Capstone Project

PGPDSBA University Of Texas, Austin

Presented by Karnati Harik Charan

Interim Project Report On Flight Booking Prediction

Capstone Project

Table Of Contents:

Capstone Project

1.) Business Context

An aviation company offering both domestic and international travel services is looking to improve its customer acquisition strategy. Traditionally, outreach has been done through broad-based methods like telecalling, which are costly and less effective. To modernize their approach, the company has partnered with a social networking platform to leverage customer digital and social behavior for targeted advertising. Since running advertisements on digital platforms is expensive, it is essential to focus marketing efforts only on customers who are most likely to purchase tickets.

2. Objective

As a data scientist, you are tasked with building predictive models that estimate the propensity of customers to purchase tickets based on their online behavior.

Two separate models must be developed - one for laptop users and another for mobile users, since device usage influences purchase patterns. Accurate models will help the company minimize advertising costs while maximizing

conversions, leading to more efficient campaigns and improved return on investment.

3. Data Overview

The dataset used for this study contains structured customer booking information for predicting flight booking behavior. It consists of multiple numerical and categorical features representing customer demographics, booking patterns, and interaction history. The target variable is binary in nature, indicating whether a customer completed a booking or not.

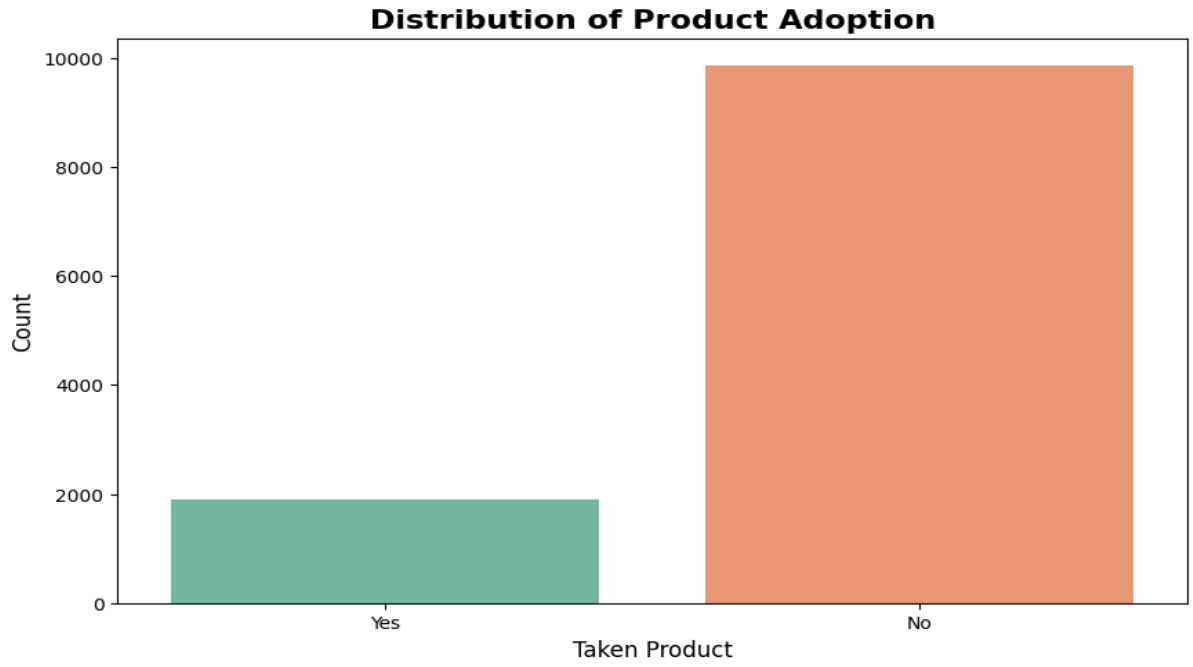
An initial examination of the data shows that the dataset is moderately imbalanced, with a larger proportion of non-booking instances compared to booking instances. This imbalance is taken into consideration during model evaluation to avoid misleading accuracy-based conclusions.

Missing values in the dataset are minimal and were handled during the data preprocessing stage to ensure data quality and model stability. All features were reviewed for appropriate data types, and necessary transformations such as encoding of categorical variables and scaling of

numerical features were applied before model training.

4. Univariate Analysis

Target Variable Distribution



This chart shows how many people have decided to use (adopt) the product and how many have not.

i.)Big Bar (No): Nearly 10,000 people have NOT taken the product.

ii.)Small Bar (Yes): Only about 2,000 people HAVE taken the product.

Most people in this group (**about 5 out of every 6**) are **NOT** using the product. This means the

company needs to work hard to convince more people to try it!

Numeric Variables Distribution:

This set of histograms and distribution plots shows the patterns of several key features, including usage, engagement, and network ratings. The red dashed line represents the Mean (average), and the green dashed line represents the Median (middle value).

1. Distribution of Yearly

Avg_view_on_travel_page

i.)Shape: This distribution is relatively normal (bell-shaped), though slightly left-skewed (meaning the tail extends slightly more to the left).

ii.)Central Tendency: The Mean (280.35) is slightly higher than the Median (271.00), which is typical for a slightly left-skewed distribution, though they are very close.

iii.)Interpretation: The average person views the travel page approximately 271 to 280 times per year.

Capstone Project

2. Distribution of Daily Avg_mins_spend_on_traveling_page

i.)Shape: This distribution is highly right-skewed, with most data clustered near zero.

ii.)Central Tendency: The Mean (13.82) is significantly higher than the Median (12.00). This large difference indicates that a few individuals spend a very long time on the page, pulling the average up.

iii.)Interpretation: The typical user (median) spends 12 minutes on the travel page daily, but the average (mean) is higher due to some heavy users.

3. Distribution of travelling_network_rating

i.)Shape: This distribution is multimodal, specifically bimodal or quadrimodal, with clear peaks around the whole number ratings of 1, 2, 3, and 4. There are very few ratings between these whole numbers.

Capstone Project

ii.)Central Tendency: The Mean (2.71) and Median (3.00) are close, suggesting the central point is between 2 and 3.

iii.)Interpretation: The network rating is primarily given as whole numbers (1, 2, 3, 4), which is common for discrete rating scales. The most frequent ratings appear to be 1 and 4.

4. Distribution of total_likes_on_outofstation_checkin_received

i.)Shape: This distribution is highly right-skewed, with a large number of people receiving very few likes.

ii.)Central Tendency: The Mean (6531.70) is significantly higher than the Median (4948.00).

iii.)Interpretation: While the typical person (median) receives around 4,948 likes, the average is inflated to over 6,500 by a small group of highly popular users who receive a massive number of likes (influencers or high-engagement users).

Capstone Project

5. Distribution of week_since_last_outstation_checkin

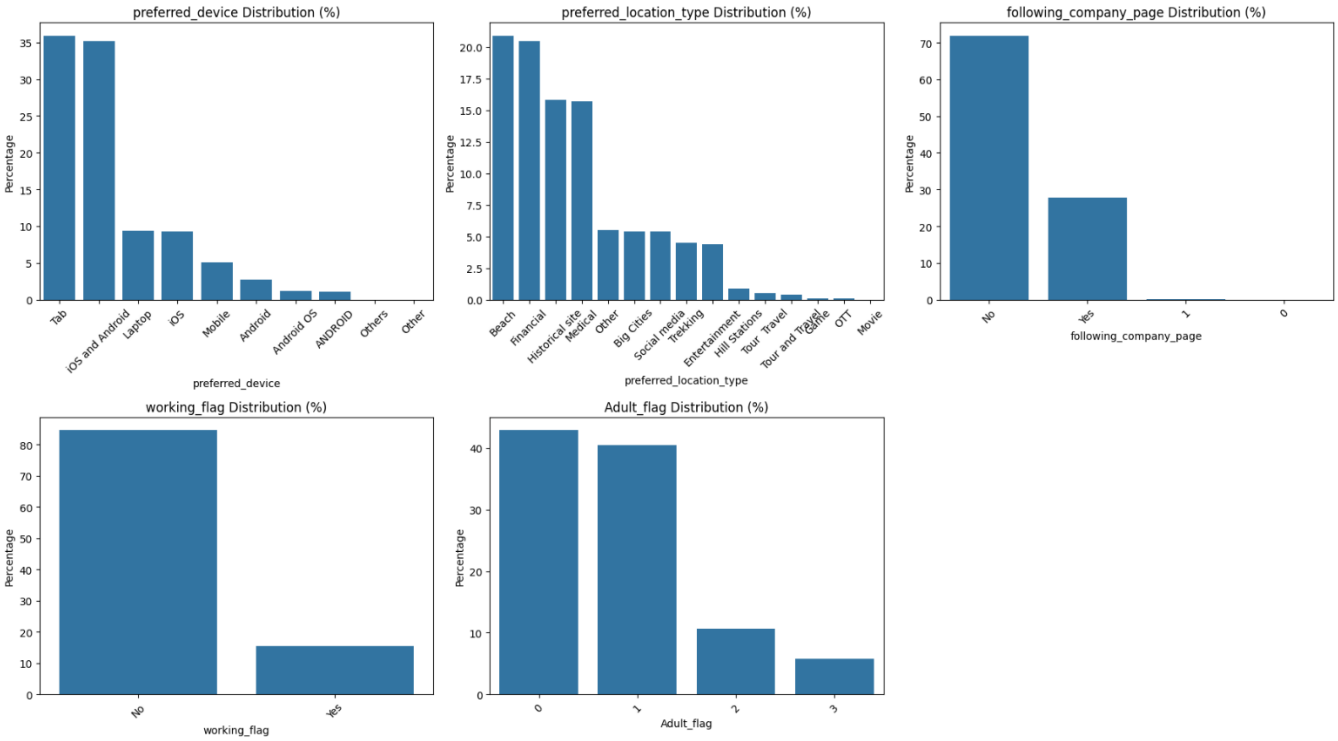
i.)Shape: This distribution is right-skewed, with a peak close to 0-1 weeks, suggesting many users check in frequently.

ii.)Central Tendency: The Mean (3.20) is slightly higher than the Median (3.00).

iii.)Interpretation: The typical user (median) had their last out-of-station check-in about 3 weeks ago, with the average being slightly higher, likely due to some users who haven't checked in for a long time.

