December 2023

Impact of Marketing Techniques on Sales Performance

TeamName: Fiinall2.txt

1- Bishoy Wadea Fathy
2- Eslam Wageh Samer
3- Peter Safwat Sahyon
4- Sara Ahdy Kamal
5- Ashraf Bahy Galal

Table Of Contents

Introduction

In a dynamic era shaped by evolving consumer behavior and changing market landscapes, the success of businesses hinges significantly on the efficacy of marketing strategies. This research addresses gaps in existing literature, aiming to deepen our understanding of the intricate relationship between marketing techniques and sales performance. The study focuses on three dimensions: the comparative effectiveness of radio, TV, and social media strategies; the impact of Instagram interactions (ads vs. hashtags) on sales; and the correlation between user demographics and marketing outcomes. By exploring these dimensions, we aim to offer practical insights for businesses to refine their marketing approaches and enhance sales outcomes. As the marketing landscape evolves, this research, employing a robust methodology, seeks to contribute valuable insights into the dynamic interplay between marketing techniques and sales. Subsequent sections will cover the literature review, theoretical framework, methodology, findings, and implications for businesses, providing a comprehensive understanding of this complex relationship.

ABSTRACT

This study investigates the nuanced connection between marketing techniques and sales, addressing gaps in current literature. Focused on the comparative effectiveness of radio, TV, and social media, the impact of Instagram interactions, and the correlation with user demographics, the research employs a comprehensive methodology.

The literature review encompasses existing research on marketing techniques, Instagram strategies, and demographic influences. The study contributes insights into optimal marketing channels, the effectiveness of Instagram approaches, and demographic considerations.

Rooted in quantitative and qualitative data analysis, the findings provide practical implications for businesses. This research enhances our understanding of marketing dynamics and offers a foundation for future investigations.

Problem description

Our primary objective is to conduct an examination and quantification of the influence of various marketing techniques on the sales performance of newly established businesses. Through this study, we aim to dive deep into the data to unearth actionable insights. These insights will be invaluable for businesses, serving as a practical guide to help them refine and optimize their marketing strategies.

Statistical Questions

1- Do businesses that invest more in modern marketing strategies experience higher sales growth compared to those relying on traditional marketing methods?

2. What is the overall impact of different marketing techniques (TV, social media, radio) on sales?
Can we rank them based on their effectiveness?

3. Is the status of the intended audience a significant factor to consider when it comes to advertising and marketing campaigns?

4. Given the best marketing technique, what attributes should you work on to get the highest profit?

Methodologies

Our research methodology involved a comprehensive process to analyze the impact of marketing techniques on sales, aiming for both depth and precision. Here is a detailed breakdown of our approach:

1-Data Collection:

Gathering datasets pertaining to various marketing techniques and their corresponding effects on sales. The datasets encompassed diverse industries, allowing for a broad understanding of the impact of marketing strategies across different sectors.

2-Data Analysis:

Using different plots for our data (e.g. Scatter plots, box plots, bar plots and linear plots) and calculating different numerical values that show the effect of each feature on the output (e.g. correlation, mean , median, standard deviation).

3-Identification of Most Effective Technique:

Determining the most impactful marketing technique through data analysis , we delve deeper into its attributes by collecting additional datasets. Statistical analyses were applied to identify key attributes, and a regression model was developed to predict ad impressions based on the quantity of each identified attribute.

4-Regression Modeling:

Constructing linear and polynomial models allowing the prediction of product sales based on investment in different marketing techniques and also allowing us to predict the impressions based on positioning and interactions that affect the impressions.

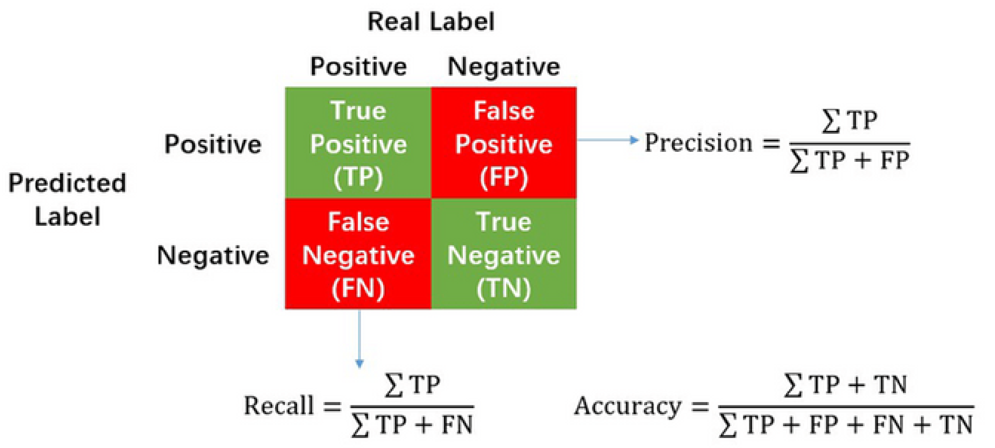
Linear regression models the relationship between the dependent variable and one or more independent variables as a linear equation.
Polynomial regression is a form of linear regression where the relationship is modeled as an nth-degree polynomial.

5-User Statistics Analysis:

Expanding our investigation to include user statistics such as gender, age, and salary. Datasets were collected to establish correlations between user characteristics and the likelihood of making a purchase based on the specific marketing technique employed.

6-Classification Models:

To enhance predictive capabilities, Studying a variety of classification models and implementing them, including logistic regression, Naïve Bayes, KNN, SVM, Decision Trees, and Random Forests. The selection of the most suitable model was based on rigorous evaluation metrics such as confusion matrix, accuracy, precision, and recall and f1-score.



The F_1 Score is given by :

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Logistic Regression:

Binary classification algorithm that models the probability of an instance belonging to a particular class.
Utilizes the logistic function for mapping input features to probabilities.
Effective for linearly separable data.

K-Nearest Neighbors (KNN):

Simple algorithm for classification and regression tasks.
Assigns a class label based on the majority class among its k-nearest neighbors.
Sensitive to outliers and requires careful selection of k.

Support Vector Machine (SVM):

Effective for both linear and non-linear classification.
Finds a hyperplane that maximally separates classes in feature space.
Handles high-dimensional data well.

Decision Trees:

Hierarchical tree structure for decision-making.
Splits data based on features to create a tree-like model.
Prone to overfitting but can be mitigated with techniques like pruning.

Random Forests:

Ensemble learning method using multiple decision trees.
Reduces overfitting by averaging predictions from different trees.
Robust and suitable for high-dimensional data

Accuracy, Precision, Recall, and F1 Score are metrics commonly used to evaluate the performance of classification models.

Accuracy:
Definition: Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances.
Interpretation: High accuracy indicates a good overall performance, but it may not be suitable for imbalanced datasets where one class dominates.

Precision:
Definition: Precision is the ratio of correctly predicted positive instances to the total instances predicted as positive.
Interpretation: Precision focuses on the accuracy of positive predictions. It is useful when the cost of false positives is high.

Recall (**Sensitivity or True Positive Rate**):
Definition: Recall is the ratio of correctly predicted positive instances to the total actual positive instances.
Interpretation: Recall measures the ability of the model to capture all positive instances. It is important when the cost of false negatives is high.

F1 Score:
Definition: F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics.
Interpretation: F1 Score is useful when there is a need to consider both false positives and false negatives. It is especially valuable in imbalanced datasets.

In summary:
Accuracy: Overall correctness of predictions.
Precision: Accuracy of positive predictions, relevant when false positives are costly.
Recall: Ability to capture all positive instances, relevant when false negatives are costly.
F1 Score: Harmonic mean of precision and recall, providing a balance between the two metrics.

7-Model Refinement:

Continuous refinement of our models was carried out, incorporating feedback loops to ensure optimal performance and relevance. This iterative process allowed us to adapt to evolving market dynamics and changes in consumer behavior.
In summary, our methodology was designed to provide a comprehensive understanding of the relationship between marketing techniques, sales outcomes, and user characteristics, employing advanced statistical analyses and predictive modeling for actionable insights.

