Application of Bayesian Modelling & Computations for Media Mix Models

Olatomiwa Akinlaja
2534648

Supervisors:
Dr. M.S. Mosia
Prof. R. Ajoodha

A research report submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

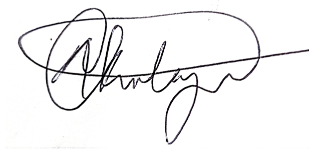
in the

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

5 June 2023

Declaration

I, Olatomiwa Akinlaja, declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.



Olatomiwa Akinlaja

5 June 2023

Abstract

Marketers seek to properly measure the effectiveness of different marketing channels. The purpose of this research report is to optimize a business's marketing budget through a combination of Bayesian computations and Media Mix Models. A Bayesian framework approach helps incorporate prior knowledge within the model description, as opposed to the traditional Frequentist media mix modelling approach that fails to account for uncertainties that might occur within multiple advertisement channels. Most of the Media Mix Modelling (MMM) research conducted fails to incorporate the knowledge domain and account for the uncertainties within a business's marketing channels before modelling. This research project attempts to solve the problem advertisers face when measuring the effectiveness of different marketing channels. The solution was achieved through the use of Bayesian methods for approximation, such as the Markov Chain Monte Carlo (MCMC), and media mix models for optimizing a marketing budget. Overall, we successfully applied Bayesian modelling to Media Mix Models for optimizing the marketing budget to generate the most sales. This was achieved by developing a Bayesian linear regression model that considers probability distributions, as opposed to training data alone. When compared to the Frequentist approach models, our Bayesian model provides reliable estimates and confidence intervals for sales based on the funds allocated to multiple marketing channels because it accounts for uncertainties and prior knowledge. The Bayesian media mix model was evaluated using Mean Square Error (MSE) and Mean Absolute Error (MAE) metrics to ensure the outcomes were reliable when compared to those generated by the Frequentist approach.

keywords:

Probabilistic Programming; Bayesian Computations; Media Mix Modelling; Bayesian Modelling; Budget Optimization

Acknowledgements

The support of the DSI-NICIS National e-Science Postgraduate Teaching and Training Platform (NETTP) towards this research report is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NETTP. A special gratitude towards my supervisors Dr. M.S. Mosia and Prof. R. Ajoodha, whose contributions has greatly assisted towards the formulation of this research report. A special thank you to Dr. H. Robertson, whose lecture material and suggestions helped towards the formatting of this research report.

Contents

List of Figures

List of Tables

List of Abbreviations

MMM	**M**edia **M**ix **M**odelling
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MSE	**M**ean **S**quare **E**rror
MAE	**M**ean **A**bsolute **E**rror
RMSE	**R**oot **M**ean **S**quare **E**rror
LR	**L**inear **R**egression
OLS	**O**rdinary **L**east **S**quares
SOM	**S**hare **O**f **M**arket
ML	**M**achine **L**earning
ROI	**R**eturn **O**n **I**nvestment

Chapter 1

Introduction

Businesses usually dedicate much of their resources to customer relations and advertising [14], [22]. There exists multiple advertising streams within a company, with the main focus being how to essentially reach as many customers as possible or acquire their engagement or interaction. It is crucial for a company to market their product or service in order to establish identity and increase brand awareness, therefore companies usually allocate a certain amount for marketing purposes.

The marketing budget of a company is a valuable resource, with the multiple advertisement streams and marketing avenues, it becomes difficult to actually measure the impact within these areas [22], [14]. Understanding how marketing activities impact sales is necessary and can help determine the market plans that will contribute towards the best outcome, with regards to sales, revenue, and returns. Additionally, one has to account for factors that impact the overall performance of the company, factors such as price, promotions, competitor activity, media activity, and economic conditions [14]. Accounting for these factors allows for effective allocation of budget across products, markets, and marketing programmes.

The allocation of the marketing budget across multiple advertising channels has proven to be one of the many problems usually experienced by businesses [10]. Failing to understand the relationship between marketing and sales could lead to major losses in sales, as discussed by an article which provided examples of major corporations which lost an average of 16% in sales, and up to 25% within 2 years [9]. It can be difficult to determine how funds can be allocated to marketing channels such as social media, TV, mail, or radio [3]. In order to avoid conjecture with regards to the allocation of the marketing budget, optimization is key. The optimization of the marketing budget is achieved through the estimation of effectiveness of the marketing channels.

In summary, the task of budget allocation optimization proves to be a complex one due to the multiple factors involved. The creation of a probabilistic model which has Bayesian computations for Media Mix Models, if declared correctly with regards to the channel parameters in question could potentially maximize customer acquisition and sales. This research report proposes Bayesian methods and computations for Media Mix Models. Furthermore, we aim to contribute towards the provision of a Bayesian based solution for marketing budget optimization.

We discuss the themes of this research project within the "Background" section 1.1, as well as methods in which authors have taken to solve the marketing funds optimization problem. In the Related Work Chapter 2, we provide a systematised discussion of the literature that addresses the problem of fund allocation for advertising. The Methodology, Chapter 3, provides an overview and description of the methods used within this project. We present the findings of the experiments performed in the Results and Discussion, Chapter 4, and discuss its significance, before concluding and discussing future work in Chapter 5.

1.1 Background

Media Mix Modelling is an analytical approach that uses historic sales data for the quantification of how marketing activity impacts sales over a period of time [21]. The use of MMM has become quite favourable with researchers who seek to improve overall budget allocation or spend for a specific business [13], [21], [16], [7]. There is great value in determining overall user interest and awareness, which has led to extensive research. The main reason would be to determine the most optimal marketing mix that would generate the highest interest from users and customers, while also being within the allocated budget.

Advertisers usually seek to understand the effectiveness of their media spend, which introduces many opportunities within MMM, for the optimization of budget allocations. MMM was developed using the traditional modelling approach based on linear regression [21], many changes have been implemented over the years to account for the uncertainties and dynamic behaviours within different advertisement channels.

Multiple studies have sought to optimize the marketing budget within different industries [21], [5]. The studies have managed to successfully create MMM optimization methods which have a noticeable impact on the business. Out of most of the studies which apply MMM only a few incorporate Bayesian modelling into their optimization solutions [17], [18], [8], [20].

The types of data used within the studies range from pharmaceutical historical data [5], to tourism data [21]. Regardless of the industry or type of data used, the studies all sought to estimate the key influences of marketing mix elements on overall sales.

Most studies aim to identify the key factors which influences business outcomes. It proved necessary to evaluate the influences the key factors have, for the sake of business optimization and determining the best marketing solutions.

An example of one of the models proposed by one of the studies is an optimal multiple regression model in Equation 1.1 [5], where the share of market (SOM) value is determined by multiple relationships between different factors, and is defined as follows:

$$\begin{aligned} SOM = {} & Constant + a_1 \times Penetration_1 + a_2 \times Penetration_2 + a_3 \times Price\_index + \\ & a_4 \times Doctors + a_5 \times Pharmacists + a_6 \times Adstock(TV_1) + a_7 \times Adstock(TV_2) \\ & + ... + a_n \times Adstock(TV_n) + b_1 \times Adstock(TV_Competitor_1) + \\ & b_2 \times Adstock(TV_Competitor_2) + ... + b_m Adstock(TV_Competitor_m) \end{aligned} \tag{1.1}$$

The different factors included the penetration levels, recommendation based on brand level, as well as marketing activities of competitors. The measure of the lagged or prolonged effects on advertising, known as adstock was also included. Generally, the SOM is determined by a linear combination of all of these factors.

A review of multiple MMM related studies, showcase the various legacy methods developed and used for the MMM. These legacy methods were used without the consideration of uncertainty or incorporation of the prior knowledge domain. Most of these legacy methods involves a basic linear model represented by Equation 1.2

$$y = \beta_1 x_1 + ... + \beta_j x_j + ... + \beta_p x_p + \epsilon. \tag{1.2}$$

This basic approach can help determine the response variable in question (y), which is represented as a linear combination of multiple factors which affects the response. These factors are influenced by the β. It is assumed that the error term (ϵ) is normally distributed, and the dependent and independent variables are standardized ($\epsilon \sim N(0, \sigma)$).

The aim of these methods is to decompose the coefficient of determination (r^2) for regression models [13]. Various real and simulated datasets were used for the sake of demonstrating the performance of these legacy methods [22], [14].

After reviewing the studies mentioned above, we can conclude that most research done on MMM incorporate the use of a traditional linear regression approach [14], they fail to account for uncertainties within the different advertising channels. They also fail to encode prior beliefs within the marketing channels of advertising [10].

This research project sought to combine Bayesian modelling together with Media Mix Models in order to account for the uncertainties, as well as providing a comparison between the Frequentist and Bayesian approach. The addition of Bayesian methods within MMMs would ensure additional value, through the incorporation of prior knowledge and making modelling assumptions more explicit.

1.2 Problem Statement

Businesses are undergoing a digital transformation. It is through this digital transformation that companies have developed digital methods to reach customers through advertising. Therefore advertisers seek to properly measure the effectiveness of the different marketing/media channels. Since the adoption of additional advertisement streams, it has become necessary to be able to measure the effectiveness of online and offline media and the contributions they have towards achieving business goals.

Marketing/Media Mix Models (MMM) is a solution implemented to achieve the goal of measuring market effectiveness. There are challenges and limitations involved with regards to the development of a model that can be used. Essentially a well defined MMM is determined by how well it can contribute towards the decision making process of a business.

1.3 Research Question

Given the multiple marketing channels of a business, with the purpose of optimising the funds allocated to each advertisement channel, while also generating peak customer interaction and optimal return on investment. This research project sought to respond to the following research questions:

- How can the relationship between advertising channels and sales be modelled?
- Can a Bayesian Media Mix Model perform better than Frequentist approaches with respect to the MSE and MAE?

1.4 Research Aims and Objectives

1.4.1 Research Aim

The aim of this research project is to apply Bayesian modelling and computations on Media Mix Models for determining the effects of advertising expenditure.

1.4.2 Objectives

The aim can be achieved by fulfilling the following objectives:

- Determine appropriate prior distributions to represent the prior knowledge domain of advertising expenditure within the parameter values of the model.
- Choose with explicit assumption, a likelihood distribution which explains the data generation process.
- Perform Bayesian inference such as MCMC, to find posterior distribution over the parameter values of the model.
- Provide a MSE and MAE comparison of the Bayesian Media Mix and Frequentist approach methods.

1.5 Limitations

It can prove to be quite difficult to implement Media Mix Models within a Bayesian framework since inference and learning in the Bayesian paradigm is mostly NP-Hard. Problems might arise when we try to forecast reasonable estimates for sales based on the funds allocated to the different marketing channels.

1.6 Overview

Marketing Mix Modelling is a statistical model used by advertisers [3]. The marketing mix is the most fundamental concept of marketing. It is essentially a statistical approach where quantified marketing activities over a period of time are linked mathematically to a dependent variable, such as sales or revenue [21]. This research project proposes a Bayesian framework combined with MMM, which should prove to be superior when compared to the usual frequentist methods used within previous studies.

We will explore Bayesian methods and discover ways to efficiently combine it with MMM in order to ensure the optimization of the market spend of a business, within multiple marketing channels. A description of the processes we intend to follow within this research project is provided in Chapter 2, as well as the data, analysis description, limitations, models and evaluation metrics involved.

Chapter 2

Related Work

This chapter provides a systematised discussion of the literature that addresses the problem of fund allocation for advertising. We focus on the application of Bayesian modelling within Media Mix Models by reviewing studies that have used Bayesian methods to estimate the effects of different media on sales, while also providing a discussion of how other studies sought to tackle our problem statement.

2.1 Media/Marketing Mix Modelling

Media/Marketing Mix Modelling is a solution used to measure the performance of different marketing channels, the measurement of the overall return on investment of advertisements is achieved through the performance of statistical analysis on historical data [16]. MMM methods are usually the most commonly used when it comes to modelling historical data for the presentation of marketing and advertising variables extracted from business metrics' functions [5], [13], [14], [18].

MMM can be compared with attribution models, which are essentially methods that provide a quantification of advertising inputs attributed to revenue [22]. While attribution modelling performs analysis on multiple channels and campaigns and how they contribute towards sales at an individual user level (A bottom-up approach) , MMM finds correlations between marketing investments and returns at a macro level through statistical modelling.

For example, over a time period, you have sales and a number of marketing activities which involves campaigns on different social media platforms for instance. The comparison of how changes occur between the different campaigns can create a picture of the impact each of the marketing activities have on the business. This would in turn ensure the optimization of returns from the funds allocated for all marketing activities involved.

