Problem statement:
Business Challenge
A leading airline industry player with offices across the globe was losing its customers consistently. This impelled the client to leverage customer analytics to minimize the gaps in the quality of service and improve service offerings to ensure long-term customer relationship.

Data science based approach:

- Data Cleasing
- Modeling
- Model Evaluation
- Model Interpretation

The above steps are very important and will help to gain insights about the historic data regarding the patterns and behavior of churners and use the inferences to predict the future occurences.

A)Data cleansing:
Preparing the data:
According to a recent study, preparing data and sampling requires 80% of the whole time and effort.
In the process of preparing the data, we have considered a random sample of 20,000 records out of 129880 records. Random sample is the sample where the probability of all records of being chosen is the same. Thus, the dataset consisted of 20,000 records of 24 fields. These are some of the fields we considered to start with and we'd prefer these to work with. In predictive+diagnostic analytics, there is one more step involved in data preparation- creating the target variable. Incase of churn analysis, it could be a binary column such as “churn” .You can fill in the values for this variable by analyzing the historical data. E.g., a value of 0/FALSE for customers who are loyal and satisfied and 1/TRUE otherwise (satisfied and disloyal, neutral and dissatisfied and disloyal, neutral and dissatisfied and loyal). This is done using tools like excel or libreoffice calc etc.
Formula:=IF(AND(B2="satisfied",D2="Loyal Customer"),0,1)
Next, we need to check for data correctness:
1.Are there any empty cells?
df=pd.read_csv('/home/pranitha/internship/airlinesdata.csv')
High level checking:

```
df.isnull().any().any()
Out[40]: True

In [42]: df.isnull().any()
Out[42]:
id                                False
satisfaction_v2                   False
Gender                            False
Customer Type                     False
.
.
.
Arrival Delay in Minutes           True
Churn                             False
dtype: bool


In [43]: df.isnull().sum()
Out[43]:
id                                0
satisfaction_v2                   0
Gender                            0
Customer Type                     0
.
.
.
Arrival Delay in Minutes          42
Churn                             0
dtype: int64
```

It is now found out that 'Arrival Delay in Minutes' is the only column with empty cells, with 42 empty cells. Now we need to handle those values

We choose to drop them and place in it a different dataframe as follows:

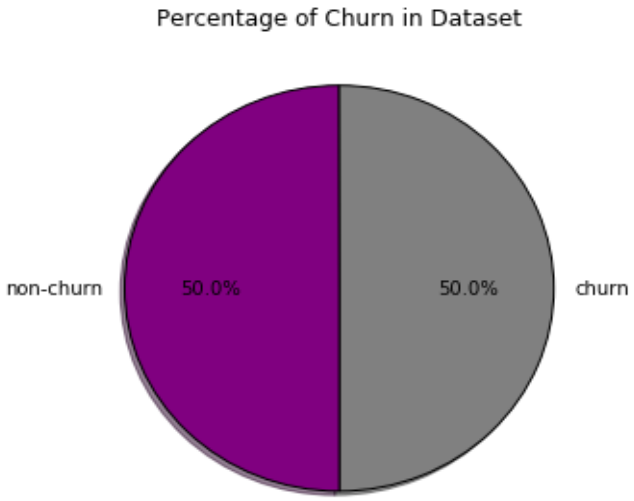
```
In [63]: data=df.dropna(how='any')

In [64]: data.shape
Out[64]: (19957, 25)
```

In [65]: data.isnull().any().any()
Out[65]: False

2.Are there any cells with ambiguous entries?Ex.California, CA, Cal,etc
Ex.df['Gender'].value_counts()
In [54]: df['Gender'].value_counts()
Out[54]:
Female 10176
Male 9823
Name: Gender, dtype: int64
check for all categorical data columns
NO AMBIGUOUS ENTRIES ARE DETECTED.
3.Are the data entries accurate.For example: Check for negative values in Flight distance etc.
NO INVALID DATA ENTRIES ARE DETECTED.

Thus, the dataset now consists of 19958 rows (inlcuding column header) and 25 columns (including new 'churn' column)

Sample characteristics:
Second step would be sample characterstics or summary, this is done to know more about the dataset.After you have loaded the dataset, you might want to know a little bit more about it. You can check attributes names and datatypes using info(). Now, we split the data based on churn and compare their means. From the above we can tell that the customers left if Flight Distance, Gate location, Departure Delay in Minutes, Arrival Delay in Minutes were more and for other attributes left when they were less. Now we can use describe() function to get an idea of the statistics.(Describe for one attrubute for example).Now we need to plot actual target variable using either pie chart or bar graph.We have exactly 50% of both churners and non churner, thus we can have more efficient insights of the data.