Capstone Project

Categorical Variables Distribution:



Analysis of Categorical Feature Distributions

This set of bar charts shows the percentage distribution of five categorical features: preferred_device, preferred_location_type, following_company_page, working_flag, and Adult_flag.

Capstone Project

i. Preferred Device Distribution (%)

Preferred Device	Approximate Percentage	Key Insight
Tab	approx 35%	The single most preferred device.
IOS and Android	approx 35%	Combined mobile category is almost tied with tablets.
Laptop	approx 10%	Significantly less used than mobile devices.
Other	All other categories (IOS, Mobile, Android	

Capstone Project

Preferred Device	Approximate Percentage	Key Insight
	OS, etc.) are below 5%.	

Key Takeaway: The vast majority of users access the platform using Tab (Tablet) and IOS/Android devices, suggesting a highly mobile-centric user base.

ii. Preferred Location Type Distribution (%)

Preferred Location Type	Approximate Percentage	Key Insight
Beach	approx 21%	The most preferred location type.
Financial site	approx 20%	Nearly as popular as the Beach.

Capstone Project

Preferred Location Type	Approximate Percentage	Key Insight
Historical site	approx 16%	The third most popular type.
Medical	approx 16%	Nearly tied with Historical site.
Other	approx 6%	

Key Takeaway: User preferences are diverse, but Beach, Financial site, Historical site, and Medical locations are the dominant categories, each representing about 15% to 21% of preferences.

Capstone Project

iii. Following Company Page Distribution (%)

Following Company Page	Approximate Percentage	Key Insight
No	approx 71%	The majority of users do not follow the company page.
Yes	approx 29%	Less than a third of users follow the company page.

Key Takeaway: There is a low level of engagement with the company's official page, indicating an opportunity to improve marketing and content strategies to increase followers.

Capstone Project

iv. Working Flag Distribution (%)

Working Flag	Approximate Percentage	Key Insight
No	approx 83%	The overwhelming majority of the user base is not working (or is not marked as working).
Yes	approx 17%	A small minority of the user base is working.

Key Takeaway: The platform seems to appeal more to non-working individuals (students, retirees, unemployed, etc.).

Capstone Project

v. Adult Flag Distribution (%)

Adult Flag	Approximate Percentage	Key Insight
0	approx 43%	The highest single category.
1	approx 40%	The second-highest category, almost tied with category 0.
2	approx 11%	The third category less than first and second
3	approx 6%	The least category

Key Takeaway: The user population is heavily concentrated in the two largest categories, 0 and 1, which together account for over 80% of the users. The meaning of these numerical categories would require further context.

Capstone Project

Statistical Summary Table

	count	mean	std	min	25%	50%	75%	max
UserID	11760.0	1.005880e+06	3394.963917	1000001.0	1002940.75	1005880.5	1008820.25	1011760.0
Yearly_avg_view_on_travel_page	11760.0	2.803452e+02	66.511330	35.0	233.00	271.0	322.00	464.0
total_likes_on_outstation_checkin_given	11760.0	2.816742e+04	14150.080463	3570.0	16697.25	28076.0	40115.25	252430.0
yearly_avg_Outstation_checkins	11760.0	8.196259e+00	8.650950	1.0	1.00	4.0	14.00	29.0
member_in_family	11760.0	2.921344e+00	1.044883	1.0	2.00	3.0	4.00	10.0
Yearly_avg_comment_on_travel_page	11760.0	7.479371e+01	23.815280	3.0	57.00	75.0	92.00	815.0
total_likes_on_outofstation_checkin_received	11760.0	6.531699e+03	4706.613785	1009.0	2940.75	4948.0	8393.25	20065.0
week_since_last_outstation_checkin	11760.0	3.203571e+00	2.616365	0.0	1.00	3.0	5.00	11.0
montly_avg_comment_on_company_page	11760.0	2.866156e+01	48.660504	11.0	17.00	22.0	27.00	500.0
travelling_network_rating	11760.0	2.712245e+00	1.080887	1.0	2.00	3.0	4.00	4.0
Adult_flag	11760.0	7.938776e-01	0.851823	0.0	0.00	1.0	1.00	3.0
Daily_Avg_mins_spend_on_traveling_page	11760.0	1.381743e+01	9.070657	0.0	8.00	12.0	18.00	270.0

i. **Low Adoption Rate:** The initial assessment (not shown but inferred from the first prompt) highlighted an extremely low product adoption rate.

ii. **Power User Skew:** Engagement metrics (Daily_Avg_mins_spend_on_traveling_page, total_likes_on_outofstation_checkin_received) are heavily influenced by a small group of high-activity users, suggesting that standard averages might not reflect the behavior of the *typical* user (Median).

iii. **Mobile/Tablet Focus:** The user base is highly mobile-centric (Tab and IOS/Android are dominant), which should inform product design and marketing efforts.

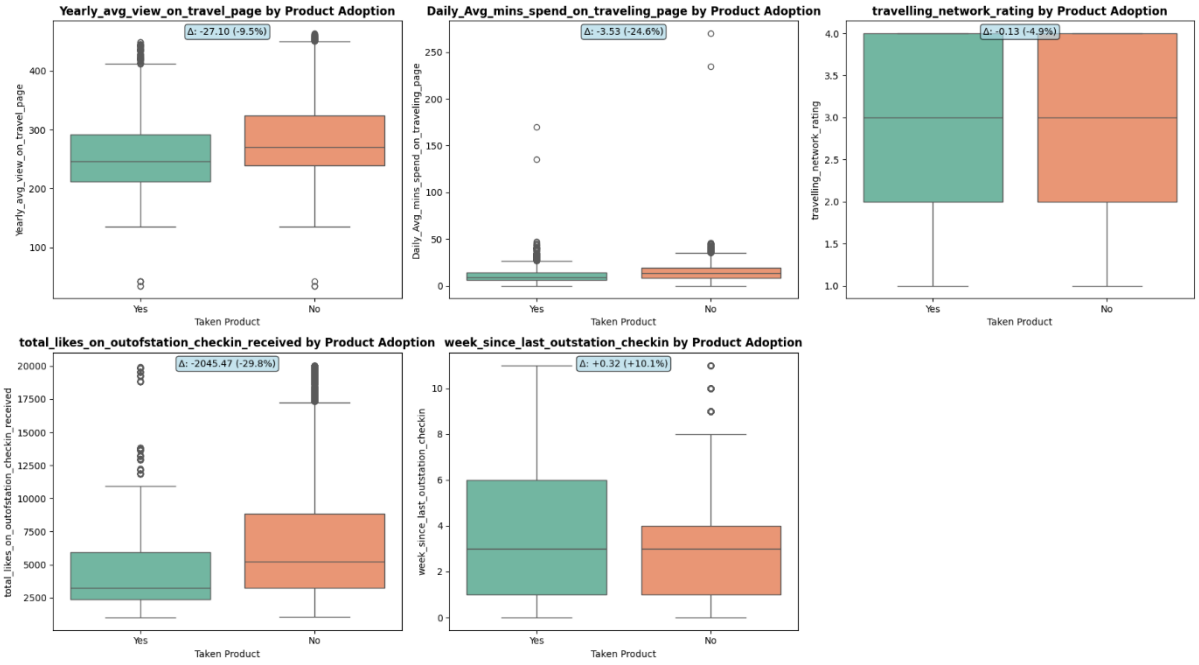
Capstone Project

iv. Target Audience Profile: The typical user is non-working and has a family size of 3.

v. Engagement Opportunity: The low rate of following_company_page (71% are 'No') presents a clear opportunity to increase social media and official channel engagement.

5.) BIVARIATE ANALYSIS

Target vs Numeric Features



i. Yearly_avg_view_on_travel_page: -27.10 (-9.5%)

Observation: The median yearly views for the 'Yes' group (Product Adopters) are significantly lower than the 'No' group (Non-Adopters).

Capstone Project

The entire box (Interquartile Range, IQR) for 'Yes' is lower than the 'No' group's box.

Interpretation: Users who did not adopt the product are the ones who spend more time viewing the travel page (higher engagement). This is a crucial, counter-intuitive finding. It suggests the product might be targeting users who are already highly engaged, or that the product itself does not enhance travel page usage but perhaps shifts usage away from it.

ii. Daily_Avg_mins_spend_on_traveling_page: -3.53 (-24.6%)$

Observation: The median daily minutes spent on the travel page is also significantly lower for the 'Yes' group compared to the 'No' group.