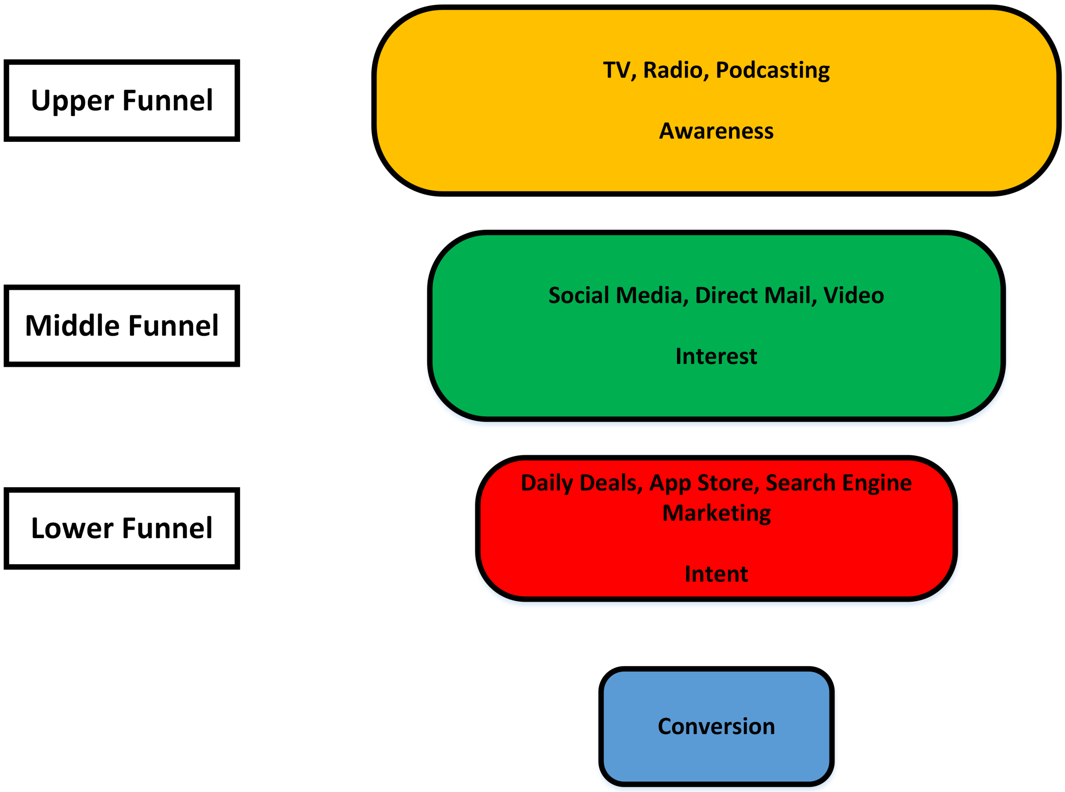


FIGURE 2.1: The Marketing Funnel for Advertising

A description of the varied and complex mixture of advertising channels within a business is illustrated within Figure 2.1. It consists of a upper, mid, and lower funnel which showcases the areas in which a business might choose to advertise. The upper funnel represents mostly offline channels which are meant to build awareness. The middle funnel consists of mostly online or a mixture of online and offline channels which build interest. The lower funnel represents the direct response channels which are usually used to measure intent of customers or consumers. Situated at the bottom of the funnel is the conversion box, which represents the turnover rate of overall customers reached as well as profits made.

In order to have an optimal allocation of funds for each marketing activity, the use of a complicated solution is required to achieve the goal. This sounds much easier theoretically but practically complicated. There are a huge amount of variables that would need to be accounted for such as promotions, market conditions, competitor activity, seasonality. There is also the likelihood of the business running multiple marketing campaigns across multiple channels. All of the variables in question would have to be tracked. The volatility of marketing spend (funds allocated) during the campaigns would have to be factored in as well.

2.1.1 Current Optimization Approach

Marketers make use of marketing response curves and marginal revenue curves similar to Figure 2.2. The response curve shows the maximum amount required to allocate to a marketing channel before revenue diminishes [1]. The information the response curves provide are translated into budget allocations for each of the marketing channels. This ensures that the most Return On Investment (ROI) is achieved. Marketing response curves and marginal revenue curves are currently used as the industry standard [1]. A Bayesian approach is then applied to the response curve to introduce uncertainty.

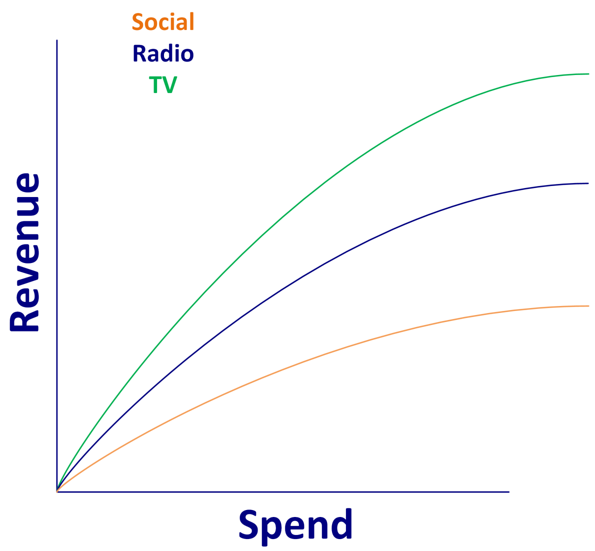


FIGURE 2.2: Response Curve Example

2.1.2 Difference between Media Mix Modelling and Marketing Mix Modelling

In order to perform Media Mix Modelling, we would have to perform Marketing Mix Modelling therefore Marketing Mix Modelling and Media Mix Modelling are essentially the same from a media perspective, if the data involved is media related [14].

2.1.3 Advantages of using Media Mix Modelling

- Encourages business growth through the transformation of marketing practices.
- Drives change within the digital market, as well as the improvement of business marketing effectiveness.
- Introduces transparency through custom models which cater for analyst subjectivity and bias.

2.2 Bayesian Computations for Media Mix Models

Studies have proposed Bayesian approaches in order to address the many challenges faced by traditional MMM approaches [19], [12]. These studies attempt to develop a model that is superior to traditional frequentist approaches in terms of both accuracy and interpretability. They mostly use both real-world and simulated marketing datasets and multiple forms of Bayesian optimization methods to achieve the outcome of budget allocation. What most of the studies fail to account for is how the Bayesian models perform when compared with the Frequentists' MSE and MAE evaluation metrics.

2.2.1 Why a Bayesian Framework Approach

The Bayesian framework approach allows for the combination of prior knowledge within the model description, this is achieved through the parameters being represented as random variables based on their posterior distribution. Another reason for choosing the Bayesian framework is due to its extreme flexibility and ability to model complex scenarios. For example, using non-normal distributions to estimate situations which require non-linear variables with unknown parameters but are represented as non-linear models [7].

A combination of domain expertise and statistical skill are the factors required during the building of a model. The knowledge from these factors are incorporated into objectives for the determination of result usefulness [11]. Bayesian models consists of defining characteristics which include the quantities parameter, this parameter contains the description of unknown quantities through probability distributions. The second characteristic includes the reallocation of probabilities, this is when the values of the parameters conditioned on the data are updated using Bayes' theorem [11].

Essentially, we sought to explore bayesian inference, in order to acquire probability distributions. A combination of probability distributions has to take place which is a form of statistical inference [11].

2.2.2 The Benefits of a Bayesian Media Mix Model

Ordinary Least Squares (OLS) have been known to be used traditionally for fitting MMM models, this frequentist method usually provided a reasonable coefficient point-wise estimate [17]. But over time Bayesian methods have become more favourable since it was first introduced by authors at Google [10]. The benefits of the Bayesian approach over previous frequentist approaches are as follows:

- Prior knowledge can be incorporated ($P(\theta)$ in Equation 2.1)
- Prior knowledge can be optimally combined with data ($P(Y|\theta)P(\theta)$ in Equation 2.1)
- There is level of certainty ($P(\theta|Y)$ in Equation 2.1)

Prior knowledge can be incorporated into the model, which can be quantified with mathematical distributions appropriately through the Bayesian approach. The Bayesian methods augments human knowledge as opposed to overruling it. Essentially, Bayesian approach provides a description of uncertainties as well as ways of dealing with and reducing uncertainties from the model outputs [17].

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \tag{2.1}$$

2.2.3 The Intractability of a Bayesian Approach

Bayesian Inference is incredibly difficult and is mostly achieved through the use of Universal Inference Engines [11]. Estimations are usually determined through the help of approximation methods such as the Monte Carlo Markov Chain [20], as closed form estimations are usually impossible to compute.

2.3 Model Specification

Within a proposed regression model where Y_t as the dependent variable represents the sales of a company, and the independent variables in the Equation 2.2 represents the budget allocated to the marketing channels involved.

- Y is a representation of overall sales
- A is a representation of advertising
- α and β represents coefficients that can be estimated.
- $_t$ is a representation of different time periods.
- ϵ_t is a representation of possible errors within estimations of Y_t

Y_t within the model:

$$Y_t = \alpha + \beta A_t + \epsilon_t, \tag{2.2}$$

is assumed to follow a normal distribution independently, the equation can also be estimated through regression.

Different patterns of the response to advertising and marketing are classified by authors within [13]. The most common four effects are current, carry-over, shape, and competitive effects. The other effects include, dynamic, content, and media effects.

We will be referring to the carry-over and shape effects description from the study.

The carry-over effect is described as the area within effects of advertising that occurs in time periods after the advertising pulse. It is usually caused by delayed customer response or delayed exposure to the ad, as well as many other delayed customer factors with regards to purchasing the product which leads to generation of sales.

The shape effect is referred to as the change in sales in response to either the high intensity of advertising or the increase of advertising investments, all within an identical period.

The basic linear model can be defined as:

$$Y_t = \alpha + \beta_1 A_t + \beta_2 P_t + \beta_3 R_t + \beta_4 Q_t + \epsilon_t, \tag{2.3}$$

and only captures some of the advertising effects by being a function of one or more predictor variables.

The multiplicative model, which allows for the interaction of effects amongst predictors:

$$Y_t = Exp(a) \times A_t\beta_1 \times P_t\beta_2 \times R_t\beta_3 \times Q_t\beta_4 \times \epsilon_t, \tag{2.4}$$

Using a logarithmic transformation, it can be further transformed to make it even simpler

$$Y_t = \alpha + \beta_1 log(A_t) + \beta_2 log(P_t) + \beta_3 log(R_t) + \beta_4 log(Q_t) + \epsilon_t. \tag{2.5}$$

The study provides a look into MMM concepts and model interpretation. It provides an overview of the advertising effects that occur, followed by a discussion of the overall strengths, limitations, and complexities of different models which capture such effects [13].

There are a few studies that incorporate Bayesian modelling within their Media

Mix Models, one of them involves the application of Bayesian methods for media mix modelling with carry-over and shape effects [10]. Different ways of modelling carry-over and shape effects of advertising through flexible functional forms of MMM are proposed by the study [10]. A general model and estimation methodology which can be applied to any MMM is investigated by the study. This is done to address the carry-over and shape effects issue.

The study applied the estimated model on simulated data, and makes use of attribution metrics Return on Ad-Spend (ROAS) to configure the optimal media mix from the simulated data. The paper specifies some general hyperparameters as well as other default variables, and choice of priors. The study also describes the impact of sample size on estimation accuracy.