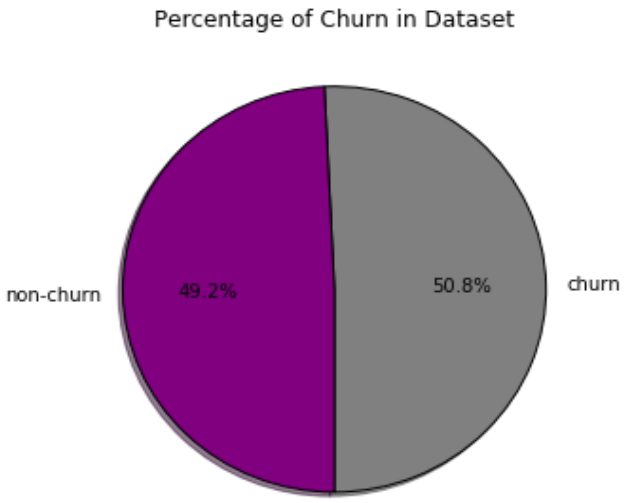
There may be data included that is not needed to improve our results. Best is that to identify by logic thinking. In this data set we have the customerID for example. As it does not influence our predicted outcome, we drop the column with the pandas “drop()” function.We need to store the data without 'id' field into a new dataframe.Next we need to convert the categorical data into numerical data using get_dummies() function.First our model needs to be trained, second our model needs to be tested. Therefore it is best to have two different dataset. As for now we only have one, it is very common to split the data accordingly. X is the data with the independent variables, Y is the data with the dependent variable. The test size variable determines in which ratio the data will be split. It is quite common to do this in a 80 Training / 20 Test ratio. Thus, training data had 15965 and test data had 3992 rows making a total of 19957 data rows.

B)Modeling

In this phase we need to train the algorithms with the train data and use the model to predict future trends. In our analysis we used three different algorithms. They are logistic regression, random forest and decision tree in that order and finally compared the results and get cumulative conclusions. This way, we can ensure that client recieves a most trusty algorithm.
The steps in the process of construction of any of the three models in general is:
1.importing required packages
2.create and fitting a model object with trianing data
3.prediction using the test data

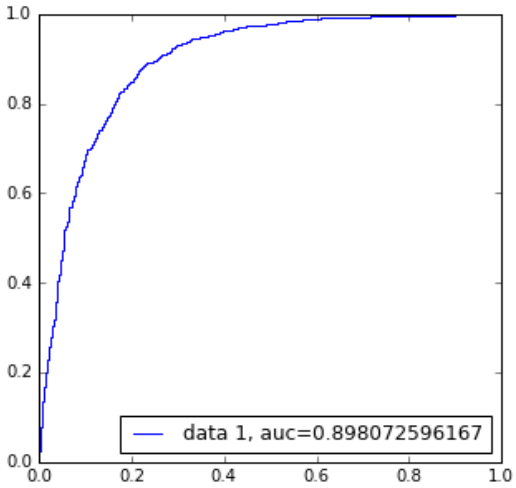
3.1pie chart
3.2confusion matrix
4.calculating accuracy score
5.identifying most influential features according to their importance(sorting according to weights in logistic regression, bargraph in random forest and from visualization of tree in decision tree).
6.Roc curve and calculation of area under curve.
The above steps may differ for some models but the basic objective of each step is the same.

C)Model evaluation:
Model evaluation is the process of choosing between models, different model types, tuning parameters, and features. Better evaluation processes lead to better, more accurate models in your applications.
Model:Logistic regression
Pie chart:



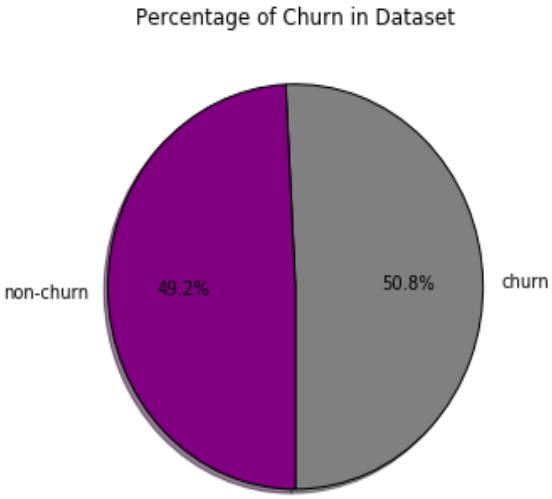
Confusion matrix:
[[1630 364]
 [336 1662]]

