CAPSTONE PROJECT - FINAL REPORT

SOCIAL MEDIA & TOURISM

JAYA PREETHI R M

1. Introduction of the business problem

- Defining problem statement:

Supervised learning classification problem with a domain specific to the social media platform.

- Need of the study/project:

The aviation company aims to transition from traditional outreach to a more efficient, targeted digital approach. The study addresses the need for leveraging machine learning, specifically supervised learning classification, to analyse the digital and social behaviour of potential customers on a social networking platform. This shift from telecalling to digital advertising seeks to enhance customer engagement and streamline marketing efforts, making the process more data-driven and effective.

- Understanding business/social opportunity:

By collaborating with a social networking platform, the company seizes the opportunity to tap into the vast digital landscape and strategically reach customers with a higher likelihood of availing domestic and international trips. The project capitalises on the potential of supervised learning to classify users based on their behaviour, enabling precise targeting for digital advertisements. This not only enhances marketing efficiency but also aligns with contemporary trends, showcasing the company's adaptability and commitment to a more personalised, tech-savvy customer approach.

- Constraints

Not enough for market analysis

a) Data Limitations:

- The dataset may not capture all nuances of customer behaviour, leading to potential gaps in understanding.
- Incomplete or biassed data may hinder the accuracy and reliability of our models.

b) Model Complexity:

- The complexity of machine learning models, such as KNN and logistic regression, may pose challenges in interpretability and implementation.
- Balancing model accuracy with the need for simplicity and ease of integration into existing systems.

c) Resource Constraints:

- Limited computational resources may impact the scalability of certain models or the ability to process large datasets efficiently.
- Time constraints for model development, testing, and deployment may influence the depth of analysis and optimization.

d) Business Dynamics:

- The dynamic nature of the aviation industry and social media trends may introduce uncertainties, requiring constant model adaptation.

- External factors, such as economic changes or global events, could influence the effectiveness of our strategies.
By acknowledging these constraints, we can proactively address challenges and refine our approach to ensure practical and impactful solutions.

- Objectives
    - Clean and Wrangle the Available Dataset
    - Analyse the Potential of the Dataset for Business Recommendations
    - Perform Exploratory Data Analysis (EDA)
    - Build Different Models & find the Best Model
    - Derive Business Insights
    - Recommend Best Practices for the Business

2. Exploratory Data Analysis

- Nature of the Data
  Here's an overview of how the data may have been collected:

1. Time:
   - Yearly and Monthly Metrics: Many variables, such as "Yearly_avg_view_on_travel_page," "yearly_avg_Outstation_checkins," "montly_avg_comment_on_company_page" indicates yearly or monthly averages, suggesting a time-based aggregation. “Weeks Since Last Outstation Check-in” indicates the duration since the last outstation check-in, providing a time-related context.

2. Frequency:

“Daily_Avg_mins_spend_on_travelin g_page“ represents the average time spent on the travel page by the user on a daily basis, indicating frequency of engagement. “Following_company_page” indicates whether the user is following the company page or not, reflecting a binary frequency metric.

3. Methodology:
   - User Interaction Metrics: Variables such as "total_likes_on_outstation_ checkin_given", "total_likes_on_outofstation_checkin_received," and "Yearly_avg_comment_on_travel_page" suggest direct user interactions on the platform, possibly through likes and comments. "preferred_location_type" captures user preferences for types of locations, providing insights into travel interests. "travelling_network_rating" measures the user's social network's affinity for travel, indicating a subjective rating. "following_company_page" and "Adult_flag" suggest whether the user follows the company page and if they are considered an adult, possibly influenced by peer preferences. The "Buy_ticket" variable is a binary outcome, indicating whether the user intends to purchase a ticket in the next month.

- Visual inspection of data (rows, columns, descriptive details):

Glimpse of Data: This is an initial glimpse into the dataset (First 5 entries of the dataset).

2.1 TABLE - Glimpse of Data

UserID	Taken_product	Yearly_avg_view_on_travel_page	preferred_device	total_likes_on_outstation_checkin_given
1000001	Yes	307.0	iOS and Android	38570.0
1000002	No	367.0	iOS	9765.0
1000003	Yes	277.0	iOS and Android	48055.0
1000004	No	247.0	iOS	48720.0
1000005	No	202.0	iOS and Android	20685.0

yearly_avg_Outstation_checkins	member_in_family	preferred_location_type	Yearly_avg_comment_on_travel_page
1	2	Financial	94.0
1	1	Financial	61.0
1	2	Other	92.0
1	4	Financial	56.0
1	1	Medical	40.0

working_flag	travelling_network_rating	Adult_flag	Daily_Avg_mins_spend_on_traveling_page
No	1	0	8
Yes	4	1	10
No	2	0	7
No	3	0	8
No	4	1	6

total_likes_on_outofstation_checkin_received	week_since_last_outstation_checkin	following_company_page	montly_avg_comment_on_company_page
5993	8	Yes	11
5130	1	No	23
2090	6	Yes	15
2909	1	Yes	11
3468	9	No	12

Rows and Columns:

The number of rows (observations) is 11760
The number of columns (variables) is 17

Describe: Following is the glimpse into the description of the dataset (It is not a comprehensive look into the dataset). The description shows us that there are some outliers in the dataset and the UserID must be dropped because it is not significant for our analysis.