2.3.1 Carry-over Effect

The adstock function is used as a means of transforming the time series representation of the media spend per respective channel, for the purpose of modelling the carry-over effect of advertising. When denoted as

$$adstock(x_{t-L+1,m,...}, x_{t,m}; w_m, L) = \frac{\sum_{l=0}^{L-1} w_m(l) x_{t-l,m}}{\sum_{l=0}^{L-1} w_m(l)}, \tag{2.6}$$

It is necessary to assume the maximum duration (L) of carry-over effect for the marketing channels. L assists with the adstock transformation weight estimation, if chosen adequately. w_m is a non-negative representation of the weight function

Different functional forms can be used for the weight function, such as the "geometric decay". An adstock effect with a geometric decay makes the assumption that that ad exposure and advertising effects are directly proportional and peak around a similar time period. α_m is a representation of the ad-effect retention rate from consecutive periods. The function is denoted as

$$w_m^g(l; \alpha_m) = \alpha_m^l, l = 0, ..., L-1, 0 < \alpha_m < 1. \tag{2.7}$$

In instances where there might be a delay with regards to the building up and peak effects of some marketing channels, θ_m represents the delay of the peak effect. The "delayed adstock" function w^d is considered and denoted as

$$w_m^d(l;\alpha_m,\theta_m) = \alpha_m^{(l-\theta_m)^2}, l = 0, ..., L-1, 0 < \alpha_m < 1, 0 \leq \theta_m \leq L-1. \tag{2.8}$$

2.3.2 Shape Effect

The curvature function is used as a means of transforming the marketing spend representation per respective channel, for the purpose of modelling the shape effect of advertising. The Hill function is the curvature function of choice [10].

$$Hill(x_{t,m};\kappa_m, S_m) = \frac{1}{1 + (x_{t,m}/\kappa_m)^{-S_m}}, x_{t,m} \geq 0 \tag{2.9}$$

The shape parameter (a.k.a the slope) is represented by $S_m \geq 0$. Since the Hill function approaches 1 as x approaches infinity, $K_m > 0$ is the half saturation point [10]. The multiplication of the Hill function with a regression coefficient β permits the allocation of different maximum effects for different marketing channels [10]. This rewrites the shape transformation $\beta Hill(x)$ as

$$\beta_m Hill(x_{t,m}) = \beta_m - \frac{\kappa_m^{S_m}\beta_m}{x_{t,m}^{S_m} + \kappa_m^{S_m}}. \tag{2.10}$$

2.3.3 Combining the Carry-over and Shape Effect

Two possible approaches are proposed by [10] for the combination of carry-over and shape effects. One is in terms of applying the shape transformation before the adstock transformation to the time series representation of media spend [10]. The other way involves performing the reverse by applying the adstock transformation before the shape transformation to the time series representation of media spend. The study models the sales y_t with the equation defined as:

$$y_t = \tau + \sum_{m=1}^{M} \beta_m Hill(x_{t,m}^*;\kappa_m, S_m) + \sum_{c=1}^{C} \gamma_c z_{t,c} + \epsilon_t, \tag{2.11}$$

where $x^*_{t,m}$ represents the adstock function in 2.6. There is also representations for baseline sales τ, the effect of control variables z_c, as well as the introduction of uncorrelated white noise ϵ_t.

2.3.4 Estimating the Bayesian Model

The model defined in 2.11 was estimated using a Bayesian approach, for the purpose of utilising prior knowledge. When comparing a Frequentist approach to a Bayesian Inference approach, the former maximises the likelihood through the most likely parameter values, which can be defined as:

$$\Phi = argmax\ell(y|X, Z, \Phi). \tag{2.12}$$

The latter which is based on the posterior distribution of the parameters, considers the model parameters as random variables, and is defined as:

$$p(\Phi|y, X) \propto \ell(y|X, Z, \Phi)\pi(\Phi). \tag{2.13}$$

The prior knowledge accumulated with related or previous MMM is necessary for estimating the model through a Bayesian approach. The study used Bayesian approach to estimate the model using MCMC algorithms. [10] emphasises how it is important to incorporate additional information through prior knowledge, because of the low informational content that is available within a single MMM dataset.

Since the study dealt with non-linear model parameters, it becomes non-trivial to obtain the parameter estimates through the minimization of the residual sum of squares or the maximization of the log likelihood. Furthermore, transformations were performed on the time series representation of the budget of a channel, the transformation was through the adstock function to model the carry-over effect, as well as the media spend through a curvature function for the modelling of advertising shape effects. In the end, due to the variance of the parameter estimates, the optimal media mix inspired by the model generated had a large variance [10].

Chapter 3

Research Methodology

This chapter provides an overview and description of the methods used within this research report. We provide the research design, a description of the Frequentist and Bayesian approach, as well as a description of how Bayesian modelling was applied in order to determine the optimal fund allocation that would generate the most sales.

3.1 Research Design

This research report is classified as quantitative, due to the analysis of mostly numeric data, we will focus on using Bayesian methods to explore Media Mix Models for the optimization of total sales, based on the multiple advertisement channels. Depending on the marketing budget of a business, marketing plays a major role in achieving growth within a business.

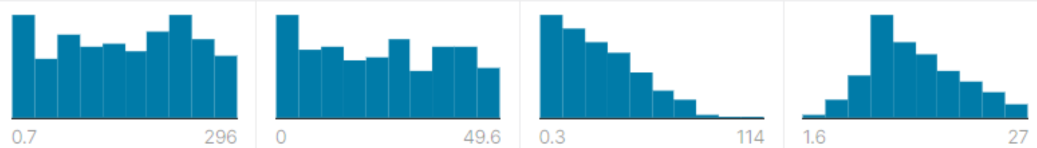
The measurement and optimization of marketing effectiveness is a key Data Science task used to grow the brand of a business while also achieving profitability. The measurement of the effectiveness of marketing is critical for profitable growth. We can measure the effectiveness of marketing by the cost to acquire a customer (CAC_{mt}), amount of money spent by a particular marketing channel x_m, divided by the total number of customers acquired through the channel y_t in

$$CAC_{mt} = \frac{\sum_{m=1}^{M} x_m}{y_t}. \tag{3.1}$$

The equation 3.1 is simply an over simplification of how to determine the optimal allocation spend for each channel to achieve the required sale amount.

3.2 Data

The data used in this research report is illustrated in the Figure 3.1, The data consists of two hundred rows and four columns, it is made open source and available on Kaggle. It consists of the amount spent for each marketing channel. The channels include TV, Radio, and Newspaper. The sales column represents the overall sales attained after the allocation funds for each advertisement channel. Each row represents a monthly instance, therefore the data is a collection of monthly budget allocations for multiple marketing channels, as well as the overall sales recorded within a span of two hundred months. We assume that the sales value is recorded in millions and the funds allocated to the different marketing channels in thousands, however, we ensured that the values were scaled, to ensure that the channels values and the sales values fall within the same range.



# TV	# radio	# newspaper	# sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

FIGURE 3.1: Overview of Data

In order to solve the optimization problem of optimizing sales. We want the model to learn a mapping from the features (TV, Newspaper, Radio) to the target (Sales). It is a regression task since the values are continuous.

3.3 Assumptions and Definitions

An explanation of the assumptions used within this reports is provided within this section, as well as definitions of relevant key terms. We discuss assumptions with regard to data production, sampling techniques, as well as model assumptions.

3.3.1 Data, Sampling, and Modelling

The advertising data used in this report is secondary data that was recorded to show the overall trend of sales over time, we assume that each record within the dataset represents a month of sales generated based on the amount allocated to each of the respective marketing channels. In order to model the data we assume that the values recorded represent actual budget allocations and use weakly informative priors to represent our beliefs about the marketing channels involved. Model parameters are treated randomly, while also inferring their posterior distribution.

3.3.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a class of algorithms which is used for sampling from probability distributions, the more steps taken equal more samples from the probability distribution, which results in a higher accuracy. The Markov chain of samples' third equilibrium should be equal to the target distribution that is required. In other words, MCMC tries to construct a set of samples that can search over the range of possible outcomes. The density of the distribution should be proportional to the total amount of time spent in each interval.

3.3.3 Uncertainties

There is necessity for building models with uncertainty. How do we know good uncertainty? Uncertainty is not so unusual and usually arises from two sources of lack of knowledge. In this report we assume that there are some uncertainties that arise within marketing data and model these accordingly using our proposed Bayesian model.

3.4 Methods

The different methods adopted within this report is discussed in this section, we provide a discussion of the Frequentist and Bayesian approach, as well as the incorporation of priors and model approximation.

3.4.1 The Frequentist Approach

Why would we want to adopt Bayesian linear regression? First of all we want to make sure there is an extension from the Frequentist approach to cover the Bayesian approach. The Frequentist approach essentially focuses on the fit of the model to the data, while the Bayesian approach aims to include a prior distribution for the model parameters, followed by the calculations of the likelihood of the data as a means of sampling the posterior for the model parameters. The prior, likelihood, posterior of the parameters, are essentially integrating uncertainty.

An Ordinary Least Squares (OLS) regression is an example of a Frequentist approach to modelling a MMM [4]. We essentially minimize a determined loss function, based on the training data alone, there are no priors modelled and no updating to posterior, all that is done is the estimation of the model parameters from the data. In the Frequentist interpretation, we can model the sales y, as linear combinations of the TV, Radio, and Newspaper features, represented as $\beta_0 + \beta_1(TV) + \beta_2(Radio) + \beta_2(Newspaper)$. An error term is included to account for random sampling noise in

$$y = \beta_0 + \beta_1(TV) + \beta_2(Radio) + \beta_2(Newspaper). \tag{3.2}$$

We used multiple standard machine learning models as a baseline in order to compare with our Bayesian model. Extra Trees, Random Forest, Gradient Boosted, Elastic Regression, Linear Regression and SVM models.

3.4.2 The Bayesian Approach

Bayesian methods are built upon Bayes' Theorem. Why would we want to use Bayesian methodologies with machine learning? Firstly it is more flexible than Frequentist approaches, secondly, there are many problems that are difficult to solve with Frequentist approaches. Bayesian methodologies can be easily updated as new information is available which enables us to solve probability problems that simple Frequentist approaches cannot solve. Bayes Theorem is defined as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{3.3}$$

where $P(A|B)$ is known as the Posterior $P(Model|NewData)$, the probability of sales given the budget allocated to the different marketing channels. $P(B|A)$ the Likelihood $P(NewData|Model)$, the likelihood probability of the budget allocated to the multiple marketing channels and their influence on overall sales. $P(A)$ the Prior $P(Model)$ probability of sales generated, and $P(B)$ the Evidence $P(NewData)$ probability of the marketing channels. The prior should have no information from likelihood, and the evidence term is usually just a standardization to measure closure and is usually independent of the model [6]. In our case we are interested in finding the probability distribution $P(sales|(TV, Newsapaper, Radio))$ through a combination of the other probability distributions [11].

3.4.3 Bayesian Model for Linear Regression

The formulation of a linear regression model which considers probability distributions instead of just the training data is necessary. Since we will be working with distributions of uncertainty, we have to formulate the predictions we are making from the perspective of probability distributions. The prediction we want to make is the response feature. We assume its conditional is Gaussian distributed (N), and that its conditional expectation ($\beta^T X$) is going to be the coefficient of the model transposed by the predictor features of the model.

As observed within equation 3.4, $\beta^T X$ is essentially predictions from the linear regression model, while σ^2 represents the variance within the estimate. The constant scalar is applied to an identity matrix $\sigma^2 I$ in order to get it to the matrix form similar to $\beta^T X$.

$$y\ N(\beta^T X, \sigma^2 I) \tag{3.4}$$

Posing the estimation problem of the parameters in a Bayesian perspective provides a simple notation which simplifies the problem, as seen in the equation 3.5.

$$P(\beta|y,X) = \frac{P(y,X|\beta)P(\beta)}{P(y,X)} \tag{3.5}$$

In equation 3.5, $P(\beta|y,X)$ represents the posterior of model parameters, it is the conditional distribution of the model parameters given the response and predictor features. $P(y,X|\beta)$ is the likelihood of the response features given the model and predictors features, the conditional distribution of the response and predictor features given the model. It is data driven which causes the likelihood to overwhelm the prior distribution as the number of sample data increases. $P(\beta)$ is the prior probability of the model parameters, $P(y,X)$ is the normalizing constant independent of the model, also known as the Evidence, the normalization constant which is independent of the model parameters.

Given a simple linear model $y = X\beta + \epsilon$ which has known data X, an $n \times p$ matrix, we assume the response variable y is generated by a linear combination of all the features. The weights on the linear combinations are given by the parameter vector β. β is essentially our goal to solve for.

There are some assumptions associated with the mode, such as errors being normally distributed with mean zero, and variance σ^2.

$$\begin{aligned} \epsilon_i \; N(0,\sigma^2) \\ y_i \; N(\beta^T x_i, \sigma^2) \end{aligned} \tag{3.6}$$

It is also assumed that each of the y_i are normally distributed, with mean $\beta^T x_i$, and variance σ^2.

We are trying to maximise the posterior probability $P(\beta|y)$ in 3.7.

$$\beta_{MAP} = argmax_\beta P(\beta|y) \tag{3.7}$$

β is an unknown parameter vector we are trying to solve for, y is some known response variable which we have. The posterior distribution is asking the question:

given that known parameter vector y, what is the probability of observing some instance of the β. The natural thing to do is to maximize the posterior, because if the β which maximises the posterior are found then it answers the question regarding which setting of β is the most likely, given the β that we observe.

