1.0000	0.5645	0.5198	0.3509	-0.0916	0.0260	-0.0850	-0.1500	-0.0734	0.5525	-0.0806	-0.1072	0.1110	-0.0610	0.3464
d37	0.6250	0.0472	0.3964	-0.1473	0.5645	1.0000	0.4729	0.7413	0.0673	-0.1336	-0.2044	-0.0073	0.0427	0.9904	-0.1557	0.0306	0.4518	-0.2717	0.3422
d43	0.2292	0.1494	0.1759	-0.1016	0.5198	0.4729	1.0000	0.4551	-0.1089	0.0552	-0.1296	-0.1665	-0.1302	0.4405	-0.1785	0.1391	0.1070	-0.1841	0.4158
d49	0.5504	0.0866	0.3406	-0.0582	0.3509	0.7413	0.4551	1.0000	-0.0845	-0.1202	-0.1788	-0.1396	-0.0505	0.7097	-0.1961	0.0645	0.2869	-0.2916	0.3544
d58	0.1535	-0.1632	0.0033	-0.1636	-0.0916	0.0673	-0.1089	-0.0845	1.0000	-0.1002	-0.0235	0.6324	-0.0489	0.0590	0.2420	-0.0993	0.2022	-0.1142	-0.0128
d71	-0.2121	-0.2093	-0.1183	0.0422	0.0260	-0.1536	0.0552	-0.1202	-0.1002	1.0000	0.2773	-0.0783	-0.1568	-0.1634	0.0584	-0.2366	-0.1419	0.3850	0.0125
d79	-0.3475	-0.1917	-0.2570	0.0342	-0.0850	-0.2044	-0.1296	-0.1788	-0.0235	0.2773	1.0000	-0.0789	-0.0933	-0.2158	0.1084	-0.1312	-0.1215	0.6417	-0.1757
d80	0.1414	-0.2886	0.1396	-0.1738	-0.1500	-0.0073	-0.1665	-0.1396	0.6324	-0.0783	-0.0789	1.0000	-0.0855	0.0048	0.1578	-0.2124	0.2746	-0.1587	0.0936
d88	0.0662	0.2409	-0.0314	-0.0907	-0.0734	0.0427	-0.1302	-0.0505	-0.0489	-0.1568	-0.0933	-0.0855	1.0000	0.0493	-0.2220	0.2153	0.0477	-0.0943	0.0391
d93	0.6247	0.0351	0.4029	-0.0892	0.1110	0.9904	0.4405	0.7097	0.0590	-0.1634	-0.2158	0.0048	0.0493	1.0000	-0.1580	0.0213	0.4582	-0.2843	0.3270
d94	-0.0973	-0.3137	-0.0892	0.0986	-0.1072	-0.1557	-0.1785	-0.1961	0.2420	0.0584	0.1084	0.1578	-0.2220	-0.1580	1.0000	-0.2975	-0.0721	0.1496	-0.1172
d103	0.0562	0.6273	-0.3033	-0.1159	-0.1072	0.0306	0.1391	0.0645	-0.0993	-0.2366	-0.1312	-0.2124	0.2153	0.0213	-0.2975	1.0000	-0.0101	-0.2348	-0.2042
d108	0.2891	-0.0498	0.2782	-0.2044	0.1110	0.4518	0.1070	0.2869	0.2022	-0.1419	-0.1215	0.2746	0.0477	0.4582	-0.0721	-0.0101	1.0000	-0.2059	0.1367
d110	-0.3846	-0.2459	-0.2022	0.0748	-0.0610	-0.2717	-0.1841	-0.2916	-0.1142	0.3850	0.6417	-0.1587	-0.0947	-0.2843	0.1496	-0.2348	-0.2059	1.0000	-0.1410
d122	0.2900	-0.1614	0.6165	-0.0459	0.3464	0.3422	0.4158	0.3544	-0.0128	0.0125	-0.1757	0.0936	-0.0391	0.3270	-0.1172	-0.2042	0.1367	-0.1410	1.0000
d131	-0.1153	-0.1038	0.0660	-0.1247	0.0542	0.0196	0.4461	0.1138	-0.1363	-0.0104	-0.0687	-0.0222	-0.1037	0.0192	-0.1708	-0.1062	0.0339	-0.1412	0.1674
d132	0.2010	-0.1205	-0.1348	-0.2248	-0.2433	0.0608	-0.1846	0.0035	0.5996	-0.0730	-0.0069	0.6000	-0.0247	0.0676	0.1364	-0.0859	0.2706	-0.1682	0.1210
d139	0.5078	0.0393	0.2109	-0.1367	0.3993	0.6429	0.3761	0.3731	0.2751	-0.1628	-0.1845	0.2415	0.0643	0.6479	-0.0092	0.0783	0.5195	-0.2897	0.2255
d141	-0.2728	-0.1359	-0.1879	0.2738	-0.0423	-0.1180	-0.0415	-0.1298	-0.1117	0.1102	0.1166	-0.1494	-0.1263	-0.1223	0.1789	-0.0987	-0.1540	0.1227	-0.1430
d146	-0.3610	-0.1741	-0.1811	0.1255	-0.0823	-0.3552	-0.2448	-0.3422	-0.1279	0.1282	0.0895	-0.1586	-0.0360	-0.3514	0.0967	-0.1862	-0.2909	0.1753	-0.2201

Figure 8 Correlation between Destination Variables

#### Result:

Variables d37 and d93 were strongly correlated (correlation value = 0.994). We decided to remove the variable d93. The remaining 23 variables were carried forward to PCA.

#### 4.6.3 Principal Component Analysis

##### Identifying the dominating features from user reviews:

Out of the 23 variables retained from Two by two correlation analysis, we observed that 12 principal components account for almost 85% variation. So, we decided to consider 12 components for our further analysis.

## Team 2: Expedia Hotel Bookings: Project Report

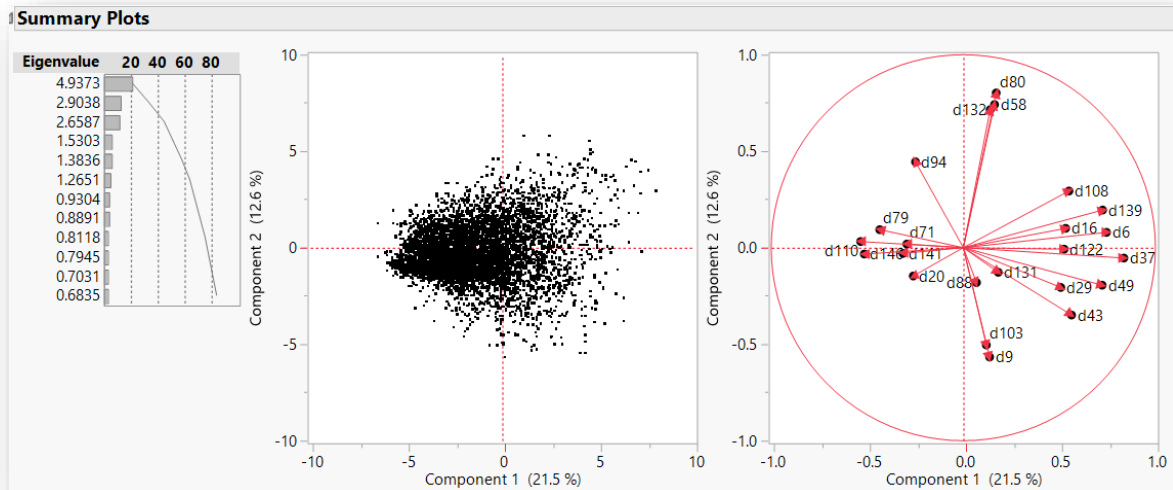


Figure 9: Principal Component Analysis of Destination Variables

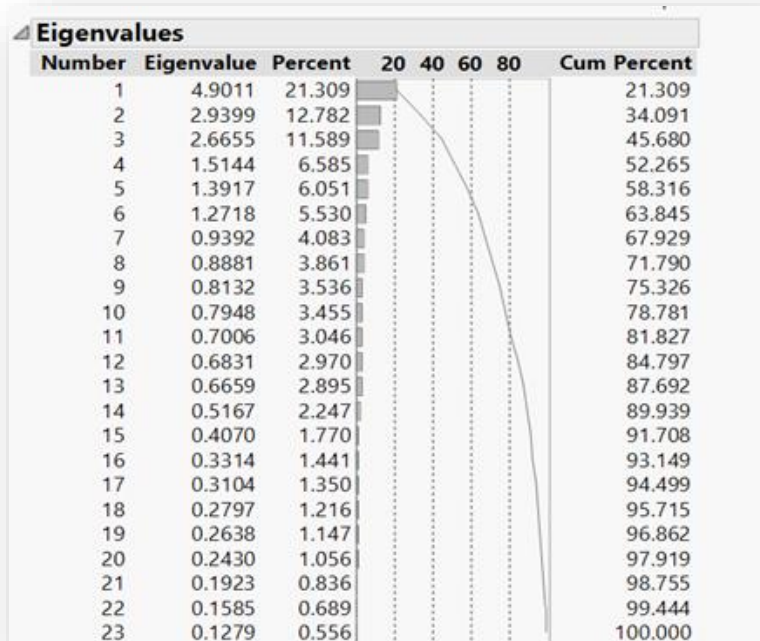


Figure 10: Principal Components of the Destination Variables showing column contribution

## **4.7 Dummy Variable Creation**

Certain models like Discriminant Analysis and Logistic Regression accept only continuous predictor variables. To retain the explanatory power of the nominal variables, we created dummy variables. We accomplished this in R.

### **Approach:**

If there were 'n' distinct values in a variable, we created  $n - 1$  dummy variables.

### **Result:**

Dummy variables were created for the following shortlisted nominal variables -

- posa\_continent
- channel
- srch\_destination\_type\_id
- hotel\_continent
- hotel\_cluster
- booking\_month
- ci\_month

## **4.8 Model Sampling**

In this step, we split the dataset into training, validation, and testing. This decision was taken because the classes in the target variable were not balanced and dominated by the non - booking event (i.e. ~90% 0's denoting non - booking event and ~10% 1's denoting booking event). The dataset was split in the following proportion to retain maximum information.

- Balanced Training: 50%
- Balanced Validation: 15%
- Unbalanced Testing: 35%

## Team 2: Expedia Hotel Bookings: Project Report

Select Column

- RowID
- is\_booking
- posa\_continent
- user\_location\_county
- is\_mobile
- is\_package
- channel
- srch\_adults\_cnt
- srch\_children\_cnt
- srch\_rm\_cnt

Specify Data Proportions

Training Set: 0.5  
Validation Set: .15  
Test Set: 0.35

☐ Don't Alter Group Proportions  
☐ Alter Proportions in Training  
☒ Alter Proportions in Both Training and Validation

Select Focal Group

0  
1

Focal Group Proportion: 0.5

☐ Balance Remaining Groups (Default is to maintain present ratios)  
☒ Balance All Groups (Focal group proportion will be ignored)  
☐ Bootstrap Augmentation (Leave unchecked for random trimming)