Numerical Results & Discussion

Note: You can find these graphs and tables in our code attached in attachments section.

	Social Media	TV	Radio	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9
...	...	...	...	...
195	38.2	3.7	13.8	7.6
196	94.2	4.9	8.1	9.7
197	177.0	9.3	6.4	12.8
198	283.6	42.0	66.2	25.5
199	232.1	8.6	8.7	13.4

200 rows × 4 columns

Figure 0: Rows from our first dataset

Description: This dataset contains the relation between total sales for some products and the investment in TV, Social Media, and Radio ADs.

Some numerical statistics from our data:

	Social Media	TV	Radio	Sales
count	200.000000	200.000000	200.000000	200.000000
mean	147.042500	23.264000	30.554000	14.022500
std	85.854236	14.846809	21.778621	5.217457
min	0.700000	0.000000	0.300000	1.600000
25%	74.375000	9.975000	12.750000	10.375000
50%	149.750000	22.900000	25.750000	12.900000
75%	218.825000	36.525000	45.100000	17.400000
max	296.400000	49.600000	114.000000	27.000000

Figure 1: numerical statistics from our data

Numerical Results & Discussion

Note: You can find these graphs and tables in our code attached in attachments section.

By analyzig our first dataset here are our results:

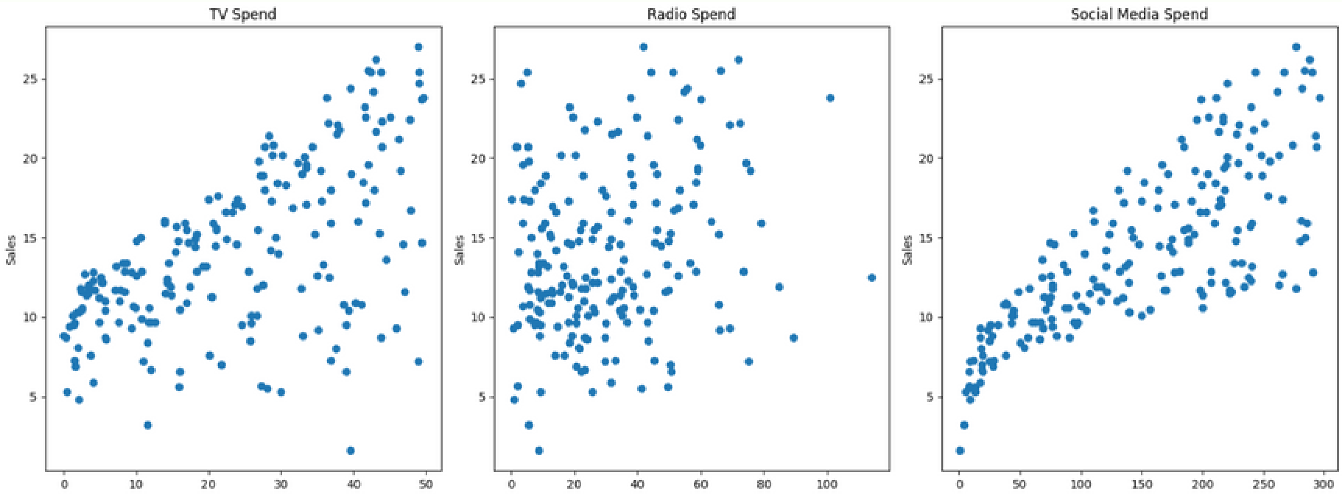


Figure 2: Scatter plot between different marketing techniques and sales

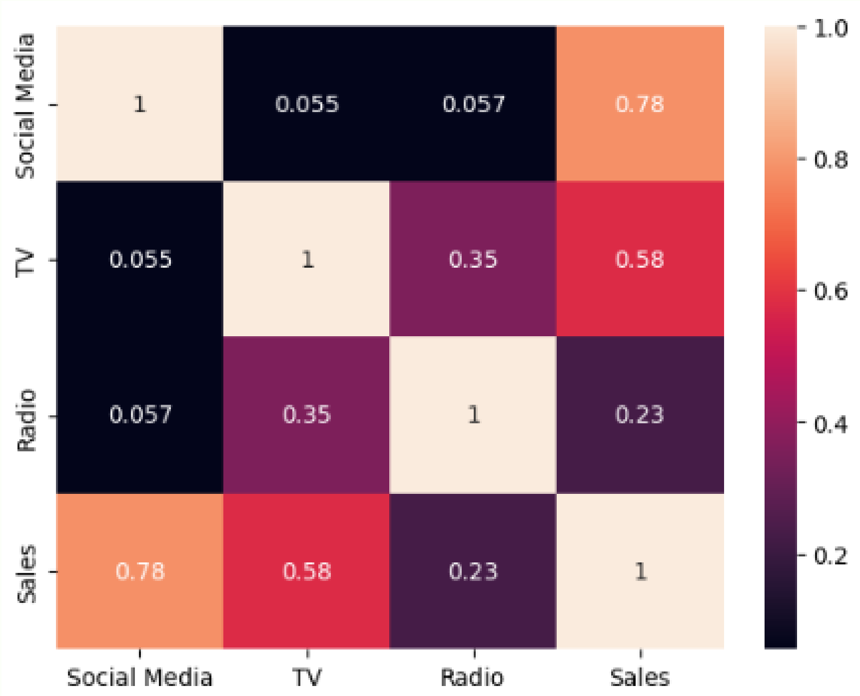


Figure 3: Different correlations

We can notice the correlations between each market technique and sales in the last row.

Numerical Results & Discussion

by applying Linear regression to our dataset we got the equation:

Sales = 0.04469599323076605 Social Media + 0.1875657018658699 TV + -0.00032275300998738 1 Radio + 3.1515267680706547

by analysis, we encountered a discrepancy between our observed correlation of social media activity and sales. We opted for polynomial regression as a resolution, leading to a more accurate outcome.

MAE	1.213745773614481
MSE	2.2987166978863796
RMSE	1.5161519375993884
R2_Score_Train	0.8856665510409361
R2_Score_Test	0.9185780903322445

Note: We calculated R2 score on both train and test data to detect overfitting and underfitting, with the shown numbers the r2 score of train and test data is high so there is no overfitting or underfitting.

Numerical Results & Discussion

Using polynomial regression to get higher accuracy we iterated over multiple polynomial degrees and used the one which has the least mean squared error:

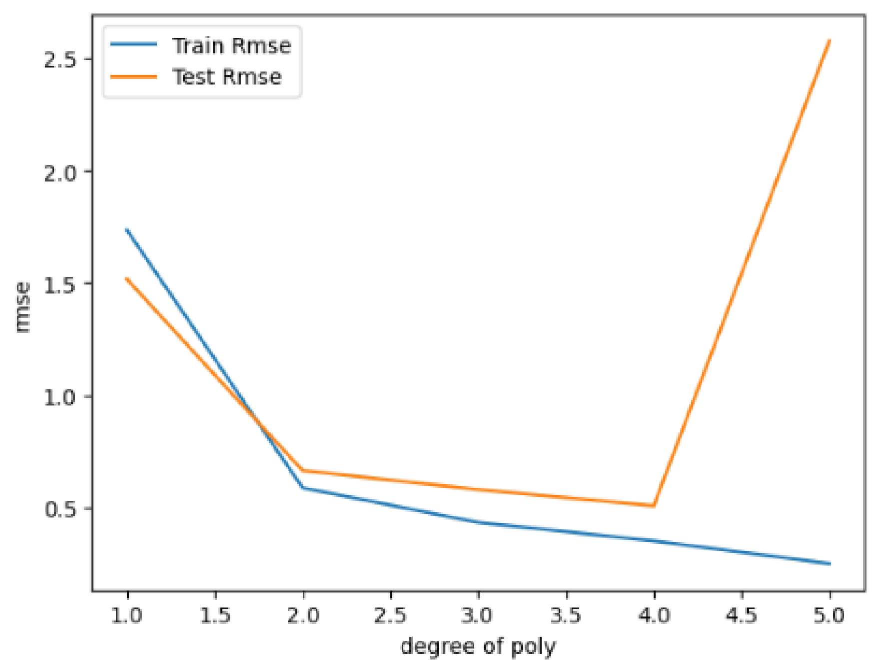


Figure 4: Model Complexity (degree)