These instances of β are called β_{map}, maximum a posteriori. The β which maximises the posterior probability. How do we solve this? We rewrite the posterior in terms of the three probabilities using Bayes' theorem, illustrated in

$$\beta_{MAP} = argmax_{\beta} \frac{P(y|\beta)P(\beta)}{P(y)} \tag{3.8}$$

Since the purpose is to optimize β, and $P(y)$ has nothing to do with β, it can therefore be ignored. This works in our favour since computing the probability of observing y would be unclear or difficult to compute. The equation can then be represented as:

$$\beta_{MAP} = argmax_{\beta} P(y|\beta)P(\beta) \tag{3.9}$$

Maximising/Minimising something is the same as doing it to the log of something. Using the log terms would be less computationally expensive, therefore the equation can be represented as:

$$\beta_{MAP} = argmax_{\beta}[\log(P(y|\beta)) \log(P(\beta))] \tag{3.10}$$

In equation 3.10, $P(y|\beta)$ which represents the likelihood can be computed based on the assumptions stated. The prior $P(y)$ which is the prior understanding of the world for β before the data is observed, can be determined based on media channel behaviour.

3.4.4 Approximation

After the introduction of the concept of Bayesian Linear Regression, we established how powerful the system is and how we incorporate priors as well as how the distribution of parameters can be estimated. However, the problem cannot be solved directly, we would need to sample the posterior in order to solve it. This can be achieved using Markov Chain Monte Carlo (MCMC).

Given that we do not know a distribution, in order to estimate the distribution, we can try to construct a number of samples that are capable of searching a possible range of outcomes. This is performed while spending an amount of time proportional to the density distribution within each interval. MCMC is a class of algorithms used for probability distribution sampling, the Markov chain of samples' equilibrium focuses on the target distribution. This means each Markov step is sampling the target distribution, more steps equals more samples which then equals more accurate assessment of the distribution. To summarize at a high level, MCMC uses samples to provide an approximation of the posterior distribution [11].

Essentially, we are trying to achieve a posterior with our Bayesian Linear Regression, and we cannot get directly to it for a continuous variable, but with MCMC we are able to sample from the distribution and if we do enough samples we are able to know what the distribution is.

3.4.5 Incorporating Priors

One of the hardest aspects of Bayesian modelling is with regards to choosing priors. We worked with around two hundred data points within the dataset, and incorporated Bayesian inference, unlike Frequentist machine learning approaches which are based entirely on the data and perform inference in the absence of priors.

3.5 Analysis

In the previous section, we introduced the fundamental concepts of Bayesian Regression, where we made some comparisons to the Frequentist approach. Then we discussed the MCMC approach, since we cannot solve for the model parameters in a regular way. Due to the fact that the solution is not closed form, and for continuous predictor features, we are in a situation where it cannot be solved for directly. The problem (posterior distribution) would have to be sampled, which led to the introduction of a MCMC approach to achieve the goal. It is important to use additional metrics to help with regards to the summarization of the distributions, since we will be sampling the distributions of the model parameters directly.

We considered multiple factors when trying to determine if Bayesian Media Modelling can perform better than Frequentist based approaches. We had to explore ways to present how the Bayesian model performed as opposed to just providing a MSE and MAE comparison alone.

3.5.1 Traceplots

Assuming we run a normal linear regression model on the training data, it simply would not be able to effectively minimise the sum of squares of the errors. Since we are getting the actual distributions of the intercept, slope and sigma terms. It is recommended to observe the traceplot in order to observe the samples along the Markov chain of the intercept, slope and sigma. Essentially, the traceplot provides a summary of the processing happening within MCMC [11].

3.5.2 Sampled Model Parameter Summary Statistics

We get to the distributions by sampling them with the MCMC approach, another powerful thing to do is to summarize the distributions statistically, as opposed to just visualizing the sample with a traceplot or kernel density smoother (to get the probability density function). We can calculate the following statistics (Expectation, Standard Deviation, 5 and 95 interval), these are not the regular confidence intervals since we are incorporating a Bayesian approach which is more intuitive. It is a credible interval, a direct measure of the uncertainty in the model parameters, e.g. there is a 95% probability that the parameter will exist within the specified interval.

This fact is the opposite with the Frequentist approach, the confidence intervals within the Frequentist approach are the uncertainty in the interval, not within the actual model parameter. It is a measure of whether or not the true parameter will be included within the interval, considering that if we re-sampled through the Frequentist approach, there is a 95% chance one of the intervals would include the true value.

3.5.3 Error Metrics

We proposed two standard metrics for evaluating better and worse allocations, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Both involve the difference between predictions and the true values. MAE deals with the average of the absolute value, while RMSE deals with the square root of the average of the squared differences. We aim to determine which of the metrics will be more interpretable within our modelling solution, since RMSE might heavily penalise larger errors [11].

Chapter 4

Results and Discussion

This chapter presents the findings of the experiments performed and discusses their significance. The chapter shows the outcomes achieved after the methodological process taken to apply Bayesian modeling to media mix models. We provide a comparison between the Frequentist approach and a Bayesian linear regression approach. We also compare multiple standard machine learning models and how they perform in regard to predicting sales.

4.1 Frequentist Approach

4.1.1 Data Exploration

We started by exploring the data to see how budget allocations are distributed within different channels, as well as the sales distribution over a period of two hundred months. Let us assume that the values recorded within the advertisement channels are in thousands, and those within sales are in millions. Before feeding the data into our different models, we ensured that the data was cleaned and transformed.

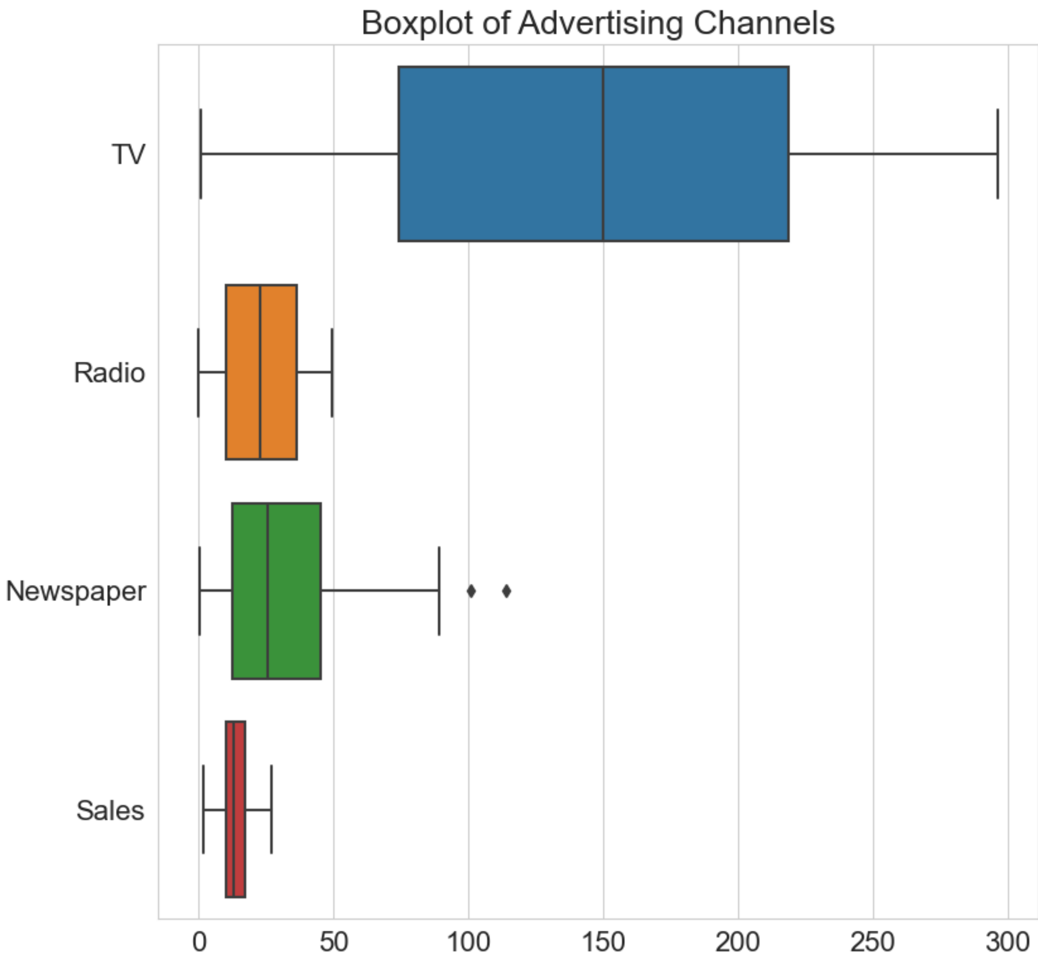


FIGURE 4.1: Box Plot of Marketing Data

Overall, we can see that the TV channels records far higher values when compared to the others, as seen in Figure 4.1.There is a clear class imbalance within the multiple marketing channels. This shows that overall, there is a higher budget allocation to the TV marketing channel when compared to Radio and Newspaper.

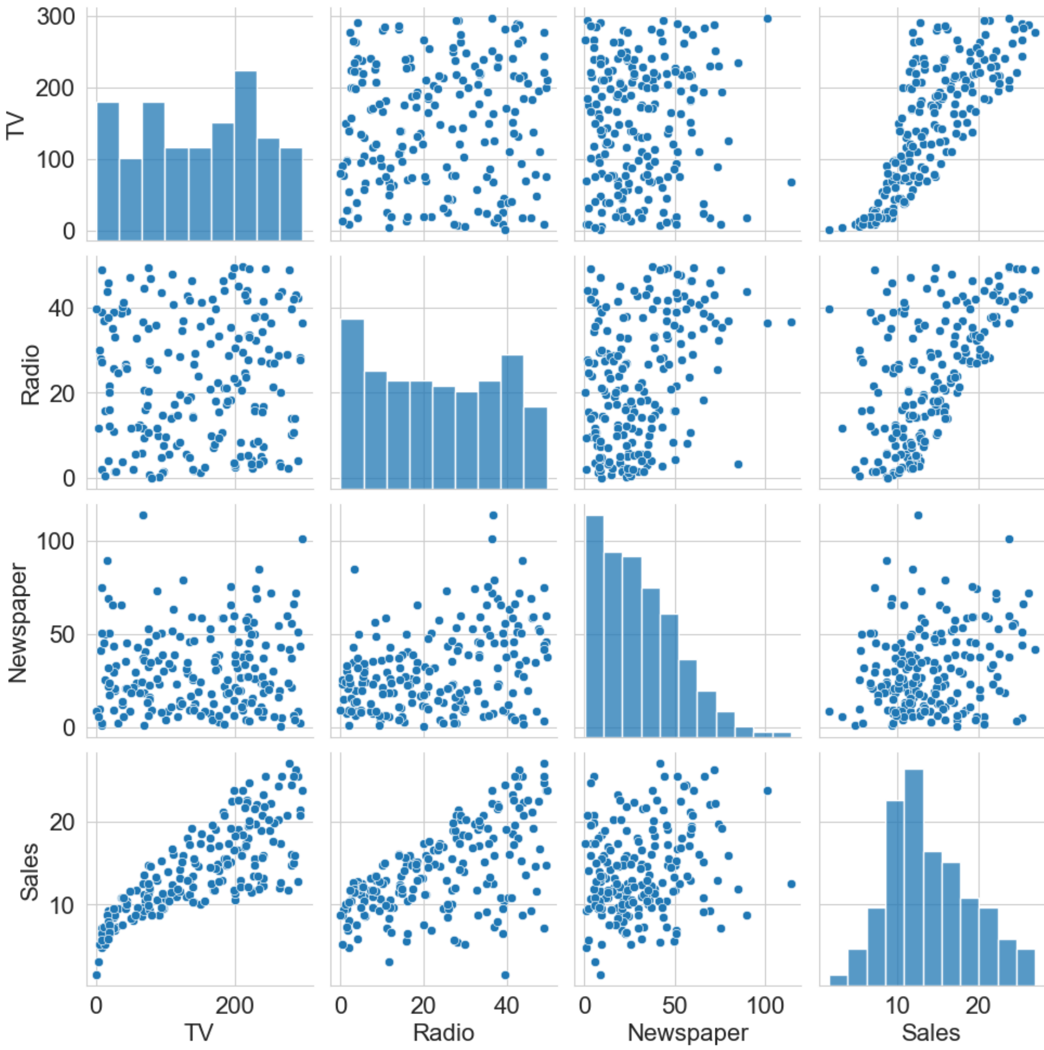


FIGURE 4.2: Pair Plot of Marketing Channels

The pairplot in Figure 4.2 shows the dispersion of the recorded data. We can observe the overall dispersion of the budget allocation with regards to the sales generated throughout a period of two hundred weeks.