Action

OK  
Cancel  
Help

Figure 11: Stratified Sampling

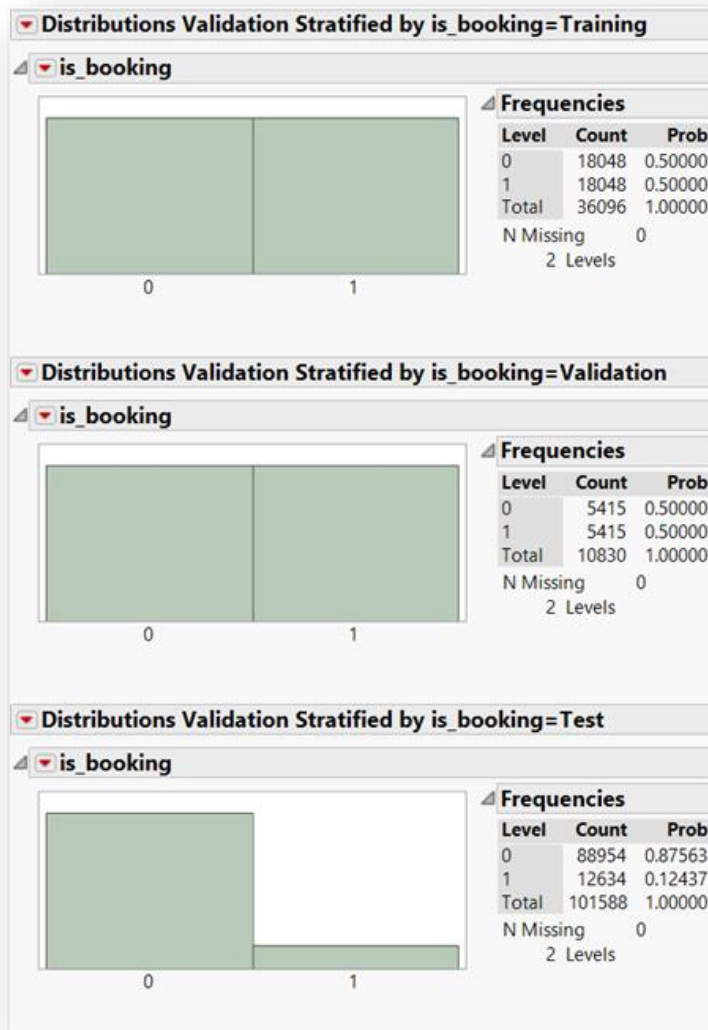
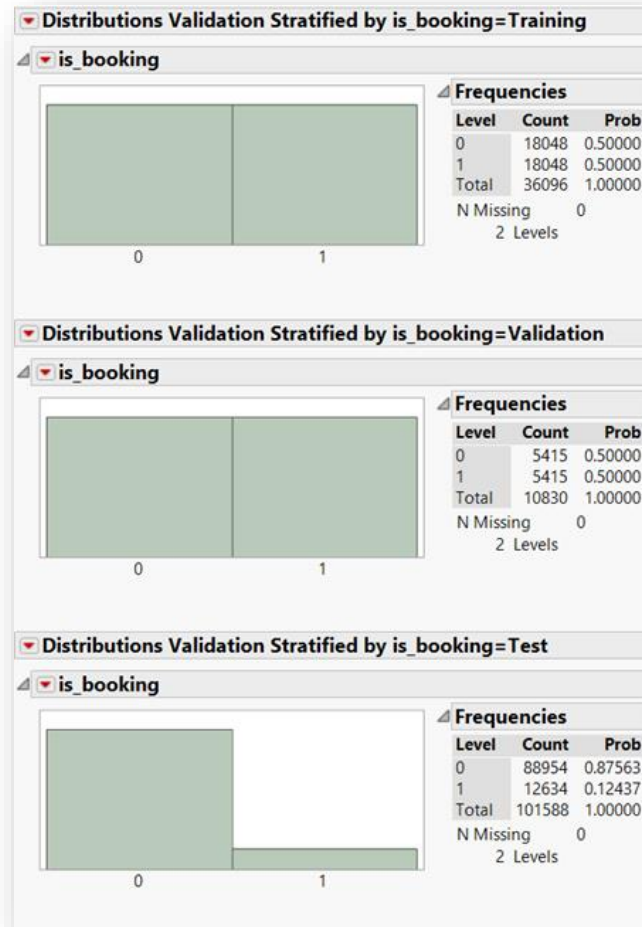


Figure 12: Training, Validation, and Testing after Stratified Sampling



## Team 2: Expedia Hotel Bookings: Project Report



VL-Expedia_Outlier Treat...		RowID	is_booking	posn_continent	user_location_co	country	is_mobile	is_package	channel	srch_adults_cnt	srch_children_cnt	srch_rm_cnt	srch_destination_type_id	hotel_continent	hotel_country	hotel_cluster
Source		36075	1	3	66	0	0	9	1	0	1	5	3	82	6	
		36076	1	3	205	0	0	5	1	1	1	1	2	198	6	
		36077	1	3	66	0	0	9	1	0	1	3	2	50	7	
		36078	1	3	66	0	0	7	2	0	1	1	2	50	7	
		36079	1	2	194	0	1	0	2	0	1	1	3	182	6	
		36080	1	3	66	0	0	9	2	1	1	6	2	50	15	
		36081	1	3	66	1	0	9	4	0	2	6	4	93	15	
		36082	1	3	66	0	0	9	2	2	1	6	2	50	7	
		36083	1	3	66	0	0	2	1	0	1	1	2	50	8	
		36084	1	3	66	0	0	2	2	3	1	1	2	50	7	
		36085	1	3	66	0	0	0	3	6	3	1	2	50	8	
		36086	1	3	66	0	0	9	2	0	1	5	2	50	6	
		36087	1	3	66	1	0	9	1	0	1	5	2	198	6	
		36088	1	3	66	0	0	0	2	0	1	4	2	50	15	
		36089	1	3	66	0	0	9	1	0	1	1	4	131	15	
		36090	1	2	3	0	0	0	3	0	3	1	3	168	6	
		36091	1	3	66	0	0	1	1	0	1	1	2	50	13	
		36092	1	3	230	0	0	9	2	1	1	1	6	70	6	
		36093	1	3	66	0	0	1	2	0	1	3	2	50	7	
		36094	1	3	66	0	1	2	1	0	1	1	2	50	13	
		36095	1	2	119	0	0	3	2	0	1	6	3	106	13	
		36096	1	3	66	0	0	0	2	0	1	4	2	50	4	
		36097	1	3	66	1	0	1	2	0	1	1	2	50	7	
D Dummy Variable (148/0)		36098	0	2	3	0	0	0	2	0	2	1	6	204	7	
Booking Month Dummy (1)		36099	0	3	66	0	0	9	1	0	1	1	2	50	13	
Check in Month Dummy (1)		36100	0	3	205	0	0	9	2	2	1	1	2	50	7	
Posn Continent Dummy (4)		36101	0	3	66	0	0	9	2	0	1	6	2	50	13	
Hotel Continent Dummy (1)		36102	0	3	66	0	0	9	2	0	1	5	2	50	15	
Channel Dummy (150/0)		36103	0	1	1	0	0	0	2	0	1	1	2	50	13	
Rows		36104	0	3	66	0	0	9	2	0	1	1	2	198	15	
All rows		36105	0	2	3	0	1	4	2	0	1	1	3	106	15	
Selected		36106	0	3	80	0	0	9	4	2	3	6	2	50	7	
Excluded		36107	0	3	66	1	0	2	2	0	1	1	4	8	7	
Hidden		36108	0	3	66	1	0	0	2	0	1	1	2	50	7	
Labelled		36109	0	3	66	1	0	0	2	0	1	1	2	50	7	

Figure 13: Snapshot of dataset after sampling



## 5. Data Modelling and Assessment

### Approach:

1. We ran several classification and regression models with different parameter settings for each of the identified modelling techniques and recorded the results.
2. We ensured that the models were not overfitted by comparing the model fit statistics like R-square, RMSE, AUC, misclassification rate between the training and validation datasets. Based on these criteria we filtered out few overfit models.
3. We varied the cut-off/ threshold limits between 0.3 and 0.7 for each of the models and recorded the corresponding confusion matrices. The confusion matrices were evaluated for each of the models to finalise on the best cut-off/ threshold value for the respective models.
4. We analysed the remaining models and selected the one with a high true negative rate and low false negative rate. The decision was taken considering the assumption that the cost of incorrectly predicting a user who is likely to book, as a user who will not (false negative), is more than the cost of incorrectly predicting a user who is not likely to book, as someone who will (false positive).

**Assumption Rationale:** If Expedia decides to give discounted offers to a true negative instance i.e. a customer who is less likely to make a hotel booking, the probability of converting the customer to make a booking increases. This would result in generating revenue for the company which otherwise would have been lost. Additionally, for a false negative instance i.e. giving discounts to a customer who is already more likely to make a booking would result in a loss of revenue for Expedia.

5. Based on these assumptions we finalised 5 models which are discussed in detail in Sections 5.1 - 5.5.

*<Refer attached file: **Team 2 Afternoon Final Excel Report.xlm** for detailed model-wise comparison of each of the executed models>*

6. Finally, ensemble models were built with the best selected models of each type. Similar approach was taken to select the best ensemble technique. Refer Section 5.6.
7. These 6 models were carried forward to Final Model Assessment. Refer Section 5.7.

## 5.1 Logistic Regression

We started with Logistic regression since the target variable was binary and we had sufficient continuous predictor variables. Logistic regression fits the probabilities for the response classes using a logistic function.

### Approach:

Analyze → Fit Model

- We built multiple Stepwise Regression models with varied stopping rules – Maximum RSquare validation, Min BIC, Min AIC and P-value threshold and variable selection in Forward and mixed direction.
- Variables with high p-value were not considered as input variables for the model.
- However, variable selection in mixed direction did not significantly improve the accuracy of the model.
- Based on the business logic, we selected Stepwise Regression – Forward selection – Max RSquare Validation since it resulted in minimum misclassification rate in training and validation dataset.

### Results:

The model fit statistics of the training and validation sets are compared to avoid model overfitting. In this case, the results are similar in both training and validation datasets.

Table 11: Training and Validation – Sensitivity and Specificity – Logistic Regression

Dataset	Sensitivity	Specificity	False Positive Rate	False Negative Rate	Accuracy
Training	67%	56%	44%	33%	61%
Validation	67%	56%	44%	33%	61%

RSquare (U)	0.0539		
AICc	47372.6		
BIC	47508.5		
Observations (or Sum Wgts)	36096		
Measure	Training	Validation	Test
Entropy RSquare	0.0539	0.0543	-0.747
Generalized RSquare	0.0961	0.0967	-1.424
Mean -Log p	0.6558	0.6555	0.6559
RMSE	0.4814	0.4814	0.4831
Mean Abs Dev	0.4638	0.4638	0.4633
Misclassification Rate	0.3867	0.3889	0.4307
N	36096	10830	101588

Figure 14: Logistic Regression Model Summary Stats

## Team 2: Expedia Hotel Bookings: Project Report

### Confusion Matrix

Training			Validation			Test		
Actual		Predicted	Actual		Predicted	Actual		Predicted
is_booking		0 1	is_booking		0 1	is_booking		0 1
0		10029 8019	0		3011 2404	0		49256 39698
1		5941 12107	1		1808 3607	1		4057 8577

Figure 15: Confusion Matrix for Logistic Regression

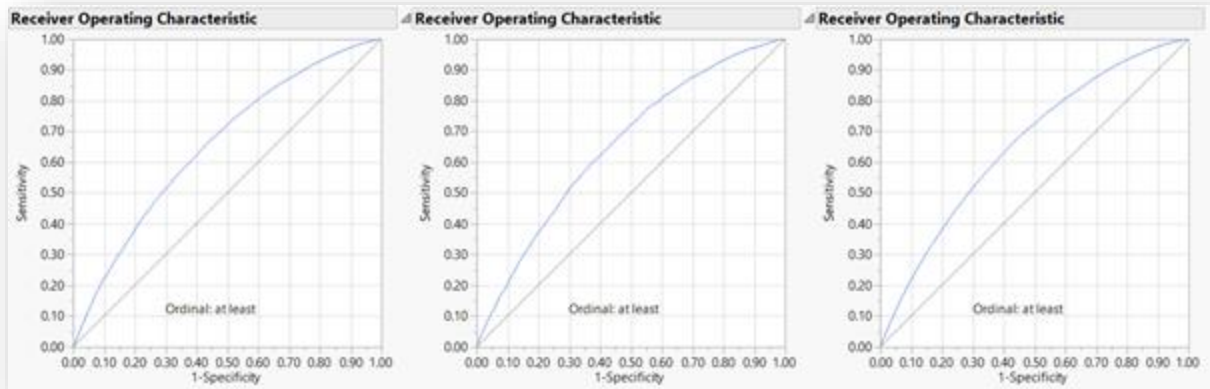


Figure 16: ROC Curve for Logistic Regression

Fit Group		
Ordinal Logistic Fit for is_booking		
Effect Summary		
Source	LogWorth	PValue
Standardize[no_chkin_days]	87.631	0.00000
Standardize[no_days_to_chkin]	55.002	0.00000
Prin1	40.880	0.00000
Standardize[srch_adults_cnt]	36.027	0.00000
Prin3	21.606	0.00000
Prin4	20.915	0.00000
Standardize[srch_rm_cnt]	18.788	0.00000
is_mobile	14.371	0.00000
srch_destination_type_id_5	13.853	0.00000
Prin6	9.972	0.00000
Standardize[srch_children_cnt]	5.028	0.00001
srch_destination_type_id_3	4.435	0.00004
factor(finalHotel\$hotel_continent)3	3.489	0.00032
srch_destination_type_id_6	2.634	0.00232
factor(finalHotel\$hotel_continent)2	2.433	0.00369
Prin7	1.945	0.01136
factor(finalHotel\$booking_month)4	1.591	0.02563
factor(finalHotel\$ci_month)8	1.227	0.05922
factor(finalHotel\$posa_continent)2	0.925	0.11895
factor(finalHotel\$ci_month)4	0.916	0.12134
factor(finalHotel\$booking_month)11	0.880	0.13190
Prin5	0.607	0.24710

Figure 17: Column Contributions for Logistic Regression

*Table 12: Training and Validation comparison for overfitting*

Dataset	Misclassification	R-Square	AUC	RMSE
Training	39%	0.0539	0.655	0.4814
Validation	39%	0.0543	0.653	0.4814

**Formula:** Please refer the Section C of Appendix

**Inference:**

Based on the Beta estimates from the prediction formula, we inferred that the top variables influencing the booking behaviour are is\_mobile, srch\_adults\_cnt, no\_chkin\_days, no\_days\_to\_chkin, srch\_rm\_cnt.