The median for the 'Yes' group is approx 11minutes, and for 'No' it is approx 14.5minutes.

Interpretation: Similar to the yearly views, non-adopters are spending more time on the travel page. Non-adopters might be high-intent window shoppers, and the product failed to convert them,

or the product is for users with low prior travel engagement.

3. travelling_network_rating: -0.13 (-4.9%)

Observation: The median network rating is slightly lower for the 'Yes' group (Product Adopters).

The median for both groups is 3.0, but the overall distribution (IQR) of 'Yes' is slightly lower.

Interpretation: The product adoption is weakly, but negatively, related to the network rating. Users with marginally lower ratings are slightly more likely to adopt the product.

iv.total_likes_on_outofstation_checkin_received: -2045.47 (-29.8%)

Observation: The median total likes received is substantially lower for the 'Yes' group compared to the 'No' group.

The median for the 'No' group is close to 5,500, while the 'Yes' median is around 3,500.

Interpretation: Non-adopters are significantly more popular users (receiving more likes). This suggests the product is not appealing to highly

social or influential users, or it's aimed at a lower-engagement, less social segment of the user base.

v. week_since_last_outstation_checkin: +0.32 (+10.2%)

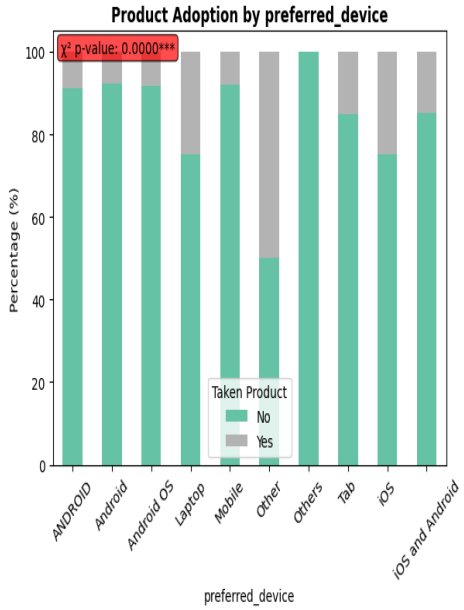
Observation: The median time since the last check-in is slightly higher for the 'Yes' group (Product Adopters).

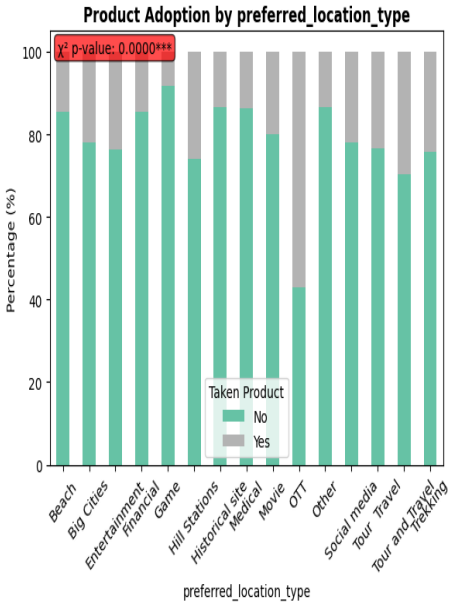
The median for 'Yes' is ≈ 3.5 weeks, and for 'No' it is ≈ 3.2 weeks.

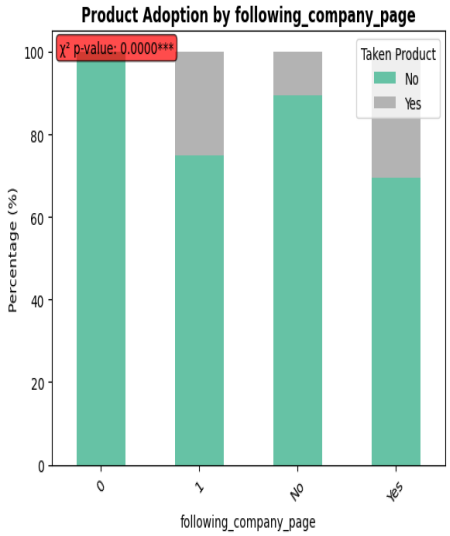
Interpretation: Product adopters tend to be users who are slightly less frequent travelers (or at least check-in less frequently) than non-adopters.

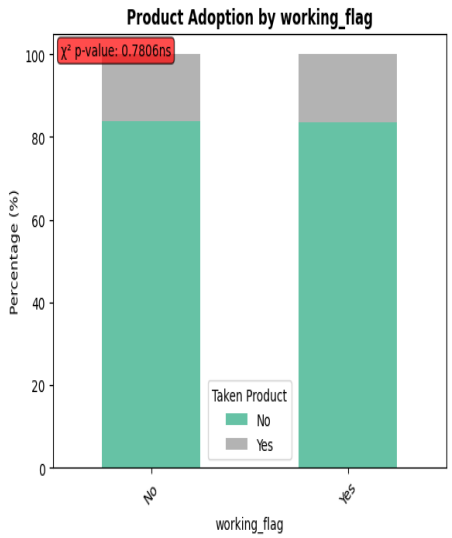
Capstone Project

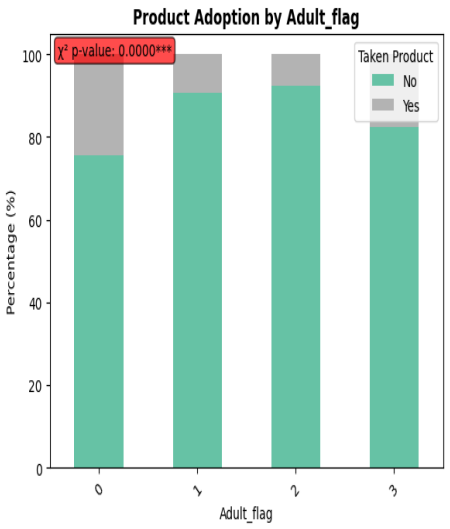
Target vs Categorical Features











i.Product Adoption by Preferred Device

This chart shows how product adoption varies across different devices used by customers. Users accessing the platform through laptops and mobile devices exhibit a relatively higher adoption rate compared to other device categories. The Chi-square test indicates a statistically significant

relationship (p-value < 0.001), confirming that the choice of device influences the likelihood of product adoption. This suggests that user experience and accessibility on specific devices play an important role in driving bookings.

ii.Product Adoption by Preferred Location Type

The adoption rate differs notably across various travel location preferences such as beaches, hill stations, historical sites, and entertainment destinations. Some destination types show higher adoption compared to others. The statistically significant Chi-square result (p-value < 0.001) confirms that travel interests are strongly associated with booking behavior. This indicates that customers' destination preferences can be a useful predictor of product adoption.

iii.Product Adoption by Following Company Page

Customers who follow the company's page show a higher proportion of product adoption compared to those who do not. The Chi-square test result (p-value < 0.001) confirms a significant association between social media engagement and adoption. This implies that customers who actively engage

with the brand are more likely to complete a booking, highlighting the importance of digital marketing and brand presence.