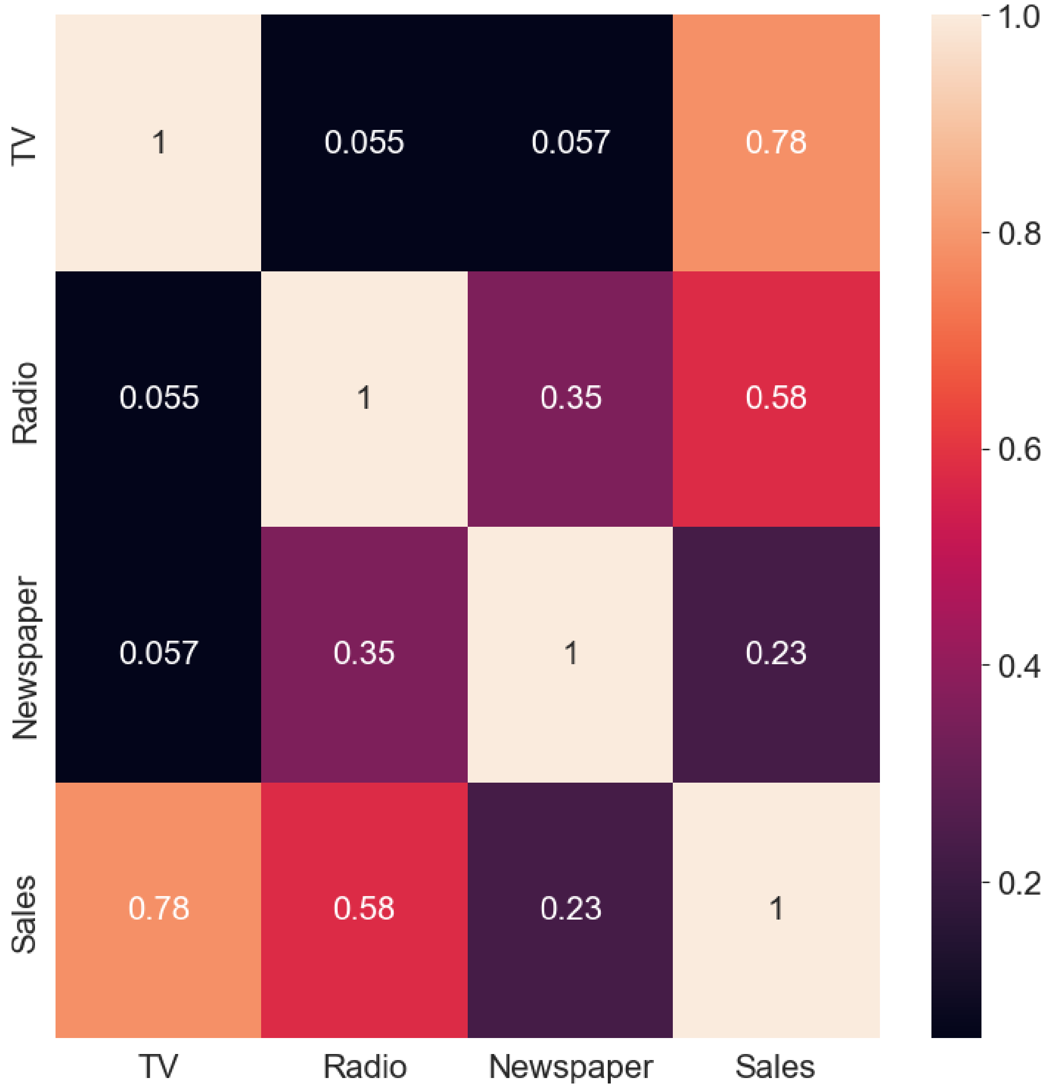


FIGURE 4.3: Correlation Heatmap of Marketing Channels

Within the correlation heatmap, we can observe that the TV and Radio channels have the most influence on the sales, as they are highly correlated with the Sales, Newspaper having the least correlation.

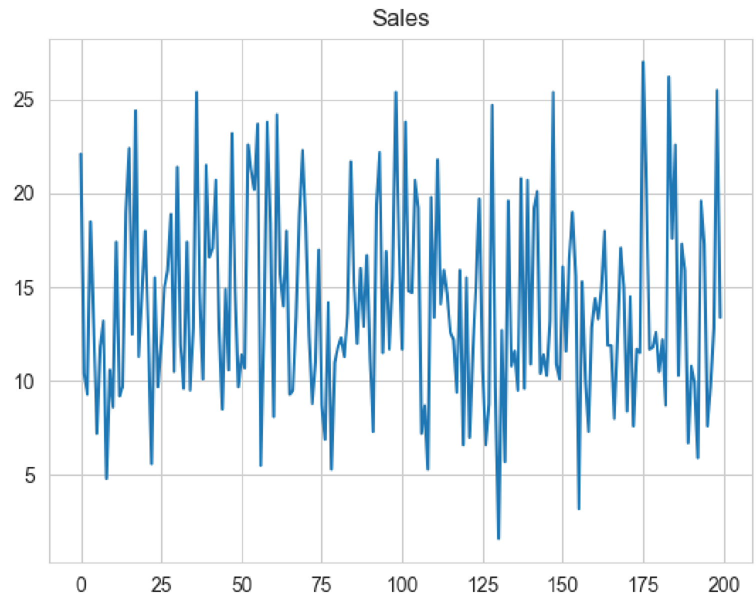


FIGURE 4.4: Sales Over a Period of 200 Months

During the period of two hundred months, Figure 4.4 shows how sales were distributed. We can see that the overall sales is stochastic in nature and would be difficult to estimate. A total of two hundred samples were recorded of the sales made on a monthly basis.

4.1.2 Ordinary Least Squares Regression Model

Since Sales can be expressed as a linear combination of the different marketing channels, we fitted the data to an Ordinary Least Squares (OLS) regression model. The summary of the models can be observed in Table 4.1

Dep. Variable:	sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Sat, 12 Nov 2022	**Prob (F-statistic):**	1.58e-96
Time:	17:33:38	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Omnibus:	60.414	**Durbin-Watson:**	2.084
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	151.241
Skew:	-1.327	**Prob(JB):**	1.44e-33
Kurtosis:	6.332	**Cond. No.**	454.

TABLE 4.1: Generated Statistics for Ordinary Least Squares Model 1

A summary Table 4.1 of the initial OLS model developed shows a high R-squared score, indicating that the OLS model performed well in determining overall sales. However, we were more interested in the other metrics generated by the model, particularly the p-value. A high p-value for the Newspaper channel signifies that it is not a significant parameter and should be removed from the model. This supports our initial belief, as discussed in the data exploration section, that the Newspaper channel does not contribute significantly to overall sales.

Dep. Variable:	sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	859.6
Date:	Sat, 04 Feb 2023	**Prob (F-statistic):**	4.83e-98
Time:	14:13:01	**Log-Likelihood:**	-386.20
No. Observations:	200	**AIC:**	778.4
Df Residuals:	197	**BIC:**	788.3
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	2.9211	0.294	9.919	0.000	2.340	3.502
TV	0.0458	0.001	32.909	0.000	0.043	0.048
Radio	0.1880	0.008	23.382	0.000	0.172	0.204

Omnibus:	60.022	**Durbin-Watson:**	2.081
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	148.679
Skew:	-1.323	**Prob(JB):**	5.19e-33
Kurtosis:	6.292	**Cond. No.**	425.

TABLE 4.2: Generated Statistics for Ordinary Least Squares Model 2

To experiment further, we modified the model to disregard the Newspaper channel as seen in Table 4.2 above. We also tested a possible disregard of the Radio channel as well to see if the model will perform better, in Table 4.2.

Dep. Variable:	sales	**R-squared:**	0.612
Model:	OLS	**Adj. R-squared:**	0.610
Method:	Least Squares	**F-statistic:**	312.1
Date:	Sat, 04 Feb 2023	**Prob (F-statistic):**	1.47e-42
Time:	14:13:01	**Log-Likelihood:**	-519.05
No. Observations:	200	**AIC:**	1042.
Df Residuals:	198	**BIC:**	1049.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

Omnibus:	0.531	**Durbin-Watson:**	1.935
Prob(Omnibus):	0.767	**Jarque-Bera (JB):**	0.669
Skew:	-0.089	**Prob(JB):**	0.716
Kurtosis:	2.779	**Cond. No.**	338.

TABLE 4.3: Generated Statistics for Ordinary Least Squares Model 3

We observed a sharp decline within the r-squared and Adj r-squared metrics recorded from model 2 in Table 4.2 to model 3 in Table 4.3, which suggests that OLS Model 2 is the fit with respect to our data, and Sales. Figures 4.5 and 4.6 provides a visual representation of how our best OLS model (Model 2) fitted our data.

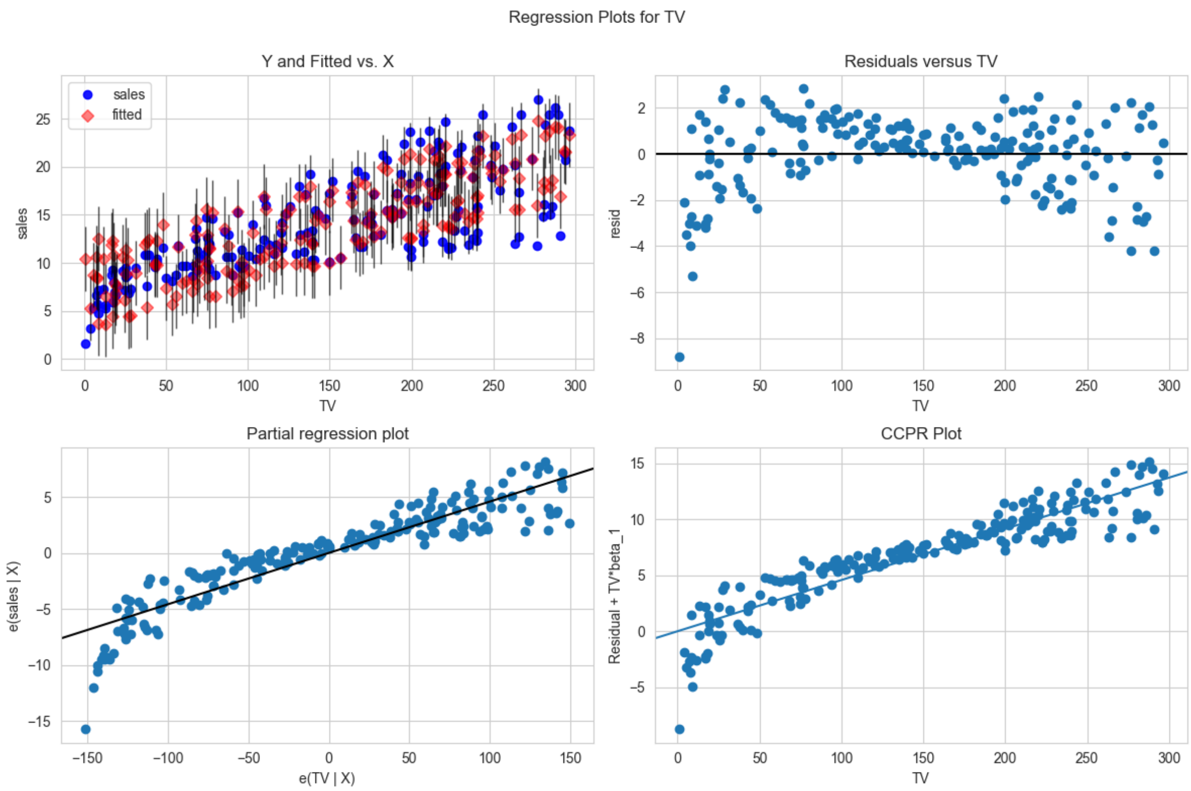


FIGURE 4.5: Regression Results for TV Marketing Channel in Model 2

Regression results against one regressor can be observed in Figures 4.5 and 4.6, fitted vs actual TV and Radio channel funds, residuals versus the TV and Radio channel funds, partial regression plot of TV and Radio channel funds, and the component and component-plus-residual (CCPR) for TV and Radio channel funds.