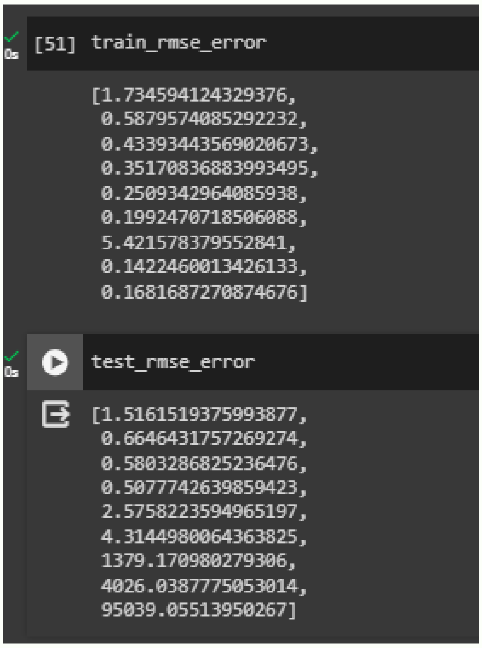


Figure 5: Different degrees errors

We observe from figure 4 that the model test RMSE is kind of correlated with train RMSE until degree of 4 , then after that the model overfits so we used degree of 4.

The R2 score of this polynomial degree is : 0.9908673417956482

Numerical Results & Discussion

After deploying our model we tested it with arbitrary values and the higher we invest in Social Media the higher sales we get and this agrees with our correlation results.

In conclusion, Polynomial regression of degree 4 has given us the best accuracy.

Now we can answer our first two questions in our statistical questions :
1- The more you invest in modern marketing strategies like Social Media the higher sales you expect.

2- We can rank our marketing techniques based on their participation in sales :
1- Social Media
2- TV
3- Radio

Numerical Results & Discussion

Note: You can find these graphs and tables in our code attached in attachments section.

	Impressions	From Home	From Hashtags	From Explore	From Other	Saves	Comments	Shares	Likes	Profile Visits	Follows	Caption	Hashtags
0	3920	2586	1028	619	56	98	9	5	162	35	2	Here are some of the most important data visua...	#finance #money #business #investing #investme...
1	5394	2727	1838	1174	78	194	7	14	224	48	10	Here are some of the best data science project...	#healthcare #health #covid #data #datascience ...
2	4021	2085	1188	0	533	41	11	1	131	62	12	Learn how to train a machine learning model an...	#data #datascience #dataanalysis #dataanalytic...
3	4528	2700	621	932	73	172	10	7	213	23	8	Here□s how you can write a Python program to d...	#python #pythonprogramming #pythonprojects #py...
4	2518	1704	255	279	37	96	5	4	123	8	0	Plotting annotations while visualizing your da...	#datavisualization #datascience #data #dataana...

Figure 0: Rows from our first dataset

Description: This dataset contains the the total interactions to posts from Home, Explore, Hashtags, Other, and so on.

By analyzig our second dataset here are our results:

	Impressions	From Home	From Hashtags	From Explore	From Other	Saves	Comments	Shares	Likes	Profile Visits	Follows
count	119.000000	119.000000	119.000000	119.000000	119.000000	119.000000	119.000000	119.000000	119.000000	119.000000	119.000000
mean	5703.991597	2475.789916	1887.512605	1078.100840	171.092437	153.310924	6.663866	9.361345	173.781513	50.621849	20.756303
std	4843.780105	1489.386348	1884.361443	2613.026132	289.431031	156.317731	3.544576	10.089205	82.378947	87.088402	40.921580
min	1941.000000	1133.000000	116.000000	0.000000	9.000000	22.000000	0.000000	0.000000	72.000000	4.000000	0.000000
25%	3467.000000	1945.000000	726.000000	157.500000	38.000000	65.000000	4.000000	3.000000	121.500000	15.000000	4.000000
50%	4289.000000	2207.000000	1278.000000	326.000000	74.000000	109.000000	6.000000	6.000000	151.000000	23.000000	8.000000
75%	6138.000000	2602.500000	2363.500000	689.500000	196.000000	169.000000	8.000000	13.500000	204.000000	42.000000	18.000000
max	36919.000000	13473.000000	11817.000000	17414.000000	2547.000000	1095.000000	19.000000	75.000000	549.000000	611.000000	260.000000

Figure 1: numerical statistics from our data

Numerical Results & Discussion

By analyzig our second dataset here are our results:

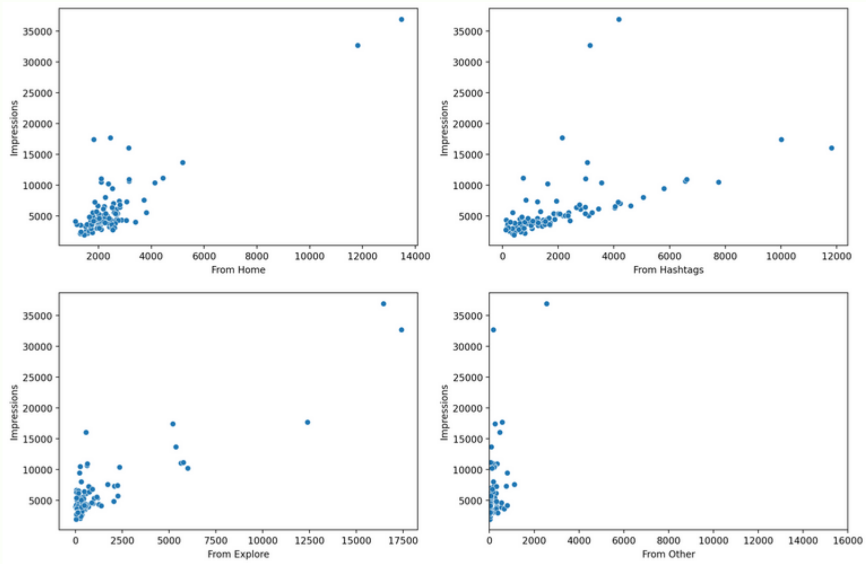


Figure 2: Scatter plot between different screens on Instagram app and impressions