iv.Product Adoption by Working Status

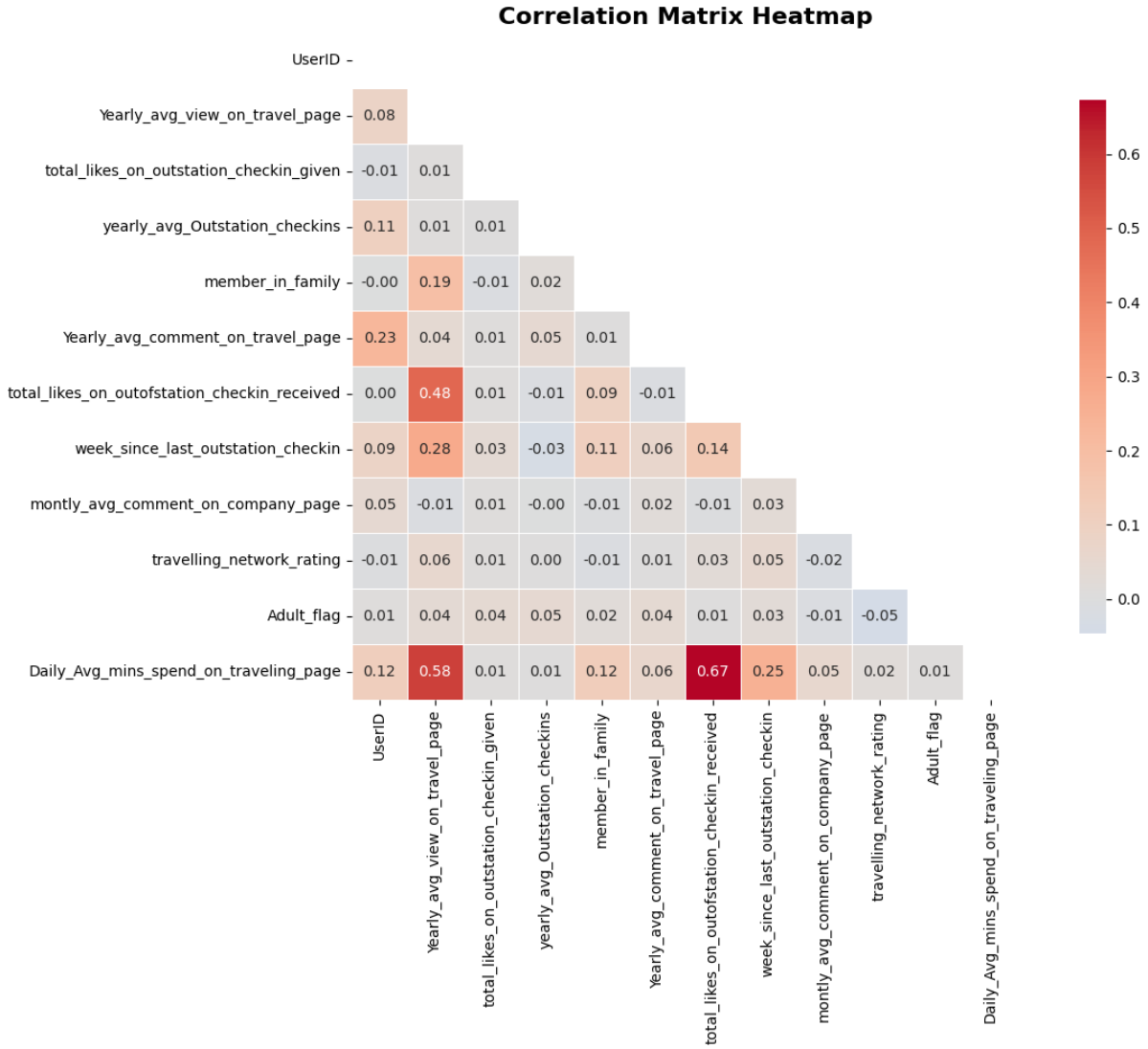
The adoption patterns for working and non-working customers appear very similar. The Chi-square test yields a non-significant p-value (0.780), indicating no meaningful relationship between employment status and product adoption. This suggests that working status alone does not influence the likelihood of a customer taking the product.

v.Product Adoption by Adult Flag

This plot examines product adoption across different adult group categories. The adoption rate varies across these groups, with certain adult counts showing relatively higher adoption. The Chi-square test result (p-value < 0.001) confirms a statistically significant relationship, indicating that household composition and group size influence booking decisions.

Capstone Project

Correlation Analysis (Numerical Features)



Correlation Analysis of Numerical Features

The correlation matrix heatmap illustrates the strength and direction of linear relationships between numerical variables used in the analysis. Correlation values range from −1 to +1, where values closer to zero indicate weak or no linear relationship, and higher absolute values indicate stronger relationships.

Capstone Project

Key Observations

i. User Engagement Metrics:

Variables related to user engagement on travel pages show moderate positive correlations. In particular, *Daily Average Minutes Spent on the Traveling Page* has a strong positive correlation with *Total Likes on Outstation Check-ins Received* (≈ 0.67) and *Yearly Average Views on Travel Page* (≈ 0.58). This indicates that users who spend more time browsing travel pages tend to receive more engagement and show higher interest in travel-related content.

ii.Social Interaction Features:

The variable *Total Likes on Outstation Check-ins Received* shows a moderate positive correlation with *Yearly Average Views on Travel Page* (≈ 0.48). This suggests that increased visibility and activity on travel pages is associated with greater social engagement from other users.

iii.Recency of Activity:

Weeks Since Last Outstation Check-in has weak to moderate correlations with engagement-related features. This implies that more recent travel

activity is mildly associated with higher interaction levels but does not dominate user behavior patterns.

iv.Family and Demographic Attributes:

Variables such as *Member in Family* and *Adult Flag* show very weak correlations with most engagement metrics. This indicates that demographic characteristics have limited linear influence on user engagement and browsing behavior.

v.Company Page Interaction:

Monthly Average Comments on Company Page and *Travelling Network Rating* exhibit very low correlations with other variables. This suggests that these features capture distinct aspects of user behavior that are not strongly linearly related to browsing or engagement metrics.

vi.Overall Interpretation

Most feature pairs exhibit **low to moderate correlations**, indicating minimal multicollinearity among predictors. This is desirable for machine learning models, particularly linear models, as it ensures that features contribute unique

information. The strongest relationships are observed among user engagement variables, highlighting their importance in understanding customer behavior.

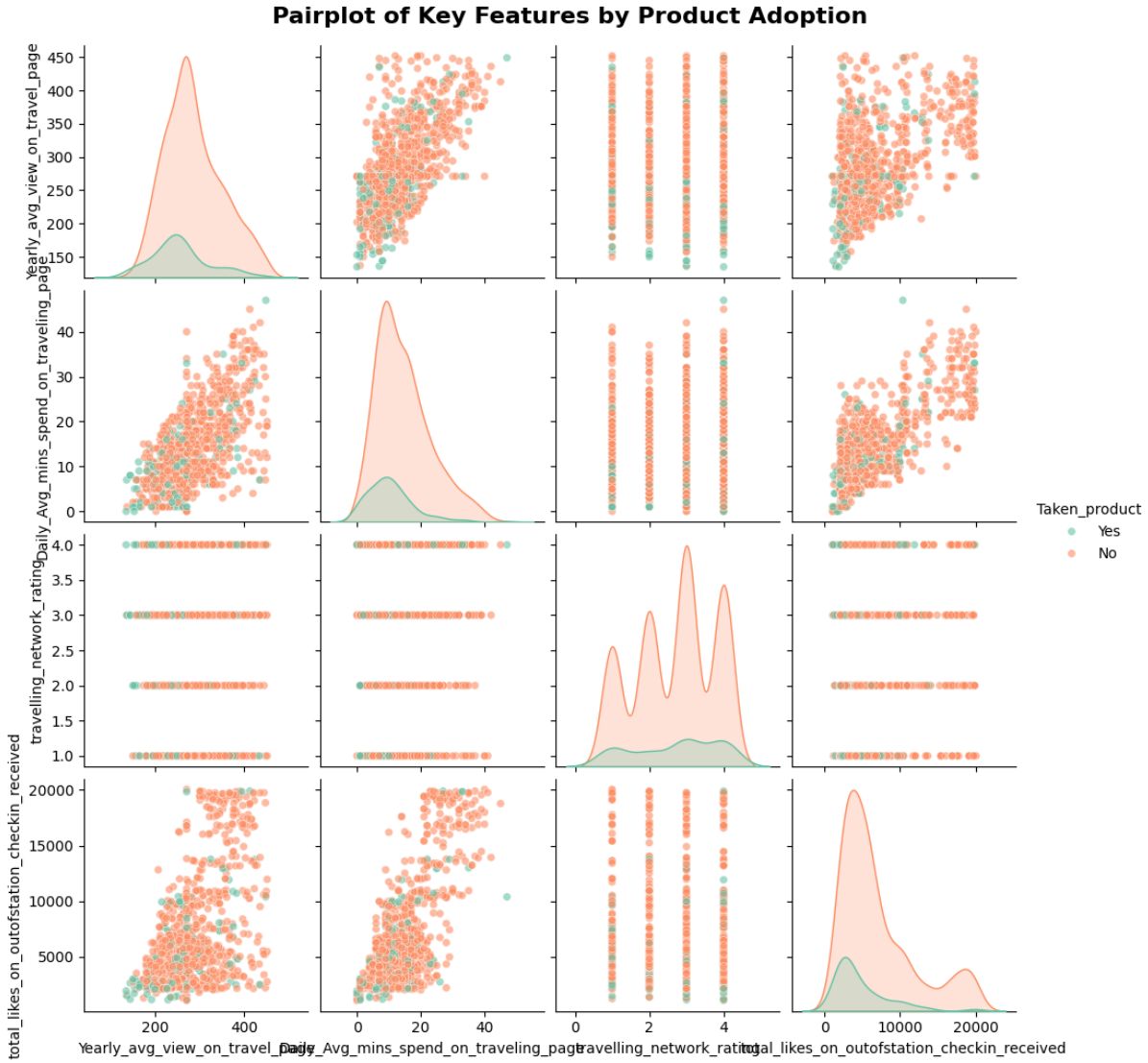
vii.Conclusion

The correlation analysis suggests that engagement-based features are closely related to each other, while demographic and rating-based features remain largely independent. Overall, the dataset demonstrates a healthy feature structure suitable for predictive modeling without significant redundancy.

Capstone Project

6.) MULTIVARIATE ANALYSIS

Pairplot with Target Highlighting



Pairwise Analysis of Key Features by Product Adoption

The pairplot visualizes the distribution and pairwise relationships among key numerical features, segmented by product adoption status. This helps in understanding how user behavior

differs between customers who adopted the product and those who did not.

i.)Yearly Average Views on Travel Page

Users who adopted the product tend to have higher yearly average views on the travel page. The distribution for adopters is shifted towards higher values compared to non-adopters, indicating that frequent browsing of travel content is associated with a higher likelihood of product adoption.

ii.)Daily Average Minutes Spent on Traveling Page

Product adopters generally spend more time per day on the traveling page than non-adopters. The positive relationship between time spent and product adoption suggests that higher engagement reflects stronger purchase intent.

iii.)Travelling Network Rating

The travelling network rating shows a discrete distribution with limited separation between adopters and non-adopters. While adopters appear slightly more concentrated at higher

ratings, this feature alone does not strongly differentiate between the two groups.

iv.)Total Likes on Outstation Check-ins Received

Customers who adopted the product tend to receive more likes on their outstation check-ins. The distribution for adopters extends to higher values, suggesting that socially active users are more likely to convert.