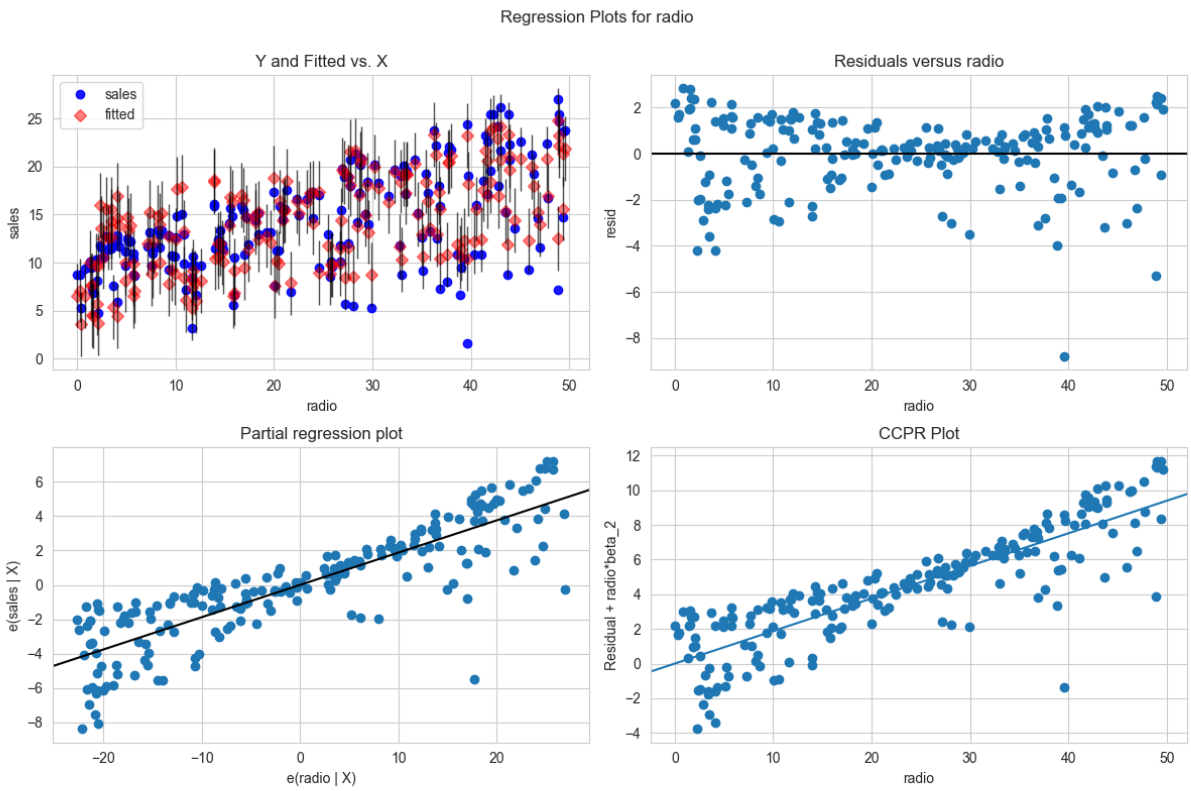


FIGURE 4.6: Regression Results for Radio Marketing Channel in Model 2

The Frequentist approach described above is typically used to forecast sales based on funds allocated to multiple channels. The OLS model enables the determination of future sales projections through predictions, with the optimal sales prediction being determined by the funds allocated to the channels that offer the best market mix, as identified in the overall model summary. However, Tables 4.2 and 4.3 demonstrate a classic case of selection bias taking place during the development of the OLS model, as the TV marketing channel ended up being more favourable compared to other channels. Nevertheless, this is justified by the model's p-values.

This approach also failed to incorporate any adstock effects within the data. Overall, it provides a reasonable justification for the allocation of funds to the channels deemed to offer the most sales. From the results of Model 2, it is clear that Newspaper marketing does not significantly affect overall sales. The high p-value (>0.005), which fails to reject the null hypothesis, justifies that there is no relationship between Newspaper marketing and sales. However, let us observe the addition of a Bayesian approach and determine whether it would provide a more robust justification of budget allocation.

4.2 Bayesian Approach

This section discusses the outcomes of the development of our Bayesian solution, as well as the outcomes of the standard machine learning models used as a basis of comparison.

4.2.1 Standard Machine Learning Approaches

We fed the data into multiple standard machine learning models to see how they would perform in terms of predicting the overall sales. Which in return would help determine the best allocation of funds to generate the maximum sales. We conducted the experiment with the Extra Trees, Random Forest, Gradient Boosted, Elastic Regression, Linear Regression, and Support Vector Machines (SVM) models.

We will utilize two standard metrics for this regression task:

- Mean Absolute Error (MAE): The mean absolute difference between the predicted and actual values.
- Root Mean Squared Error (RMSE): The square root of the mean squared difference between the predicted and actual values.

The mean absolute error is more interpretable, but the root mean squared error penalizes larger errors more heavily. Either one may be appropriate depending on the situation. We provide both metrics for comparison sake.

	MAE	RMSE
Linear Regression	1.402312	1.697063
ElasticNet Regression	1.39501	1.68948
Random Forest	0.57204	0.713065
Extra Trees	0.40344	0.520245
SVM	4.481912	5.233421
Gradient Boosted	0.935081	1.188162

TABLE 4.4: Evaluation Metrics for the Standard Machine Learning Models

An overview of the metrics used to measure the standard ML models' performance is provided within Table 4.4. The lower the value the better the model performed. Since it might be difficult to determine which model performed well at first glance, a visual comparison is shown in Figure 4.7.

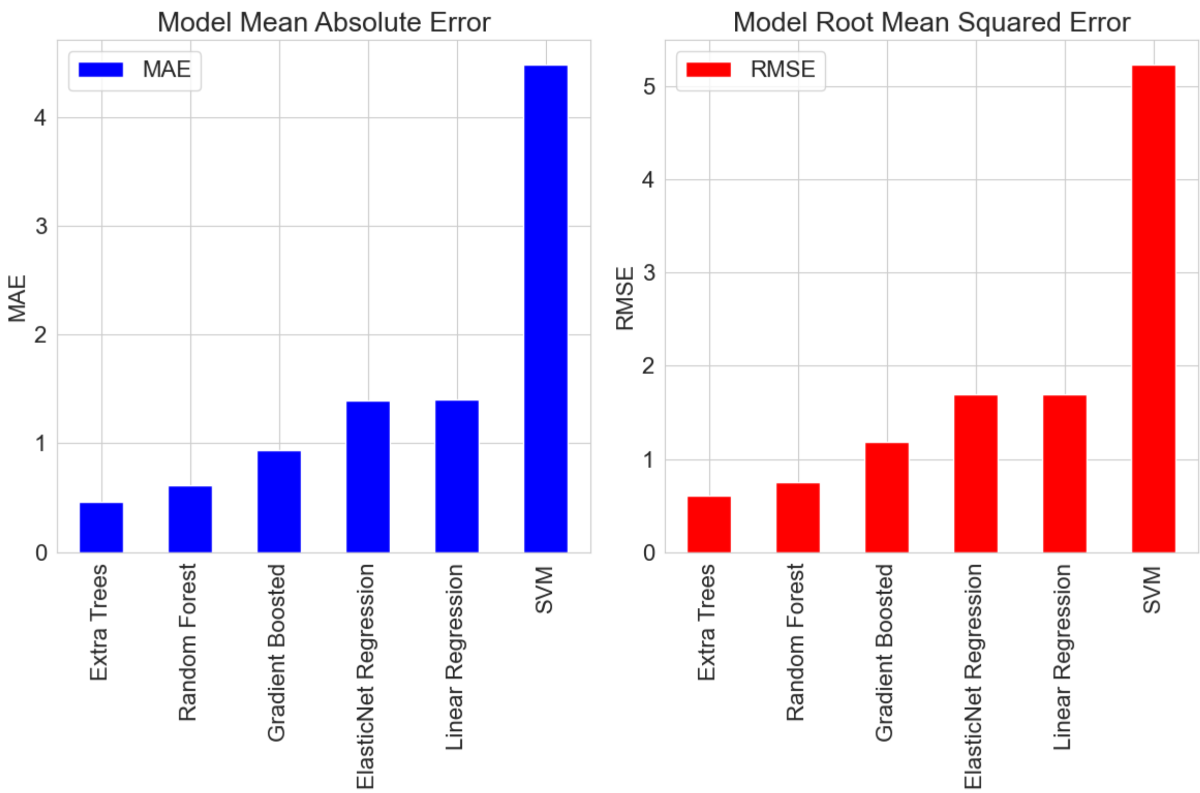


FIGURE 4.7: Visual Comparison of Standard Machine Learning Models' Metrics

Overall, the top three performers, with the lowest MSE, and RMSE scores were Extra Trees, Random forest, and the Gradient Boosted models. The SVM model performed significantly worse than the other standard ML models.

4.2.2 Bayesian Regression

We implemented a Bayesian Linear Regression model using PyMC3 [15] and Bambi [2]. The model is built using sales as a linear combination of the advertisement channels (Sales $\sim$ TV + Radio + Newspaper). To approximate the posterior for each of the model parameters, we employed a Markov Chain Monte Carlo (MCMC) algorithm to draw samples from the posterior, while utilizing a normal distribution for the data likelihood.

In order to represent our beliefs within the different marketing channels, Bayesian inference requires one to specify prior probability distributions. However, we lacked sufficient information to formulate well-defined priors and opted to use "weakly informative" priors that prevent the model from exploring completely pathological parts of the parameter space. For example, when defining a prior on the distribution of human weight, a value of 1000 kilograms should be assigned a probability of 0. We explicitly set priors to disable Bambi's default behaviour of scaling priors to the data to ensure that they remain weakly informative.

4.2.3 Hyper-parameters

We have established that sales is represented as a linear combination of the marketing channels. Our Bayesian model is built with this logic in mind, we use a normal distribution for the data likelihood. Markov Chain Monte Carlo is then used to draw samples from the posterior to approximate the posterior for each of the model's parameters. The model draws two thousand samples and is fine tuned five hundred times.

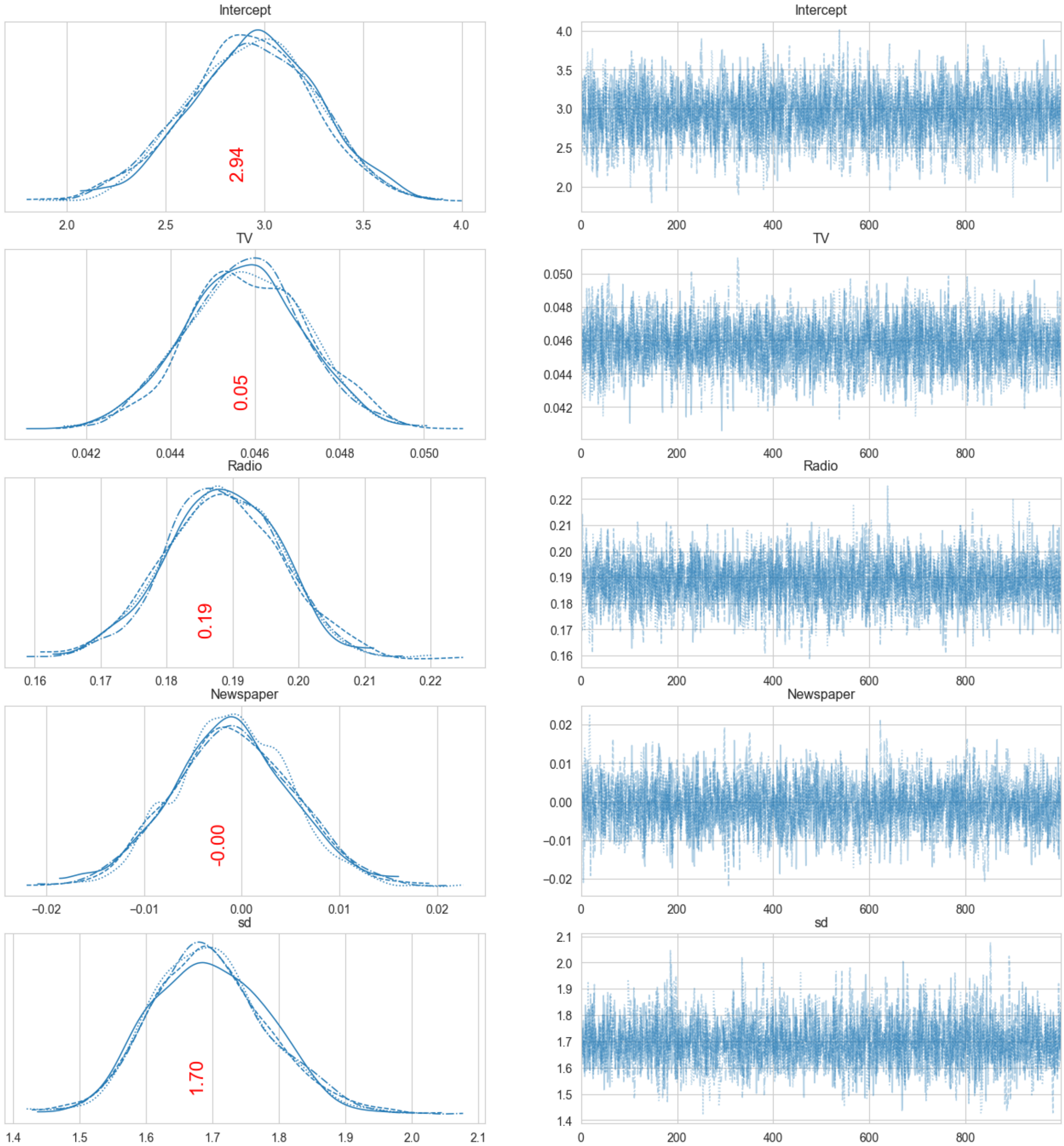


FIGURE 4.8: Traceplot of all the Marketing Channel Samples

A traceplot of all samples is provided in Figure 4.8 as a means of examining the trace results. The traceplot's left side illustrates the marginal posterior, where the x-axis displays the variable's values and the y-axis displays the probability of the variable as determined through sampling. The multiple lines denote two chains of Markov Chain Monte Carlo used in the analysis. By observing the left side, we can identify a range of values for each weight. Conversely, the right side represents the sample values drawn during the sampling process.

