Correlating ROI, CPC, and Acquisition Costs Across Channels for Insightful Advertising

Khizar Mohamed Zubair Sait
Department of Artificial Intelligence and Data Science
Chaitanya Bharathi Institute of Technology
Hyderabad, India

Gatla Vijayendher
Department of Artificial Intelligence and Data Science
Chaitanya Bharathi Institute of Technology
Hyderabad, India

K. Sai Karthikeya
Department of Artificial Intelligence and Data Science
Chaitanya Bharathi Institute of Technology
Hyderabad, India

Dr. Satya Kiranmai Tadepalli
Asst. Professor, Dept. of AI&DS
Chaitanya Bharathi Institute of Technology
Hyderabad, India

Shobarani Salvadi
Asst. Professor, Dept. of AI&DS
Chaitanya Bharathi Institute of Technology
Hyderabad, India

***Abstract*—In today's world, digital advertising has become a dominant force in the economic market. The current landscape is highly competitive and cost-oriented, offering a vast array of platforms for posting ads. To navigate this complexity and maximize Return on Investment (ROI), leveraging data analysis is essential. The project aims at building a complete easy understanding to the advertiser who aims to publish his advertisement keeping in mind all different attributes related to it. The project uses different Data Science techniques and machine learning models in order to analyses and find the performance of Advertisement. The project uses a Dataset which includes different fields such as Acquisition Cost, ROI , customer segment, date of posting , target audience age ,etc. The project is performed using R. The project uses different pre-processing steps in order to make the data more easily understandable. Exploratory Data Analysis is performed on the data such that use full insights can be taken from it. Finally, Regression model, Clustering model and Time Series Analysis is performed.**

Keywords—Digital Advertisement, Cost Optimization, Acquisition Cost, ROI, Regression, Clustering, Time Series Analysis.

I. INTRODUCTION

The growth of digital advertising in today's age is massive. According to a PwCs Global Entertainment & Media outlook report, advertising grew at a stunning 22.6% in 2021. The forecast calls for advertising growth of CAGR(Compound Annual Growth Rate) at a rate of 6.6% through 2026. This era, it is vital for a business owner to have a proper understanding and tools while publishing an advertisement in the digital space.

Digital advertising offers unique advantages over traditional advertising channels, such as real-time tracking, precise targeting, and cost-effectiveness. However, to fully leverage these benefits, it is essential for marketers to evaluate the performance of their digital advertising campaigns. With multiple available platforms in advertising such as Google AdSense Facebook AdSence, Email advertising ,Twitter Ads ,etc. It is of high importance that the business owner understands which platform will provide the best results based on the type and product of advertisement.

The advertising has also changed with the advancement of technology. Fields like Data Science and Machine Learning ,have made advertisers focus more on profitability and cost benefit oriented while publishing an Ad. Also, techniques like Personalizing of ads have been one of the major breakthroughs in advertising by which the advertisement is shown to the user based on his/her interest by collecting personal data points. The use of data driven decision can help the advertiser to help is Ad reach to the correct set of target audience. The project analyses and carries out exploratory data analysis to gain insights from the available data. Once insights are taken the project aims to build machine learning models which can help make impact full predictions.

II. LITERATURE SURVEY

The changing landscape of development of digital advertising has been credited to numerous writers. The following list of recent methods and techniques that have been used and developed and put out in papers:

The paper [1] explains the impact of paid advertising on a business's return on investment and profitability in digital marketing. It explains us the different methods which is used in modern digital marketing, such as social media marketing, email marketing, pay-per-click advertising and search engine optimization(SEO).Additionally, it aids in our comprehension of the various techniques used to calculate return on investment (ROI) in digital advertising, including cost per click (CPC), conversion rate, ROAS (return on ad spend), ROI (return on investment), and marketing mix modeling. It also discusses the difficulties with paid advertising, which can assist us in overcoming such difficulties.

