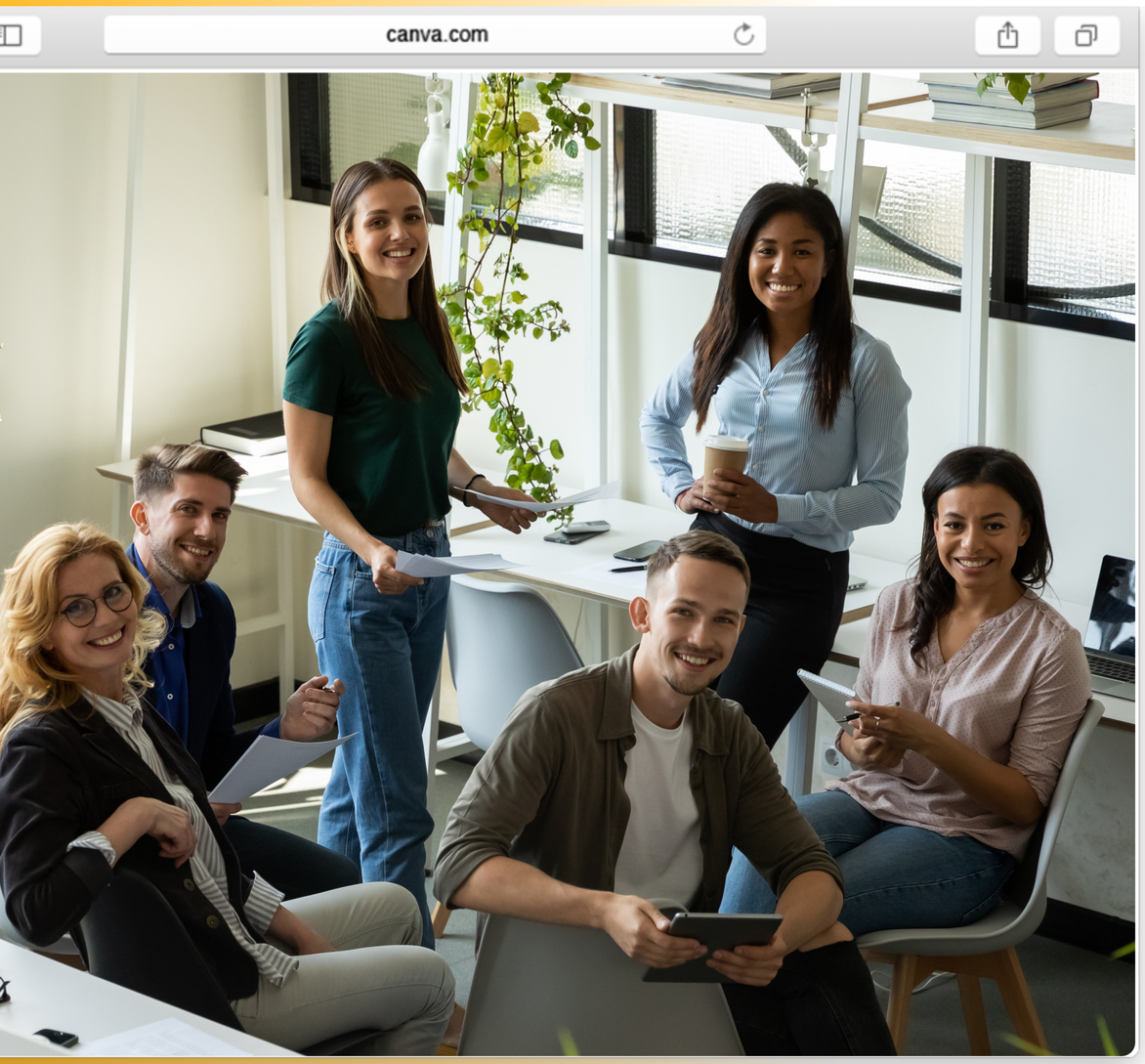
Great Learning & UT Austin

CAPSTONE PRESENTATION

Social Media_Tourism_Project

Submitted By- Gunjar Fuley
Batch- PGPDSBA Online Nov_A 2020
Email- gforgunjaar@gmail.com