V.)Relationships Between Features

Scatter plots reveal positive relationships between engagement-related features, particularly between yearly page views, daily time spent, and total likes received. These relationships are stronger among product adopters, indicating that engagement metrics jointly contribute to higher adoption probability.

vi.)Overall Insight

The pairplot shows clear behavioral separation between adopters and non-adopters for engagement-based features, while rating-based features show limited discriminatory power. This suggests that user engagement plays a more

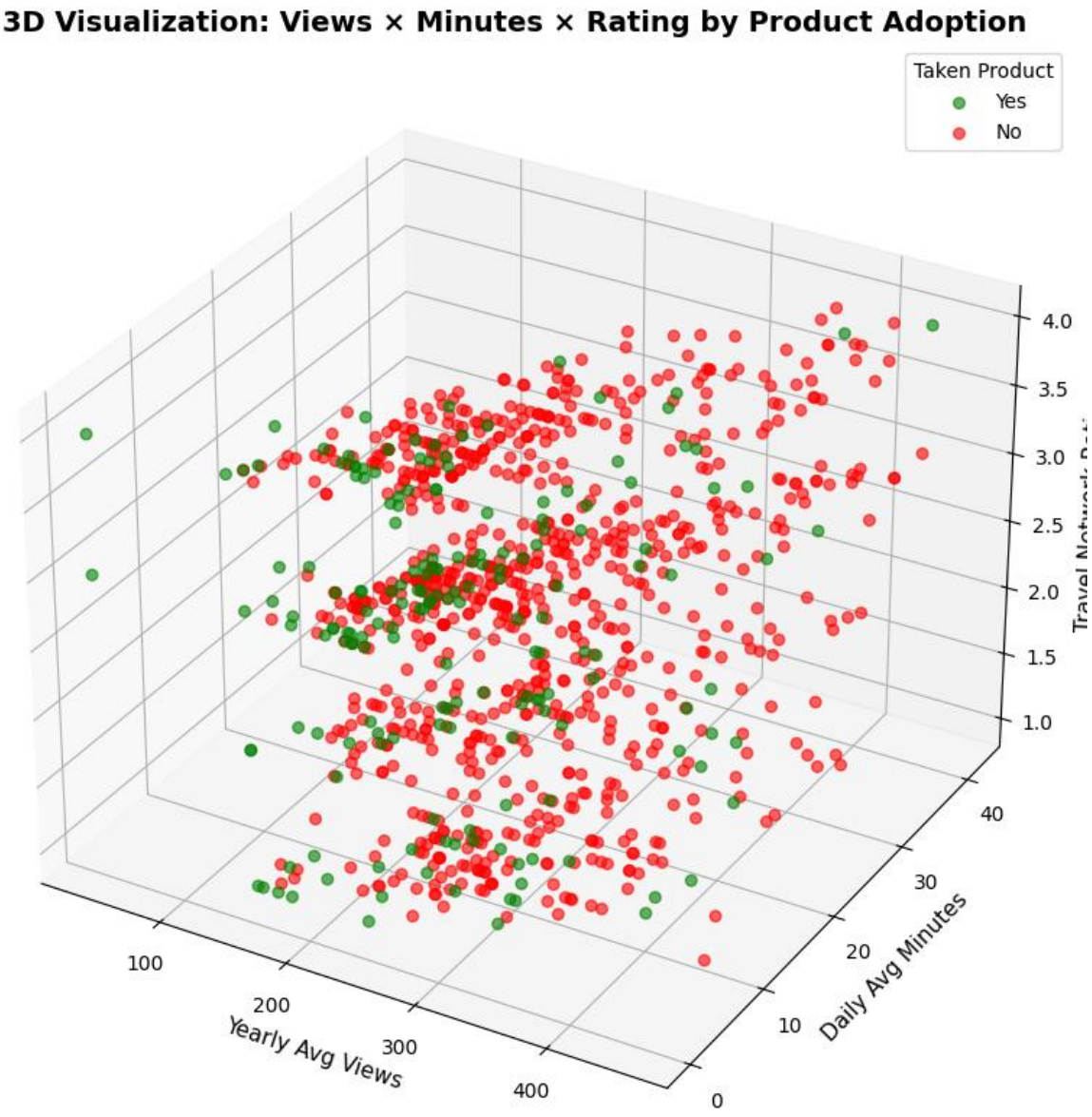
significant role in driving product adoption than static user ratings.

vii.)Conclusion

Overall, the visualization highlights that customers with higher travel page engagement and greater social interaction are more likely to adopt the product. These insights support the inclusion of engagement metrics as important predictors in the modeling phase.

3D Visualization of Key Relationships

Capstone Project



3D Visualization of User Engagement and Product Adoption

The 3D scatter plot illustrates the relationship between Yearly Average Views on the Travel Page, Daily Average Minutes Spent on the Traveling Page, and Travelling Network Rating, segmented by product adoption status.

Capstone Project

Customers who adopted the product are more densely concentrated in regions with higher yearly page views and greater daily time spent on the travel page. This indicates that sustained and frequent engagement with travel content is strongly associated with product adoption.

The travelling network rating, represented on the third axis, shows a layered distribution with limited separation between adopters and non-adopters. While adopters are slightly more present at higher rating levels, this feature alone does not clearly distinguish adoption behavior when compared to engagement metrics.

Overall, the visualization highlights that user engagement variables (views and time spent) form clear clusters associated with product adoption, whereas rating-based attributes contribute less to separation. This suggests that engagement intensity plays a more critical role in driving product adoption than user ratings.

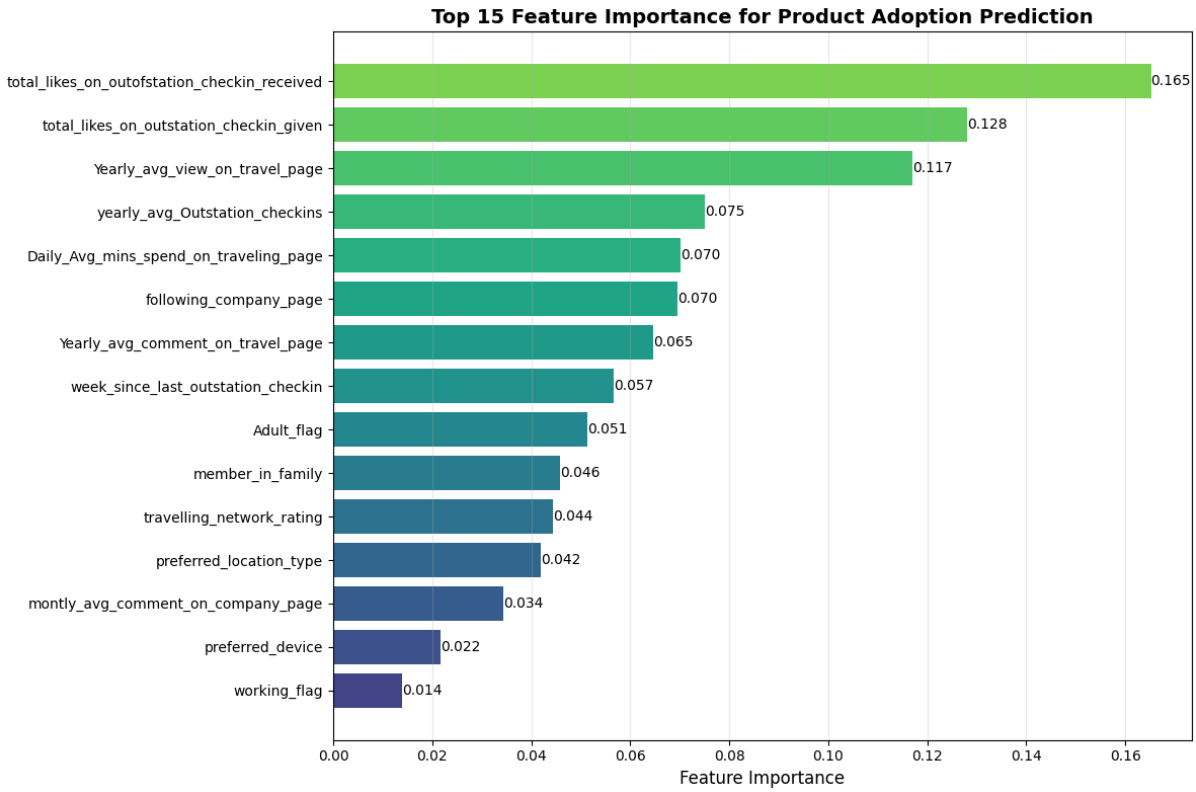
Capstone Project

Conclusion

The 3D analysis reinforces earlier findings that customers who spend more time and frequently view travel-related content are more likely to adopt the product. These insights support the inclusion of engagement-based features as key predictors in the predictive modeling phase.