2.2 TABLE - Description table

	Yearly avg view on travel page	total likes on out station check in given	Yearly avg comment on travel page	total likes on out of station check in received	week since last out station check in	monthly avg comment on company page	travelling network rating	Adult flag	Daily Avg mins spend on travel page
count	11179	11379	11554	11760	11760	11760	11760	11760	11760
mean	280.83	28170.4	74.79	6531.7	3.2	28.66	2.71	0.79	13.8
std	68.18	14385	24.026	4706.6	2.62	48.66	1.08	0.85	9.07
min	35	3570	3	1009	0	11	1	0	0
25%	232	16380	57	2940.7	1	17	2	0	8
50%	271	28076	75	4948	3	22	3	1	12
75%	324	40525	92	8393.2	5	27	4	1	18
max	464	252430	815	20065	11	500	4	3	270

- Understanding of attributes (variable info, renaming if required):

I have made the description of the dataset in the previous section. From the table we can see that most of the data variables are integer based and some variables are categorical object based.

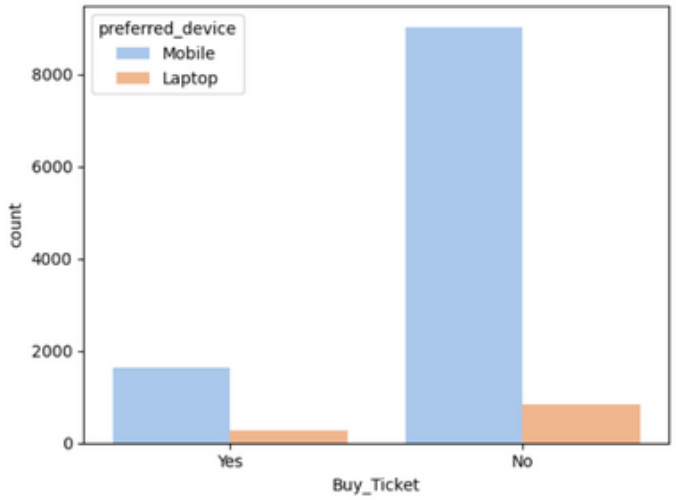
2.3 TABLE - Variable Information

Column	Count	Dtype
UserID	11760	integer
Taken_product	11760	object
Yearly_avg_view_on_travel_page	11179	float64
preferred_device	11707	object
total_likes_on_outstation_checkin_given	11379	float64
yearly_avg_Outstation_checkins	11685	object
member_in_family	11760	object
preferred_location_type	11729	object
Yearly_avg_comment_on_travel_page	11554	float64
total_likes_on_outofstation_checkin_received	11760	integer
week_since_last_outstation_checkin	11760	integer
following_company_page	11657	object

montly_avg_comment_on_company_page	11760	integer
working_flag	11760	object
travelling_network_rating	11760	integer
Adult_flag	11760	integer
Daily_Avg_mins_spend_on_traveling_page	11760	integer