Roc curve:



Model:Random forest

Pie chart:

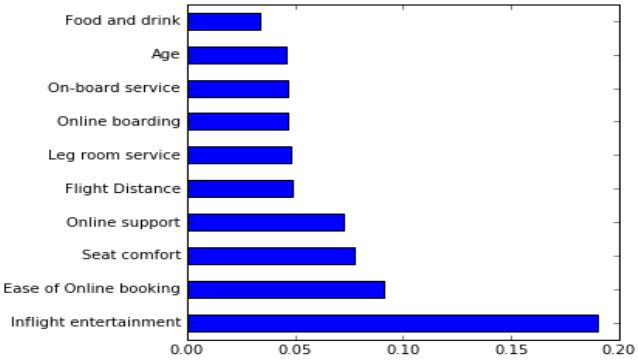


Confusion matrix:

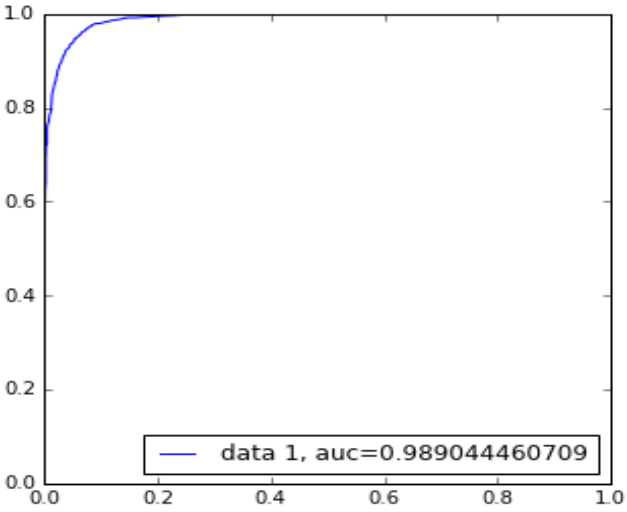
[[1870 124]

 [93 1905]]

Bar graph:

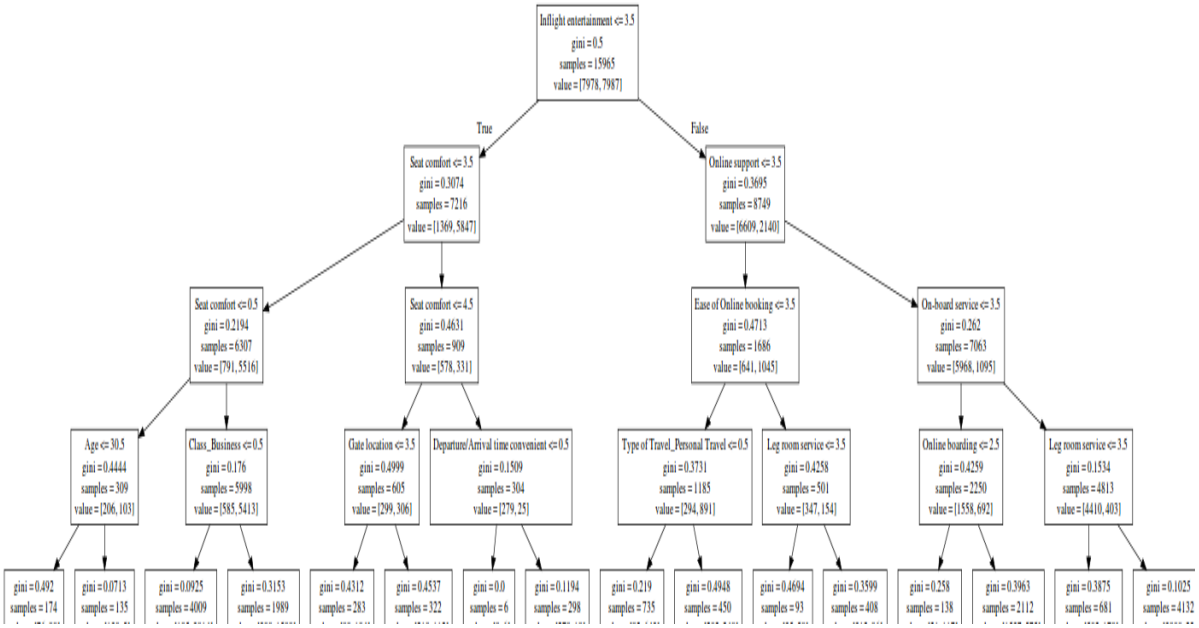


Roc curve:

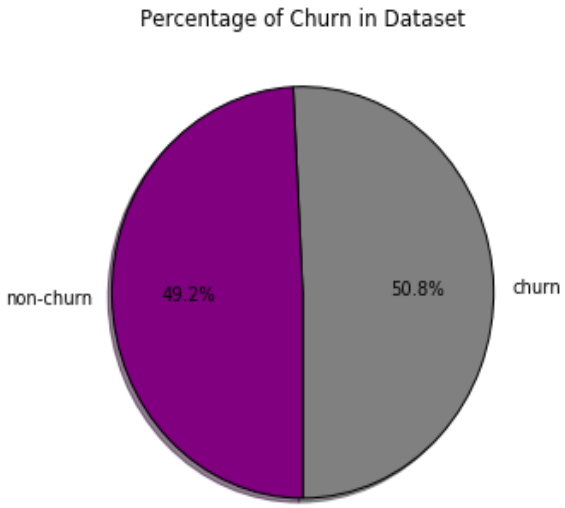


Model:Decision Tree

Visualization of decision tree

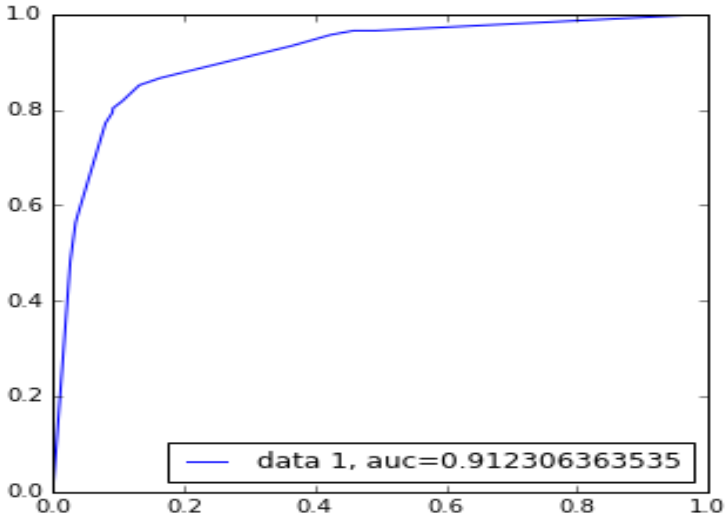


Pie chart:



Confusion matrix:
[[1731 263]
 [297 1701]]

Roc curve:



Model comaprision:
Now that we have model fit statistics, we compare them on the basis of the following statistics:
1.Accuracy
2.Precision
3.Confusion matrix
4.Auc score

5.Features(in the first 3 places of importance)

Model	Accuracy score	Precision	Confusion matrix	auc score	Features
Logistic regression	0.824649298597	0.82	700/3992	0.898071843157099	Gender_Male,Type of Travel_Business travel,Class_Eco
Random forest	0.945641282565	0.95	217/3992	0.98904446070945 6	Inflight entertainment,Ease of Online booking,Seat comfort
Decision tree	0.859719438878	0.86	556/3992	0.91230636353505	Inflight entertainment,Seat comfort,Online support

D)Model interpretation and conclusions:

From the above comparision, we can tell that Random forest algorithm is most accurate which is also evident in confusion matrix where only 217 out of 3992 were wrongly predicted, whereas this number is higher in other two models.

However, we need to consider all the possible inferences all the models give us. They are as follows:

1.Concentration on inflight entertainment is very important as it is given highest importance in both random forest and decision tree model. It is suggested to maintain the rating among customers more than 3.5 in case of inflight entertainment (from decision tree).

2.Next in importance is the online services. Online services must be made more accessible and efficiently working. Any humane procedures should be converted into online as much as possible. Again its important customer is satisfied and gives a rating of 3.5 or above.

3.Next we need to concentrate on seat comfort.

4.Now, coming to customer side features. It is found that male, business travellers(opposed to personal travellers) of economic class are more vulnerable to churn. Thus, this should be kept in mind and proper advertising and offers must be provided for these segment of customers.

Thus, for future prediction we employ Random forest algorithm.