## 5.2 Discriminant Analysis

Linear Discriminant Analysis (LDA) extracts the discriminant function in which the coefficients of linear combination explain the discriminative ability of a variable. Since our target variable is categorical - binary and to understand which variables are contributing to the classification group, we went ahead with this classification technique.

**Approach:**

Analyze → Multivariate Methods → Discriminant

- We built multiple models with different combinations of continuous variables
- Based on the business logic, we selected LDA with all continuous variables since it resulted in minimum misclassification rate in training and validation dataset.

**Results:**

The model fit statistics of the training and validation sets are compared to avoid model overfitting. In this case, the results are similar in both training and validation datasets.

*Table 13: Training and Validation – Sensitivity and Specificity – Discriminant Analysis*

Dataset	Sensitivity	Specificity	False Positive Rate	False Negative Rate	Accuracy
Training	67%	55%	45%	33%	61%
Validation	67%	56%	44%	33%	61%

Discriminant Scores					
Score Summaries					
Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	36096	13995	38.7716	0.05508	47283.4
Validation	10830	4206	38.8366	0.05425	
Test	101588	43731	43.0474	-0.7566	

Training			Validation			Test		
Actual	Predicted		Actual	Predicted		Actual	Predicted	
is_booking	0	1	is_booking	0	1	is_booking	0	1
0	10000	8048	0	3015	2400	0	49306	39648
1	5947	12101	1	1806	3609	1	4083	8551

Figure 18: Overall Statistics – Discriminant Analysis

Table 14: Training and Validation comparison for Overfitting- Discriminant Analysis

Dataset	Misclassification	R-Square	AUC
Training	39%	0.055	0.656
Validation	39%	0.054	0.656

**Formula:** Please refer the Section C of Appendix

**Inference:**

Based on the Beta estimates from the prediction formula, we inferred that the top variables influencing the booking behaviour are is\_mobile, is\_package, srch\_adults\_cnt, srch\_children\_cnt, srch\_rm\_cnt, no\_chkin\_days, no\_days\_to\_chkin.

### 5.3 Neural Network

Neural network is a black box algorithm that has the ability to learn and determine the function based on the sample inputs. In order to achieve maximum accuracy, we tweaked the number of functions at the Learning Layer and Classification Layer.

**Approach:**

Analyze → Modeling → Neural

We tried different combination of n functions in the two layers and we chose function (NTanH(5) NLinear(5) NGaussian(5) NTanH2(2) NLinear2(2) NGaussian2(2)) since it resulted in minimum misclassification rate in training and validation dataset.

## Team 2: Expedia Hotel Bookings: Project Report

### Results:

The model fit statistics of the training and validation sets are compared to avoid model overfitting. In this case, the results are similar in both training and validation datasets.

Table 15: Training and Validation – Sensitivity and Specificity – Neural Network

Dataset	Sensitivity	Specificity	False Positive Rate	False Negative Rate	Accuracy
Training	66%	60%	40%	34%	63%
Validation	65%	58%	42%	35%	62%

Model NTanH(4)NLinear(4)NGaussian(4)NTanH2(8)NLinear2(8)NGaussian2(8)

Training

is\_booking

Measures	Value
Generalized RSquare	0.1219548
Entropy RSquare	0.0691939
RMSE	0.4763719
Mean Abs Dev	0.4547055
Misclassification Rate	0.373504
-LogLikelihood	23288.621
Sum Freq	36096

Confusion Matrix

Actual	Predicted	
is_booking	0	1
0	10746	7302
1	6180	11868

Validation

is\_booking

Measures	Value
Generalized RSquare	0.1010164
Entropy RSquare	0.0568321
RMSE	0.480454
Mean Abs Dev	0.458343
Misclassification Rate	0.3843952
-LogLikelihood	7080.1578
Sum Freq	10830

Confusion Matrix

Actual	Predicted	
is_booking	0	1
0	3166	2249
1	1914	3501

Test

is\_booking

Measures	Value
Generalized RSquare	-1.399614
Entropy RSquare	-0.736844
RMSE	0.4807566
Mean Abs Dev	0.4569331
Misclassification Rate	0.4055499
-LogLikelihood	66259.942
Sum Freq	101588

Confusion Matrix

Actual	Predicted	
is_booking	0	1
0	52088	36866
1	4333	8301

Figure 19: Neural Network Summary Stats



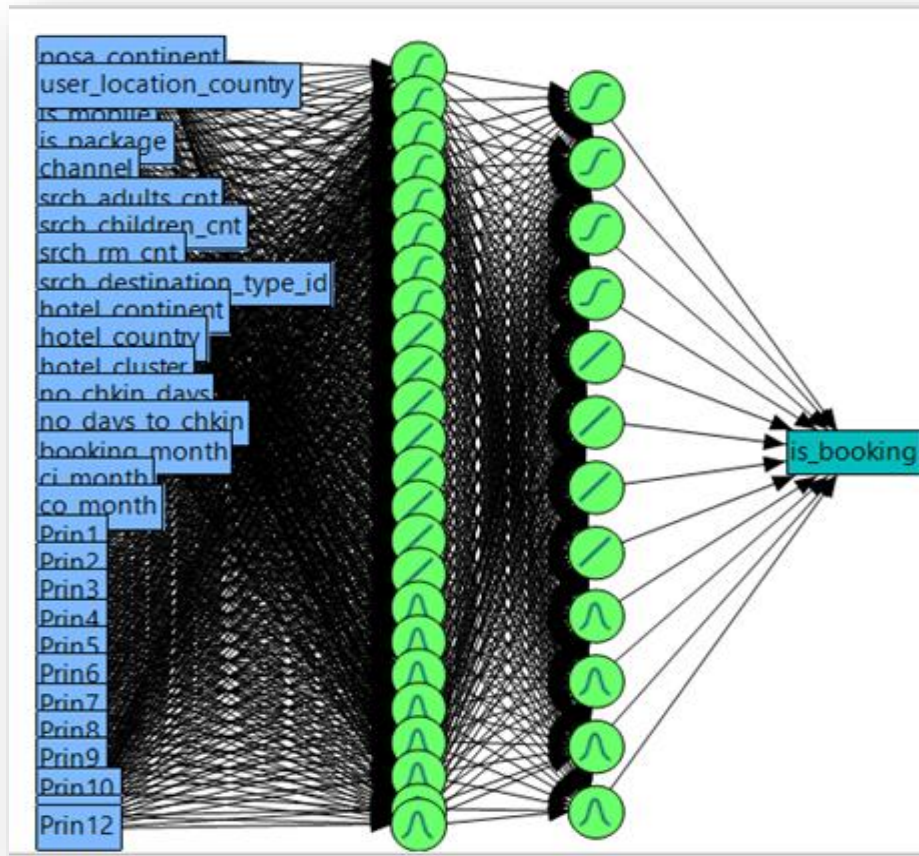


Figure 20: Neural Network Model

Table 16: Training and Validation Comparison for Overfitting – Neural Network

Dataset	Misclassification	R-Square	AUC	RMSE
Training	37%	0.069	0.677	0.476
Validation	38%	0.056	0.661	0.480

**Formula:** Please refer the Section C of Appendix

## 5.4 Decision Trees

Decision tree is the classification technique which identifies the path with highest likelihood of success based on highest information gain, calculated at every step. Since we have several nominal predictor variables we went ahead with this technique.

**Approach:**

Analyze → Modeling → Partition → Select Decision Tree under Method → Go

- We followed this approach to avoid model overfit.

## Team 2: Expedia Hotel Bookings: Project Report

- This approach resulted in a Tree with 25 splits that resulted in validation R-square value better than what the next 10 splits would obtain.

### Results:

The model fit statistics of the training and validation sets are compared to avoid model overfitting. In this case, the results are similar in both training and validation datasets.

Table 17: Training and Validation – Sensitivity and Specificity – Decision Tree

Dataset	Sensitivity	Specificity	False Positive Rate	False Negative Rate	Accuracy
Training	64%	58%	42%	36%	61%
Validation	63%	60%	40%	37%	61%

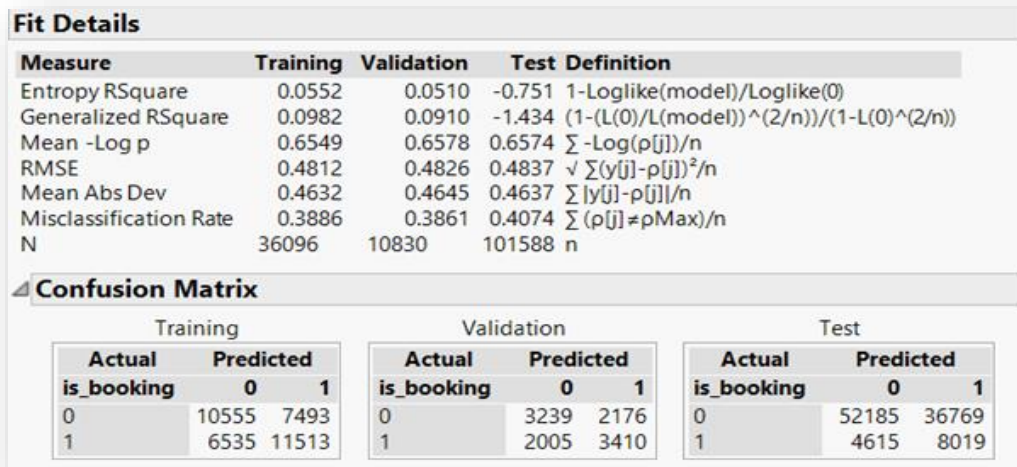


Figure 21: Overall Statistics – Decision Tree

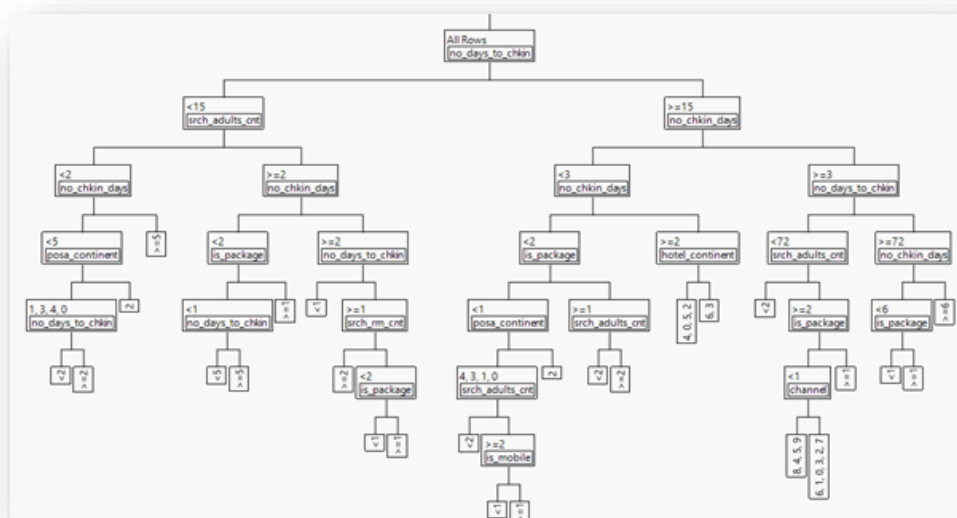


Figure 22: Small Leaf Report



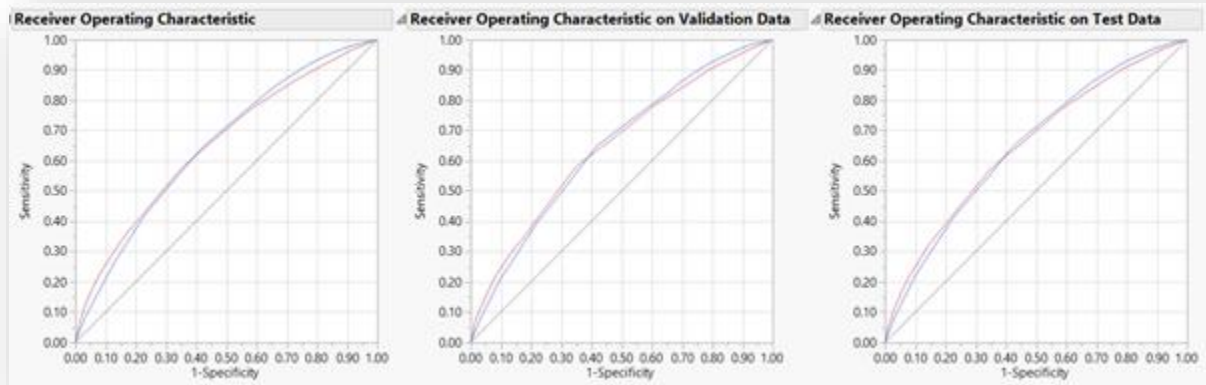


Figure 23: ROC Curve – Decision Tree

Table 18: Training and Validation Comparison for Overfitting – Decision Tree

Dataset	Misclassification	R-Square	AUC	RMSE
Training	39%	0.055	0.652	0.481
Validation	39%	0.051	0.647	0.482