The biggest startups allocate

75%

TO SOCIAL MEDIA MARKETING

Current Scenario

- Go-Go AIR is a multinational aviation organization
- Brand Go-Go AIR is degrading significantly due to the aggressive cold calling campaigns
- The company decided to reach out to the masses using social media marketing campaigns
- Go-Go Air decided to collaborate with a social media platform for Ad campaigns
- The analytics team was instructed to come up with a model which will predict whether customer will buy the ticket or not.

Final Problem Statement

Improving the effectiveness of Social Media Campaign for higher revenue through an increase in sales of tickets

Image representation of Social Media Websites

Understanding the data

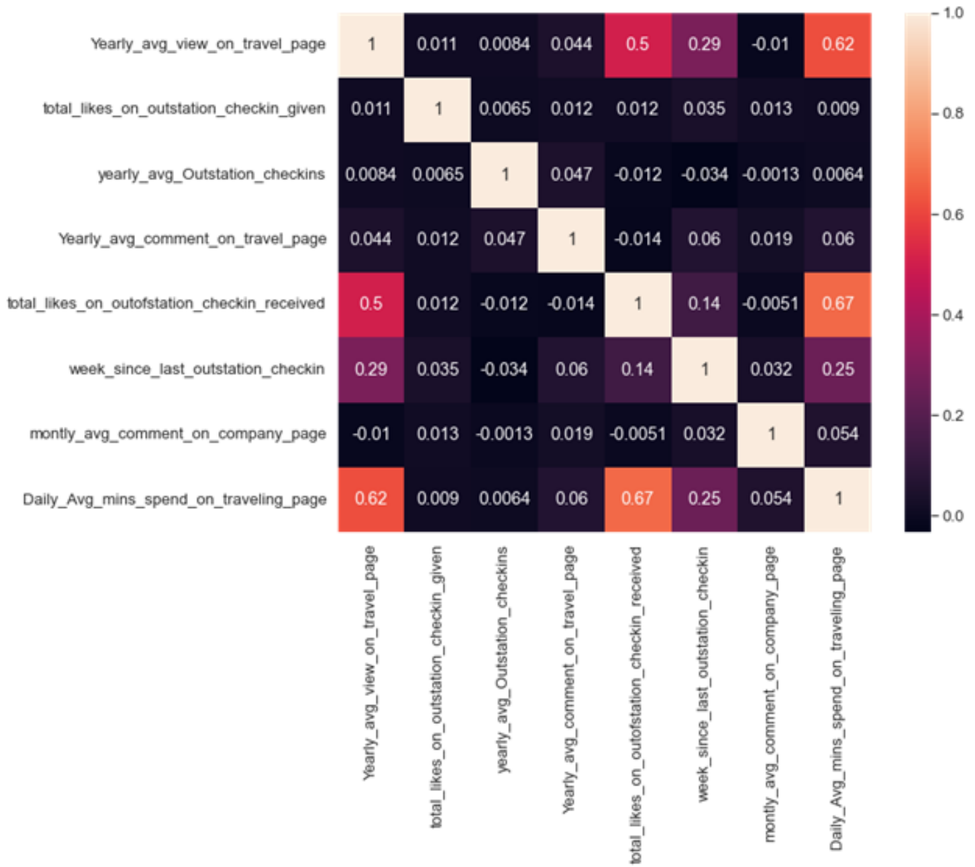
```
RangeIndex: 11760 entries, 0 to 11759
Data columns (total 17 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   UserID                                  11760 non-null  int64
 1   Taken_product                           11760 non-null  object
 2   Yearly_avg_view_on_travel_page          11179 non-null  float64
 3   preferred_device                        11707 non-null  object
 4   total_likes_on_outstation_checkin_given 11379 non-null  float64
 5   yearly_avg_Outstation_checkins          11685 non-null  object
 6   member_in_family                        11760 non-null  object
 7   preferred_location_type                 11729 non-null  object
 8   Yearly_avg_comment_on_travel_page       11554 non-null  float64
 9   total_likes_on_outofstation_checkin_received 11760 non-null  int64
 10  week_since_last_outstation_checkin      11760 non-null  int64
 11  following_company_page                  11657 non-null  object
 12  montly_avg_comment_on_company_page      11760 non-null  int64
 13  working_flag                            11760 non-null  object
 14  travelling_network_rating               11760 non-null  int64
 15  Adult_flag                              11760 non-null  int64
 16  Daily_Avg_mins_spend_on_traveling_page  11760 non-null  int64
dtypes: float64(3), int64(7), object(7)
```