The challenge of optimizing many ad groups in internet advertising is covered by Marco Gigli and Fabio Stella in their study [2], which also suggests a parametric Bayesian regression model. It emphasizes on the shortcomings of the current methods and the ineffective use of data. The study offers experimental findings on both real-world and synthetic data to assess the effectiveness of their suggested approach. They discussed the findings of their comparison between the bootstrapped Thompson sampling strategy and their parametric model with full-fledged Bayesian inference. When the numb

979-8-3315-1905-6/24/$31.00 ©2024 IEEE
DOI 10.1109/ICACCTech65084.2024.00141

er of campaigns and ad groups is kept constant and the daily budget is increased, the average percentage of conversions lo st as a result of not employing the parametric Bayesian techni que can reach 40% in some conditions. Relative regret for bo th GPs and Bootstrapped TS declines quite consistently.

The study by Amit Verma and Veena Vemuri [3] pr ovided an explanation of the development and future prospec ts of the digital marketing sector in India.It explains the grow ing trends of many platforms, including email marketing, sea rch engine optimization, social media marketing, and website marketing. It additionally illustrates which digital marketing platforms yield the highest return on investment. With a ROI of 28%, email marketing performed best.

The study conducted in the paper shows [4] that ther e arises in purchase intention due to the impact of digital mar keting .It also studies different perception factors of digital m arketing .The Model was tested on different test like z test be ta coefficients etc and it was noted that digital marketing had a positive correlation with the purchase decision made by the people .The objective of the study was to find a correlation b etween the purchase intention of a product with the digital ad vertisement viewed by the person. It was found to have a pos itive correlation between both the factors.

In the paper [5] it focuses on various KPIs to be use d while evaluating the performance of digital advertisements such as including click-through rate (CTR), conversion rate, cost per acquisition (CPA), return on ad spend (ROAS), and customer lifetime value (CLV). By analysing key metrics and KPIs, it allows the advertiser to understand the market and al so the insights which will help make more suitable and effect ive decisions for better profitability. The paper also highlight s the importance of advanced analytics techniques, such as A /B testing, multivariate testing, and attribution modelling, wh ich can help marketers identify the most effective strategies f or optimising ad performance.

The paper [6] highlights the use and implementation s of different technologies which are increasing the growth of digital marketing . The use of Data Driven decisions along wi th the use of CRM and other Data Mining and Data analysis t echniques combined with Machine Learning have helped the growth of the market. The paper has not much focus specific ally on each sub domain the way each proportionally impacts the growth ie how proportionally the data driven decisions im pact and how the usage of CRM systems impacts the growth of the segment .

The Study in the paper [7] focuses on demonstrating the applications of big data in the digital marketing space. Th e study uses a qualitative approach which uses secondary sou rces of data that were derived from different sources. Miles a nd Huberman interactive models were used in order to perfor m the analysis of the project.The usage of big data have enab led to understand the customer more efficiently and optimize accordingly to the environment rapidly to expand and grow i n the marketing space.

In the paper [8] explore the tools and techniques of data science which can be applied in digital marketing. They examine various methods of analysis, performance metrics, a nd their practical uses. Their work aims to demonstrate how data mining, combined with data science, has transformed th e landscape of digital marketing. The authors identify nine ke y areas that warrant in-depth study and analysis, emphasizing the need for broader data sources and a more comprehensive data exploration.

III. Proposed Work

A. About the Dataset

The dataset which is used in this project is downloaded fr om Kaggle called as 'marketing campaign dataset'. The datas et consists of 2,00,000 rows and 16 columns. The dataset con tains attributes such as ROI, Conversion Rate , Acquisition C ost , Location , Channels Used etc.

B. Exploratory Data Analysis

To understand the dataset thoroughly and understanding the data different plots are plotted so that inferences can be made. The plots are made using the ggplot2 package is used to for different data visualization.

The below heat map Fig 1 helps us understand the correlation between the different numerical attributes present in the dataset such as clicks, CTR, CPA, ROI, Acquisition cost, CPC, Impressions. We can observe that clicks and CTR has a positive correlation. Clicks and acquisition cost has almost no correlation. Clicks and Cost per click has a negative correlation of approximate -0.5. This attributes the higher number of clicks an Ad gets the cost per click on that Ad reduces.