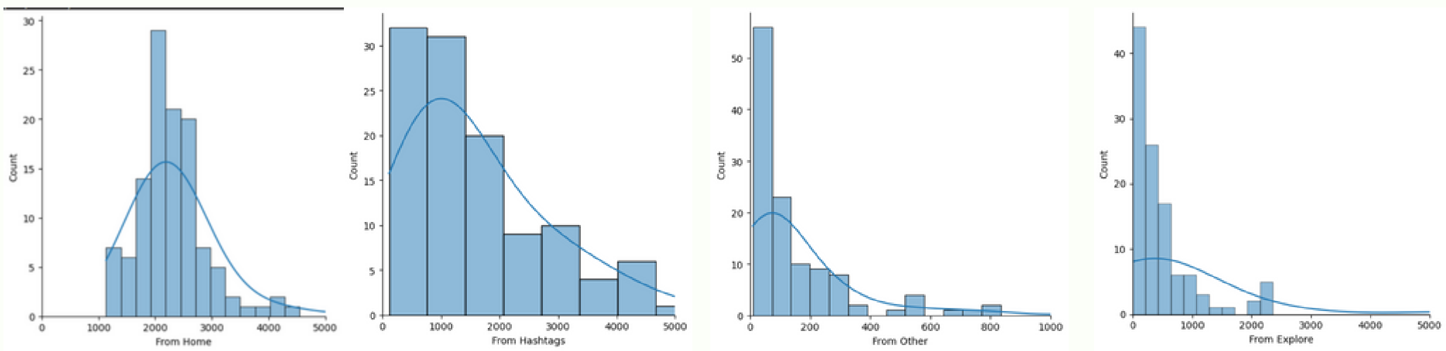


Figure 3: Distribution of views from different screens

Numerical Results & Discussion

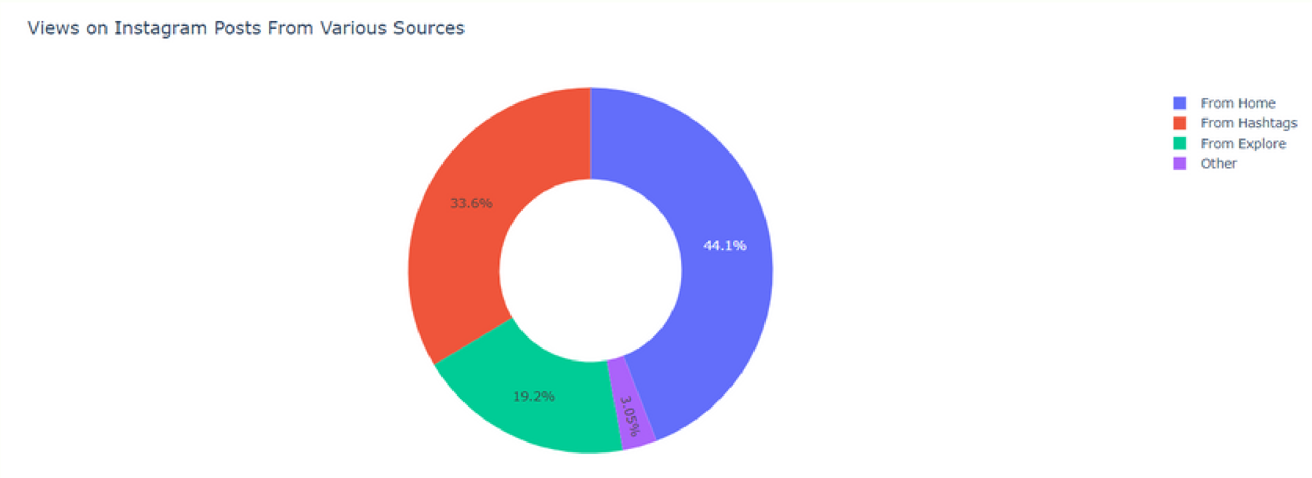


Figure 3: pie chart between different view screens

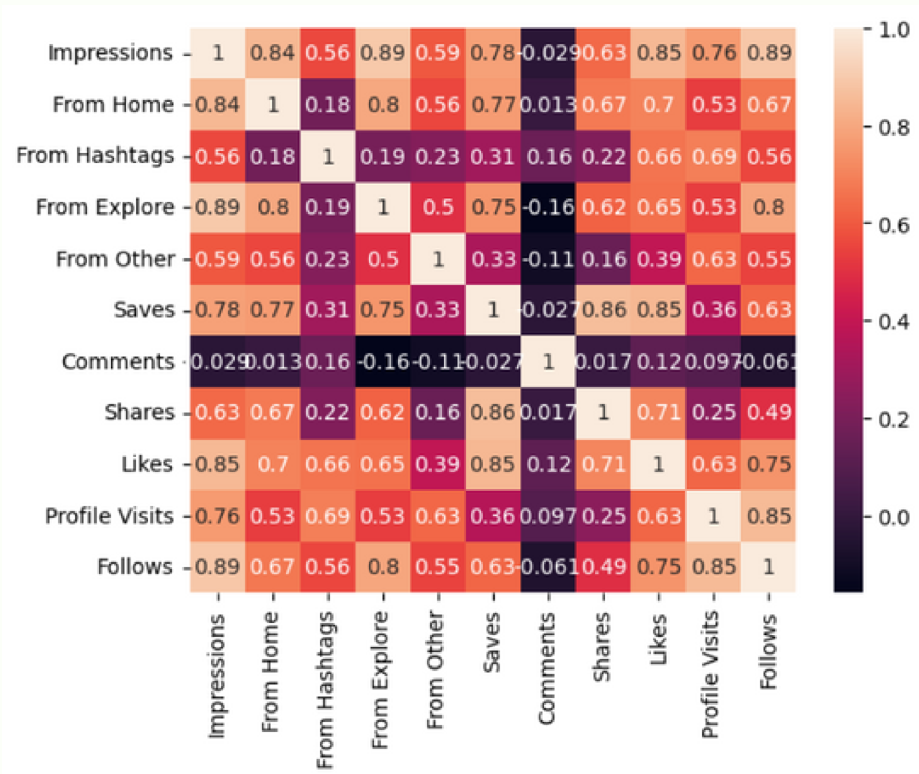


Figure 4: Different correlations

Numerical Results & Discussion

Note: You can find these graphs and tables in our code attached in attachments section.

Note: the larger the word the more common it is used

Figure 5: Most commonly used words in ADs captions

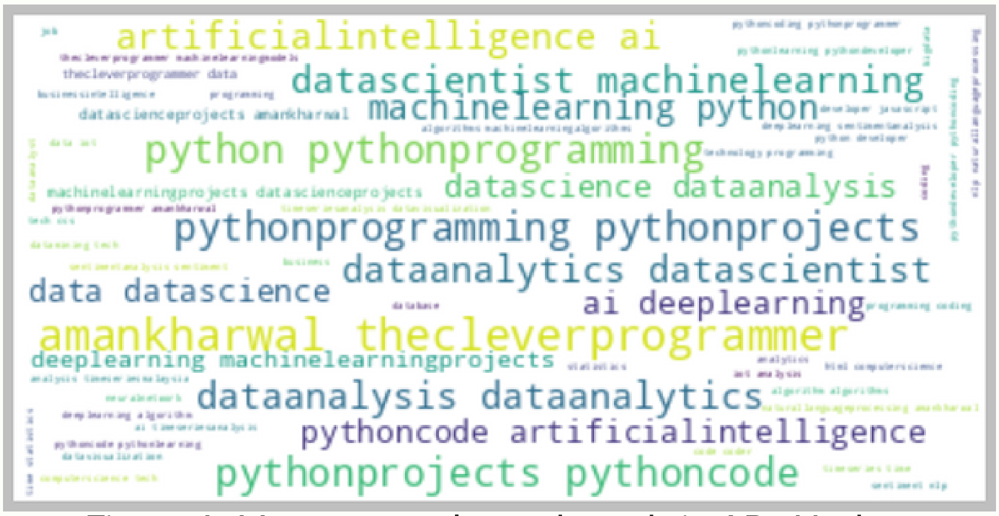


Figure 6: Most commonly used words in ADs Hashtags

Numerical Results & Discussion

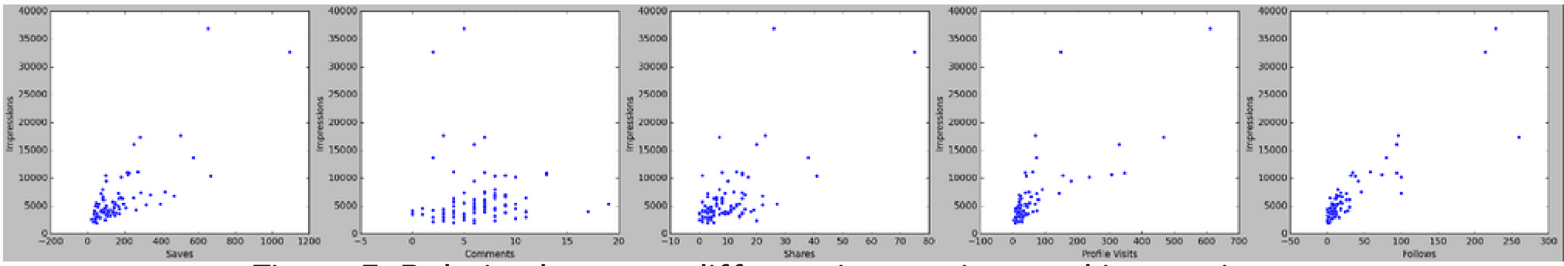


Figure 7: Relation between different interactions and impressions

By applying Linear Regression on our dataset we obtained that :

MAE	0.002195121345446386
RMSE(normalized)	0.00349616768504986
R2_Score_Train	0.9993693795723955
R2_Score_Test	0.9984465023274626

Numerical Results & Discussion

In conclusion: Views from explore and follows are the most two factors that are best correlated to impressions on ADs.