**Formula:** Please refer the Section C of Appendix

**Inference:**

Based on the tree generated by the model the variables with highest information gain are no\_chkin\_days, srch\_adults\_cnt, no\_chkin\_days, posa\_continent, is\_package, hotel\_continent and search\_adult\_cnt.

## 5.5 Bootstrap Forest

This technique creates multiple trees and averages the predicted values of each tree to get the final predicted probabilities. To validate the results of decision tree we went ahead with bootstrap forest technique.

**Approach:**

Analyze → Modeling → Partition → Select Bootstrap instead of Decision Tree under Method

- We built several bootstrap forest models with number of trees set to 25, 50, 75, 100
- Based on the outcomes of the models we selected the bootstrap forest with 100 trees as it resulted in minimum misclassification rate in training and validation dataset.

## Results:

The model fit statistics of the training and validation sets are compared to avoid model overfitting. In this case, the results are similar in both training and validation datasets.

Table 19: Training and Validation – Sensitivity and Specificity - Bootstrap Forest

Dataset	Sensitivity	Specificity	False Positive Rate	False Negative Rate	Accuracy
Training	69%	61%	39%	31%	65%
Validation	66%	59%	41%	44%	62%

### Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.0805	0.0603	-0.733	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.1407	0.1069	-1.391	$(1 - (L(0)/L(\text{model}))^{(2/n)}) / (1 - L(0)^{(2/n)})$
Mean -Log p	0.6374	0.6514	0.6509	$\sum -\text{Log}(p[j]) / n$
RMSE	0.4723	0.4793	0.4796	$\sqrt{\sum (y[j] - p[j])^2 / n}$
Mean Abs Dev	0.4595	0.4663	0.4655	$\sum  y[j] - p[j]  / n$
Misclassification Rate	0.3539	0.3777	0.4086	$\sum (p[j] \neq p\text{Max}) / n$
N	36096	10830	101588	n

### Confusion Matrix

Training			Validation			Test		
Actual	Predicted		Actual	Predicted		Actual	Predicted	
is_booking	0	1	is_booking	0	1	is_booking	0	1
0	10947	7101	0	3187	2228	0	51666	37288
1	5674	12374	1	1862	3553	1	4221	8413

Figure 24: Overall Statistics – Bootstrap Forest

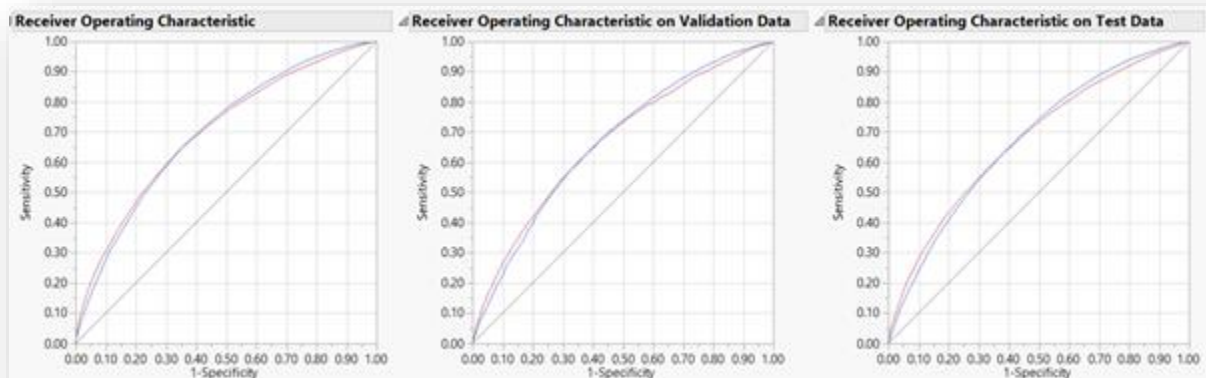


Figure 25: ROC Curve – Bootstrap Forest

*Table 20: Training and Validation Comparison for Overfitting – Bootstrap Forest*

Data set	Misclassification	R-Square	AUC	RMSE
Training	35%	0.080	0.700	0.472
Validation	38%	0.060	0.668	0.479

**Formula:** Please refer the Section C of Appendix

**Inference:**

Based on the tree generated by the model the variables with highest information gain are no\_days\_to\_chkin, srch\_adults\_cnt, no\_chkin\_days, posa\_continent

## 5.6 Ensemble Model

To improve the accuracy of the predicted probabilities we tried different ensemble modelling techniques.

**Approach:**

Analyze → Fit Model

- The predicted probabilities of best selected models were fed as input to the different modelling techniques.
- We tried ensemble model with Decision Tree, Neural Network, Logistic, Discriminant analysis and Bootstrap Forest
- Based on the outcome of different ensemble models we selected Ensemble Decision tree as it resulted in minimum misclassification in training and validation dataset.
- Similar trend was observed in the individual model assessment as Decision Tree resulted in best model among various models.

**Results:**

The model fit statistics of the training and validation sets are compared to avoid model overfitting. In this case, the results are similar in both training and validation datasets.

*Table 21: Training and Validation – Sensitivity and Specificity – Ensemble Model*

Data set	Sensitivity	Specificity	False Positive Rate	False Negative Rate	Accuracy
Training	69%	59%	41%	31%	64%
Validation	66%	58%	42%	34%	62%

Overall Statistics

Measure	Training	Validation	Test Definition	
Entropy RSquare	0.0853	0.0658	-0.725	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.1488	0.1162	-1.371	$(1 - (L(0) / L(\text{model}))^{2/n}) / (1 - L(0)^{2/n})$
Mean -Log p	0.6340	0.6476	0.6479	$\sum -\text{Log}(p[j]) / n$
RMSE	0.4709	0.4775	0.4793	$\sqrt{\sum (y[j] - p[j])^2 / n}$
Mean Abs Dev	0.4522	0.4585	0.4578	$\sum  y[j] - p[j]  / n$
Misclassification Rate	0.3565	0.3781	0.4112	$\sum (p[j] \neq p_{\text{Max}}) / n$
N	36096	10830	101588	n

Confusion Matrix

Training			Validation			Test		
Actual	Predicted		Actual	Predicted		Actual	Predicted	
is_booking	0	1	is_booking	0	1	is_booking	0	1
0	10725	7323	0	3155	2260	0	51295	37659
1	5546	12502	1	1835	3580	1	4112	8522

Figure 26: Overall Statistics – Ensemble

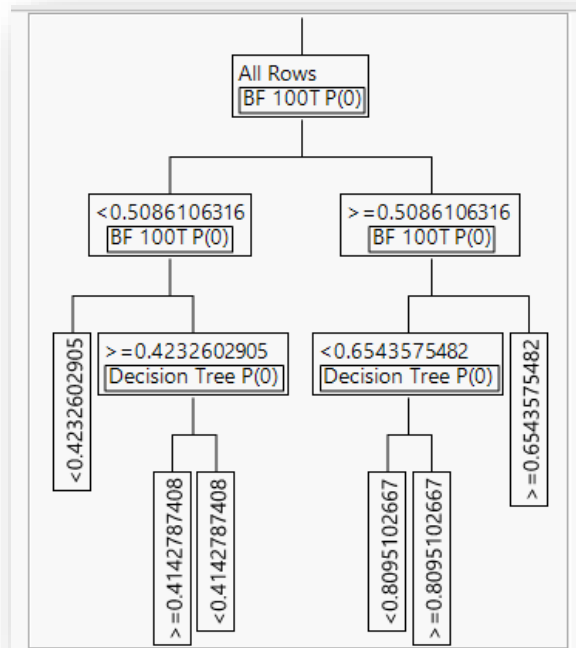


Figure 27: Small Tree Report -Ensemble Decision Tree

## Team 2: Expedia Hotel Bookings: Project Report

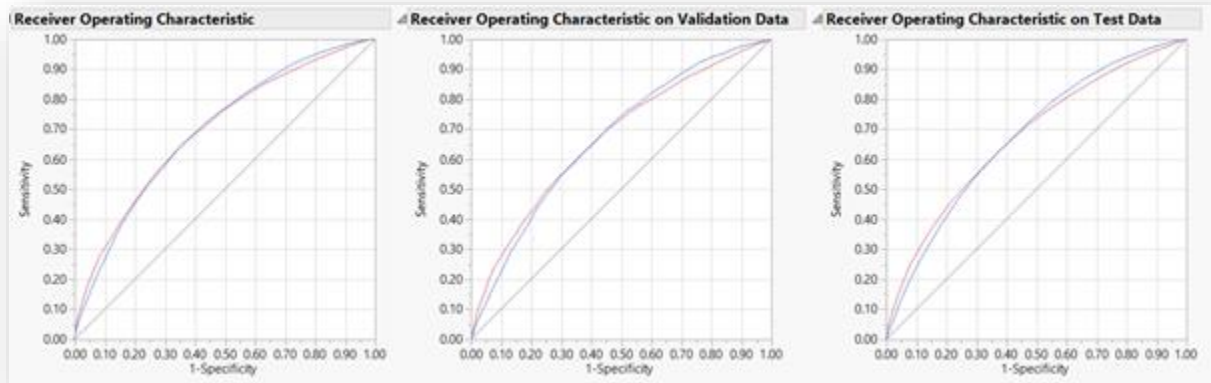


Figure 28: ROC Curve - Ensemble

Table 22: Training and Validation Comparison for Overfitting

Dataset	Misclassification	R-Square	AUC	RMSE
Training	36%	0.0853	0.696	0.470
Validation	38%	0.0658	0.645	0.477

**Formula:** Please refer the Section C of Appendix

### Inference:

Based on the prediction formula the variables with highest information gain are Boosted Forest Probabilities and Decision Tree Probabilities

## 5.7 Final Model Comparison

All the models had approximately 60% accuracy. To validate the robustness of the models that we had finalized, we checked for the seasonality trend in the test data of each of the models. We observed that the model predictions replicated the seasonality trend (booking trend/ check-in trend) followed in the entire population.