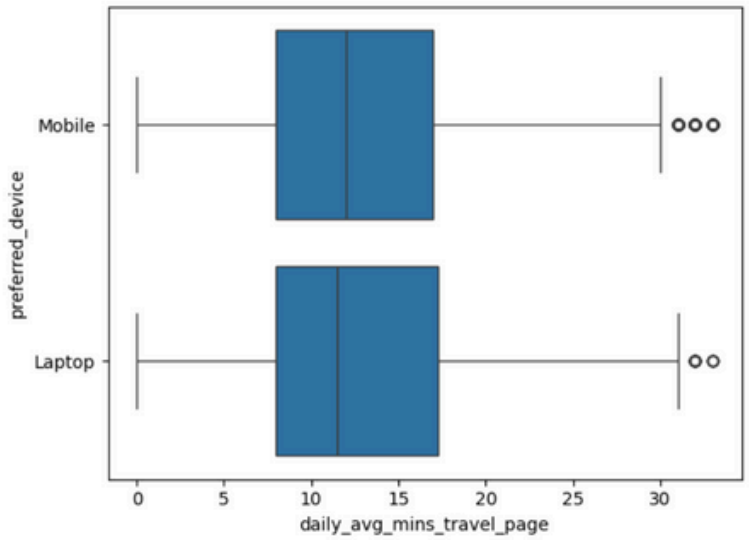
3. Data Visualization

3.1 DIAGRAMS - Buy Ticket & Preferred Device



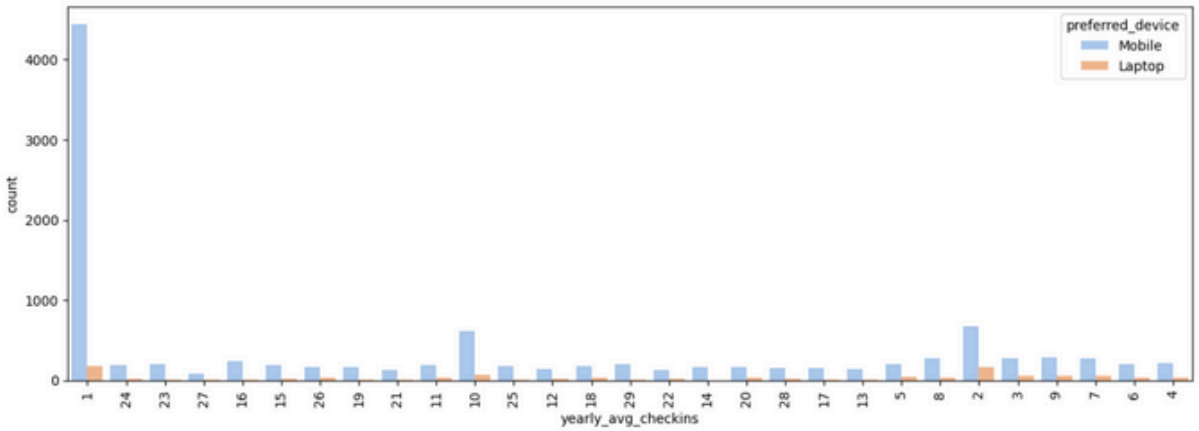
Buy Ticket & Preferred Device
This graph shows the graphical representation of the consumer's device preference in tandem with whether they bought a ticket from the company or not. We can notice that Majority of consumers prefer mobile phones and have never bought tickets from the company. We can also see that people who purchase tickets are in the minority compared to non-purchasers.

3.2 DIAGRAMS - Minutes on social media page & Preferred Device



Daily Minutes spent on the travel page & Preferred Device
This graph shows the graphical representation of the consumer's device preference in tandem with how much time they spend on the company's page. We can see that it is very similar in both the cases, where an average consumer spends about 7-17 minutes on the travel page.

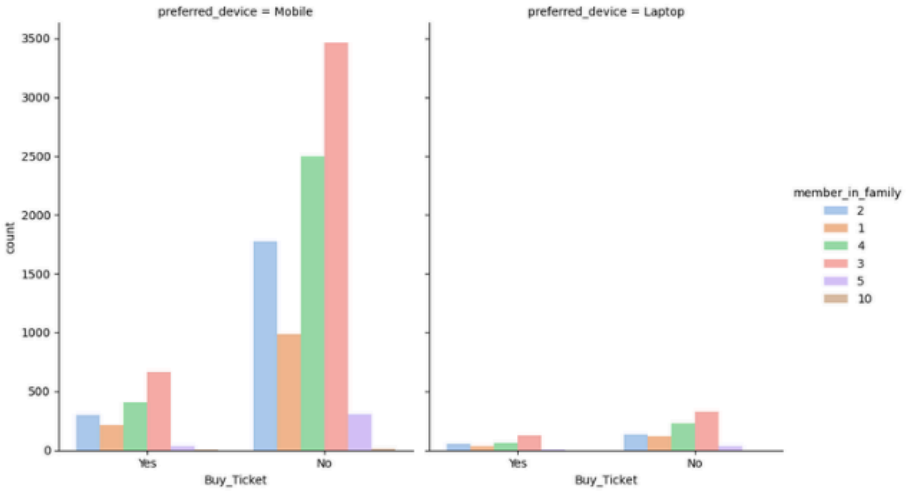
3.3 DIAGRAMS - Yearly average checkin & Preferred Device



Yearly Average checkins & Preferred Device

This graph shows the graphical representation of the consumer's device preference in tandem with yearly average checkins by each consumer. Majority of consumers only travel once a year.

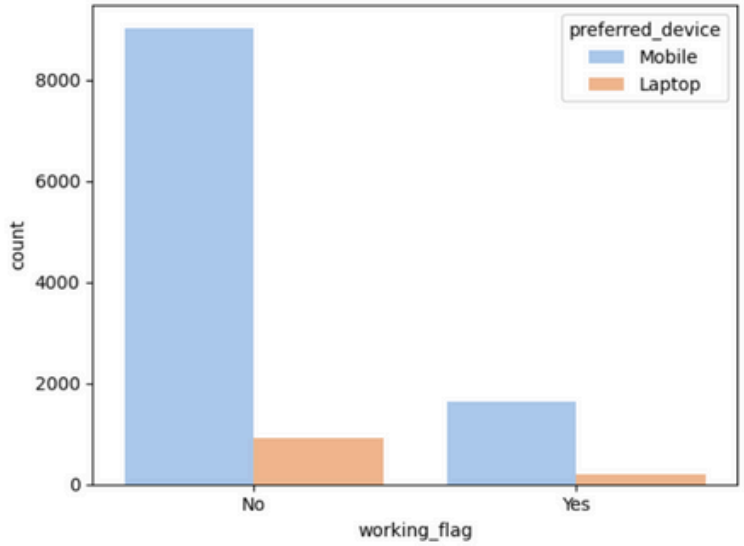
3.4 DIAGRAMS - Buy Ticket & Preferred Device & Yearly average checkin