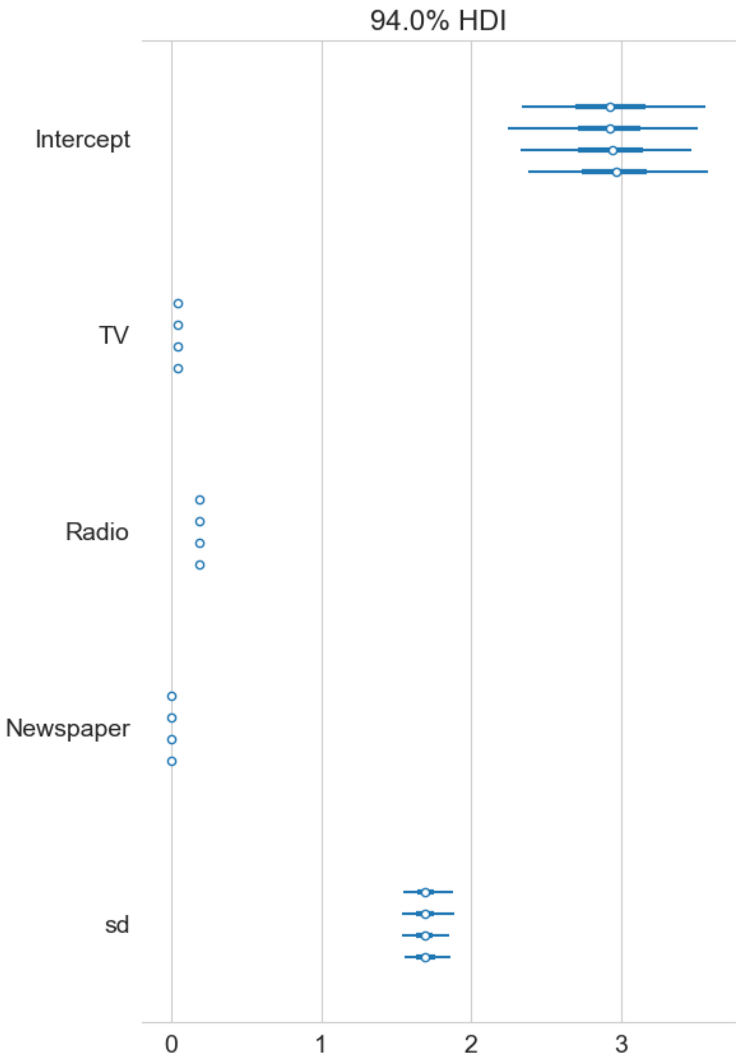


FIGURE 4.9: Forest Plot of all Marketing Channels

Another method built into PyMC3 for examining trace results is the forestplot [15] which shows the distribution of each sampled parameter. This allows us to see the uncertainty in each sample. From the forest plot in Figure 4.9, we can see the most likely value of the parameter (the dot) as well as the 95% credible interval for the parameter. The "Intercept" and "Radio" channels have larger uncertainty compared to the other variables.

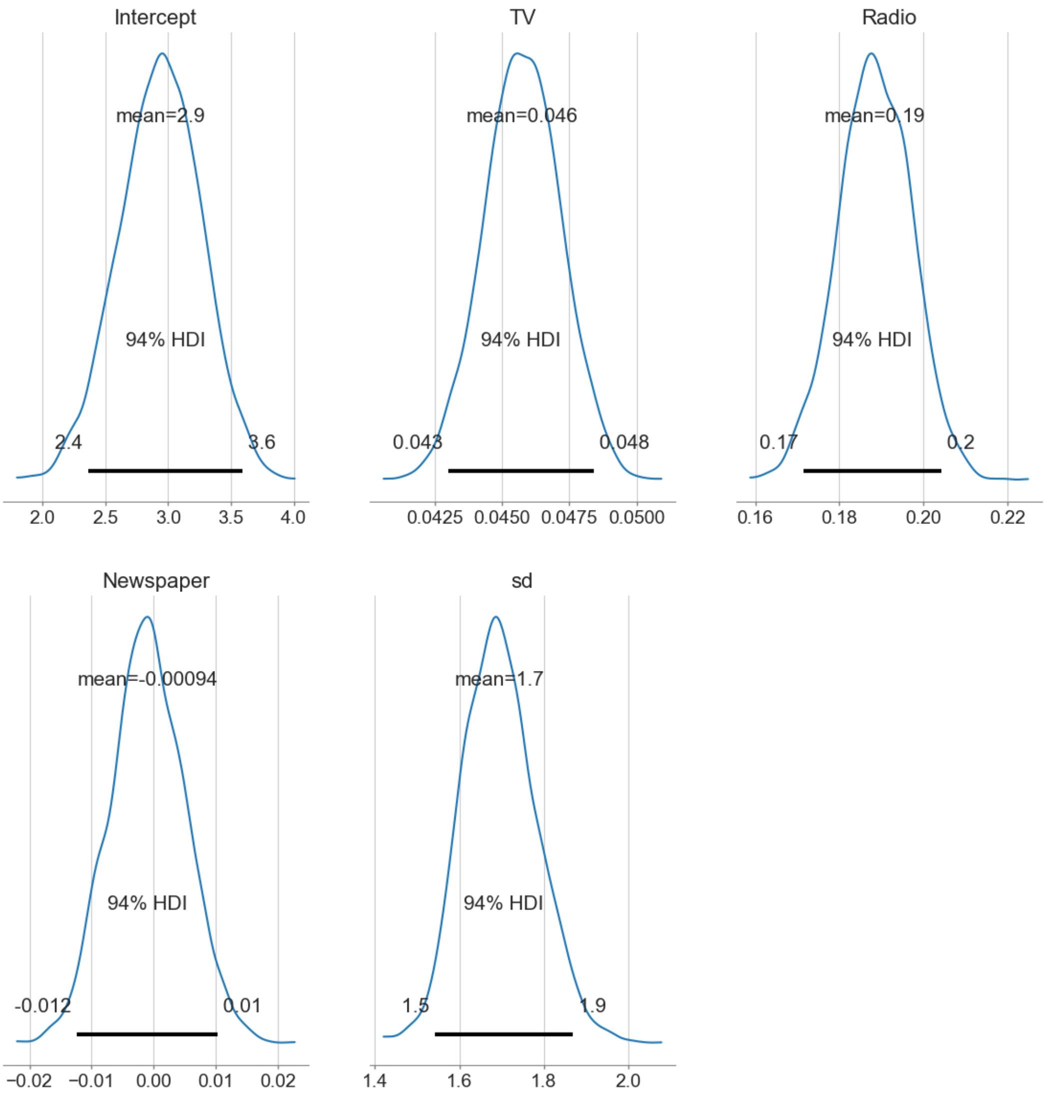


FIGURE 4.10: Posterior Distribution of all Marketing Channel Parameters

The posterior distribution for all the model parameters is illustrated by Figure 4.10. The plots enable us to visualize how the model result represents a distribution of parameter values rather than a single value. We can make the following inferences regarding the features in our dataset based on the sign and location of the mean weights in Figure 4.10:

- Advertisements within the Newspaper channel is negatively related to the overall (monthly) sales, or does not contribute at all, therefore the budget allocated for Newspaper has the least contribution in terms of overall sales.
- Advertisements within the Radio and TV channels have a positive relation to the overall (monthly) Sales, Radio has a higher positive relation than TV. The budget allocations to these two channels have a great contribution. (We can deduce that more can be allocated to Radio)

The new Bayesian formula generated from the model can be represented as

$$\begin{aligned} Sale = {} & (2.94 \times Intercept) + (0.05 \times TV) + (0.19 \times Radio) + \\ & (-0.00 \times Newspaper) + (0.53 \times sdlog) + (1.70 \times sd). \end{aligned} \tag{4.1}$$

We can evaluate the Bayesian model using the mean of the model parameters, by evaluating the MCMC trace.

The MSE and RMSE scores for the Bayesian model is significantly higher than that of the other standard ML models, but since the Bayesian model provides estimates within a confidence interval, it provides an edge over the other methods. Observing the predicted value as as well as its respective confidence will help make better informed decisions with regard to budget allocation.

Making Predictions

By predicting sales, we can input the budget allocation for each channel and predict the projected sale value for the month. The allocations with the highest predicted sales are essentially the optimal budget allocations which would generate the most sales every month.

	TV	Radio	Newspaper	True Sales	Avg Est	5% Est	95% Est
Test obs. 1	67.8	36.6	114	12.5	12.81	10.1	15.66
Test obs. 2	117	33.4	38.7	17.1	17.3	14.61	20.03

TABLE 4.5: Test Observations for the Different Marketing Channels and their Respective Estimates from the Bayesian Model

New Observations

In an attempt to determine the optimal funds to allocate to each marketing channel, that will generate the most sales, we fed new observations of possible fund allocations within the marketing channels into the model. We assumed there is a set amount allocated for advertising and distributed the amount within the different channels to see which pair allocation would predict the highest sales.

We observed the predictions generated for the sales when a specific amount (simulating a specified budget) is allocated within the different channels. Firstly, we split the budget evenly amongst the channels, then unevenly amongst the channels to observe the predicted sales. The highest sales amount generated determines the optimal budget allocations to the marketing channels. Assuming the advertising budget is 300k, we observed the total sales generated in millions.

	TV	Radio	Newspaper	Avg Est	5% Est	95% Est
New obs. 1	100	100	100	23.34	20.72	26.12
New obs. 2	200	50	50	18.53	15.9	21.36
New obs. 3	50	200	50	39.97	37.23	42.71
New obs. 4	50	50	200	11.52	8.86	14.56
New obs. 5	150	150	0	35.17	32.35	38.06
New obs. 6	250	50	0	20.87	17.98	23.56
New obs. 7	50	250	0	49.45	46.54	52.31

TABLE 4.6: New Observations for the Different Marketing Channels and their Respective Estimates from the Bayesian Model

4.2.4 Bayesian and Frequentist Model Performance with respect to MSE and MAE

The Bayesian model provides a reliable estimate of the future projection of sales based on the amount allocated to different advertising channels. It can determine the optimal allocation of the marketing budget that would generate the most sales on a monthly basis while ensuring the best allocation of the marketing budget.

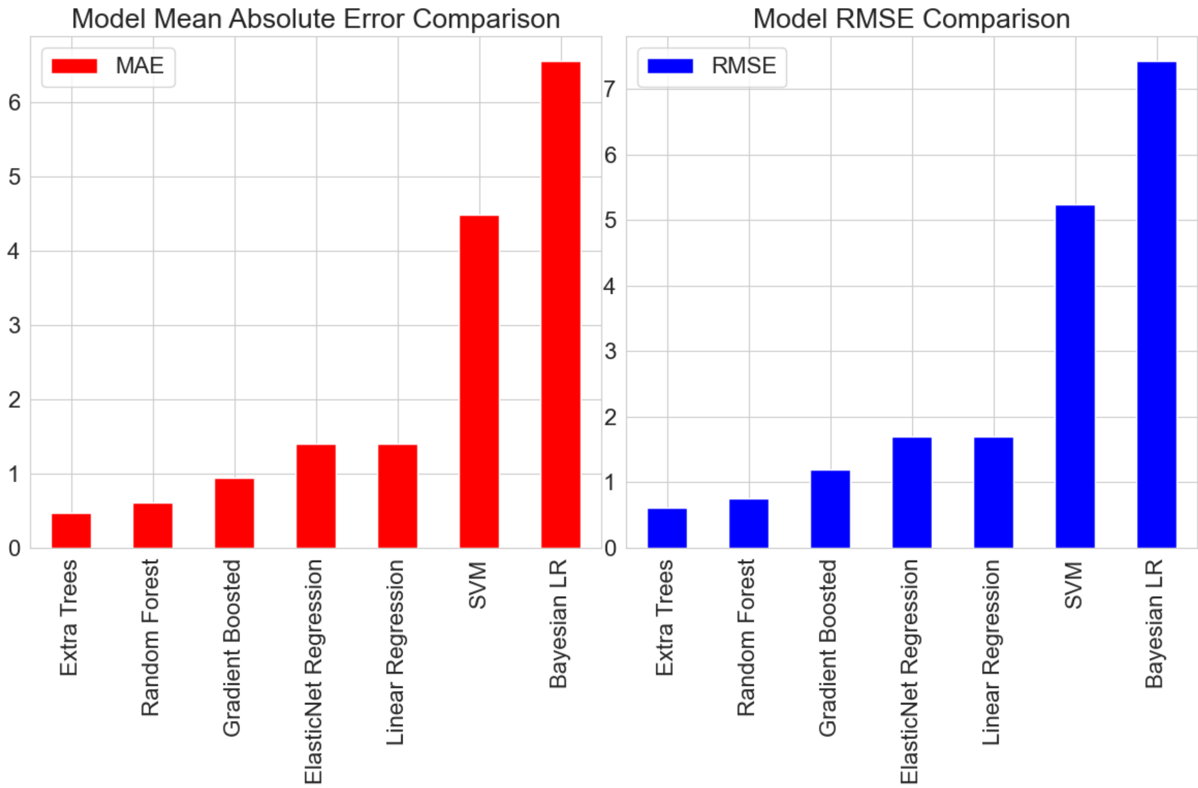


FIGURE 4.11: Visual Comparison of Bayesian Model vs Frequentist approaches' Metrics

The overall model performances in terms of the MSE and MAE metrics can be observed in Figure 4.11 as well as in Table 4.7. We can see that overall the Bayesian model has a higher MAE and MSE when compared to the others recorded for the baseline Frequentist approach models. In terms of metrics, this suggests that the Bayesian model did not perform better than other Frequentist approaches, which answers our research question in Chapter 1.