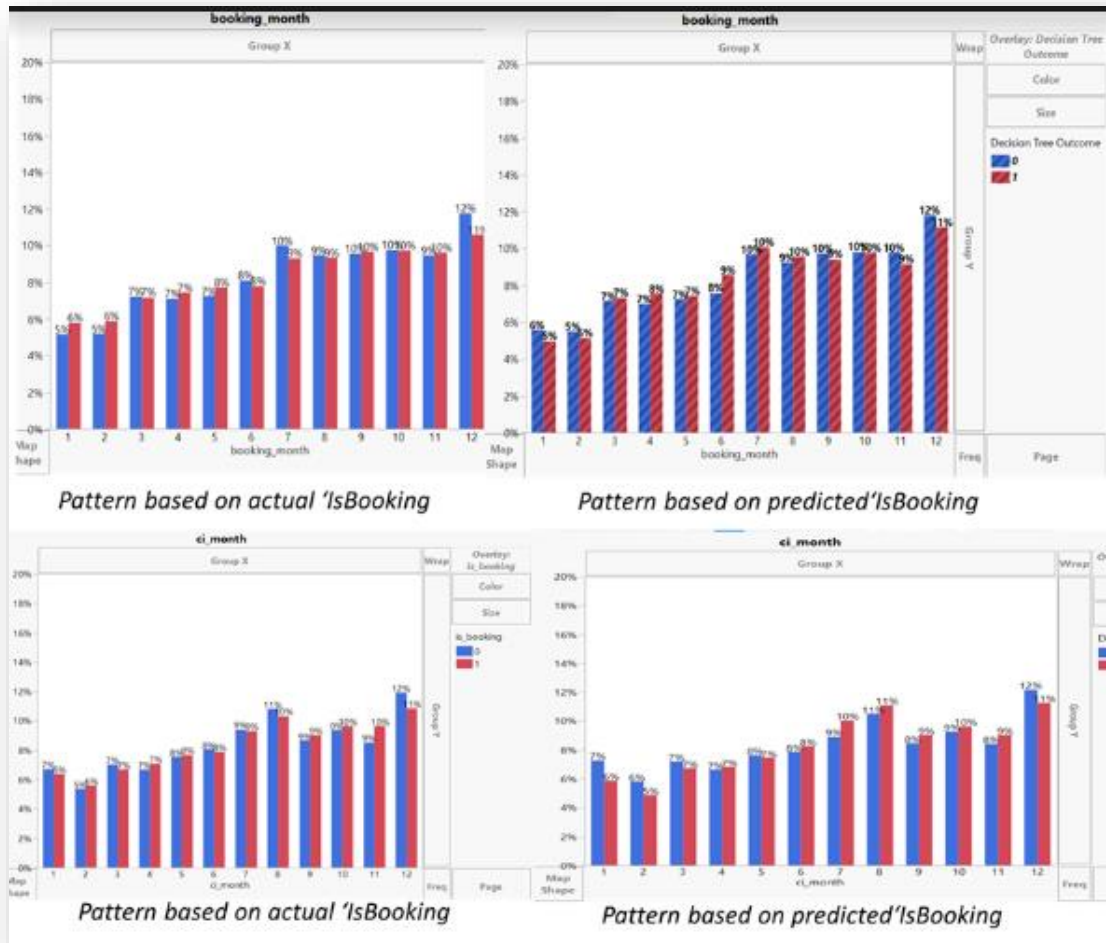


Figure 29: Seasonality Trend Actual vs Predicted

However, this is the maximum model accuracy that can be achieved with the available attributes in the dataset without overfitting the models. To further improve the model performance, we would require additional data pertaining to user web activities and demographic details.

## Team 2: Expedia Hotel Bookings: Project Report

Below is the table which summarizes the model performances for the 6 finalized models.

Table 23: Model Assessment – All model comparison

Modelling Technique	Overall Model Accuracy	Sensitivity (TPR)	Specificity (TNR)	False Negative Rate (FNR)	False Positive Rate (FPR)
Logistic Regression	57%	44%	56%	32%	44%
Linear Discriminant Analysis	57%	45%	55%	46%	45%
Neural Network	59%	42%	58%	34%	42%
Decision Tree	59%	41%	59%	37%	41%
Bootstrap Forest	59%	42%	58%	33%	42%
<b>Ensemble Model-Decision Tree</b>	<b>62%</b>	<b>37%</b>	<b>63%</b>	<b>39%</b>	<b>37%</b>

### Decision:

All the finalized 6 models generated similar results in terms of the model accuracy.

To make a definitive decision on the best possible model, we need the actual cost functions to be defined by the business for false positive and false negative predictions. However, this has not been provided by the business. Based on the assumptions made in Section 5.1 that the cost of incorrectly predicting a user who is likely to book, as a user who will not (false negative), is more than the cost of incorrectly predicting a user who is not likely to book, as someone who will (false positive).

Hence, we chose **Ensemble Decision tree** model with the **highest true negative rate** and **lowest false negative rate** since these would have the least cost impact for the business.

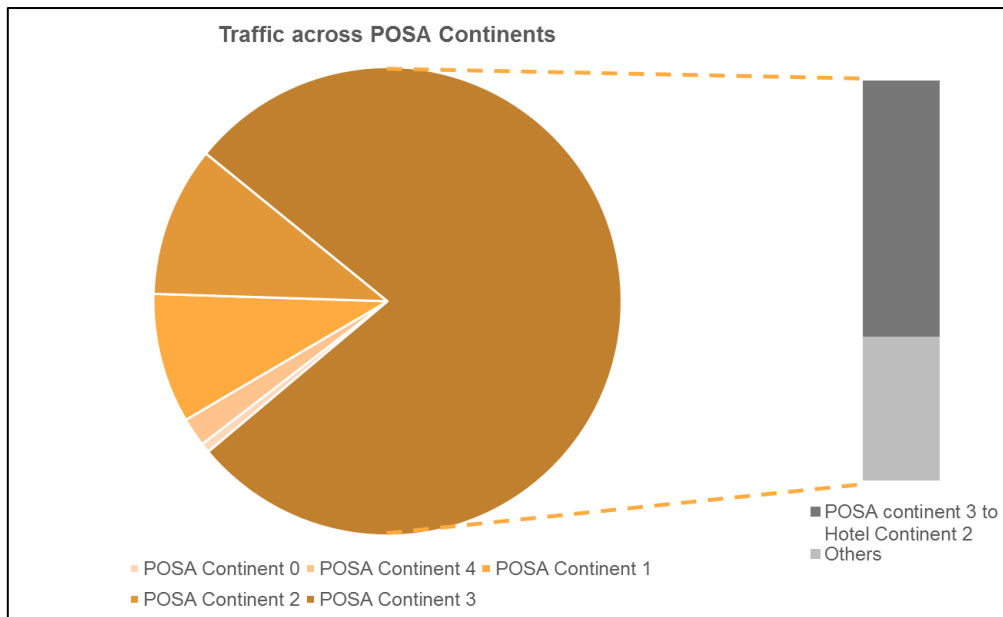


## 6. Analysis and Communication to Business

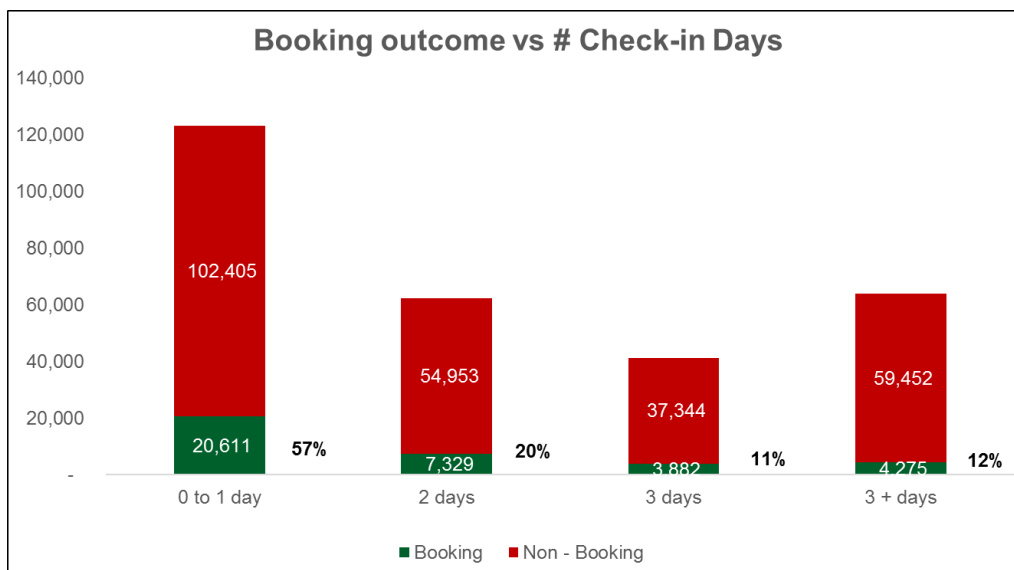
### 6.1 Key Findings

Below are some of our key findings based on the variables which significantly influenced the booking behaviour:

- About 78% of the overall traffic is generated from POSA continent 3. This is considerably higher than the traffic received from each of the other continents. 64% of the searches made from POSA continent 3 are to destination continent 2.



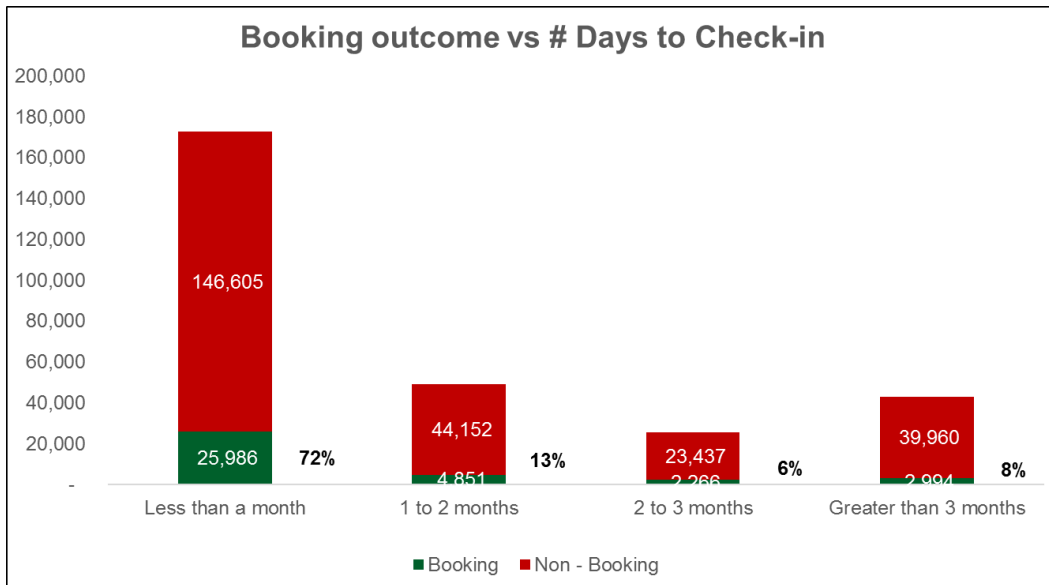
- Approximately 88% of the customers visiting Expedia portals prefer hotel and package bookings for at most 3 days.



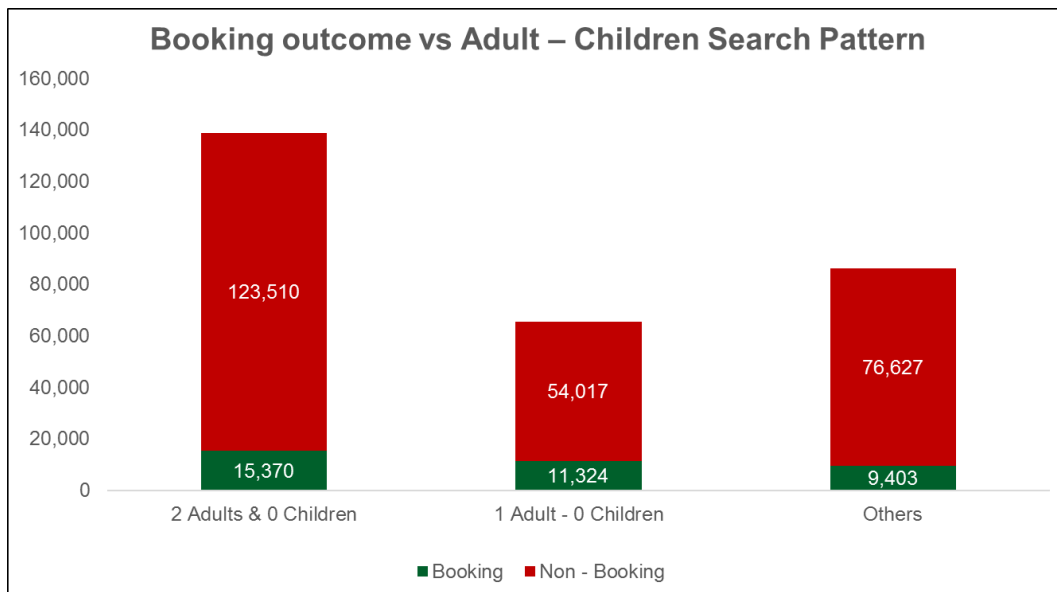


### Team 2: Expedia Hotel Bookings: Project Report

- About 85% of all the customers make a booking before 2 months of the intended date of travel.