Yearly Average checkins & Preferred Device & Buy Ticket

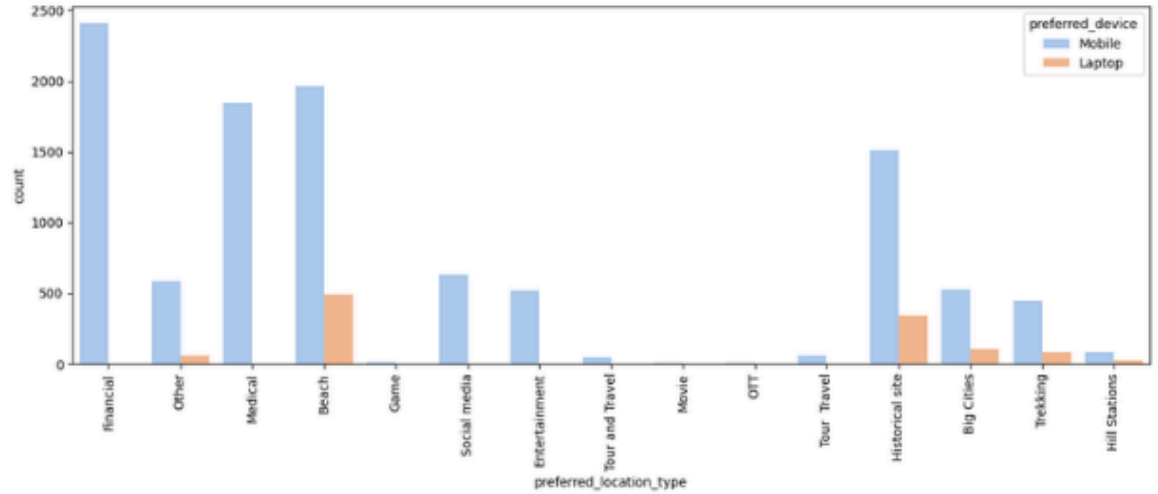
This graph shows the graphical representation of the consumer's device preference in tandem with whether the consumer has purchased a ticket or not and the number of family members. Majority of consumers have 3 family members followed by consumers with 4 and 2 family members, showing that most consumers have small family.

3.5 DIAGRAMS - Working Flag & Preferred Device



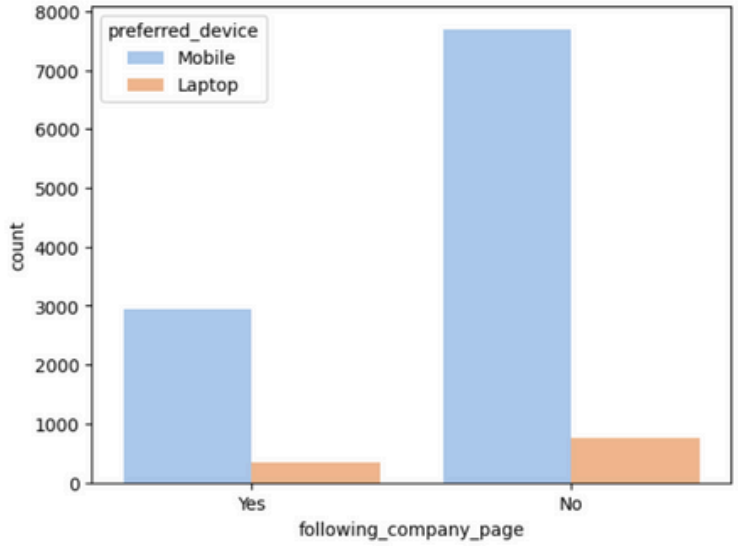
Working Flag & Preferred Device
This graph shows the graphical representation of the consumer's device preference in tandem with whether they are working or not. In both type of consumers, majority of consumers are not working individuals.

3.6 DIAGRAMS - Preferred Location Type & Preferred Device



Preferred location type & Preferred Device
This graph shows the graphical representation of the consumer's device preference in tandem with their preferred location for vacation. Among mobile users, we can see that most consumers prefer to visit financial, beach and medical sites as their vacation destination. On the other hand, laptop users prefer Beach and historical sites. On both instances, Beach is predominantly preferred as vacation spot.