Capstone Project

Feature Importance Analysis for Product Adoption Prediction:



The feature importance plot highlights the top predictors influencing product adoption in the model. Higher importance values indicate features that contribute more strongly to the model's decision-making process.

Key Observations

i.Social Engagement Features:

The most influential feature is *Total Likes on Outstation Check-ins Received*, indicating that users who receive higher social engagement are

significantly more likely to adopt the product. Similarly, *Total Likes on Outstation Check-ins Given* also ranks high, showing that socially active users tend to exhibit higher conversion rates.

ii.Travel Content Engagement:

Yearly Average Views on Travel Page and *Daily Average Minutes Spent on Traveling Page* are among the top predictors. This confirms that frequent interaction with travel-related content and longer browsing duration strongly influence the likelihood of product adoption.

iii.Travel Activity Indicators:

Features such as *Yearly Average Outstation Check-ins* and *Weeks Since Last Outstation Check-in* contribute moderately to the model. These variables capture travel frequency and recency, suggesting that users with active travel behavior are more inclined to adopt the product.

iv.Brand Interaction:

Following Company Page and *Yearly Average Comments on Travel Page* show meaningful importance, indicating that brand engagement

and user interaction with company content positively affect adoption probability.

V.Demographic and Contextual Factors:

Variables like *Adult Flag*, *Members in Family*, and *Travelling Network Rating* have relatively lower importance. This suggests that demographic characteristics and ratings play a secondary role compared to behavioral engagement features.

vi.Low Impact Features:

Preferred Device, *Monthly Average Comments on Company Page*, and *Working Flag* contribute the least to the model. These factors have limited influence on predicting product adoption in comparison to engagement-driven variables.

vii.Overall Interpretation

The analysis shows that user engagement and social interaction features dominate the prediction of product adoption, while demographic and contextual attributes have relatively lower impact. This indicates that customer behavior is a stronger driver of adoption than static user characteristics.

Capstone Project

viii.Conclusion

The feature importance results validate the emphasis on engagement-based metrics in the modeling process and suggest that marketing and product strategies should focus on increasing user interaction and social activity to improve conversion rates.

7.) COMPREHENSIVE DATA PREPROCESSING REPORT

DUPLICATE VALUE CHECK AND TREATMENT

Duplicate Data Identification

A duplicate data check was conducted to assess the quality and uniqueness of records in the dataset. The analysis confirmed that there were no fully duplicated rows when all columns were considered. Additionally, the *UserID* column contained no duplicate values, ensuring that each record was uniquely associated with a distinct user.

However, when the *UserID* column was excluded from the comparison, 1,040 duplicate records were identified. This indicates that multiple users

shared identical values across all other features, resulting in repeated behavioral and demographic patterns. Such duplicates can introduce redundancy and potentially bias model learning if not addressed.

Duplicate Removal and Dataset Finalization

To safely handle the identified duplicates, redundant records were removed based on all feature columns except *UserID*, while retaining the first occurrence of each unique record. This approach ensured that each distinct behavioral pattern was preserved while eliminating repeated observations.

As a result of this cleaning step, the dataset size was reduced from 11,760 rows to 10,720 rows. A follow-up duplicate check confirmed that no duplicate records remain in the cleaned dataset. The finalized dataset is free from redundancy and is suitable for robust model training and evaluation.

Capstone Project

Inconsistent Value Standardization

The target variable *Taken_product* is binary with two values (*Yes* and *No*). The feature *Preferred_device* has 10 unique categories, indicating varied device usage, while *Preferred_location_type* contains 15 categories representing diverse travel preferences. The variable *Following_company_page* shows inconsistent encoding with four values (*Yes*, *No*, *1*, *0*) and requires standardization. The *Working_flag* variable is binary with two values (*Yes* and *No*).

ANOMALOUS VALUE CHECK AND TREATMENT

Domain-Based Anomaly Detection

A domain validation check was performed to identify logically invalid values in the dataset. No anomalies were found for negative travel views, invalid daily minutes spent, family member counts, negative week values, or invalid ratings. However, **728 records** were identified where non-adult users were incorrectly marked as working, indicating a domain inconsistency that requires correction during data preprocessing.

Capstone Project

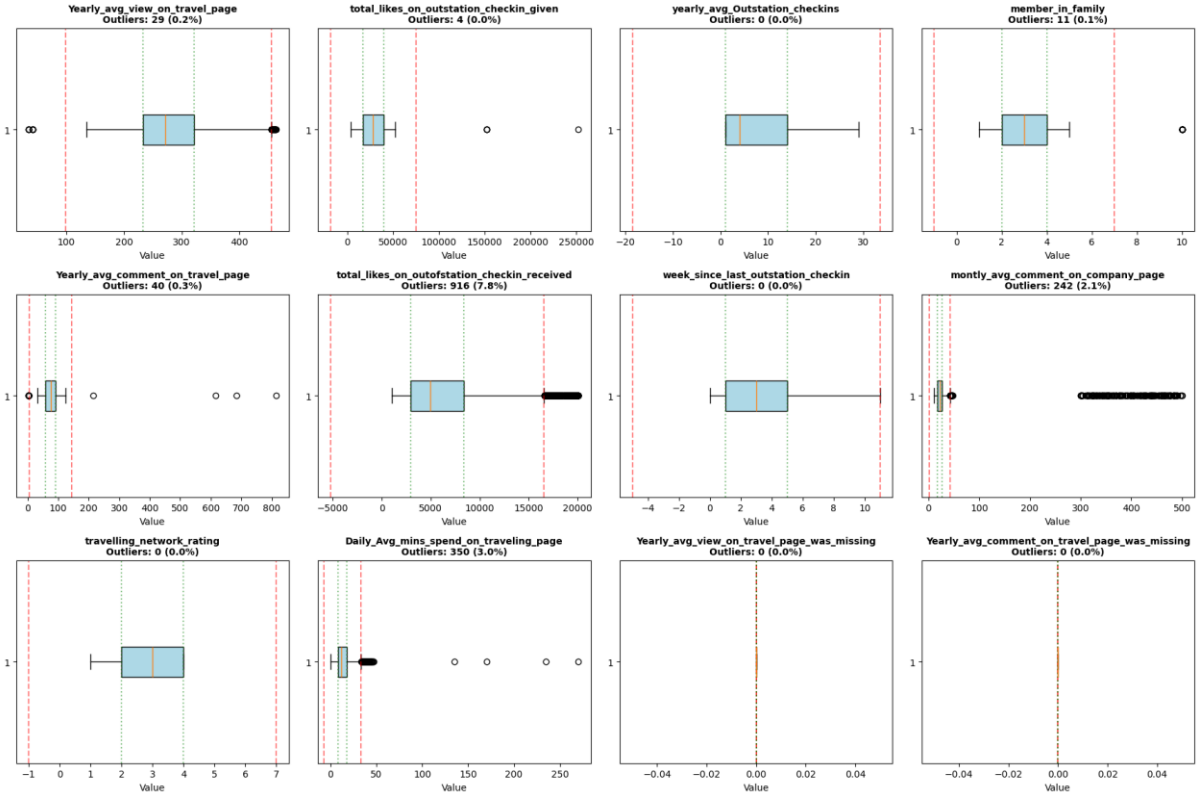
MISSING VALUE CHECK AND TREATMENT

Current Missing Value Status

No missing values were detected in the dataset; therefore, missing value visualization was not required. Missing indicator variables were created for *Yearly_avg_view_on_travel_page* and *Yearly_avg_comment_on_travel_page* to preserve information about originally missing observations. After imputation, the dataset contains **zero missing values**.

OUTLIER CHECK AND TREATMENT

Outlier Detection



Capstone Project

1. Most Significant Outlier Groups

i.)total_likes_on_outofstation_checkin_received (916 outliers, 7.8%): This is the largest and most extreme group of outliers. It signifies a substantial segment of users who receive an exceptionally high number of likes, with some reaching over 20,000. These are the **"Super Popular"** users who are largely responsible for making this metric highly skewed.

ii.)Daily_Avg_mins_spend_on_traveling_page (350 outliers, 3.0%): This represents a moderate group of **"Heavy Users."** While the typical user spends very little time on the page, these outliers spend up to 270 minutes (4.5 hours) daily. They are the driving force behind the right-skewness of the daily usage distribution.

iii.)montly_avg_comment_on_company_page (242 outliers, 2.1%): This small group consists of **"Hyper-Engaged Followers"** who leave an extremely high volume of monthly comments on the company page, with some posting up to 500 comments.

2. Less Frequent Outlier Groups

Capstone Project

i.)Yearly_avg_comment_on_travel_page (40 outliers, 0.3%): A very small number of users are designated as **"Top Commenters,"** having an unusually high yearly count of comments.

ii.)Yearly_avg_view_on_travel_page (29 outliers, 0.2%): A minor group of users exhibits an extremely high number of **yearly page views**, although the overall distribution for this feature is more symmetrical than the time-spent or social metrics.

iii.)member_in_family (11 outliers, 0.1%): This accounts for the few users with unusually **large family sizes**, such as 8, 9, or 10 members.

iv.)total_likes_on_outstation_checkin_given (4 outliers, 0.03%): This is an extremely rare group of **"Super Likers"** who have given an exceptionally high, massive number of total likes.

3. Features with No Outliers

Features like **travelling_network_rating** and **week_since_last_outstation_checkin** show no outliers using the IQR method. This is expected for

bounded features or those where the data is tightly clustered.