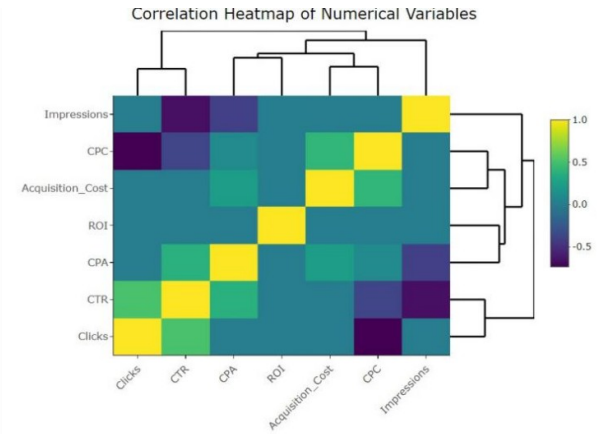

Fig. 1. Heat Map

The below density plot Fig 2 Shows a correlation between CTR (click through rate) with respect to the channels used. The mean peak of all the graphs are approximately 15-20. YouTube and website have the maximum CTR with approximately 80 clicks.

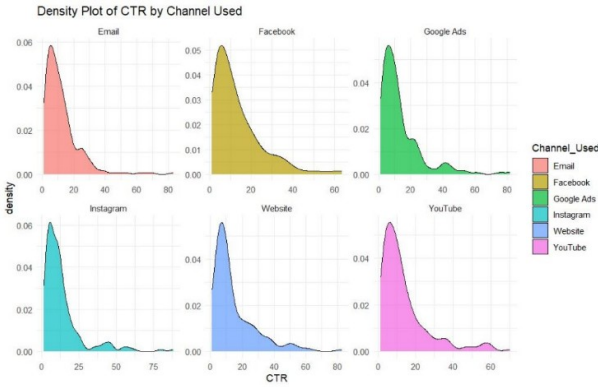

Fig. 2. Density Plot of CTR by channels used

C. Model Building and Evaluation

Several models were trained on the dataset in order to add ress different queries and evaluation which has to be done on evaluating the performance of the Advertisement. The model s are built keeping in mind to address the query which a cam paigner looks out for before publishing the Ad such that it is more profitable for the campaigner.

The models are build using the CARET packages in R [9] CARET helps us to streamline the model building of comple x classification and regression model. The package has sever al tools and functions which helps us in performing different steps required in model building. Also additionally a package E1071 [10] which was introduced by functions of department of statistics, probability theory group for building an SVR m odel.

1) Models

K means: It is a clustering model that divides the dataset into K distinct clusters. The K means clustering is done with respect to impressions and clicks [11]. This model will help us analyse how the ad is performing in real time as profitability would increase when a viewer seeing the ad goes to the desired website or app. Impression is calculated as the number of times the Ad is shown where as number of clicks shows the number of people clicked on the ad to get redirected to the desired location by the advertiser.

Regression: Regression is a technique which helps us correlate a dependent variable to one or more independent variable numerically that is done with the help of different statistical methods [12]. There are two types of regression model built in order for the best performance. The model is trained using all the numerical parameters with respect to predicting the Return on Investment (ROI). The first model is a simple multi-class linear regression model. Simple Linear Regression uses to discover a best fitting line by reducing least square error. The other regression model uses a Support Vector Regression (SVR) technique in order for building the model. SVR works on the similar principals of Support Vector Machine (SVM). The major benefit of using SVR is that it is a non-parametric technique. The SVR allows us to construct non linear model easily without changing the explanatory variables. The kernel functions are the one which the SVR depends upon. The other major advantage is SVR is very flexible on distribution of the underlying parameters and the relationship between the dependent and the independent variables.

Time Series Analysis: Time series analysis is performed using ARIMA. ARIMA (An auto regressive integrated moving average) is a statistical analysis model [13] that uses time series to understand the dataset and to analyse the data. In simple words it helps to measure events that happen over a period of time.

2) Parameters for Model Building

Mean Squared Error (MSE): MSE calculates the average squared difference between predicted and observed values, providing a measure of the model's accuracy by quantifying prediction errors.