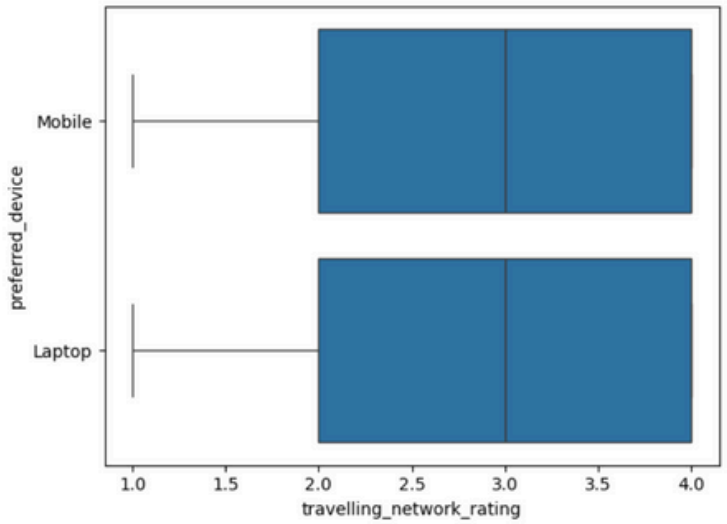
3.7 DIAGRAMS - Following company page & Preferred Device



Following Company page & Preferred Device

This graph shows the graphical representation of the consumer's device preference in tandem with whether they are following the company's social media page or not. In both type of consumers, majority of consumers are not following the company's social media page and a minority of individuals are following company page.

3.8 DIAGRAMS - Travelling friends & Preferred Device



Travelling network rating & Preferred Device

This graph shows the graphical representation of the consumer's device preference in tandem with how many friends they have who are also interested in travelling. In both type of consumers, majority of consumers have two to four friends in their social circle who are also interested in travelling.

4. Data Cleaning and Preprocessing

Various preprocessing measures were taken:

- Null variable treatment: Following null variables were treated by imputing mean for numerical variables and mode for non-numerical variables.

4.1 TABLE - Null variable treatment

Column	null variable before treatment	null variable after treatment
UserID	0	0
Taken_product	0	0
Yearly_avg_view_on_travel_page	581	0

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

- Outlier treatment using Interquartile method: Following description table shows the results.

4.2 TABLE - Outlier treatment

	Yearly avg view on travel page	**total likes on outstation check in given**	**Yearly avg comment on travel page**	**total likes on out of station check in received**	**week since last out station check in**	**monthly avg comment on company page**	**travelling network rating**	**Adult flag**	**Daily Avg mins spend on travelling page**
count	11731	11756	11720	10844	11760	11518	11760	11080	11410
mean	280.71	28119.69	74.84	5527.43	3.2	22.45	2.71	0.66	13.04
std	65.81	13858.92	21.17	3312.61	2.62	6.88	1.08	0.67	7.33
min	135	3570	31	1009	0	11	1	0	0
25%	233	16695	58	2845	1	17	2	0	8
50%	275	28170.48	74.79	4668	3	22	3	1	12
75%	321	40111.25	92	6844	5	27	4	1	17
max	455	52512	125	16567	11	42	4	2	33

- Dropped UserID variable and renamed variables: Dropped User ID because it is a descriptive variable and changed variable names because they were too lengthy

4.3 TABLE - Variable treatment

Original Variables	**Renamed Variables**

UserID	-
Taken_product	Buy_Ticket
Yearly_avg_view_on_travel_page	yearly_avg_views
preferred_device	preferred_device
total_likes_on_outstation_checkin_given	total_likes_checkin
yearly_avg_Outstation_checkins	yearly_avg_checkin
member_in_family	member_in_family
preferred_location_type	preferred_location_type
Yearly_avg_comment_on_travel_page	yearly_avg_comments
Total_likes_on_outofstation_checkin_ received	total_likes_user_outofstation
week_since_last_outstation_checkin	week_since_checkin
following_company_page	following_company_page
Montly_avg_comment_on_company_ page	montly_avg_comment
working_flag	working_flag
travelling_network_rating	travelling_network_rating
Adult_flag	Adult_flag
Daily_Avg_mins_spend_on_traveling_ page	daily_avg_mins_travvel_page

- Logical changes
  Various logical changes like changing repetition, unnecessary entries and entries that don't format with the other endings were changed.

5. Model Building & Validation

Following are the models built separately for mobile and laptop users and their model accuracy metrics.

