****

**Machine Learning (DSCI-6003)**

**Predicting Sales Using Linear Regression**

Professor: Md Moinuddin Bhuiyan

**Group Members:**

Krishna Sai Manas Karthik Bhiravajosyula

Pavan Reddy Boyapally

Jagadeesh Chandra Bose Mende

**ABSTRACT**

This study utilizes a dataset encompassing TV, radio, and newspaper advertising expenditures, alongside corresponding sales figures, to investigate their predictive relationship. Employing a regression model, we aim to discern the influence of each advertising platform on sales outcomes. Through thorough analysis, we seek to identify which advertising channels exhibit the strongest impact on sales and provide insights into optimizing marketing strategies for enhanced sales performance. This study contributes to the understanding of the effectiveness of different advertising mediums in driving sales, offering valuable guidance for businesses aiming to allocate their advertising budgets effectively.

Moreover, we delve into the consumer behaviour aspects associated with different advertising mediums. By analysing consumer preferences, attitudes, and responses to various advertising campaigns, we aim to uncover deeper insights into the mechanisms driving sales. Understanding consumer behaviour is crucial for tailoring advertising strategies that resonate with target audiences, ultimately leading to improved sales outcomes.

**INTRODUCTION**

In our project, we're diving into the world of advertising expenditures across various platforms like TV, radio, and newspaper, and how they affect sales. Through some in-depth regression analysis, our goal is to figure out which advertising channels really pack a punch when it comes to driving sales. We're excited to uncover some fresh insights into how different marketing strategies play out in the real world and help businesses make smarter decisions about where to invest their advertising budgets for maximum impact on sales. It's all about finding those hidden gems in the data that can make a big difference for businesses looking to boost their bottom line.

Furthermore, our study aims to explore the temporal dynamics of advertising effectiveness. By analysing seasonal variations and trends in consumer behaviour, we seek to provide nuanced insights into the optimal timing and frequency of advertising campaigns. Understanding these temporal aspects can empower businesses to capitalize on peak sales periods and maximize the return on their advertising investments.

**METHODS**

In this study, we employed linear regression analysis to analyse our dataset. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In our case, the dependent variable is the sales figures, while the independent variables include advertising expenditures on TV, radio, and newspaper. By fitting a linear regression model to the data, we aimed to understand the impact of each advertising platform on sales outcomes.

**DATA PREPROCESSING**

In this step we import all necessary libraries needed for our model and handle any missing values in our data.

**A white rectangular object with a black border

Description automatically generated**

**DATA INFORMATION**

The advertising dataset captures the sales revenue generated with respect to advertisement costs across multiple channels like radio, tv, and newspapers.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Exploratory Data Analysis**

**A screen shot of a graph

Description automatically generated**

**A screen shot of a graph

Description automatically generated**

**A screen shot of a graph

Description automatically generated**

This three scatter plot visualizations depict the relation between advertising and sales to see any linear growth in our model in all the three platforms which are ‘TV’, ‘Radio’, ‘Newspaper’.

We can see direct linear sales in Tv, Radio more than Newspaper from scatter plot which shows that our model is definitely a linear model and can offer valuable insights to advertising can lead to more sales in all the three platforms.

**A screen shot of a graph

Description automatically generated**

The box plot in our model detects any anomalies or outliers in our data and show distribution of data across the Data set. Our newspaper data in box-plot suggests there are two outliers or data points which can be ignored.

**Model Training**

In Model Training we import linear Regression classifier from sklearn.linear\_model and also split our Data into 70% training and 30% testing form sklearn.model\_selection to import train\_test\_split

and also convert any categorical variables into numerical variables using one hot code encoding

**A screenshot of a computer program

Description automatically generated**

**MODEL EVALUATION**

In model evaluation we print out y\_pred or predicted labels for our test model(X\_test) and also find mean squared error, calculate R-squared error which shows how much variance between actual and predicted target variable

**A screenshot of a computer

Description automatically generated**

**A white screen with text

Description automatically generated**

**Results and explanation**

**Mean Squared Error (MSE):** My MSE is approximately 3.798, which means, on average, my model's predictions are off by about 3.798 million dollars squared. Lower MSE values indicate better performance, so this is a reasonably good result.

**R-squared Score:** My R-squared score is approximately 0.861, indicating that approximately 86.1% of the variancein the salescan be explained by the advertising budgets in my model. This is a good fit for a regression model.

**Adjusted R-squared Score:** My adjusted R-squared score is approximately 0.851. Adjusted R-squared takes into account the number of predictors in the model, and it penalizes the addition of unnecessary predictors. This value suggests that my model is still performing well even after considering the number of predictors.

**DATA VISUALIZATION**

A screen shot of a computer

Description automatically generated

As this is Regression model, we fit the model with the best line by plotting the actual line vs predicted data points and from the visual there is slight variation of predicted data points from actual points due to model’s accuracy and the model is overlapped with some predicted points to exact actual points which shows model performs significantly better on unseen data and can be used for model deployment

**Conclusion**

Based on the analysis conducted, it appears that the linear regression model built to predict sales based on advertising budgets (TV, radio, and newspaper) has yielded promising results. With a mean squared error of approximately 3.798 and an R-squared score of around 0.861, the model demonstrates a good level of accuracy in capturing the relationship between advertising expenditures and sales revenue. Furthermore, visual examination of the actual versus predicted values suggests that the model effectively captures the underlying trends, as evidenced by the alignment of data points along the diagonal. Overall, the findings suggest that the constructed model can serve as a valuable tool for businesses in making informed decisions regarding their advertising strategies to optimize sales performance.