Root Mean Squared Error (RMSE): RMSE quantifies the average error between predicted and observed values, offering a measure of model accuracy.

R-Squared (R2): R-Squared helps in gauging the goodness-of-fit of regression model. R-squared is a indicator of how well predictive models align with observed data.

Mean Absolute Error: Mean absolute error calculates the error between the paired observation expressing the same phenomena. For time series analysis the calculation of RMSE or R-squared is not a fruitful parameter due to there occurs multiple repetition of the same data. Whereas in the case of Mean absolute error it considered the absolute error between the predicted and actual values which does not influences to the outliers in this case of huge repeated data.

Silhouette Score: The silhouette score helps us measuring the goodness of a clustering technique. The value ranges from 0 to 1 where 1 denotes a perfect fit of clusters.

IV. Results

A. Regression

The Regression model helps us to numerically predict the value. The ROI being the most important term used in evalua ting the performance of a advertisement is used.

The Support Vector Regression helps us fit the multi-dim ensional data which is present in our dataset. A radial kernel i s used for fitting the model. The Support Vector Regression model is built in 2 stages. First a subset of 1000 records is tak en in order to analyse the performance of the model. In the 1 000 records subset the parameters that are used while training the model are as shown in the table I and the performance of the model is as shown in the graph Fig 3.

TABLE I. Parameters of SVR subset Model

Parameters	Value
SVM-TYPE	Eps-regression
SVM-Kernel	radial
Cost	1
gamma	0.2
epsilon	0.1
Number of Support Vectors	112

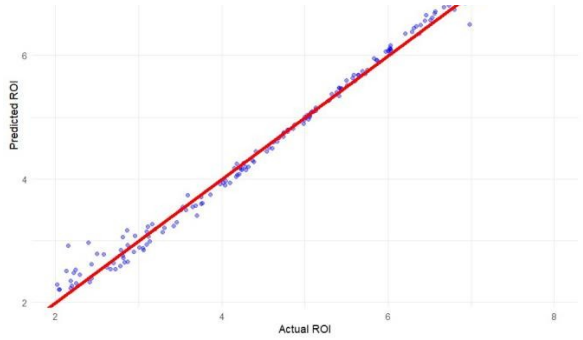

Fig. 3. Actual vs Predicted SVR subset model

As there is an increase in performance using the SVR model the data is then trained using the complete dataset. The parameters used in the SVR is as shown in the Table II and the performance is shown as in the graph Fig 4

TABLE II. PARAMETERS OF SVR MODEL

Parameters	Value
SVM-TYPE	Eps-regression
SVM-Kernel	Radial
Cost	1
gamma	0.2
epsilon	0.1
Number of Support Vectors	23

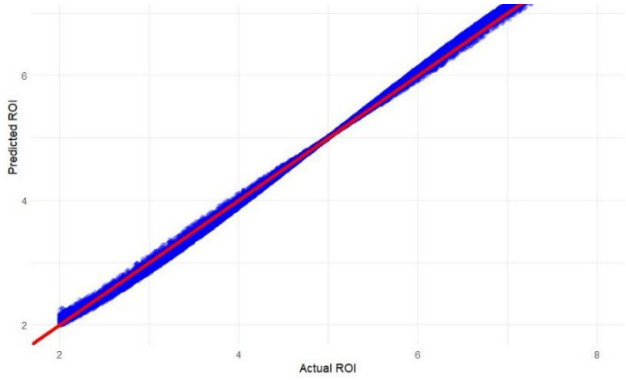

Fig. 4. Actual vs Predicted SVR model

The SVR is a clearly effective model with low RMSE values which are in the desired range of a model.

TABLE III. EVALUATION METRICS FOR SVR MODEL

Model	RMSE	MSE	R-Squared
Support Vector	0.160	0.026	0.991
Regression with Subset			
Support Vector Regression	0.0717	0.005	0.998

B. Clustering

From the given Fig 5 below we can observe that the clusters are audience engagement. The clusters are divided into zones. Each zone talks about the client under which category he wants to target the ads. By looking the graph, we can also conclude that to reach the minimum impression count (2000 according to the graph) the clicks should be ranging from 500-600. The **Silhouette Score** for the clustering algorithm comes out to be 0.5112601.