Now we can answer our third question:
The attributes we should take into consideration are where we put our Ads and which interaction types contribute to more impressions.

Numerical Results & Discussion

Note: You can find these graphs and tables in our code attached in attachments section.

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
...	...	...	...	...	...
395	15691863	Female	46	41000	1
396	15706071	Male	51	23000	1
397	15654296	Female	50	20000	1

Figure 0: Rows from our dataset

Description: This dataset contains the relation between user personal information and whether he purchased or not which would help us identify the targeted audience.

By analyzig our third dataset here are our results:

	Age	EstimatedSalary	Purchased
count	400.000000	400.000000	400.000000
mean	37.655000	69742.500000	0.357500
std	10.482877	34096.960282	0.479864
min	18.000000	15000.000000	0.000000
25%	29.750000	43000.000000	0.000000
50%	37.000000	70000.000000	0.000000
75%	46.000000	88000.000000	1.000000
max	60.000000	150000.000000	1.000000

Figure 1: numerical statistics from our data

Numerical Results & Discussion

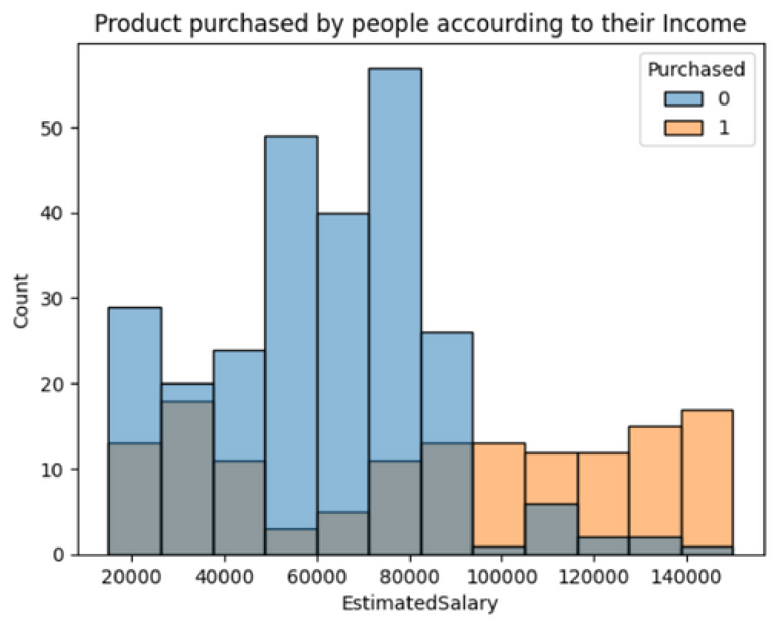


Figure 2: Distributions of salaries and whether the user purchased or not.

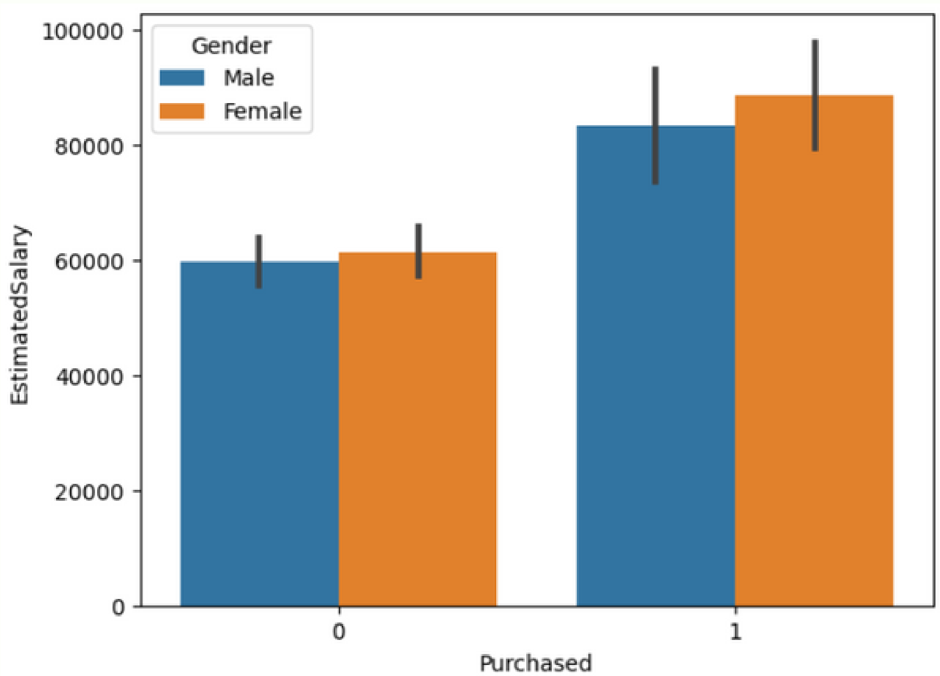


Figure 3: Relation User gender and purchase process given his/her salary

Numerical Results & Discussion

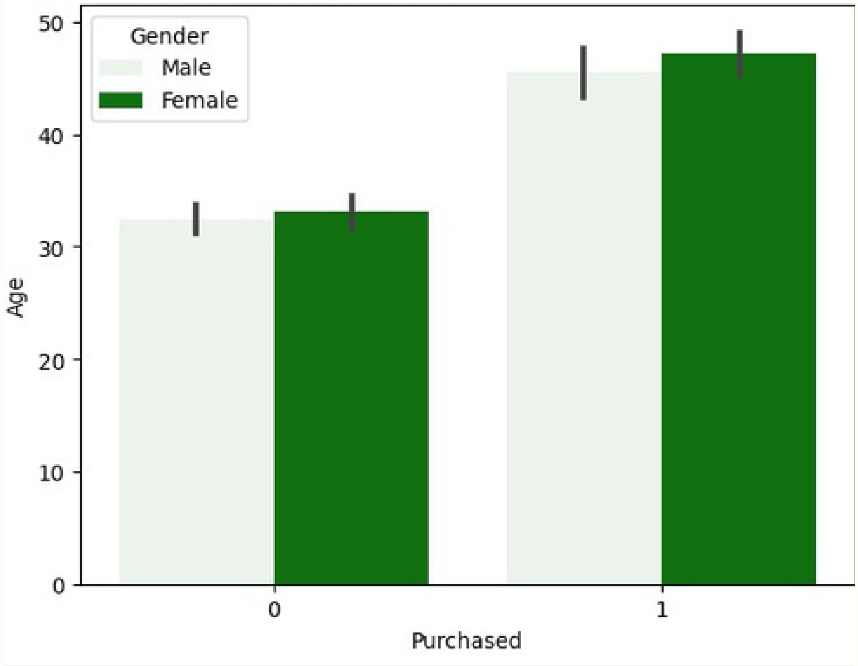


Figure 4: Distributions of salaries and whether the user age.

