**Additional Questions**

**Question 1, Part A:** What makes the classification problem difficult in this task?

The main problem faced during solving the problem was:

**Solution 1: Part A:**

Our dependent variable i.e. **has\_booking** was unevenly distributed as shown in Figure 1 and Figure 2

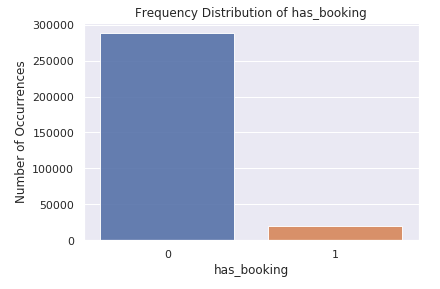


Figure : Frequency of “has\_booking” (Yes/No)

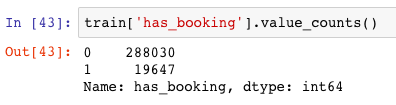


Figure : Absolute frequency of “has\_booking”

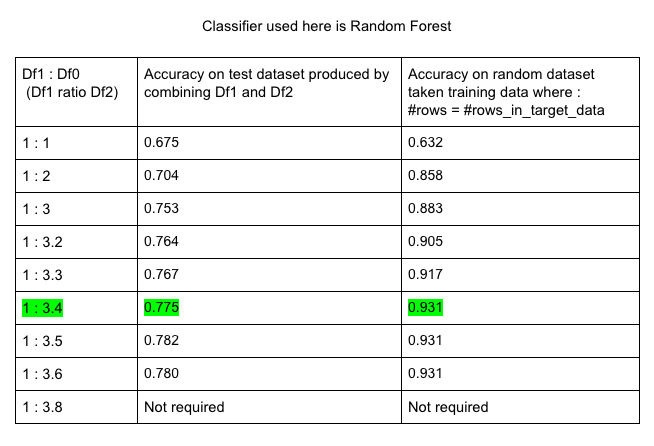
If we observe, in order to build an efficient classifier, we need properly to build a proper training dataset out of the given training set.

**Question 1, Part B**: How do you handle that?

**Solution 1, Part B:**

In order to tackle the above problem, I followed the following steps:

1. Extracted the data with “has\_booking” =1 from given “case\_study\_bookings\_train.csv” and named the dataframe as *df1*.
2. Extracted the data with “has\_booking” =0 from given “case\_study\_bookings\_train.csv” and named the dataframe as *df0*.
   1. Decided the ratio in which df0 and df1 should be to prepare the final dataset. Using hit and trials I analyzed the following table:



* 1. In order to avoid over-fitting. I finalized to resume with ***df1 : df0 = 1 : 3.4.***

1. Combined both the dataframe i.e. df0 and df1 and named it as “*data\_equal*”.
2. Randomized the rows in the “*data\_equal*” and constructed the final dataset to train my classifier

**Question 2:** Evaluate and compare at least 3 classification algorithm for this task.

**Solution 2:**

Chosen classification algorithms:

1. Random Forest
2. Decision Tree
3. Naïve Bayes
4. Logistic Regression

Predefined parameters:

1. Test\_size = 25%
2. Train\_size = 75%
3. As decided above Df1: Df0 = 1 : 3.4

Table for classification algorithm comparison according to the dataset provided is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameters** | **Random Forest** | **Decision Tree** | **Naive Bayes** | **Logistic Regression** |
| Accuracy on test\_set  ACC = (TP+TN)/(TP+FP+FN+TN) | 0.774 | 0.772 | 0.596 | 0.773 |
| Predicted accuracy on target set  ACC = (TP+TN)/(TP+FP+FN+TN) | 0.930 | 0.920 | 0.568 | 0.938 |
| Sensitivity  TPR = TP/(TP+FN) | 0.986 | 0.972 | 0.553 | 1.0 |
| Specificity  TNR = TN/(TN+FP) | 0.058 | 0.087 | 0.744 | 0.0 |
| Precision or Positive predictive value  PPV = TP/(TP+FP) | 0.779 | 0.784 | 0.881 | 0.773 |
| Negative predictive value  NPV = TN/(TN+FN) | 0.565 | 0.476 | 0.326 | NAN |
| False positive rate  FPR = FP/(FP+TN) | 0.941 | 0.912 | 0.255 | 1 |
| False negative rate  FNR = FN/(TP+FN) | 0.013 | 0.027 | 0.446 | 0 |
| False discovery rate  FDR = FP/(TP+FP) | 0.220 | 0.215 | 0.118 | 0.226 |

On the basis of above data, I can say, there is close competition between Random Forest and Decision Tree but Random forest is little better.

**Naïve Bayes** has very low accuracy (59.6%) and other parameters reveals it is not performing good at classification task.