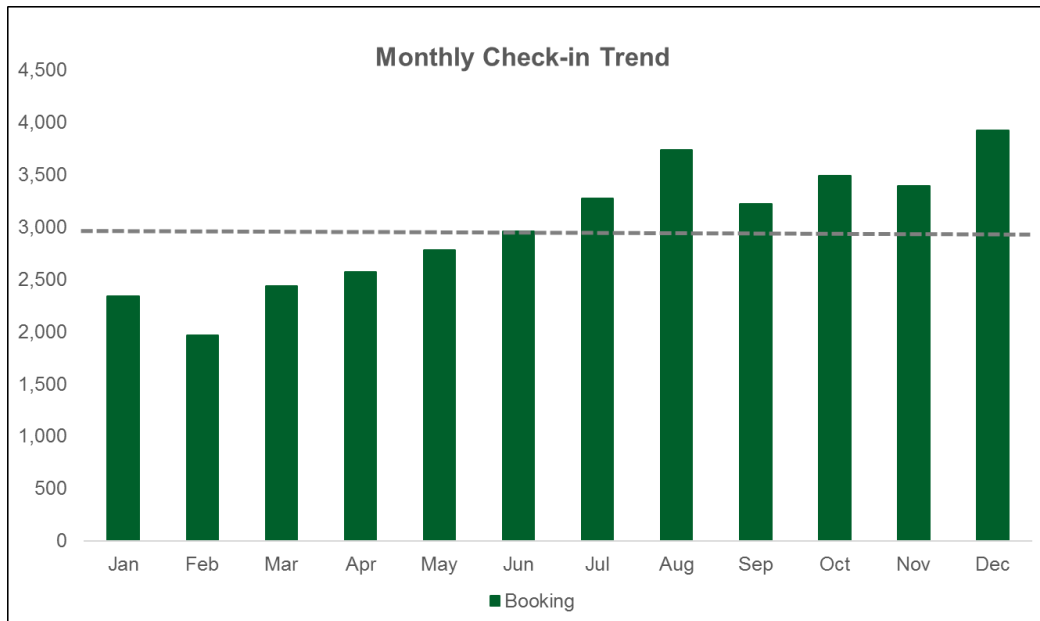


- Out of all the bookings made 74% of the customers' book duo, couple or solo trip packages without any child.



## Team 2: Expedia Hotel Bookings: Project Report

- Check-in rate in the second half of the year is more than the average yearly check-in rate. Customers travel more during second half of the year.



- First half of the year experiences significantly lower traffic volume about 41% of the annual traffic as compared to traffic volume in the second half of the year which experiences about 59% traffic volume. In spite of the low traffic in the first of the year the conversion rate is much higher as compared to the second half of the year.

Time Period	# Booking Customers	Overall Traffic	Response Rate
1 <sup>st</sup> Half Year	15,203	117,034	0.129
2 <sup>nd</sup> Half Year	20,894	173,217	0.121
Overall	36,097	290,251	0.124

Lift (Conversion  $\rightarrow$  H1) = 1.06

Lift (Conversion  $\rightarrow$  H2) = 0.96

H1 experiences low traffic volumes to the Expedia sites (41%) but has a higher conversion rate.  
H2 experiences high traffic volumes to the Expedia sites (59%) but has a lower conversion rate.

## 6.2 Insights

- With 10% increase in booking conversions in the priority sector  $x(x = POSA \text{ continent: } 3 + \text{Hotel continent: } 2 + \text{Adult Search Count: } 2 \text{ Check In Date: } 2 \text{ months of the search date} + \text{Intent of Stay: at most 3 days})$  would increase the overall conversion rate to 16% from the current 10%.
- Promotional efforts in the first 2 quarters of the year are more likely to yield better conversion as compared to the last 2 quarters (Based on Lift analysis)

### **6.3 Business Recommendations**

- To improve the overall accuracy of the model to predict the booking outcome, Expedia should consider capturing user web activity metrics such as session duration, bounces, time spent on booking page, hit rate, traffic source, number of unique searches, search depth.
- To gather demographic details of users which would help in improving the model accuracy, Expedia should provide promotional offers to make first time users register on their portal.
- For predicted 'Non-Booking' customers: Design customized time bound discount deals to entice customers to make a booking in the current user sessions.
- For predicted 'Booking' customers: Design value add packages (add on meal plans, travel plans etc.) and incentivize customers on their next booking by making them join the Expedia loyalty program.
- The marketing team can look at designing customized packages for customers searching from POSA continent 3 to Hotel continent 2 with at most 2 adults and an intended duration of stay of at most 3 days.
- Run Online/TV/Radio advertisement and promotional campaigns during the first 6 Months of the year to drive more traffic to the site. With a higher conversion rate in the first half of the year, the higher traffic is expected to yield better results in terms of conversions.
- Design hotel packages with complimentary deals that cater to 88% of the customers who prefer booking for at most 3 days.
- Design customised packages for duos, couples and solo travellers which account for 74% of the bookings made on the site.

## 7. Appendix

### A: Data Dictionary

Table 24: Data Dictionary

Variable Name	Description	Variable Type
date_time	Timestamp	Nominal
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp)	Nominal
posa_continent	ID of continent associated with site name	Nominal
user_location_country	The ID of the country the customer is located	Nominal
user_location_region	The ID of the region the customer is located	Nominal
user_location_city	The ID of the city the customer is located	Nominal
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated	Continuous
user_id	ID of user	Nominal
is_mobile	1 when a user connected from a mobile device, 0 otherwise	Continuous
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise	Continuous
channel	ID of a marketing channel	Nominal
srch_ci	Checkin date	Nominal
srch_co	Checkout date	Nominal
srch_adults_cnt	The number of adults specified in the hotel room	Continuous
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room	Continuous
srch_rm_cnt	The number of hotel rooms specified in the search	Continuous
srch_destination_id	ID of the destination where the hotel search was performed	Nominal
srch_destination_type_id	Type of destination	Nominal
hotel_continent	Hotel continent	Nominal
hotel_country	Hotel country	Nominal
hotel_market	Hotel market	Nominal
is_booking	1 if a booking, 0 if a click	Nominal
cnt	Number of similar events in the context of the same user session	Continuous
hotel_cluster	ID of a hotel cluster	Nominal
srch_destination_id	ID of the destination where the hotel search was performed	Nominal
d1-d149	latent description of search regions	Continuous
No_chkn_days	Check In Period	Continuous
No_days_to_chkin	Number of days to Check In	Continuous
Boooking_month	Month of Booking	Nominal
ci_month	Check In Month	Nominal
co_month	Check Out Month	Nominal

B: Principal Component Analysis for Destination Variables

Table 25: Principal Component Analysis for Destination Variables

Principal Components	Formula
Prin 1	$2.66771716668449 * :d6 + 0.214966456346219 * :d9 + 2.47025619369227 * :d16 +$ $-1.08155625259951 * :d20 + 2.61760554701825 * :d29 + 2.7012411188548 * :d37$ $+2.09086845868001 * :d43 + 2.73337326870686 * :d49 + 0.550657411015797 * :d58 +$ $-1.18013198212 * :d71 + -1.87456786010172 * :d79 + 0.679643835320841 * :d80$ $+0.287470009276379 * :d88 + -1.04988549439225 * :d94 + 0.335006115578626 * :d103$ $+1.7010563185047 * :d108 + -2.68238940305404 * :d110 + 2.75748280822681 * :d122$ $+0.569232899376271 * :d131 + 0.591970789585106 * :d132 + 2.38929978709572 * :d139$ $+-1.42057143452972 * :d141 + -3.59521012685701 * :d146 + 21.1248157409503$
Prin 2	$0.378244906746747 * :d6 + -1.27231940750194 * :d9 + 0.648160940680355 * :d16 +$ $-0.82099546274194 * :d20 + -1.34721304880174 * :d29 + -0.231193785240403 * :d37$ $+-1.72356895005778 * :d43 + -0.985736109838978 * :d49 + 3.35474383825191 * :d58$ $+0.0806685743812364 * :d71 + 0.452791638783805 * :d79 + 4.16609571580963 * :d80$ $+-1.03537180096618 * :d88 + 2.5148142938763 * :d94 + -1.99498589361692 * :d103$ $+1.23660375607926 * :d108 + 0.173101111465512 * :d110 + -0.0847539627220704 * :d122$ $+ -0.512933291302947 * :d131 + 4.14116921463712 * :d132 + 0.838971104269668 * :d139$ $+ -0.203366707384322 * :d141 + -0.237783713005117 * :d146 + 15.5733561757398$
Prin 3	$-0.619716175866101 * :d6 + -1.56645797020179 * :d9 + 2.71879008182926 * :d16$ $+1.39997861425495 * :d20 + 3.35940877783545 * :d29 + 0.637908292897842 * :d37$ $+1.50054248393253 * :d43 + 0.809665054970646 * :d49 + -1.22580286050988 * :d58$ $+2.29062532853005 * :d71 + 1.45215321673214 * :d79 + -0.960436281965368 * :d80 +$ $-2.28030385318666 * :d88 + 1.00568956365783 * :d94 + -2.81866270399542 * :d103 +$ $-0.403877109157455 * :d108 + 2.55916776694047 * :d110 + 3.17257913651785 * :d122$ $+0.904513340907274 * :d131 + -2.22988179612566 * :d132 + -0.21922746809424 * :d139$ $+ 1.22763653962038 * :d141 + 1.44400000872533 * :d146 + 25.5434372456096$
Prin 4	$-0.0486056057280046 * :d6 + 0.483573118536397 * :d9 + -3.67351555424604 * :d16 +$ $-1.57226224292176 * :d20 + 1.18226528273246 * :d29 + 1.32823860377407 * :d37$ $+1.78968778581157 * :d43 + 0.930433730988004 * :d49 + 0.710563805103596 * :d58$ $+2.67123216434263 * :d71 + 4.69458352062757 * :d79 + -0.549358683708943 * :d80 +$ $-1.00349105721009 * :d88 + 0.366457883424669 * :d94 + 1.15060327352962 * :d103$ $+0.940166069823842 * :d108 + 4.20156446688845 * :d110 + -2.29413325599174 * :d122$ $+-0.0934545404666437 * :d131 + 1.42833810403814 * :d132 + 1.74554211215522 * :d139$ $+ 0.883880195156398 * :d141 + -3.24471657704398 * :d146 + 24.5107561791101$
Prin 5	$1.69108297847674 * :d6 + 0.0935353987624918 * :d9 + -0.202916029678019 * :d16$ $+3.92116633182013 * :d20 + 1.78124849721454 * :d29 + 1.5698399161555 * :d37 +$ $-1.71120109957159 * :d43 + 1.18534586526915 * :d49 + 0.046594919324298 * :d58 +$ $-1.07609273908303 * :d71 + -0.485795817754179 * :d79 + -1.3520534201301 * :d80$ $+0.0948094929187919 * :d88 + 2.80130852982038 * :d94 + 0.0938689519582992 * :d103$ $+-0.0182712843612909 * :d108 + -0.139006481161141 * :d110 + -1.68584114277665 *$ $:d122 + -4.02583035593508 * :d131 + -0.333397537697454 * :d132 + 0.676976585763239$