- The number of data points or the rows was 11760 and the number of features or variables were 17
- 9 variables were numeric and 7 variables were categorical in nature
- Null values were present in the dataset
- The presence of outliers was also identified
- The percentage of Users buying tickets is 16.12. The percentage of Users not buying tickets is 83.88

Observations from the data

- The number of prospects is less on Laptop and more on mobiles or tablets
- Most families have a number of family members as 3
- The prospects are highly interested in visiting a beach.
- It is followed by financial destinations and historical sites respectively
- Most of the customers are not following the company page
- Working customers are more likely to take the product

Correlation among the features



- High correlation of 0.67 between “Daily average minutes spend on travelling page” and “total likes on outstation checkin received”
- ·The moderate correlation of 0.62 between “Daily average minutes spend on travelling page” and “yearly average view on travel page”
- The low correlation of 0.5 between “yearly average view on travel page” and “total likes on outstation checkin received”

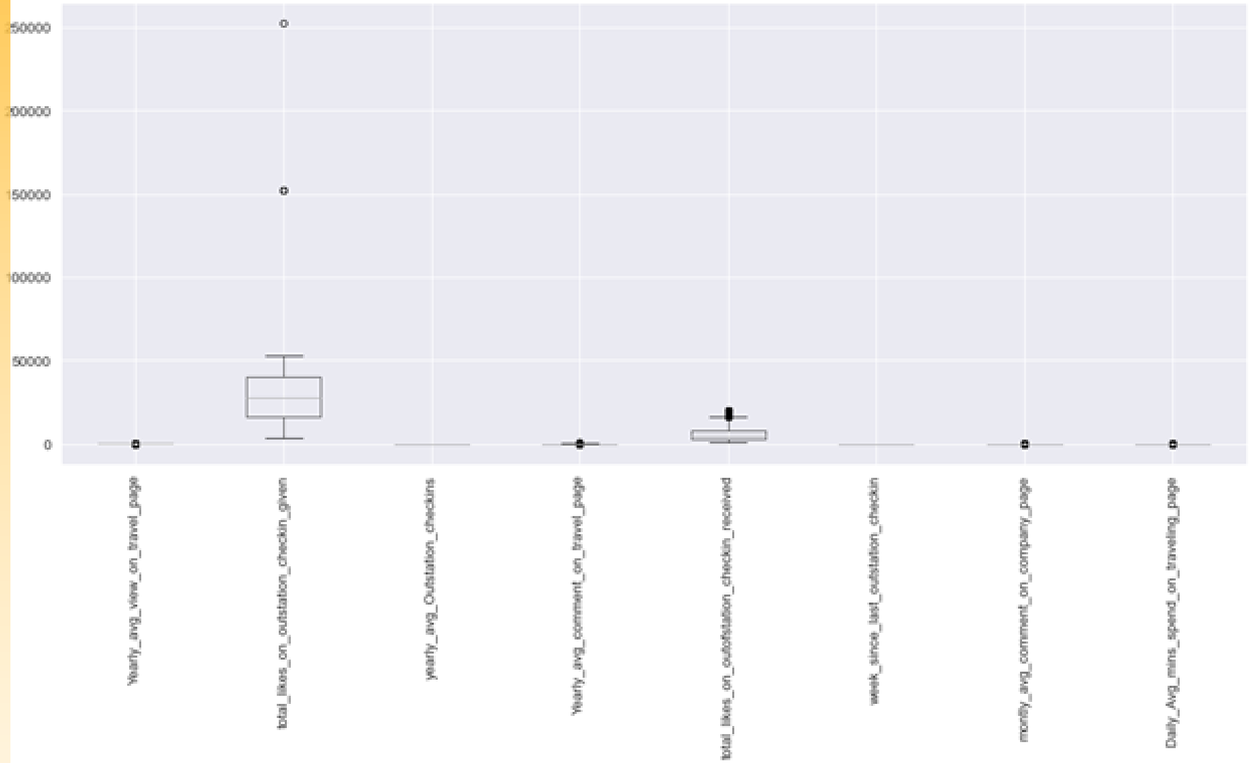
Missing Value Treatment

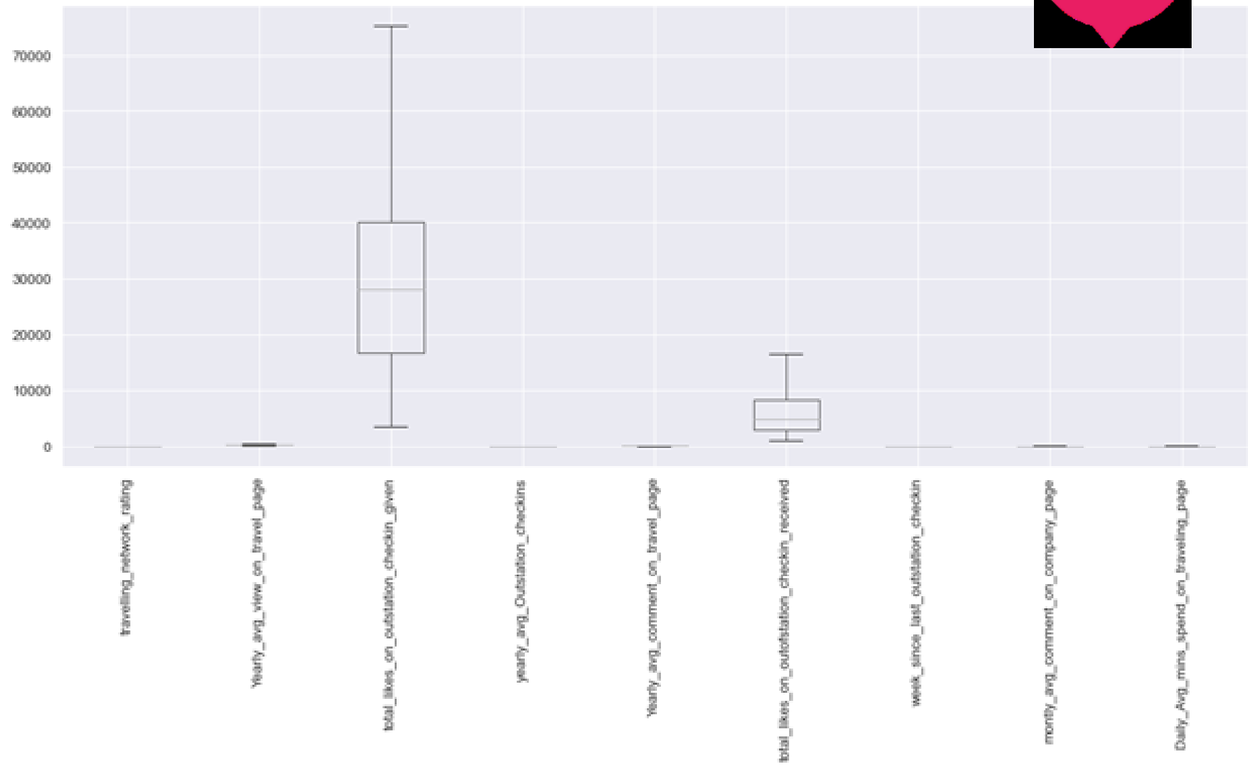
UserID	0
Taken_product	0
Yearly_avg_view_on_travel_page	581
preferred_device	53
total_likes_on_outstation_checkin_given	381
yearly_avg_Outstation_checkins	75
member_in_family	0
preferred_location_type	31
Yearly_avg_comment_on_travel_page	206
total_likes_on_outofstation_checkin_received	0
week_since_last_outstation_checkin	0
following_company_page	103
montly_avg_comment_on_company_page	0
working_flag	0
travelling_network_rating	0
Adult_flag	0
Daily_Avg_mins_spend_on_traveling_page	0