5.1 TABLE - Model Building & Validation

MOBILE						
MODEL	TRAINING			TEST		
	ACCURACY	PRECISION	RECALL	ACCURACY	PRECISION	RECALL
LOGISTIC REGRESSION	87%	73%	29%	87%	70%	26%
LOGISTIC REGRESSION - GRIDSEARCHCV	88%	73%	28%	87%	70%	26%

KNN	**98%**	**98%**	**92%**	**97%**	**93%**	**87%**
DECISION TREE	100%	100%	100%	97%	90%	89%
RANDOM FOREST	100%	100%	100%	98%	100%	87%
GRADIENT BOOSTING	100%	100%	100%	99%	100%	92%
SMOTE	98%	98%	92%	97%	93%	87%
LAPTOP						
MODEL	**TRAINING**			**TEST**		
	ACCURACY	**PRECISION**	**RECALL**	**ACCURACY**	**PRECISION**	**RECALL**
LOGISTIC REGRESSION	**86%**	**82%**	**56%**	**86%**	**79%**	**59%**
LOGISTIC REGRESSION - GRIDSEARCHCV	86%	82%	56%	86%	84%	52%
KNN	95%	94%	85%	84%	73%	59%
DECISION TREE	100%	100%	100%	97%	91%	97%
RANDOM FOREST	100%	100%	100%	98%	97%	93%
GRADIENT BOOSTING	100%	100%	100%	98%	100%	95%
SMOTE	95%	94%	85%	84%	73%	59%

Here are the models built for both the best model and a brief explanation on the validation of those models:

- **Logistic regression:**
  Parameters: solver='newton-cg', max_iter=100, penalty='none', verbose=True, n_jobs=2

  Pros:
  Logistic Regression provides easily interpretable coefficients, allowing you to understand the impact of each feature on the predicted probability and it is less prone to overfitting, making it a good choice when the dataset is limited.

  Cons:
  It assumes linearity, which might not capture complex relationships. It may not perform well if the true relationship is highly non-linear and is also Sensitive to outliers in the data.

  Fit to Context
  - Mobile: the model has 87% accuracy in both test and train, the precision is roughly 70% and recall is roughly 20%. It might not be the best way to analyze the variables due to poor accuracy of the model.

- Laptop: the model has 86% accuracy in both test and train, the precision is 82% and recall is roughly 56% in train and the precision is 79% and recall is roughly 56% in train.

I chose the Logistic regression model as the final model for Laptop users. This is because compared to other models, the logistic regression model was not overfitted and it also had good recall to predict the future customers who would buy tickets. It is also the best at interpretability.

- **KNN**
  Parameters: n_neighbors=5

  Pros:
  KNN makes no assumptions about the underlying data distribution, making it versatile and It can capture complex relationships in nonlinear datasets.

  Cons:
  It is Sensitive to Outliers: Sensitive to outliers and noisy data, which can significantly impact results and it is not easy to interpret

  Fit to Context
  - Mobile: The model has high accuracy indicates overall good performance on the test set. It has 98% accuracy, 98% precision and 95% recall in training. 93% precision test implies that when the model predicts a customer will buy a ticket and 87% recall suggests the model captures 87% of customers who actually bought a ticket.
  - Laptop: The model has high accuracy indicates overall good performance on the test set. It has 95% accuracy, 94% precision and 85% recall in training. 84% precision test implies that when the model predicts a customer will buy a ticket and 73% recall suggests the model captures 59% of customers who actually bought a ticket.

I chose the KNN model as the best model for analysing mobile users. This was because this model was not overfitted and it had good precision and recall. Even Though, it was difficult to interpret compared to other models. This model will be apt to interpret results for mobile users.

6. Model Interpretation

Logistic Regression - Laptop Users

Classification Report

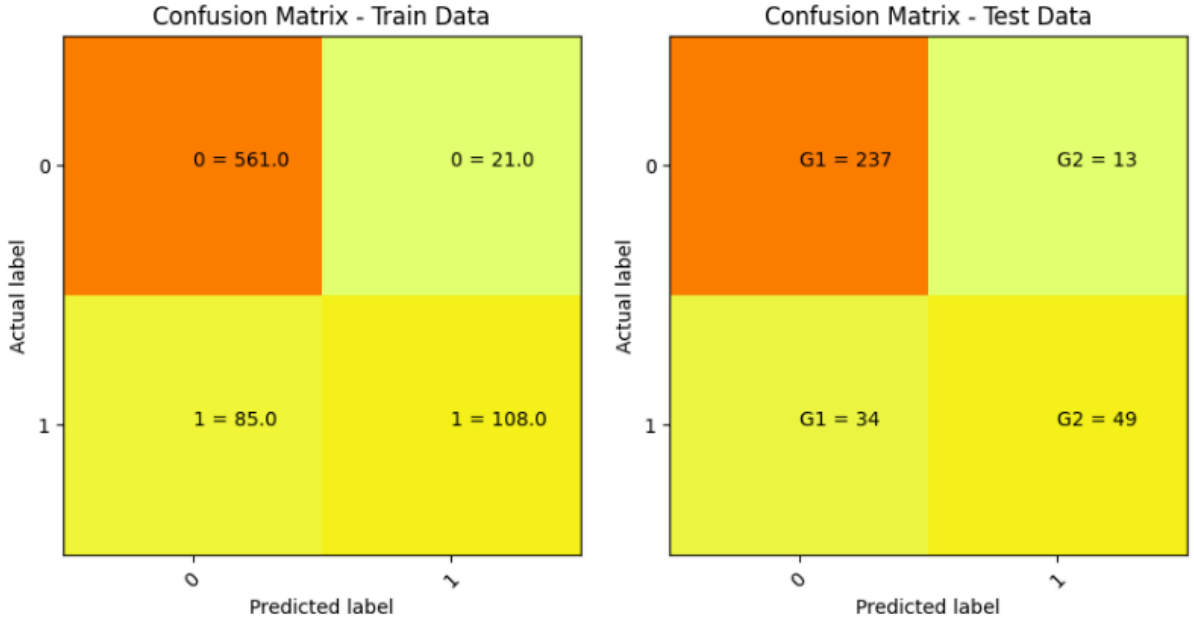
5.2 TABLE - Logreg classification report - train and test

```
Test Set:
Accuracy: 0.8589
Classification Report:
              precision    recall  f1-score   support

          No       0.87      0.95      0.91       250
         Yes       0.79      0.59      0.68        83

    accuracy                           0.86       333
   macro avg       0.83      0.77      0.79       333
weighted avg       0.85      0.86      0.85       333


Training Set:
Accuracy: 0.8606
Classification Report:
              precision    recall  f1-score   support

          No       0.87      0.96      0.91       582
         Yes       0.82      0.56      0.67       193

    accuracy                           0.86       775
   macro avg       0.84      0.76      0.79       775
weighted avg       0.86      0.86      0.85       775
```