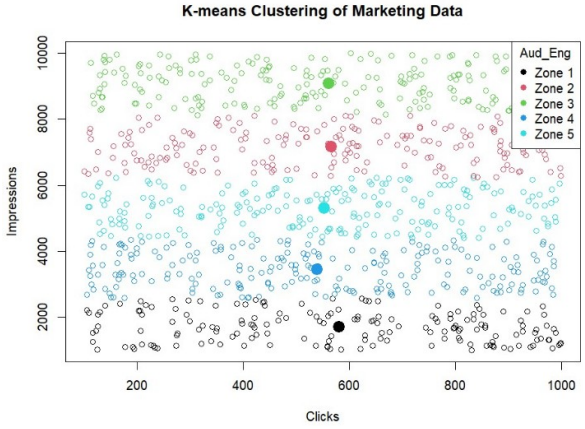

Fig. 5. K-Means Clustering

C. Time Series Analysis

1) Time Series Analysis with respect to Customer Segment

The plot Fig 6 is plotted using ARIMA. The ARIMA models uses a time series in order to correlate and find the future values based upon the previous values. The above line charts show relation with regards to Date and different market segments.

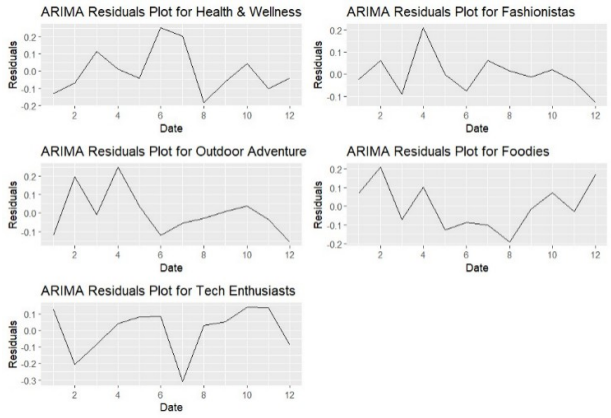

Fig. 6. Time Series for Customer Segment

The inference from the analysis can be made as follows:

- We can see a dip in advertising in Tech Enthusiast from June to Aug and then steep increase after that. Comparing with the

industry trends we see that most of the big tech companies launch there product in the month of September which can be attributed for such a graph.

- In the Fashionistas there is a steep rise in the graph in the beginning of the financial year, then there is step down at the end of the year might be attributed to year end clearance.
- The graph of outdoor adventure has multiple local minima and local maxima. This might be due to the influence of the seasonal rise in the outdoor adventure.

The table IV helps us analyse the Absolute mean error of the different customer segment. This helps us analyse the performance of the model.

TABLE IV. MEAN ABSOLUTE ERROR FOR DIFFERENT CUSTOMER SEGMENT

Customer Segment	Mean Absolute Error
Health and Wellness	0.104
Fashionistas	0.061
Outdoor Adventures	0.088
Foodies	0.103
Tech Enthusiasts	0.114

2) Time Series Analysis with respect to channel used

The below plots Fig 7 helps us analyze and understand the performance of different channels used throughout the year. The inference from the analysis can be made as follows,

The prediction of which channel is used best for publishing the ad in an year cannot be done because each and every channel has its own ups and downs in specific months. But the observation of which channel is used best for publishing the ads in a specific month can be made.

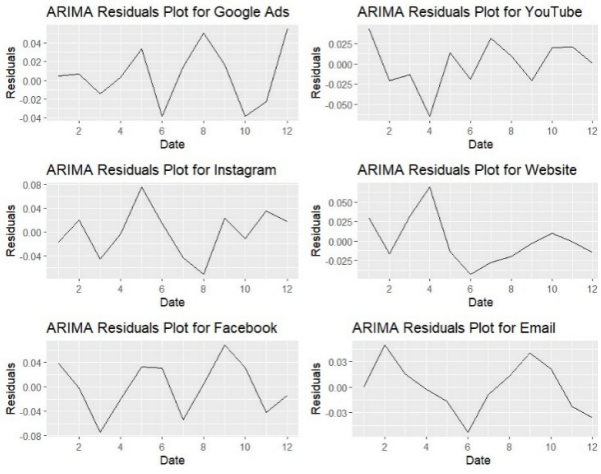

Fig. 7. ARIMA with respect to channel used

The table V helps us analyse the Absolute mean error of the different channel used. This helps us analyse the performance of the model.