**Logistic Regression** is performing outstanding in case of accuracy on both the datasets. But, if we look at the confusion matrix below:

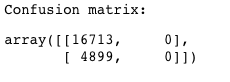


Figure : Confusion matrix for Logistic Regression

It reveals that logistic regression, is not able to predict any True Negative (TN) nor any False Negative(FN) value. Sensitivity is 1. Specificity is 0. Hence, we cannot categorize it as a good classifier for this dataset.

Hence, **Random Forest** is best classification algorithm according to me for this dataset.

**Question 3:** Propose at least three 3 features that are significant to predict booking

**Solution 3:**

As both the dependent and independent variables are categorical so in order to select the significant features we will go for **Chi-Square Test**

Result of Chi-Square test are as follows:

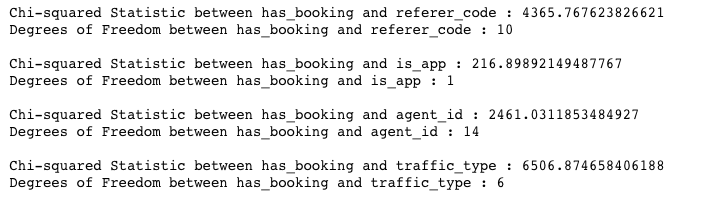


Figure : Chi-Square Statistics and Degree-of-Freedom

Now, we will look at Chi-Square Distribution table to check cut off for a p-value of 0.05.

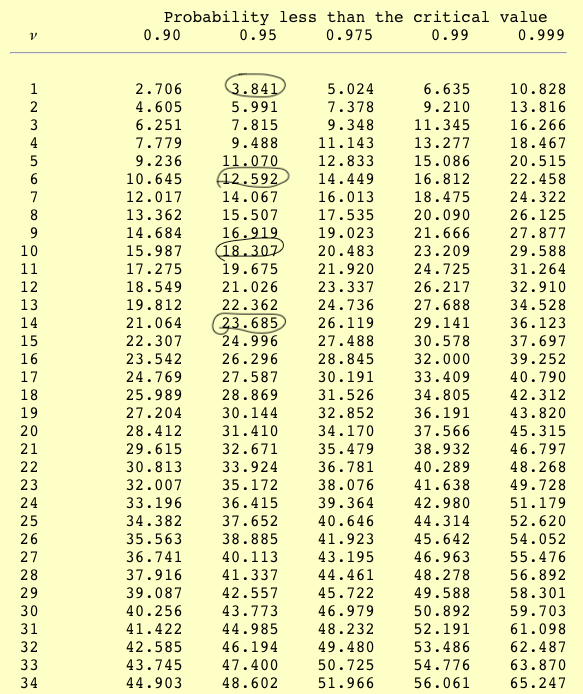


Figure : Chi-Square Distribution table with circled values according to degree of freedoms at 5% Significance

Hypothesis in case of Chi square test:

**Null hypothesis:** Assumes that there is no association between the two variables.

**Alternative hypothesis:**Assumes that there is an association between the two variables.

As we can see our Chi-Square statistics value is more than p-value at particular degree of freedom. So we can say **all the evidence are against null hypothesis.**

**Hence, all the** variables i.e. referer\_code, is\_app, agent\_id, traffic\_type **are significant to predict** has\_booking**.**

By looking at Chi-Square statistics, here is the strength of association is decreasing order:

1. traffic\_type (Most Significant)
2. referer\_code
3. agent\_id
4. is\_app

**Question 4:** We can spot a very significant action type. What might this action refer to?

**Solution 4:**

The most significant action type must be the one which yield maximum amount of booking.

Steps taken:

#### Merge the training set of booking and action based on same ‘ymd','user\_id' and 'session\_id' and prepare a new dataframe.

#### Generate the contingency table from fields: ‘has\_booking’ and ‘action\_id’.

#### Find the ‘action\_id’ against which there are maximum count of ‘has\_booking’=1

#### Result: 2142

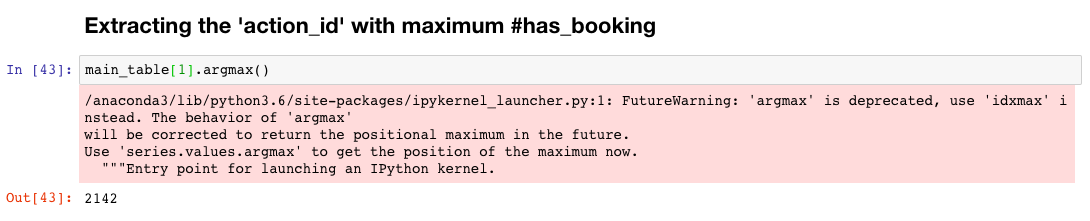


Figure : Snapshot of solution from Significant\_Action.ipynb

According to me action\_id: 2142, should belong to **“HOTELS”.**