The model has an accuracy of 86% on the test set, with a precision of 82% and recall of 56%. The model has an accuracy of 86% on the test set, with a precision of 79% and recall of 59%. There is slight overfitting in recall rates in the test model, the low recall indicates that the model may not identify all potential customers (false negatives). Precision is pretty good, meaning that when it predicts a customer will buy a ticket, it's correct about 82% of the time.

CONFUSION MATRIX

5.3 DIAGRAMS - Logreg confusion matrix - train and test

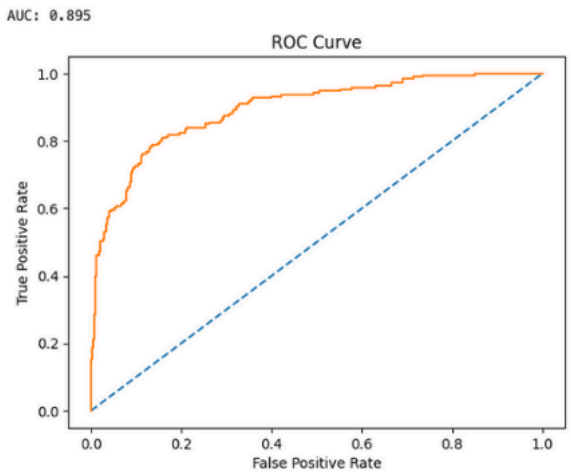


From the confusion matrix, there are high cases of true negatives followed by true positives. High True Negatives and true positives suggest that the model is good at identifying customers who are not likely and likely to buy tickets respectively. This is due to the majority of customers not buying tickets from the train dataset.

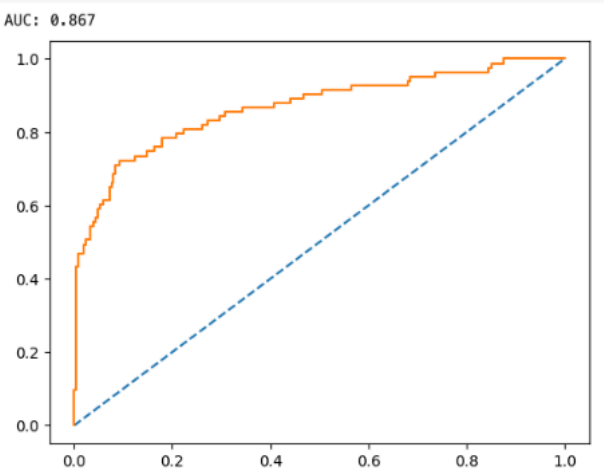
ROC-AUC CURVE

5.4 DIAGRAMS - Logreg ROC-AUC CURVE - train and test

TRAIN MODEL



TEST MODEL



The AUC score for the train is 89% and for the test is 86%. The curve itself is much kinked with the test model compared to the train model. However, they both are broadly covering the graph area. The ROC curves being kinked might indicate that the model has a threshold at which it is making predictions, and this threshold might be affecting the balance between true positives and false positives. A ROC curve covering a broad area is generally a positive sign, as it means the model is able to trade off between sensitivity and specificity effectively.