We can argue that although the Bayesian model did not perform better than the Frequentist approaches with regards to MSE and MAE, it offers a more robust approach for modelling the relationship between advertising channels and sales when compared to the baseline models. The additional value the Bayesian model offers when it comes to modelling uncertainty and incorporating past beliefs is unmatched, and is worth further fine tuning in order to reduce its respective MSE and MAE parameter outputs.

	MAE	RMSE
Linear Regression	1.402312	1.697063
ElasticNet Regression	1.39501	1.68948
Random Forest	0.57204	0.713065
Extra Trees	0.40344	0.520245
SVM	4.481912	5.233421
Gradient Boosted	0.935081	1.188162
Bayesian LR	6.551473	7.422632

TABLE 4.7: Evaluation Metrics for the Bayesian and Frequentist approaches

4.3 Summary

In this chapter, we looked at using multiple models as well as Bayesian modelling to predict sales based on three advertisement channel factors (TV, Newspaper, and Radio). The problem was phrased as a machine learning task, opposed to specifying the probabilities for the Bayesian network, which is extremely difficult for continuous variables. PyMC3 and Bambi helped in terms of incorporating our beliefs, specifying priors and likelihoods.

The ability to predict overall sales can help determine the best amount to allocate to each advertisement channel in order to generate the most interactions or sales within a monthly basis. In addition to the standard machine learning models that learn from observations, we successfully applied Bayesian modelling and regression to create a model mapping the features (funds allocated to the marketing channels) to the target (sales).

The advantage of Bayesian regression is that, if we use sensible priors, we can still get a decent estimate with a few samples, and the final weights are not a single number, but a distribution composed of every sample drawn during the sampling run. We can then make predictions using all the sampled weights to form a distribution of expected values rather than a single answer.

Overall, the Bayesian model performed well. During our exploration of the different models, we discovered that the Newspaper channel has little to no contribution to the overall sales generated, which means more funds can be allocated to the TV and Radio channels. The model provides a realistic estimate of the sales to expect when funds are allocated to the different marketing channels. The Bayesian inference approach produced intuitive estimates for the model weights and gave sales predictions that align with our expectations regarding sales margins based on the funds allocated to the marketing channels.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This research aimed to apply Bayesian modelling and computations on Media Mix Models in order to determine the effects of advertising expenditure. Overall, we succesfully modelled the relationship between the different advertising channels and sales in Chapter 4. We successfully developed a model that combined Bayesian modelling and Media Mix Modelling (Section 4.2.2), the model provides reliable sales estimates which is essentially a prediction, but can help determine the optimal fund allocations within the different marketing channels that led to the sale prediction. We addressed the gaps within previous MMM solutions which fail to incorporate the knowledge domain and uncertainty.

We developed multiple standard ML models, and provided MSE and MAE comparisons of how they performed with regards to predicting the overall sales (Chapter 4). We also incorporated priors and likelihoods into our model and used MCMC for approximations. During the data exploration process, we discovered that the newspaper channel offered no contribution or correlation to the overall sales, this was confirmed by our models as well. Since the newspaper channel offered little contribution to the overall sales, more funds could be allocated to the other channels (TV and Radio) in order to generate better sales.

5.2 Future Work

It might be crucial to have an in depth exploration of the different priors and distributions that can be applied to the Bayesian model. The choice of prior distributions can have a significant impact on the results of the analysis. Since it can prove to be quite difficult to determine which likelihood function and priors to choose, future work can be done to determine the outcomes of different likelihood functions and how they influence the outcome of the sales prediction. Since the mse and rmse do not really capture the true performance of the Bayesian model, an explanation of more suitable evaluation metrics for the Bayesian model would help with its comparison with other standard ML models.

Training a Bayesian model is computationally intensive and requires a lot of time and resources to converge. We might be able to generate even better results for the Bayesian model if we increase the amount of samples drawn as well and the amount of fine tuning steps. This would require a huge amount of computational power to achieve, and can be considered for future work and model improvement.

Appendix A

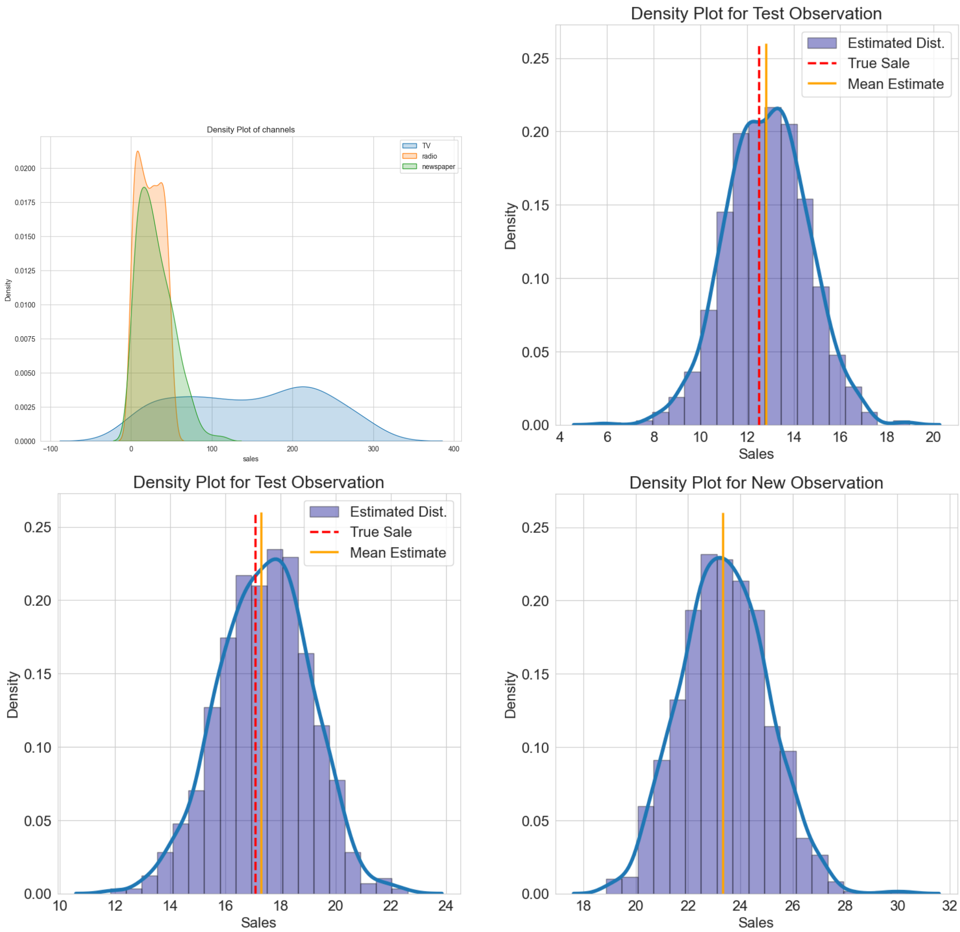
Appendix

A.1 GitHub Repository Link

Source codes for all the models developed for this research project can be accessed within the GitHub Repository. The repository contains the notebook, and dataset used within this research project, and is open source and available for editing and improvement.

A.2 Figures

A.2.1 Density Plots



(A) Density plots

Bibliography

[1] 1749. *Maximising ROI with Optimal Budget Allocation Across Channels*. English. Feb. 2023. URL: https://1749.io/resource-centre/f/using-bayesian-decision-making-to-optimise-media-budgets.

[2] Tomás Capretto et al. "Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python". In: *Journal of Statistical Software* 103.15 (2022), 1–29. DOI: 10.18637/jss.v103.i15. URL: https://www.jstatsoft.org/index.php/jss/article/view/v103i15.

[3] David Chan and Mike Perry. "Challenges and opportunities in media mix modeling". In: *Google Research* (2017).

[4] Aiyou Chen et al. "Bias correction for paid search in media mix modeling". In: *arXiv preprint arXiv:1807.03292* (2018).

[5] Galyna Chornous, Yana Fareniuk, et al. "Marketing mix modeling for pharmaceutical companies on the basis of data science technologies". In: *Access Journal* 2.3 (2021), pp. 274–289.

[6] Matteo Colombo and Peggy Seriès. "Bayes in the brain—on Bayesian modelling in neuroscience". In: *The British journal for the philosophy of science* (2020).

[7] Deloitte. *The future is modeled: A how-to-guide for Advanced Marketing Mix Models*. Tech. rep. Deloitte, 2021.

[8] Peter I Frazier. "A tutorial on Bayesian optimization". In: *arXiv preprint arXiv:1807.02811* (2018).

[9] Adam Gelzinis et al. *What happens to sales when brands stop advertising for long periods?* Tech. rep. Ehrenberg-Bass Corporate Sponsorship Program, 2018.

[10] Yuxue Jin et al. "Bayesian methods for media mix modeling with carryover and shape effects". In: *Google Research* (2017).

[11] Osvaldo A. Martin, Ravin Kumar, and Junpeng Lao. *Bayesian Modeling and Computation in Python*. Boca Raton, Dec. 2021. ISBN: 978-0-367-89436-8.

[12] Edwin Ng, Zhishi Wang, and Athena Dai. "Bayesian Time Varying Coefficient Model with Applications to Marketing Mix Modeling". In: *arXiv preprint arXiv:2106.03322* (2021).

[13] Sandeep Pandey, Snigdha Gupta, and Shubham Chhajed. "Marketing Mix Modeling (MMM)-Concepts and Model Interpretation". In: *Sandeep Pandey, Snigdha Gupta, Shubham Chhajed* (2021).

[14] José Francisco Sá Marques Rocha. "Media mix modeling: a case study on optimizing television and digital media spend for a retailer". PhD thesis. 2020.

[15] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: *PeerJ Computer Science* 2 (2016), e55.

[16] Yunting Sun et al. "Geo-level bayesian hierarchical media mix modeling". In: *Google Research* (2017).

[17] Benjamin Vincent. *Bayesian Media Mix Modeling for Marketing Optimization*. PyMC Labs. Sept. 2021. URL: https://www.pymc-labs.io/blog-posts/bayesian-media-mix-modeling-for-marketing-optimization/.

[18] Yueqing Wang et al. "A hierarchical Bayesian approach to improve media mix models using category data". In: *Google Research* (2017).

[19] Yueqing Wang et al. *A Hierarchical Bayesian Approach to Improve Media Mix Models Using Category Data*. Tech. rep. Google Inc., 2017.

[20] Andrew G Wilson and Pavel Izmailov. "Bayesian deep learning and a probabilistic perspective of generalization". In: *Advances in neural information processing systems* 33 (2020), pp. 4697–4708.

[21] Michael J Wolfe Sr and John C Crotts. "Marketing mix modeling for the tourism industry: A best practices approach". In: *International Journal of Tourism Sciences* 11.1 (2011), pp. 1–15.

[22] Kaifeng Zhao, Seyed Hanif Mahboobi, and Saeed R Bagheri. "Revenue-based attribution modeling for online advertising". In: *International Journal of Market Research* 61.2 (2019), pp. 195–209.