dtype: int64

- Among 7 variables NULL values were identified
- In the categorical variables, we have used mode for the missing value treatment
- In the case of numerical variables, we have used the median imputation method for the missing value treatment

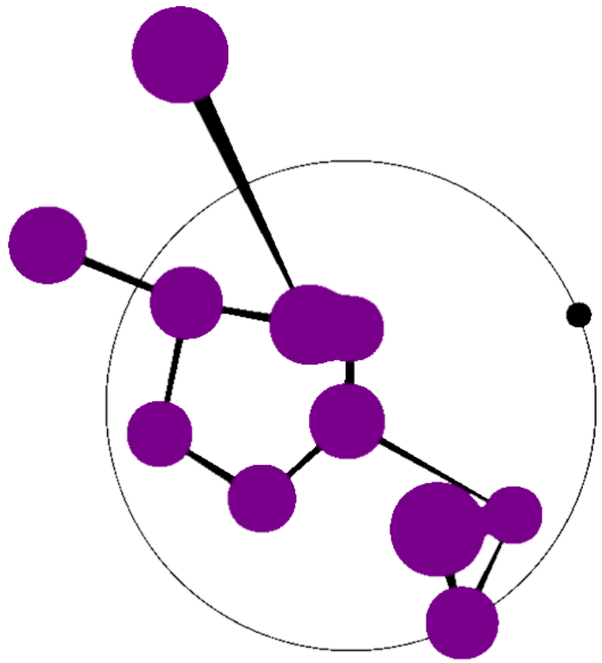
Outlier Treatment





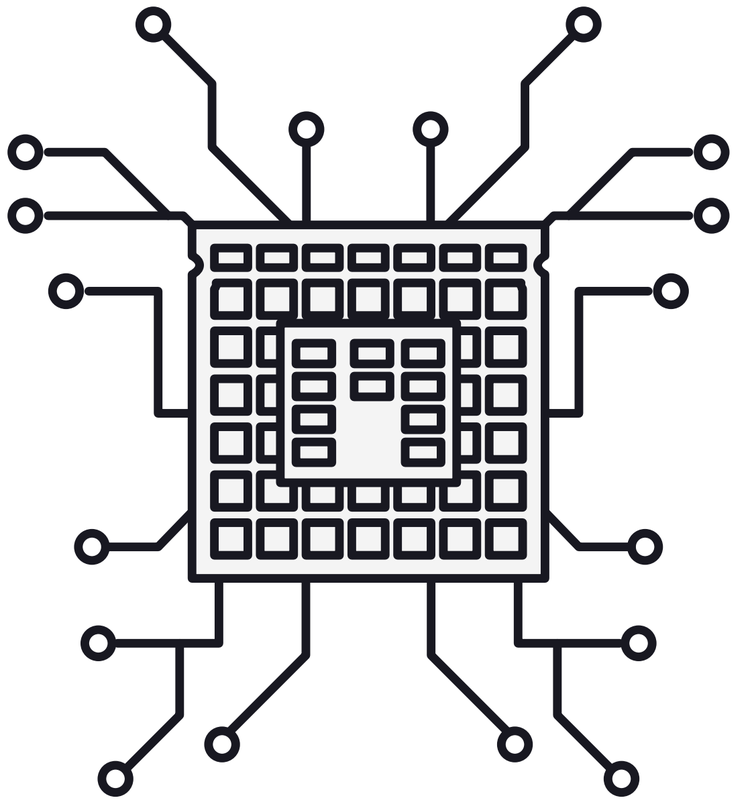
- Treatment was done on the features where the outliers were present
- The left image represents the features before treatment
- The right image represents the variables after treatment