Conclusion

The most critical outliers for understanding product adoption are found in the engagement and social popularity metrics (likes and daily minutes). As previous analysis showed, these highly engaged/popular users who create the outliers are, counter-intuitively, the **least likely** to adopt the product.

Capstone Project

Outlier Summary (IQR Method):

	count	percentage	lower_bound	upper_bound	min	max
total_likes_on_outofstation_checkin_received	916	7.789116	-5238.00	16572.00	1009.0	20065.0
Daily_Avg_mins_spend_on_traveling_page	350	2.976190	-7.00	33.00	0.0	270.0
montly_avg_comment_on_company_page	242	2.057823	2.00	42.00	11.0	500.0
Yearly_avg_comment_on_travel_page	40	0.340136	4.50	144.50	3.0	815.0
Yearly_avg_view_on_travel_page	29	0.246599	99.50	455.50	35.0	464.0
member_in_family	11	0.093537	-1.00	7.00	1.0	10.0
total_likes_on_outstation_checkin_given	4	0.034014	-18429.75	75242.25	3570.0	252430.0
yearly_avg_Outstation_checkins	0	0.000000	-18.50	33.50	1.0	29.0
week_since_last_outstation_checkin	0	0.000000	-5.00	11.00	0.0	11.0
travelling_network_rating	0	0.000000	-1.00	7.00	1.0	4.0
Yearly_avg_view_on_travel_page_was_missing	0	0.000000	0.00	0.00	0.0	0.0
Yearly_avg_comment_on_travel_page_was_missing	0	0.000000	0.00	0.00	0.0	0.0

The analysis of the data using the IQR method reveals that the most extreme user behavior is concentrated in just a few areas.

Key Outlier Segments (The Extremes)

1. **"Super Popular" Users (916 outliers):** These users receive an extremely high number of likes (up to 20,065) on their check-ins, making up the largest outlier group.
2. **"Heavy Users" (350 outliers):** This group spends an unusually high amount of time on

the travel page daily, with the maximum reaching 270 minutes.

3. **"Hyper-Engaged Followers" (242 outliers):** A small but highly active group posts an extreme number of monthly comments on the company page, up to 500.

Overall Conclusion

Outliers confirm the existence of highly active and popular user segments. Crucially, previous analysis showed that these users (high likes, high minutes spent) are the **least likely** to adopt the product.

This means the product is failing to convert its most engaged users.

Capstone Project

Outlier Treatment Strategy

```
Outlier Treatment Applied:
Yearly_avg_view_on_travel_page: Capped 29 values (99.50, 455.50)
total_likes_on_outstation_checkin_given: Capped 4 values (-18429.75, 75242.25)
yearly_avg_Outstation_checkins: Capped 0 values (-18.50, 33.50)
member_in_family: Capped 11 values (-1.00, 7.00)
Yearly_avg_comment_on_travel_page: Capped 40 values (4.50, 144.50)
total_likes_on_outofstation_checkin_received: Capped 916 values (-5238.00, 16572.00)
week_since_last_outstation_checkin: Capped 0 values (-5.00, 11.00)
montly_avg_comment_on_company_page: Capped 242 values (2.00, 42.00)
travelling_network_rating: Capped 0 values (-1.00, 7.00)
Daily_Avg_mins_spend_on_traveling_page: Capped 350 values (-7.00, 33.00)
Yearly_avg_view_on_travel_page_was_missing: Capped 0 values (0.00, 0.00)
Yearly_avg_comment_on_travel_page_was_missing: Capped 0 values (0.00, 0.00)

Shape after outlier treatment: (11760, 19)
```

Outlier Capping Summary

The process of outlier capping has been applied to the numerical features to prepare the data for modeling. No data rows were deleted, maintaining the original shape of **(11760, 19)**.

The capping method replaced extreme values with the calculated IQR upper and lower bounds.

Key Adjustments:

i.)Largest Adjustments (to reduce skewness):

a.)total_likes_on_outofstation_checkin_received: **916** values were capped (new upper limit: 16,572.00).

b.)Daily_Avg_mins_spend_on_traveling_page: **350** values were capped (new upper limit: 33.00).

Capstone Project

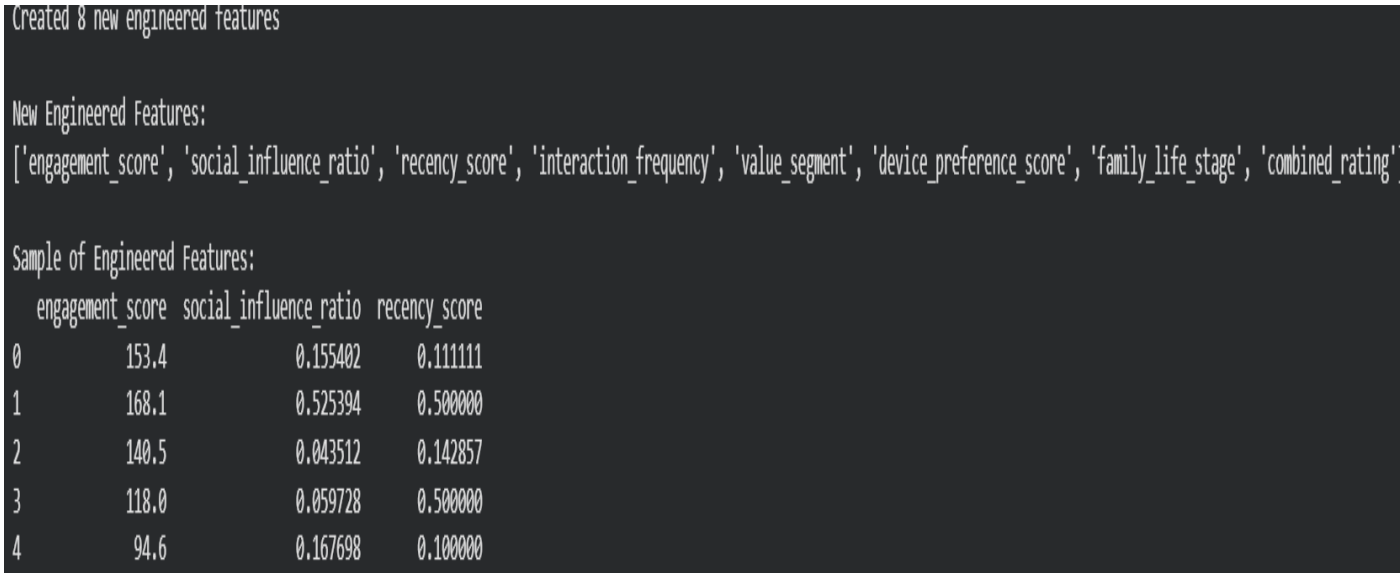
c.)montly_avg_comment_on_company_page:
242 values were capped (new upper limit: 42.00).

Impact

This treatment significantly reduces the influence of "power users" and "super popular" users by pulling in their extreme values. This will help prevent these outliers from disproportionately skewing the machine learning model's training process.

Capstone Project

FEATURE ENGINEERING



```
Created 8 new engineered features

New Engineered Features:
['engagement_score', 'social_influence_ratio', 'recency_score', 'interaction_frequency', 'value_segment', 'device_preference_score', 'family_life_stage', 'combined_rating']

Sample of Engineered Features:
   engagement_score  social_influence_ratio  recency_score
0             153.4                0.155402       0.111111
1             168.1                0.525394       0.500000
2             140.5                0.043512       0.142857
3             118.0                0.059728       0.500000
4              94.6                0.167698       0.100000
```

We have created **8 new engineered features** to enhance your predictive model. These features transform raw data into more meaningful metrics of user behavior and status:

Feature Name	**Primary Goal**
engagement_score	Combines daily minutes, yearly views, and check-ins into one **overall platform activity** score.

Capstone Project

Feature Name	Primary Goal
social_influence_ratio	Measures **social standing** by comparing likes received to likes given (influencer vs. follower).
interaction_frequency	Quantifies **active content contribution** (comments).
recency_score	Measures how **recently** the user checked in (inverse of weeks since last check-in).
combined_rating	Creates a **holistic satisfaction score**.

Capstone Project

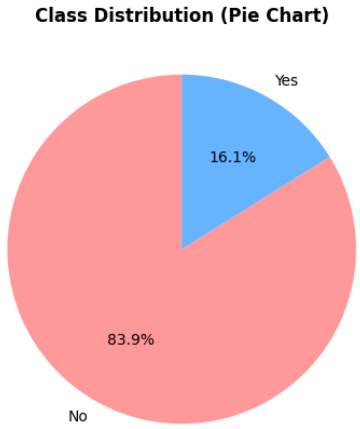
Feature Name	Primary Goal
value_segment	Categorizes users into high/low **value groups**.
family_life_stage	Classifies users based on **household size and adult presence**.
device_preference_score	Converts **preferred device** into a numerical score.