**Team 2: Expedia Hotel Bookings: Project Report**

	* :d139 + 3.59493148251115 * :d141 + 0.630043954764514 * :d146 + 14.6774099007017
<b>Prin 6</b>	-1.03652643656179 * :d6 + 0.13835800959663 * :d9 + -2.55897591845324 * :d16 +1.98957155229787 * :d20 + -0.157277628781774 * :d29 + -0.465703470639753 * :d37 +2.61121058419797 * :d43 + 0.032582637054352 * :d49 + 0.548459735066051 * :d58 + -0.736376808945294 * :d71 + -2.21204208589362 * :d79 + 0.318483750705362 * :d80 +-4.17554511057141 * :d88 + 2.06323721376916 * :d94 + 0.441055059723846 * :d103 +-0.793354758515776 * :d108 + -3.58893057176148 * :d110 + -1.41551482923387 * :d122 + 2.61728667862419 * :d131 + 1.12314648091634 * :d132 + 0.812170414871177 * :d139 + 4.01970344914173 * :d141 + 0.0318198857932859 * :d146 + (-0.681916655190113)
<b>Prin 7</b>	-2.00971044063121 * :d6 + -0.746884699038451 * :d9 + -0.105010781527536 * :d16 +1.39840163245355 * :d20 + -0.352017324142788 * :d29 + 0.609559838671974 * :d37 +-0.770069550361092 * :d43 + -0.514423512081691 * :d49 + -1.22943614394089 * :d58 +-1.95006408550846 * :d71 + 0.848729449175782 * :d79 + -0.521476828535345 * :d80 +4.59374877277998 * :d88 + -2.02979674153863 * :d94 + -0.898334265194339 * :d103 +3.11644983988548 * :d108 + -0.0798000978695419 * :d110 + -2.16905551150197 * :d122 + 1.88197923660778 * :d131 + 0.312736649348771 * :d132 + 1.78614775291116 * :d139 + 3.37285786241101 * :d141 + 0.981767012879923 * :d146 + 12.2215879006194
<b>Prin 8</b>	0.509874790849138 * :d6 + -0.185708184763254 * :d9 + -1.57415860322003 * :d16 + -3.41795903515651 * :d20 + 4.39207478439889 * :d29 + 0.838292850049183 * :d37 +0.874970810464906 * :d43 + -1.00650150559697 * :d49 + 0.507004146122381 * :d58 +0.6236606190743 * :d71 + -1.56020471678437 * :d79 + -0.951871550503482 * :d80 +1.53457332683701 * :d88 + 1.01396925092947 * :d94 + -0.454569681451722 * :d103 +-0.702699116448766 * :d108 + -0.955261821894424 * :d110 + -2.73663700299041 * :d122 + -0.0697105626740348 * :d131 + 0.559633272167619 * :d132 + 0.986250482056138 * :d139 + -1.1270417884327 * :d141 + 10.4420560365576 * :d146 + 16.2824959425975
<b>Prin 9</b>	-1.50963690162894 * :d6 + 0.420250121756305 * :d9 + 0.444020920039752 * :d16 +1.88609648187668 * :d20 + 3.44335231312906 * :d29 + -0.925321065597396 * :d37 +1.89375334062762 * :d43 + -3.58792261248392 * :d49 + 2.36241156941412 * :d58 + -0.189430444553032 * :d71 + 0.166023917811924 * :d79 + 1.67972844184993 * :d80 +3.73029737179153 * :d88 + 1.48982829222267 * :d94 + 1.14817038959832 * :d103 + -1.40641007278096 * :d108 + 0.902981245555826 * :d110 + 3.32247236594989 * :d122 +-0.323644999155702 * :d131 + -1.77807927697726 * :d132 + 1.65029856756425 * :d139 + 0.263350088483139 * :d141 + -2.36429346618098 * :d146 + 27.6871002184881
<b>Prin 10</b>	0.569580620536338 * :d6 + 0.127557737782724 * :d9 + -0.554130173165171 * :d16 +1.2239675140801 * :d20 + -0.757761627985552 * :d29 + 0.0375655226850382 * :d37 +0.388536391795451 * :d43 + 0.415365952714432 * :d49 + 0.42222026258095 * :d58 +5.72214688847052 * :d71 + -1.74829361860232 * :d79 + 1.16655395183246 * :d80 +1.82792237947153 * :d88 + -4.38500050329671 * :d94 + -0.136335120666662 * :d103 +-0.694132778663665 * :d108 + -1.49698780459827 * :d110 + 1.92141260775824 * :d122 + -0.90479755879558 * :d131 + 2.61395715206254 * :d132 + -0.471974056557116 * :d139 + 3.22311673400001 * :d141 + 1.42876001842507 * :d146 + 21.2992635420808
<b>Prin11</b>	-0.63896973731389 * :d6 + 0.104945760337038 * :d9 + -0.315606815453884 * :d16 +5.25231039087903 * :d20 + 2.13838371411101 * :d29 + 0.107304238663979 * :d37 +

## Team 2: Expedia Hotel Bookings: Project Report

	$  \begin{aligned}  &-0.0991948426685751 * :d43 + -0.0611231704777121 * :d49 + 1.17617913623673 * :d58 \\  &+ -0.842461782967928 * :d71 + 1.36979506025208 * :d79 + 1.31868845296857 * :d80 + \\  &-2.91295301741109 * :d88 + -4.11952790459984 * :d94 + 0.52173607858494 * :d103 \\  &+ 0.627309503814185 * :d108 + 0.335929167282916 * :d110 + -1.59289972488544 * :d122 \\  &+ -0.0639440205809273 * :d131 + 0.381055607456775 * :d132 + 0.616451794376715 * \\  &:d139 + -4.23058928877483 * :d141 + 4.25616375669926 * :d146 + 7.00159544607107  \end{aligned}  $
Prin12	$  \begin{aligned}  &2.0417755456024 * :d6 + -0.496043262369735 * :d9 + -2.37029676264447 * :d16 \\  &+ 2.1552201653989 * :d20 + -0.939327923247586 * :d29 + 0.487850840244194 * :d37 \\  &+ 0.229044768212575 * :d43 + 2.43557238435636 * :d49 + 0.616657196871921 * :d58 + \\  &-1.51303058707235 * :d71 + 1.64672032899149 * :d79 + -0.614839451843789 * :d80 \\  &+ 3.65405859083527 * :d88 + 0.291120223365562 * :d94 + -1.5330749108529 * :d103 + \\  &-3.88350104097917 * :d108 + 0.191210528732758 * :d110 + -0.481390226290544 * :d122 \\  &+ 1.74666929767008 * :d131 + 1.98468046475629 * :d132 + -0.29452336920506 * :d139 \\  &+ -1.65284514971769 * :d141 + -1.25451535685244 * :d146 + 5.33070413683414  \end{aligned}  $

### C: Formulas for Data Models

Table 26 : Model Prediction Formulas

Model	Formula
Logistic Regression	$  \begin{aligned}  &(-0.25637755050589) + 0.263010841164538 * :is\_mobile + 0.465132377765032 * \\  &:is\_package + 0.192953104757185 * :srch\_adults\_cnt + 0.0812278703382229 * \\  &:srch\_children\_cnt + -0.257440685739362 * :srch\_rm\_cnt + 0.106967152591187 \\  &* \\  &:no\_chkin\_days + 0.00351525889594283 * :no\_days\_to\_chkin + 0 * \\  &: \\  &\quad \text{Name("Standardize[srch\_adults\_cnt]")} \quad + \quad 0 \quad *: \\  &\quad \text{Name("Standardize[srch\_children\_cnt]")} \\  &\quad ) \quad + \quad 0 \quad *: \quad \text{Name("Standardize[srch\_rm\_cnt]")} \quad + \quad 0 \quad *: \\  &\quad \text{Name("Standardize[no\_chkin\_days]")} \quad + \quad 0 \quad *: \\  &\quad \text{Name("Standardize[no\_days\_to\_chkin]")} \\  &+ 0.0313357237397279 * : \text{Name("factor(finalHotel\$booking\_month)2")} \\  &+ 0.031056268286681 * : \text{Name("factor(finalHotel\$booking\_month)3")} + \\  &- 0.0996161589983117 * : \text{Name("factor(finalHotel\$booking\_month)4")} \\  &+ 0.0409866224447976 * : \text{Name("factor(finalHotel\$booking\_month)5")} + \\  &- 0.0371083935697817 * : \text{Name("factor(finalHotel\$booking\_month)6")} \\  &+ 0.0173782315746718 * : \text{Name("factor(finalHotel\$booking\_month)7")} + \\  &- 0.0317624036099242 * : \text{Name("factor(finalHotel\$booking\_month)8")} \\  &+ 0.0072222963124231 * : \text{Name("factor(finalHotel\$booking\_month)9")} + \\  &- 0.00449097911714238 * : \text{Name("factor(finalHotel\$booking\_month)10")} + \\  &- 0.0441465593287692 * : \text{Name("factor(finalHotel\$booking\_month)11")} + \\  &- 0.0284980998849066 * : \text{Name("factor(finalHotel\$booking\_month)12")} \\  &+ 0.0137300507046119 * : \text{Name("factor(finalHotel\$ci\_month)2")} + - \\  &0.00712501595660631 \\  &* : \text{Name("factor(finalHotel\$ci\_month)3")} + 0.0606299902757661 * \\  &: \text{Name("factor(finalHotel\$ci\_month)4")} + 0.00901560155785674 * \\  &: \text{Name("factor(finalHotel\$ci\_month)5")} + -0.0120295908014195 * \\  &: \text{Name("factor(finalHotel\$ci\_month)6")} + 0.00724628263617867 * \\  &: \text{Name("factor(finalHotel\$ci\_month)7")} + 0.0743326591119852 * \\  &: \text{Name("factor(finalHotel\$ci\_month)8")} + 0.0385962587199403 * \\  &: \text{Name("factor(finalHotel\$ci\_month)9")} + -0.0341242749129865 * \\  &: \text{Name("factor(finalHotel\$ci\_month)10")} + -0.0409333723062703 * \\  &: \text{Name("factor(finalHotel\$ci\_month)11")} + -0.00216782018592299 *  \end{aligned}  $

## Team 2: Expedia Hotel Bookings: Project Report

	: Name("factor(finalHotel\$ci_month)12") + -0.0581477174119829 * : Name("factor(finalHotel\$posa_continent)1") + -0.0243493246896268 * : Name("factor(finalHotel\$posa_continent)2") + -0.0943228474523115 * : Name("factor(finalHotel\$posa_continent)3") + -0.163213026944452 * : Name("factor(finalHotel\$posa_continent)4") + -0.211662033644979 * : Name("factor(finalHotel\$posa_continent)2") + 0.0271021148557416 * : Name("factor(finalHotel\$hotel_continent)3") + -0.187479735792066 * : Name("factor(finalHotel\$hotel_continent)4") + -0.427567622818525 * : Name("factor(finalHotel\$hotel_continent)5") + -0.165774689552419 * : Name("factor(finalHotel\$hotel_continent)6") + 0.00735112203560069 * : Name("factor(finalHotel\$channel)1") + 0.0301092735631709 * : Name("factor(finalHotel\$channel)2") + 0.0684624687553128 * : Name("factor(finalHotel\$channel)3") + 0.0949973498678308 * : Name("factor(finalHotel\$channel)4") + -0.00995662484316062 * : Name("factor(finalHotel\$channel)5") + -0.102343946753656 * : Name("factor(finalHotel\$channel)6") + 0.00804278868874867 * : Name("factor(finalHotel\$channel)7") + 0.169438173849098 * : Name("factor(finalHotel\$channel)8") + 0.0115405070472176 * : Name("factor(finalHotel\$channel)9") + -0.269314341714603 * : Name("factor(finalHotel\$channel)10") + -0.169096364096571 * : srch_destination_type_id_3 + -0.019014512271118 * : srch_destination_type_id_4 + -0.260900594688338 * :srch_destination_type_id_5 + -0.0449927467971164 * : srch_destination_type_id_6 + 0.448781562185659 * : srch_destination_type_id_7 +0.317061008212028 * :srch_destination_type_id_8 + 14.1381444745515 * : srch_destination_type_id_9 + 0.0663939551501143 * :Prin1 + 0.00113495913929855 * : Prin2 + -0.0687222185103806 * :Prin3 + 0.0855993634004558 * :Prin4 + -0.0141550309267993 * :Prin5 + 0.069957445871334 * :Prin6 + 0.0311437446720656 * : Prin7 + 0.00370712925885308 * :Prin8 + -0.00302299903541665 * :Prin9 +0.0294536545063129 * :Prin10 + 0.0302123150424024 * :Prin11 + - 0.0227590942245762 * :Prin12 + 0.0947494767047758 * :hotel_cluster 6 + 0.049604691324046 * :hotel_cluster 7 + -0.029824505986909 * :hotel_cluster 13
Discriminant Analysis	: Name("LDA SqDist[0]") + -2.05646655308167 * :is_mobile + -2.77607599854534 * :is_package + -2.71709642154418 * :srch_adults_cnt + -0.612608258181547 * : srch_children_cnt + -5.73443385960748 * :srch_rm_cnt + -1.45179861064154 * : no_chkin_days + -0.0207979647187918 * :no_days_to_chkin + - 28.2209912240147 * : Name("factor(finalHotel\$booking_month)2") + -29.9323461126951 * : Name("factor(finalHotel\$booking_month)3") + -31.8620440272104 * : Name("factor(finalHotel\$booking_month)4") + -34.0096622905557 * : Name("factor(finalHotel\$booking_month)5") + -34.7126215955659 * : Name("factor(finalHotel\$booking_month)6") + -37.0330897379068 * : Name("factor(finalHotel\$booking_month)7") + -38.0372666635245 * : Name("factor(finalHotel\$booking_month)8") + -38.3679511042638 * : Name("factor(finalHotel\$booking_month)9") + -40.2222306887646 * : Name("factor(finalHotel\$booking_month)10") + -39.2482533049539 * : Name("factor(finalHotel\$booking_month)11") + -42.1301027045161 * : Name("factor(finalHotel\$booking_month)12") + -33.2224284183281 * : Name("factor(finalHotel\$ci_month)2") + -32.2350794630811 * : Name("factor(finalHotel\$ci_month)3") + -28.3150242150662 * : Name("factor(finalHotel\$ci_month)4") + -26.073692815843 * : Name("factor(finalHotel\$ci_month)5") + -24.3750198759026 *