MODEL BUILDING AND INTERPRETATION



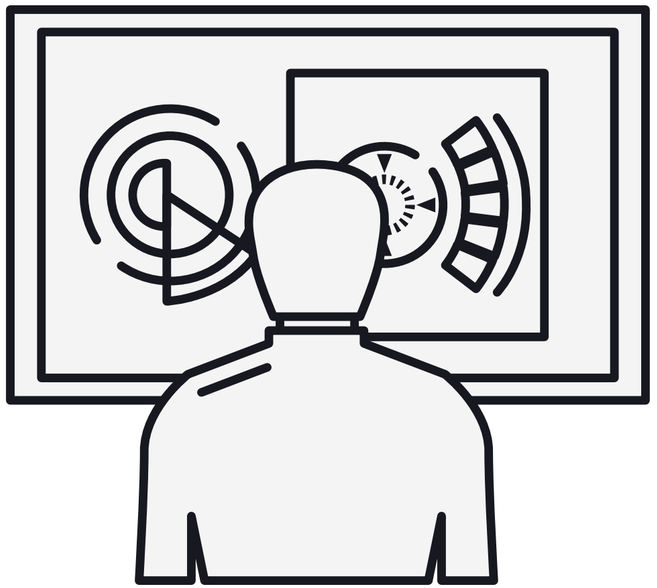
- The dataset was further divided into two parts based on Login devices i.e. either Mobile or Laptop
- The feature engineering was applied and 4 features were removed for better results
- Before building the model the data was scaled using the Standard Scalar
- the dataset was split into train & test datasets in the ratio of 80:20.
- 80% is the Train Set and 20% is the Test Set.

MODEL BUILDING AND INTERPRETATION



- The dataset was further divided into two parts based on Login devices i.e. either Mobile or Laptop
- The feature engineering was applied and 4 features were removed for better results
- Before building the model the data was scaled using the Standard Scalar
- the dataset was split into train & test datasets in the ratio of 80:20.
- 80% is the Train Set and 20% is the Test Set.

MODEL BUILDING



This is a classification problem. Therefore after the split, below mentioned Machine Learning Algorithms were applied separately on Laptop & Mobile devices:

- Logistic Regression
- Linear Discriminant Analysis
- Naive Bayes Model
- Decision Tree Classifier
- Random Forest Classifier
- K- Nearest Neighbour
- Model tuning (Bagging)
- Model tuning (Adaboosting & Gradient Boosting)

INSIGHTS FROM ANALYSIS (LAPTOP)

	LR Train	LR Test	LDA Train	LDA Test	NB Train	NB Test	CART Train	CART Test	RFC Train	RFC Test	Bagging Train	Bagging Test	Ada Boosting Train	Ada Boosting Test	KNN Train	KNN Test	Gradient Boosting Train	Gradient Boosting Test
Precision	0.739	0.678	0.734	0.667	0.707	0.682	1.0	0.953	1.0	0.987	1.0	0.955	0.890	0.840	0.969	0.851	0.951	0.885
Recall	0.756	0.696	0.770	0.703	0.813	0.784	1.0	0.959	1.0	1.000	1.0	1.000	0.876	0.818	1.000	1.000	0.966	0.939
F1 Score	0.747	0.687	0.752	0.684	0.756	0.730	1.0	0.956	1.0	0.993	1.0	0.977	0.883	0.829	0.984	0.919	0.959	0.911
Accuracy	0.737	0.718	0.739	0.712	0.731	0.742	1.0	0.961	1.0	0.994	1.0	0.979	0.881	0.850	0.983	0.922	0.957	0.919
AUC Score	0.830	0.824	0.829	0.822	0.815	0.823	1.0	0.961	1.0	1.000	1.0	0.999	0.962	0.943	1.000	1.000	0.992	0.943

- Logistic Regression- Logistic Regression has come up with poor accuracy
- Linear Discriminant Analysis- The LDA has also not performed well.
- Naïve Bayes Model- This has also not shown very poor performance.
- Decision Tree Classifier- This model has performed reasonably well.
- Random Forest Classifier- The performance is the best amongst all the models. The accuracy is 100% for the train and 99.4% for the test data sets. The precision is also 100% for the training dataset and 98.7% testing dataset.
- K- Nearest Neighbour- This model has performed fairly but not up to the mark.