TABLE V. MEAN ABSOLUTE ERROR OF DIFFERENT CHANNELS

Customer Segment	Mean Absolute Error
Google Ads	0.025
YouTube	0.023
Instagram	0.031
Facebook	0.034
Email	0.023
Website	0.023

Mean Absolute Error measures the error for the same phenomena in general sense. Here the residuals calculated for the ROI have numerous overlapping for the same ad over the month. This is because the ads have been run simultaneously on different platforms and also re ran again to attract the target no of customers in the month. So, taking the MSE, RMSE might mislead us for a huge numeric value which doesn't account for the auto-regressive model that we have built here. This is because the auto-regressive model nature is to regress the values based on the same values encountered in the past intervals. There might be numerous fluctuations in the RMSE and MSE values. Hence the Mean absolute error resolves the cases where the multiple ad runs with 0 returns by introducing a absolute (modulus) and takes out the unique ad metric values for the ROI.

V. DISCUSSIONS

Our models focuses on various different aspects from Time Series Analysis, Regression and Classification covering various different aspects of digital advertisements . Our models can be improved further using different specialized advanced architectures which focuses on a specific niche.

Time Series analysis can be improved with incorporating different models such as LSTM with ARIMA such as a model build in the study [14]. The study incorporates the usage of recurrent neural networks and long term short memory models , which allows them to remember the previous history enabling for a better forecasting as decisions are made on both current circumstances and previous experiences. Similarly, there can be a cost reduction on the advertising spend by incorporating previous experiences with our model along with LSTM's in order for robust learning decisions in the near future such as, implementation in the [15].

We can also incorporate effective customer targeting and segmentation with the combination of our classification model trained along with behavioral analysis of the customer as in the study[16].

The Click through Rate can be enhanced by using further advanced machine learning models such as a

framework implemented in [17] , The framework enhances improving of click through rate by implement random forest models depending on various possible outcomes and circumstances that changed over the period of time.

VI. Conclusion

In conclusion our study delves into different aspects of w hich makes an advertisement most success full and cost benef it to the organization. Through data analysis we have highlig hted how different variables such as CTR, mean age, CPC, i mpressions are important factor in determining the channel to be used. Also, the correlation between different numerical va lues such as CTR, CPA, acquisition, ROI ,impressions, etc. is understood and its importance has been highlighted.

The study also tries to implement different machine learni ng models which are specifically tailored such as the models help in giving the most important information to the advertise r. The different models and there evaluations have been com pared in order to select the best possible working model. The Regression models have been deployed with respect to ROI a s it is the most important factor for any successful advertisem ent. Further the time series analysis also help us understand t he working of advertisement at different duration with respec t to different channels and customer segments.

Finally, the study concludes by showing the importance a nd the need of data driven decision in today's changing econo mic sphere. It also tells how data science is useful in understa nding what are the better social media channels which give a better ROI and which type of ads should be published at what time of a year. If a new publisher wants to publish an ad in a social media channel, then this study will be helpful for his w ork.

VII. Future Scope

We aim to enhance our advertising model's effectiveness through real-time analysis. This involves extracting data from real-time sources via APIs such as the model proposed in the study [18] , followed by a cleaning process to ensure accuracy. Utilizing real time analysis tools, we can monitor KPI's like ROI and CPC allowing for dynamic adjustments to advertising strategies based on current performance metrics.Additionally, implementing statistical scaling of attributes will provide deeper insights into advertising performance, while large learning models can enhance forecasting accuracy. Ultimately, real time analysis will enable better Ad personalization by tailoring content to user behaviour, leading to increased engagement and conversion rates.