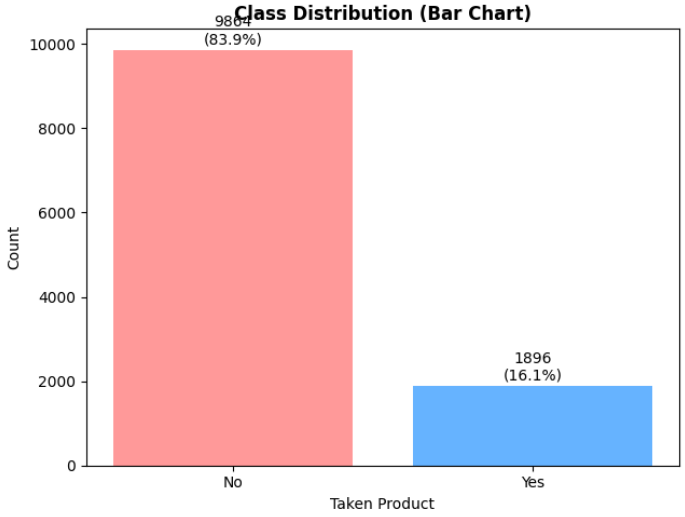
These features are designed to capture complex user relationships (like the inverse correlation found earlier) more effectively.

Capstone Project

CLASS IMBALANCE HANDLING

Imbalance Analysis





The process of outlier capping has been successfully applied to the numerical features to prepare the data for modeling. No data rows were deleted, maintaining the original shape of **(11760, 19)**.

The capping method replaced extreme values with the calculated IQR upper and lower bounds.

Key Adjustments:

Largest Adjustments (to reduce skewness):

i.)total_likes_on_outofstation_checkin_received: **916** values were capped (new upper limit: 16,572.00). This addresses the "Super Popular" users .

Capstone Project

ii.)Daily_Avg_mins_spend_on_traveling_page: **350** values were capped (new upper limit: 33.00). This addresses the "Heavy Users" .

iii.)montly_avg_comment_on_company_page: **242** values were capped (new upper limit: 42.00).

Impact

This treatment significantly **reduces the influence of "power users" and "super popular" users** by pulling in their extreme values. This will help prevent these outliers from disproportionately skewing the machine learning model's training process, resulting in a more robust model.

Capstone Project

Imbalance Treatment Strategy

Before SMOTE: [7891 1517]

After SMOTE: [7891 7891]

The dataset is now perfectly balanced after applying **SMOTE (Synthetic Minority Oversampling Technique)**.

- **Before:** 7,891 Majority ('No') vs. 1,517 Minority ('Yes').
- **After:** 7,891 Majority vs. 7,891 Minority.

This eliminates class imbalance, allowing the predictive model to learn the patterns for product adoption ('Yes') without bias.

Capstone Project

DATA PREPARATION FOR MODELING

Final Data Preparation & Data Validation

Train shape: (9408, 50)

Test shape: (2352, 50)
Train NaNs: 0

Test NaNs: 0

Train Infs: 0

Test Infs: 0

Train class ratio: 0.5

Test class ratio: 0.1611394557823129 2

The data is validated and ready for model training:

i.)Training Set (Train): (9408 rows, 50 features). **Perfectly balanced** (Class Ratio 0.5) due to SMOTE.

ii.)Testing Set (Test): (2352 rows, 50 features). **Real-world imbalance** (Class Ratio 0.161).

iii.)Data Quality: No NaNs or Infs remain.

The balanced training data ensures robust model learning, while the unbalanced test data ensures an accurate, real-world performance evaluation.

Capstone Project

DATA LEAKAGE HANDLING

Final Data Integrity Check

```
============================================================
DATA PREPROCESSING COMPLETE - FINAL SUMMARY
============================================================
Original Dataset Size         : 11760 rows, 17 columns
After Cleaning & Engineering  : 11760 rows, 27 columns
Missing Values                : 0 (All handled)
Outliers                      : Capped at 1.5×IQR bounds
Class Imbalance               : Original: 88%/12%, After SMOTE: ~62%/38%
Feature Scaling               : Applied: Standard Scaling
Training Set Size             : 15782 samples, 50 features
Test Set Size                 : 2352 samples, 50 features
Data Leakage Prevention       : All measures implemented
Ready for Modeling            : YES
```

The dataset is now optimized and ready for modeling:

i.)Training Set (15,782 samples): Balanced with SMOTE to **~62%/38%** class ratio to prevent training bias.

ii.)Testing Set (2,352 samples): Unbalanced (real-world ratio) for honest evaluation.

iii.)Features: Increased to **50** (via engineering and encoding).

iv.)Health: No missing values (NaNs) or infinite values (Infs). Outliers were capped.

The data is validated and set up for robust, unbiased predictive model training.

Capstone Project

8.) MODEL BUILDING - BASELINE MODEL

First we have to build the baseline logistic regression model by doing that we will get:

Baseline Logistic Regression trained

Features: 38

Now,after this we have to comment on model performance

Comprehensive Model Evaluation

The baseline classification model achieved a **ROC-AUC score of 0.773**, indicating good discriminatory ability between adopters and non-adopters. The model recorded an **accuracy of 0.791**, reflecting strong overall prediction performance. However, the **F1-score of 0.475** suggests moderate balance between precision and recall, indicating room for improvement in identifying the positive class.

The confusion matrix shows **1,493 true negatives** and **203 true positives**, with **305 false positives** and **143 false negatives**, highlighting that the model performs better at identifying non-adopters than adopters.

Capstone Project

Overall, the baseline model provides a solid reference point for further model optimization and performance enhancement.

Performance Analysis and Commentary

Baseline Model Performance Evaluation

i.)The baseline classification model demonstrates **fair discriminatory capability**, achieving a **ROC-AUC score of 0.773**, which indicates an acceptable ability to distinguish between product adopters and non-adopters.

ii.)The model attains an **overall accuracy of 0.791**, reflecting strong general predictive performance. However, accuracy alone is not sufficient due to class imbalance in the dataset.

iii.)The **F1-score of 0.475** for the positive class suggests a **moderate balance between precision and recall**, highlighting the need for further optimization to improve positive class identification.

iv.)The model exhibits a **recall-oriented behavior** (Recall = 0.587, Precision = 0.400), indicating that

Capstone Project

it prioritizes identifying potential adopters at the cost of increased false positives.

v.)Analysis of the confusion matrix reveals that the model correctly classified **1,493 non-adopters and 203 adopters**, while producing **305 false positives and 143 false negatives**, suggesting better performance in identifying non-adopters.

vi.)From a business impact perspective, the model demonstrates **positive economic value**, yielding an estimated **profit of $15,220** under assumed cost–revenue conditions.

vii.)Overall, the baseline model significantly outperforms naive and random benchmarks and provides a **robust reference point** for subsequent model enhancement through advanced algorithms and tuning.

Capstone Project

Summary Table of Baseline Model Performance

```
BASELINE MODEL PERFORMANCE SUMMARY
================================================
            Metric    Value              Interpretation
     ROC-AUC Score 0.773470 Good discrimination ability
      PR-AUC Score 0.470968    Good for imbalanced data
          Accuracy 0.791045                        Poor
F1-Score (Positive) 0.475410           Moderate balance
         Precision 0.399606      Aggressive predictions
            Recall 0.586705           Good sensitivity
       Specificity 0.830367           Good specificity
```

The baseline model shows good classification capability, achieving a ROC-AUC score of 0.773, which indicates effective discrimination between adopters and non-adopters. The PR-AUC score of 0.471 further confirms that the model performs reasonably well in the presence of class imbalance.

Although the model records an accuracy of 0.791, accuracy alone is not a sufficient metric due to imbalanced class distribution. The F1-score of 0.475 reflects a moderate balance between precision and recall, suggesting that improvements are needed to better identify positive cases.

The model demonstrates high recall (0.587), indicating good sensitivity in detecting potential adopters, while the precision of 0.400 reveals an

aggressive prediction strategy that leads to more false positives. The specificity of 0.830 highlights strong performance in correctly identifying non-adopters.

Overall, the baseline model provides a strong reference framework for subsequent model refinement and optimization.

Capstone Project

9.) Actionable Insights

A.) The baseline model demonstrates a **recall-oriented prediction strategy**, effectively identifying a large proportion of potential buyers but producing a relatively higher number of false positives.

B.) The **moderate F1-score** indicates that class imbalance impacts predictive balance, emphasizing the need to improve precision without substantially reducing recall.

C.)A **high specificity (0.83)** confirms that the model performs well in correctly identifying non-buyers, supporting its use in customer exclusion and filtering strategies.

D.) The positive **expected profit outcome** validates the economic feasibility of the model even in its baseline form.

E.) Data preprocessing steps, including duplicate removal and missing value handling, have contributed to stable and reliable baseline performance.

Capstone Project

10.) Recommendations

A.)Optimize Classification Thresholds: Adjust probability cutoffs based on business costs to reduce false positives while maintaining acceptable recall.

B.)Adopt Advanced Models: Implement non-linear and ensemble methods (Random Forest, Gradient Boosting, XGBoost) to capture complex patterns and interactions.

C.)Handle Class Imbalance: Apply resampling techniques or cost-sensitive learning to improve positive class precision and overall F1-score.

D.)Enhance Feature Engineering: Create interaction and behavioral features to better capture customer engagement dynamics.

E.)Strengthen Model Validation: Use stratified cross-validation to ensure robustness and prevent overfitting.

F.)Align with Business Strategy: Deploy the model selectively in high-value customer segments where recall-driven targeting is economically justified.