KNN model - Mobile users

Classification Report

5.5 TABLE - KNN classification report - train and test

```
Test Set:
Accuracy: 0.9706
Classification Report:
              precision    recall  f1-score   support

          No       0.98      0.99      0.98      2710
         Yes       0.93      0.87      0.90       486

    accuracy                           0.97      3196
   macro avg       0.95      0.93      0.94      3196
weighted avg       0.97      0.97      0.97      3196


Training Set:
Accuracy: 0.9848
Classification Report:
              precision    recall  f1-score   support

          No       0.99      1.00      0.99      6322
         Yes       0.98      0.92      0.95      1134

    accuracy                           0.98      7456
   macro avg       0.98      0.96      0.97      7456
weighted avg       0.98      0.98      0.98      7456
```

The test model has an overall accuracy of 97%, precision is 93% and recall is 87%. Train model on the other hand, has accuracy of 98%, precision and recall are 98% and 97% respectively. Of all instances predicted as positive, 93% were actually positive. This indicates a low false-positive rate. Of all actual positive instances, the model correctly identified 87%. This indicates a good ability to capture positive instances. Therefore, the model is showing high accuracy and good balance between precision and recall on both training and test sets.

Confusion Matrix

5.6 TABLE - KNN confusion matrix - train and test

```
Confusion Matrix - Test Set:
[[2678   32]
 [  62  424]]

Confusion Matrix - Training Set:
[[6297   25]
 [  88 1046]]
```

Both sets show a high number of true positives and true negatives, indicating that the model is effectively capturing both classes. Therefore, the model can predict consumers who would buy tickets or not accurately.

7. Final interpretation

Here are the combined interpretations of the model building and EDA.

- From the KNN model for Laptop users - total_likes_checkin - 0.192272& total_likes_user_outofstation - 0.189018 are the two variables that are statistically significant in predicting whether or not a mobile user would purchase a ticket for the company. When there is an increase in the total likes on each check in, the odds of a mobile user buying a ticket increases by approximately 19.23%. When the total likes on out of station posts from a customer increases, the odds of a mobile user buying a ticket increase by approximately 18.90%.
- From the Logistic regression for Mobile Users - preferred_location_type_9 , yearly_avg_checkins_24 - 1.23 , following_company_page_1 - 1.02, Adult_flag_0.0 - 0.81, member_in_family_3 - 0.40, member_in_family_4 - 0.30 are the most statistically significant in predicting whether or not a laptop user would purchase a ticket for the company.
    - There is a high correlation between what type of location the consumer would like to travel to and whether they purchase a ticket or not.
    - The higher the yearly average check-ins, the more the purchase of tickets by a consumer.
    - If a consumer follows the company page, it is highly likely for them to purchase a ticket.
    - If they are not adult users, they could purchase tickets in future.
    - Families with 3 or 4 members is more likely to purchase a ticket from the company.
- From EDA:
    - There is a high volume of users who have never purchased a ticket.
    - On a daily basis users spend close to 7-17 minutes on the company's social media page.
    - Majority of company’s social media viewers are not employed
    - Majority of consumers prefer Beach destinations for tourism.
    - Most viewers do not follow the company’s social media page.
    - There are 3 to 4 friends in the majority of the user’s friend circle who would also like to travel.
    - Most users only travel once a year.

8. Business Recommendation

- Encourage more users to follow the company's social media page by running **targeted campaigns** highlighting the benefits, promotions, and exclusive content available to followers. Engaging content and incentives can increase the number of followers, leading to a broader reach.
- Increase digital ad spend on users who have a **higher likelihood of purchasing tickets**, especially those with a higher yearly average check-ins, followers of the company page, and families with 3 or 4 members. Create loyalty programs or subscription services that cater to users who travel once a year. Provide exclusive perks, discounts, or early access to travel packages to encourage these users to choose the company consistently for their annual trips.
- Implement strategies to **increase daily user engagement** on the social media page, aiming for a 10% increase in average daily engagement time.
- Design targeted promotions and discounts for **non-adult users**. Aim for a 15% increase in the number of followers, leading to a corresponding increase in potential ticket buyers. Develop age-specific promotions, highlighting benefits that resonate with non-adult users.
- Tailor marketing efforts and travel packages to align with user preferences for **beach destinations**. Highlight beach-related promotions, travel guides, and exclusive offers to attract and retain users interested in beach vacations.Focus marketing efforts on beach destinations, aiming for a 10% increase in ticket purchases to these locations.
- Introduce **family-focused promotions** aiming for a 20% increase in ticket purchases from users with 3 or 4 family members. Develop and promote travel packages specifically designed for families, considering the average family size of three members. Create family-friendly itineraries, accommodations, and activities that cater to the needs and preferences of small families. Highlight the benefits of family travel and the unique experiences offered by the company.
- **Leverage social networks** by introducing referral programs. Encourage users to refer friends by offering incentives such as discounts, travel credits, or exclusive group travel experiences. Harness the power of word-of-mouth marketing within the travel community. Aim for a 15% increase in ticket purchases by encouraging users to travel with their friends.