Additionally, the project could be extended for use of digital advertising agencies to enhance their businesses. The project can also be extended which can help in personalising the advertisement which is seen to be more effective in attracting the customer. The project could be used as a reference to build a specific large learning models which are very much use full for efficient forecasting. Further the project could also be improved by handling the large amount of dataset. This handling of dataset could be done by formulation of a statistical scaling of the different attributes. Additionally, the project could extend to creating different models with respect to different attributes such as creating a regression model for number of impression or number of conversions.

References

[1] Ra' Almestarihi, A Ahmad, R Frangieh, I Abu-AlSondos, K Nser, and Abdulkrim Ziani. "Measuring the ROI of paid advertising campaigns in digital marketing and its effect on business profitability". In: Uncertain Supply Chain Management 12.2 (2024), pp. 1275–1284

[2] Marco Gigli and Fabio Stella. "Multi-armed bandits for performance marketing". In: International Journal of Data Science and Analytics (2024), pp. 1–15.

[3] Amit Verma and Veena Vemuri. "Digital marketing is the necessity in current scenario". In: International Journal of Health Sciences 10 (2022).

[4] MR RAM BABU CHERUKUR et al. "A Study On Impact Of Digital Marketing In Customer Purchase In Chennai". In: The journal of contemporary issues in business and government 26.2 (2020), pp. 967–973.

[5] Dave Chaffey and Mark Patron. "From web analytics to digital marketing optimization: Increasing the commercial value of digital analytics". In: Journal of Direct, Data and Digital Marketing Practice 14 (2012), pp. 30–45.

[6] Hisham Noori Hussain, Tariq Tawfeeq Yousif Alabdullah, E Ries, and KanaanAbdulkarim M Jamal. "Implementing Technology for Competitive Advantage in Digital Marketing". In: International Journal of Scientific and Management Research 6.6 (2023), pp. 95–114.

[7] Erislan, E. (2024). Application of Big Data Analytics for Decision Making in Digital Marketing. Return: Study of Management, Economic and Bussines, 3(1), 22-27.

[8] Jose Ramon Saura. "Using data sciences in digital marketing: Framework, methods, and performance metrics". In: Journal of Innovation & Knowledge 6.2 (2021), pp. 92–102.

[9] URL: https : / / cran . r - project . org / web / packages / caret /caret.pdf.

[10] URL: https : / / cran . r - project . org / web / packages / e1071 /index.html.

[11] URL: https : / / www . rdocumentation . org / packages / stats /versions/3.6.2/topics/kmeans.

[12] URL: https : / / www . rdocumentation . org / packages / stats /versions/3.6.2/topics/glm.

[13] URL: https://www.rdocumentation.org/packages/forecast/versions/8.22.0/topics/Arima.

[14] Lucas, Morais., Gecynalda, Soares, da, Silva, Gomes. (2024). Neural Network-Enhanced Decision Support: Investigating Prediction Intervals for Real-Time Digital Marketing Return on Investment Data. doi: 10.5753/brasnam.2024.2232

[15] Moon, H., Lee, T., Seo, J., Park, C., Eo, S., Aiyanyo, I. D., ... & Park, K. (2022). Return on Advertising Spend Prediction with Task Decomposition-Based LSTM Model. Mathematics, 10(10), 1637.

[16] Kunekar, P., Usman, M., Veena, C. H., Singla, A., Anute, N., & Polke, N. (2024, May). Enhancing Advertising Initiatives: Using Machine Learning Algorithms to Engage Targeted Customer. In 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) (pp. 1433-1437). IEEE.

[17] Gudipudi, R., Nguyen, S., Bein, D., & Kurwadkar, S. (2023). Improving Internet Advertising Using Click–Through Rate Prediction. Applied Human Factors and Ergonomics International.

[18] J., Ignatius., S., Selvakumar., Spandana, Jsn., Subasri, Govindarajan. (2022). Data Analytics and Reporting API - A Reliable Tool for Data Visualization and Predictive Analysis. Information Technology and Control, 51(1):59-77. doi: 10.5755/j01.itc.51.1.29467