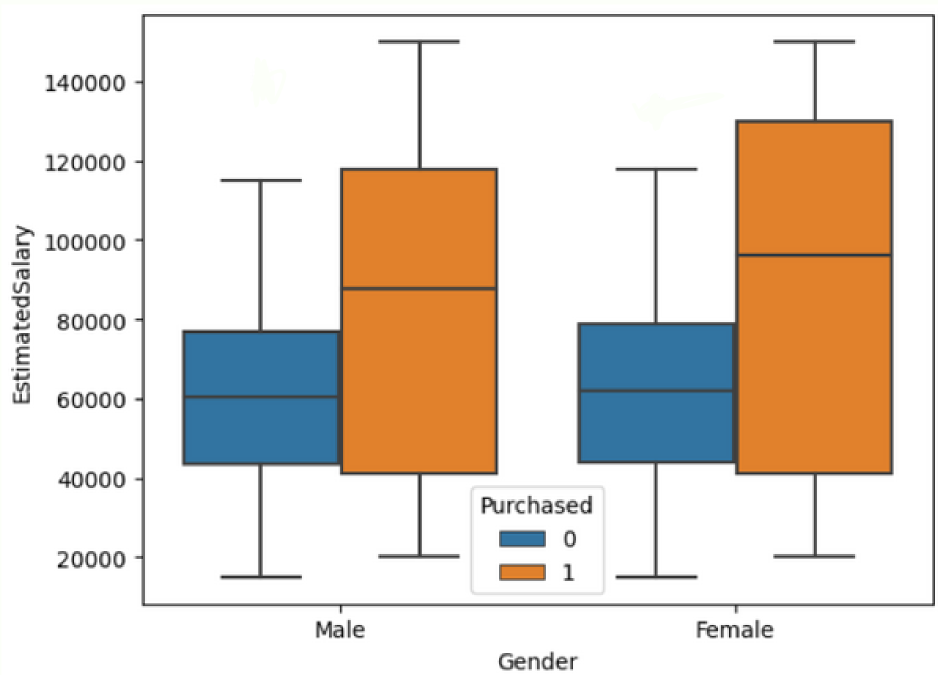


Figure 5: Boxplot between gender and estimated salary after dropping outliers

Numerical Results & Discussion

Figure 6: Correlations between age , estimated salary and purchased

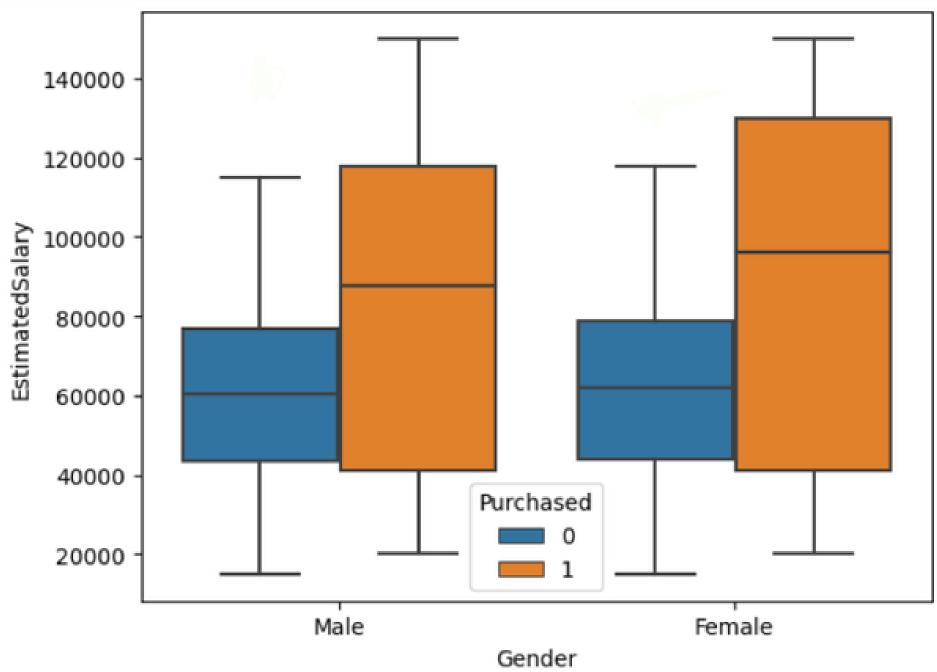


Figure 7: Boxplot between gender and estimated salary after dropping outliers

Numerical Results & Discussion

Note: You can find these graphs and tables in our code attached in attachments section.

The ratio of the females that purchased to the total people who purchased:
purchases from female count/Purchases count 0.538462
purchases from male count/Purchases count 0.461538

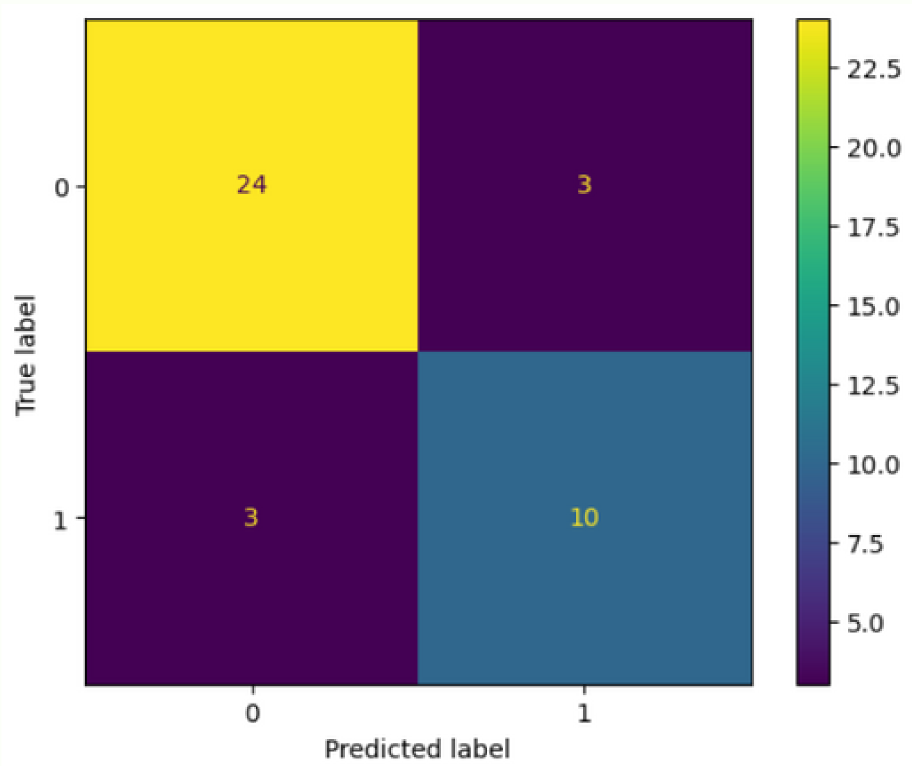


Figure 8: Confusion matrix using logistic regression

```
Train_accuracy score: 0.8305555555555556
Test_accuracy score: 0.85
classifcation report :
                precision    recall  f1-score   support

           0       0.89      0.89      0.89        27
```

Figure 9: accuracy, recall, f1 score, and precision

Numerical Results & Discussion

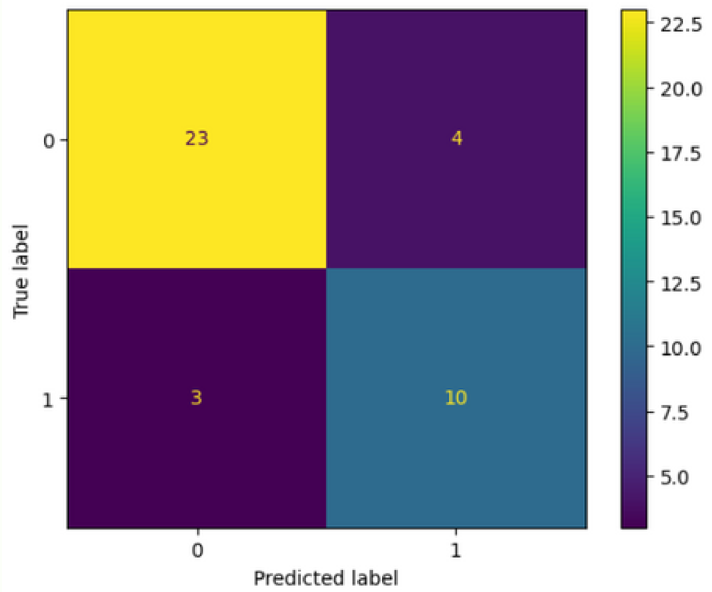


Figure 10: Confusion matrix using SVC Linear

```
Train_accuracy score: 0.8388888888888889
Test_accuracy score: 0.825
classifcation report :
              precision    recall  f1-score   support

           0       0.88      0.85      0.87        27
```