## Team 2: Expedia Hotel Bookings: Project Report

	:Name( "factor(finalHotel\$ci_month)6" ) + -23.9573419377922 * :Name( "factor(finalHotel\$ci_month)7" ) + -22.8324132049581 * :Name( "factor(finalHotel\$ci_month)8" ) + -22.2993241955416 * :Name( "factor(finalHotel\$ci_month)9" ) + -21.0226673635745 * :Name( "factor(finalHotel\$ci_month)10" ) + -21.8336268607058 * :Name( "factor(finalHotel\$ci_month)11" ) + -21.8302978533129 * :Name( "factor(finalHotel\$ci_month)12" ) + -326.214064980081 * :Name( "factor(finalHotel\$posa_continent)1" ) + -330.261244284578 * :Name( "factor(finalHotel\$posa_continent)2" ) + -321.007078599417 * :Name( "factor(finalHotel\$posa_continent)3" ) + -379.795209761236 * :Name( "factor(finalHotel\$posa_continent)4" ) + -231.008021114415 * :Name( "factor(finalHotel\$hotel_continent)2" ) + -221.139492121338 * :Name( "factor(finalHotel\$hotel_continent)3" ) + -250.461859699118 * :Name( "factor(finalHotel\$hotel_continent)4" ) + -222.996165552708 * :Name( "factor(finalHotel\$hotel_continent)5" ) + -225.632871222538 * :Name( "factor(finalHotel\$hotel_continent)6" ) + -15.8732636192711 * :Name( "factor(finalHotel\$channel)1" ) + -16.3019416074251 * :Name( "factor(finalHotel\$channel)2" ) + -17.5208482304109 * :Name( "factor(finalHotel\$channel)3" ) + -13.4364238655995 * :Name( "factor(finalHotel\$channel)4" ) + -25.9527215309305 * :Name( "factor(finalHotel\$channel)5" ) + -11.1599310852742 * :Name( "factor(finalHotel\$channel)6" ) + -14.5714207489691 * :Name( "factor(finalHotel\$channel)7" ) + -14.5918022630688 * :Name( "factor(finalHotel\$channel)8" ) + -15.4462470908594 * :Name( "factor(finalHotel\$channel)9" ) + 11.3138600790606 * :Name( "factor(finalHotel\$channel)10" ) + -1.53134474413045 * :srch_destination_type_id_3 + -2.40115780969363 * :srch_destination_type_id_4 + -3.37374399112673 * :srch_destination_type_id_5 + -3.76851727033875 * :srch_destination_type_id_6 + 19.5786664299834 * :srch_destination_type_id_7 +3.89213014945846 * :srch_destination_type_id_8 + -25.8110547577469 * :srch_destination_type_id_9 + 0.584865990885476 * :Prin1 + - 0.585221146142366 * :Prin2 + 0.252008155242915 * :Prin3 + 0.853472119707215 * :Prin4 + -0.176267726058248 * :Prin5 + 0.902745649183465 * :Prin6 + - 0.545055202243872 * :Prin7 + -0.732365891834555 * :Prin8 + -0.307040738201724 * :Prin9 + -0.743377848685219 * :Prin10 + -0.206559717300091 * :Prin11 + - 0.816160023729845 * :Prin12 + -9.26699524124486 * :hotel_cluster 6 + -6.61210713517 * :hotel_cluster 7 + -6.44055740430759 * :hotel_cluster 13 + 322.453746333077
Neural Network	1 / (1 + Exp( 0.410331498622102 + 0.184404732133816 * :H1_1 2 + - 0.134875202723846 * :H1_2 2 + 0.223581745777969 * :H1_3 2 + - 0.288369448091826 * :H1_4 2 + -0.181590480259179 * :H1_5 2 + 0.177488868737016 * :H1_6 2 + - 0.172137714255289 * :H1_7 2 + -0.207650507481413 * :H1_8 2 + 0.122680694715008 * :H1_9 2 + -0.0523167883911894 * :H1_10 2 + -0.0203712281026968 * :H1_11 2 +0.147035226962768 * :H1_12 2 + 0.157965297454916 * :H1_13 + - 0.139571282708237 * :H1_14 + -0.100122668056767 * :H1_15 ))
Decision Tree	If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin < 15, If( Is Missing( :srch_adults_cnt )   :srch_adults_cnt < 2,

## Team 2: Expedia Hotel Bookings: Project Report

	<pre> If( Is Missing( :no_chkin_days )   :no_chkin_days &lt; 5,     If(         :posa_continent == 1   :posa_continent == 3       :posa_continent == 4           :posa_continent == 0,         If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin &lt; 2,             0.278327840916746,             0.347806104024996         ),         :posa_continent == 2, 0.507859020318271,         0.334459027503625     ),     0.514189420717438 ), If( Is Missing( :no_chkin_days )   :no_chkin_days &lt; 2,     If( Is Missing( :is_package )   :is_package &lt; 1,         If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin &lt; 5,             0.391232268785734,             0.466034856322824         ),         0.812251012679396     ),     If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin &lt; 1,         0.358207131288613,         If( !Is Missing( :srch_rm_cnt ) &amp; :srch_rm_cnt &gt;= 2,             0.414278740839995,             If( Is Missing( :is_package )   :is_package &lt; 1,                 0.543962258995532,                 0.687748078975227             )         )     ) ), If( Is Missing( :no_chkin_days )   :no_chkin_days &lt; 3,     If( Is Missing( :no_chkin_days )   :no_chkin_days &lt; 2,         If( Is Missing( :is_package )   :is_package &lt; 1,             If(                 :posa_continent == 4   :posa_continent == 3   :posa_continent == 1   :posa_continent == 0,                 If( Is Missing( :srch_adults_cnt )   :srch_adults_cnt &lt; 2,                     0.353834824340348,                     If( Is Missing( :is_mobile )   :is_mobile &lt; 1,                         0.457630952113878,                         0.570668297724096                     )                 )             )         )     ) ) </pre>
--	--

## Team 2: Expedia Hotel Bookings: Project Report

	), :posa_continent == 2, 0.593429652911947,  0.460818121217461 , If( Is Missing( :srch_adults_cnt )   :srch_adults_cnt < 2,  0.480016592691959, 0.80951026672815 ) , Choose( Match( :hotel_continent, 4, 1, 0, 1, 5, 1, 2, 1, 6, 2, 3, 2, 3 ),  0.541229102933117, 0.625204220844176, 0.567559521786637 ) , If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin < 72, If( Is Missing( :srch_adults_cnt )   :srch_adults_cnt < 2, 0.549025632173191, If( Is Missing( :is_package )   :is_package < 1, Choose( Match( :channel, 8, 1, 4, 1, 5, 1, 9, 1, 6, 2, 1, 2, 0, 2, 3, 2, 2, 2, 7, 2, 3 ) , 0.604711780228994, 0.713522398231008, 0.64591967832177 ) , 0.752412037581512 ) , If( Is Missing( :no_chkin_days )   :no_chkin_days < 6, If( Is Missing( :is_package )   :is_package < 1, 0.67182213195741, 0.794614089153945 ) , 0.826394636308416 ) ) ) )
Bootstrap Forest	If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin < 15, If( Is Missing( :srch_adults_cnt )   :srch_adults_cnt < 2, If( Is Missing( :no_chkin_days )   :no_chkin_days < 5,

## Team 2: Expedia Hotel Bookings: Project Report

	<pre> If(     :posa_continent == 1   :posa_continent == 3   :posa_continent == 4       :posa_continent == 0,     If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin &lt; 2,         0.278327840916746,         0.347806104024996     ),     :posa_continent == 2, 0.507859020318271,     0.334459027503625 ),     0.514189420717438 ), If( Is Missing( :no_chkin_days )   :no_chkin_days &lt; 2,     If( Is Missing( :is_package )   :is_package &lt; 1,         If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin &lt; 5,             0.391232268785734,             0.466034856322824         ),         0.812251012679396     ),     If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin &lt; 1,         0.358207131288613,         If( !Is Missing( :srch_rm_cnt ) &amp; :srch_rm_cnt &gt;= 2,             0.414278740839995,             If( Is Missing( :is_package )   :is_package &lt; 1,                 0.543962258995532,                 0.687748078975227             )         )     ) ), If( Is Missing( :no_chkin_days )   :no_chkin_days &lt; 3,     If( Is Missing( :no_chkin_days )   :no_chkin_days &lt; 2,         If( Is Missing( :is_package )   :is_package &lt; 1,             If(                 :posa_continent == 4   :posa_continent == 3   :posa_continent == 1   :posa_continent == 0,                 If( Is Missing( :srch_adults_cnt )   :srch_adults_cnt &lt; 2,                     0.353834824340348,                     If( Is Missing( :is_mobile )   :is_mobile &lt; 1,                         0.457630952113878,                         0.570668297724096                     )                 )             )         )     ) ) </pre>
--	--

## Team 2: Expedia Hotel Bookings: Project Report

	<pre> :posa_continent == 2, 0.593429652911947, 0.460818121217461 ), If( Is Missing( :srch_adults_cnt )   :srch_adults_cnt &lt; 2, 0.480016592691959, 0.80951026672815 ) ), Choose( Match( :hotel_continent, 4, 1, 0, 1, 5, 1, 2, 1, 6, 2, 3, 2, 3 ), 0.541229102933117, 0.625204220844176, 0.567559521786637 ) ), If( Is Missing( :no_days_to_chkin )   :no_days_to_chkin &lt; 72, If( Is Missing( :srch_adults_cnt )   :srch_adults_cnt &lt; 2, 0.549025632173191, If( Is Missing( :is_package )   :is_package &lt; 1, Choose( Match( :channel, 8, 1, 4, 1, 5, 1, 9, 1, 6, 2, 1, 2, 0, 2, 3, 2, 2, 2, 7, 2, 3 ), 0.604711780228994, 0.713522398231008, 0.64591967832177 ), 0.752412037581512 ) ), If( Is Missing( :no_chkin_days )   :no_chkin_days &lt; 6, If( Is Missing( :is_package )   :is_package &lt; 1, 0.67182213195741, 0.794614089153945 ), 0.826394636308416 ) ) ) ) ) </pre>
Ensemble Model	<pre> If ( Is Missing ( : Name ("BF 100T P (0)") )   : Name ("BF 100T P (0)") &lt; 0.508610631641504, If ( </pre>

**Team 2: Expedia Hotel Bookings: Project Report**

	<pre> Is Missing (: Name ("BF 100T P (0)"))  : Name ("BF 100T P (0)") &lt; 0.423260290493545, 0.287240215229038, If (     ! Is Missing (: Name ("DT P (0)")) &amp;: Name ("DT P (0)")     &gt;=     0.414278740839995,     0.429395813471597,     0.501404018826244     ) ), If (     Is Missing (: Name ("BF 100T P (0)"))  : Name ("BF 100T P (0)") &lt;     0.654357548191557,     If (Is Missing (: Name ("DT P (0)"))  : Name ("DT P (0)") &lt;     0.80951026672815,         0.617592148070065,         0.744664460038701     ),     0.812025927116643     ) ) </pre>
--	---