INSIGHTS FROM ANALYSIS (MOBILE)

	LR Train	LR Test	LDA Train	LDA Test	NB Train	NB Test	CART Train	CART Test	RFC Train	RFC Test	Bagging Train	Bagging Test	Ada Boosting Train	Ada Boosting Test	KNN Train	KNN Test	Gradient Boosting Train	Gradient Boosting Test
Precision	0.710	0.712	0.708	0.716	0.658	0.662	1.0	0.987	1.0	0.997	1.0	0.994	0.799	0.795	0.984	0.975	0.853	0.847
Recall	0.706	0.733	0.702	0.732	0.739	0.764	1.0	0.985	1.0	0.996	1.0	0.992	0.797	0.811	1.000	0.998	0.822	0.829
F1 Score	0.708	0.723	0.705	0.724	0.696	0.710	1.0	0.986	1.0	0.996	1.0	0.993	0.798	0.803	0.992	0.986	0.837	0.838
Accuracy	0.709	0.717	0.707	0.719	0.678	0.685	1.0	0.986	1.0	0.996	1.0	0.993	0.799	0.800	0.992	0.986	0.840	0.839
AUC Score	0.777	0.783	0.776	0.782	0.757	0.765	1.0	0.986	1.0	1.000	1.0	1.000	0.888	0.886	1.000	0.997	0.930	0.886

- Logistic Regression- Logistic Regression has come up with poor accuracy
- Linear Discriminant Analysis- The LDA has also not performed well.
- Naïve Bayes Model- This has also not shown very poor performance.
- Decision Tree Classifier- This model has performed reasonably well.
- Random Forest Classifier- The performance is the best amongst all the models. The accuracy is 100% for the train and 99.6% for the test data sets. The precision is also 100% for training dataset and 99.7% testing dataset. The number of False positives are 6 and False Negatives is 8.
- K- Nearest Neighbour- This model has performed fairly but not up to the mark.

MODEL TUNING

LAPTOP

- After applying the model tuning technique bagging to the model the performance received was good. Accuracy for the training dataset was 100% and 97.9% for the test. In the case of boosting the Gradient boosting gave an accuracy of 95.7% on the train and 91.9% on the test datasets.

MOBILE

- After applying the model tuning technique bagging to the model the performance received was good. Accuracy for the training dataset was 100% and 99.3% for the test. In the case of boosting the Gradient boosting gave an accuracy of 84.0% on the train and 83.9% on the test datasets.

BUSINESS RECOMMENDATIONS

- The budget allocation for campaigns should be in a ratio of 75:25 for Laptop and Mobile respectively considering the traffic and probability of buying
- Social media campaigns, if aligned with photos related to the beach may attract higher traffic
- In social media campaign videos, there should be a reminder given to the customers to follow the page
- By following the social media page the customers will get the latest updates, promotions, discounts, and other offers launched by the company. This will increase the sale of the travel ticket.
- ·Working people have a high probability of buying the product therefore campaigns should address their concerns

BUSINESS RECOMMENDATIONS

LAPTOP

- The final recommendation to business shall be to move ahead with Random Forest Classifier, where on the test dataset the False Negatives were 0 and False Positives were only 2.

MOBILE

- The recommendation to the business shall be to move ahead with Random Forest Classifier for mobile devices. The number of False Positives and False Negatives is very high in the case of K-Nearest Neighbor.

GO GO AIR

THANK YOU