Figure 11: accuracy, recall, f1 score, and precision

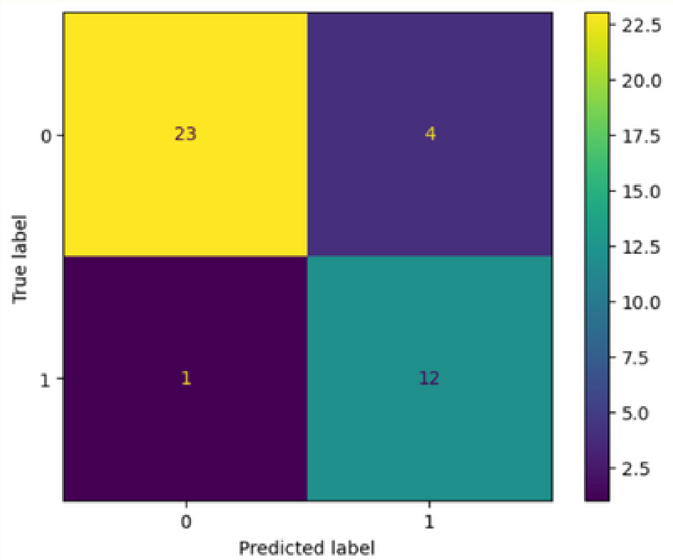


Figure 12: Confusion matrix using SVC polynomial

```
Train_accuracy score: 0.9027777777777778
Test_accuracy score: 0.875
classifcation report :
              precision    recall  f1-score   support

           0       0.96      0.85      0.90        27
```

Figure 13: accuracy, recall, f1 score, and precision

Numerical Results & Discussion

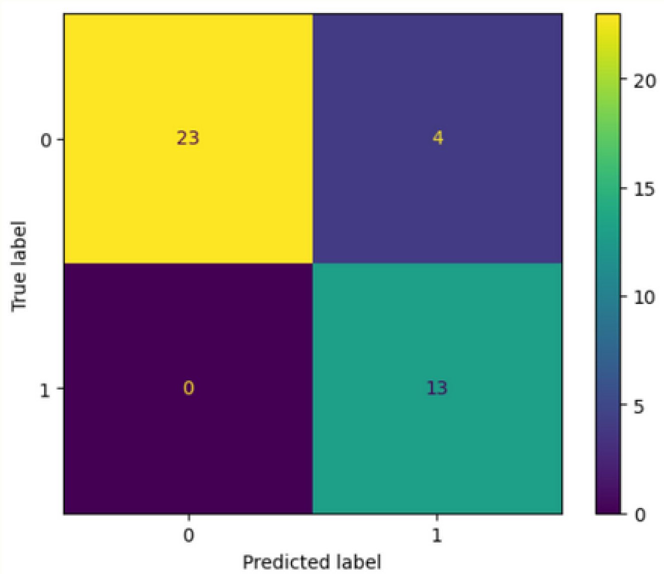


Figure 14: Confusion matrix using SVC RPF

```
Train_accuracy score: 0.9138888888888889
Test_accuracy score: 0.9
classifcation report :
              precision    recall  f1-score   support

           0       1.00      0.85      0.92        27
```

Figure 15: accuracy, recall, f1 score, and precision

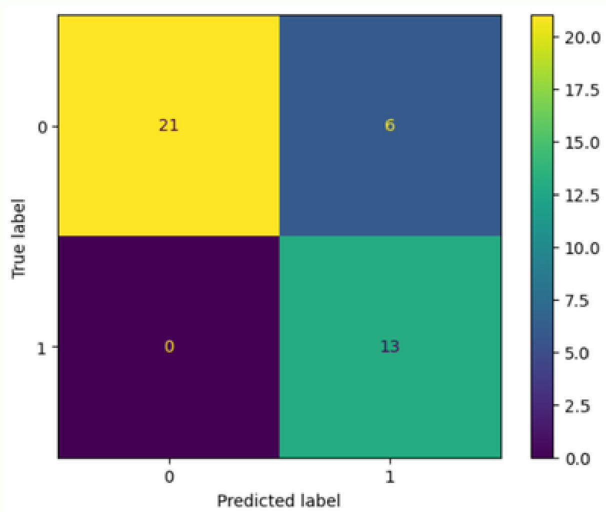


Figure 16: Confusion matrix using KNN 3N

```
Train_accuracy score: 0.9277777777777778
Test_accuracy score: 0.85
classifcation report :
              precision    recall  f1-score   support

           0       1.00      0.78      0.88        27
```

Figure 17: accuracy, recall, f1 score, and precision

Numerical Results & Discussion

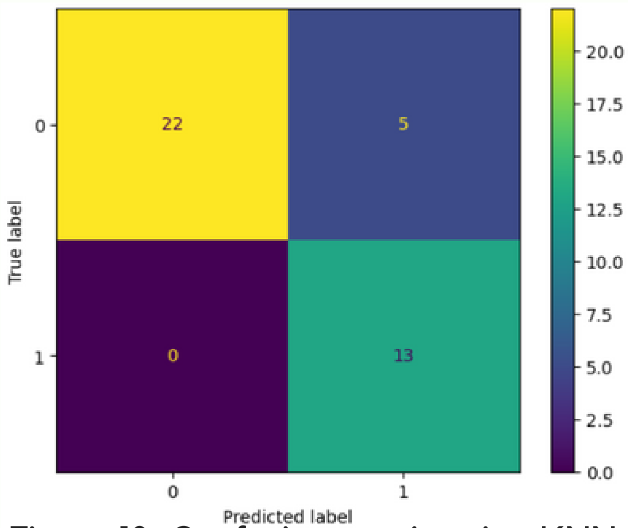
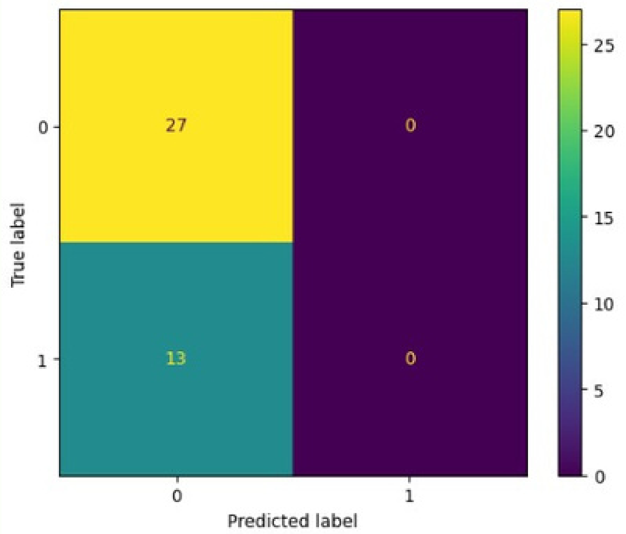


Figure 18: Confusion matrix using KNN 5N

```
Train_accuracy score: 0.9222222222222223
Test_accuracy score: 0.875
classifcation report :
              precision    recall  f1-score   support

           0       1.00      0.81      0.90        27
```

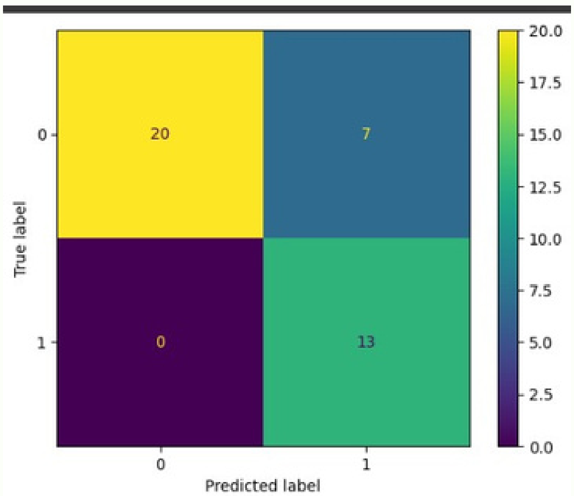
Figure 19: accuracy, recall, f1 score, and precision




```
Train_accuracy score: 0.6388888888888888
Test_accuracy score: 0.675
classifcation report :
              precision    recall  f1-score   support

           0       0.68      1.00      0.81        27
```

Figure 21: accuracy, recall, f1 score, and precision

Figure 20: Confusion Matrix using naive bayes MultinomialNB



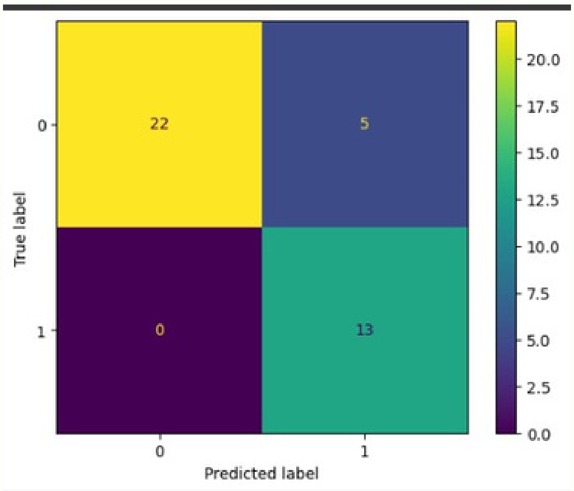

```
Train_accuracy score: 0.9972222222222222
Test_accuracy score: 0.825
classifcation report :
              precision    recall  f1-score   support

           0       1.00      0.74      0.85        27
```

Figure 23: accuracy, recall, f1 score, and precision

Figure 22: Confusion Matrix using naive bayes Descision trees

Numerical Results & Discussion

Note: You can find these graphs and tables in our code attached in attachments section.




```
Train_accuracy score: 0.9972222222222222
Test_accuracy score: 0.875
classifcation report :
              precision    recall  f1-score   support

           0       1.00      0.81      0.90        27
```

Figure 25: accuracy, recall, f1 score, and precision

Figure 24: Confusion Matrix using naive bayes random forests classifier

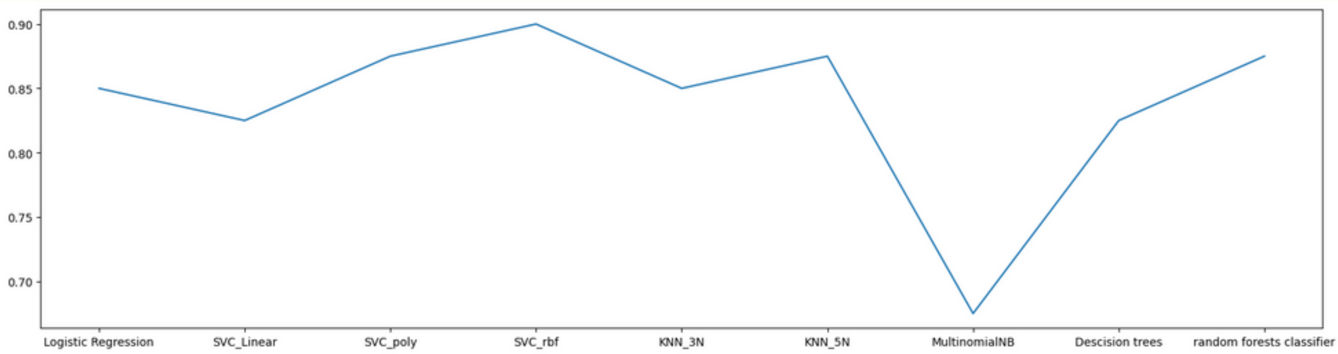


Figure 26: effect of different models on the accuracy

In conclusion, From our evaluation metrics we conclude the model with the best results was support vector classifier with RBF kernel

Concluding Section

In conclusion, our research probability project has provided valuable insights into the dynamics of advertising across various platforms. Through a meticulous examination of three distinct data sets, we determined that social media emerges as the preeminent channel for effective advertising.

The second facet of our study delved into the nuances of constructing impactful Instagram ads, emphasizing the importance of impressions counts. This segment furnished practical guidelines for advertisers aiming to enhance engagement on this prominent social media platform.

In our final exploration, we addressed the strategic dimension of ad direction by harnessing the advantages of user data stored on websites. The resultant predictive model not only optimizes the delivery of targeted ads but also aligns content with user interests, thereby enhancing the overall advertising experience.

In the rapidly evolving landscape of digital marketing, our research provides actionable insights for advertisers seeking to maximize their impact. The convergence of empirical data and analytical methodologies offers a comprehensive understanding of the efficacy of advertising channels, the intricacies of content creation, and the strategic deployment of ads based on user data. These findings are poised to inform and guide advertising strategies in an increasingly data-driven environment.

Key Outcomes

Social Media Effectiveness: The research highlighted social media as the most effective advertising platform, based on a comprehensive analysis of TV, radio, and social media channels. This conclusion was drawn from a detailed examination of user engagement and response metrics.

Instagram Ad Strategies: The study provided practical guidelines for constructing impactful ads on Instagram, emphasizing the significance of impressions counts. This insight equips advertisers with specific strategies to enhance engagement and visibility on this popular social media platform.

Strategic Ad Direction: Leveraging user data stored on websites, the research developed a predictive model for effectively directing ads. This approach optimizes the delivery of targeted content to users, aligning advertisements with their interests and preferences. The strategic use of user data enhances the relevance and impact of advertising efforts.

Comprehensive Understanding: The research outcomes offer advertisers a comprehensive understanding of the evolving digital marketing landscape. By combining empirical data and analytical methodologies, advertisers can make informed decisions regarding channel selection, content creation, and the strategic deployment of ads based on user data.

Future Work

Advanced Analytics Techniques: Explore more advanced analytics techniques beyond regression and polynomial regression. Machine learning algorithms, such as neural networks or ensemble methods, could be employed to uncover intricate patterns and relationships within advertising data.

Cross-Platform Analysis: Extend the research to encompass a more extensive range of advertising platforms. Investigate how advertising strategies and user engagement metrics vary across different social media platforms, search engines, and emerging digital channels.

User Privacy and Ethical Considerations: Address the ethical implications of leveraging user data for targeted advertising. Explore frameworks and strategies to ensure user privacy and data protection, considering the evolving landscape of privacy regulations.

Consumer Behavior Analysis: Deepen the understanding of consumer behavior in response to advertising. Investigate factors influencing user interactions with ads, such as emotional responses, brand perceptions, and the impact of social and cultural variables.

Global Market Analysis: Extend the research to include a global perspective, considering cultural and regional variations in advertising effectiveness. This could involve studying the impact of cultural nuances on user preferences and responses to advertisements.

Additional Questions For Subsequent Investigations

Influence of Ad Format on Engagement:

1- Does the format of advertisements (e.g., image, video, carousel) have a significant impact on user engagement?
Hypothesis: Advertisements presented in video format might result in higher engagement compared to static image ads due to increased interactivity.
Temporal Trends in User Engagement:

2- How do user engagement patterns vary over different times of the day, days of the week, or seasons?
Hypothesis: User engagement may peak during specific hours or days, influencing the optimal timing for ad placement.
Comparative Analysis of Social Media Platforms:

3- Are there significant differences in user engagement metrics across various social media platforms (e.g., Instagram, Facebook, Twitter)?
Hypothesis: Certain platforms may be more effective for specific demographics or types of content.

References

1- Hands-On-Machine-Learning Reference:
https://powerunit-ju.com/wp-content/uploads/2021/04/Aurelien-Geron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-OReilly-Media-2019.pdf

2- code for analyzing the first dataset:
https://colab.research.google.com/drive/1r-X5gGBFrn_R4pAMrPamWIaPIfiIELB3?usp=sharing

3-code for analyzing the second dataset :
https://colab.research.google.com/drive/1foytFT-RTKq3JKvO_tBhyC26aSFPEtD9?usp=sharing

4-code for analyzing the third dataset :
https://colab.research.google.com/drive/1rqAbVA4rS-EDXs1r8A-BU8VQhTLAXcRt?usp=sharing

Note: all the graphs we collected are from these codes which we analyzed ourselves