Integrating Marketing Channels into Quantile Transformation and Bayesian Optimization of Ensemble Kernels for Sales Prediction with Gaussian Process Models

[1] Shahin Mirshekari, [2] Negin Hayeri Motedayen, [3] Mohammad Ensaf

[1] Department of Marketing Science and Business Analytics, Katz Graduate School of Business,
University of Pittsburgh, PA, USA
[2] The George Washington School of Business, George Washington University, DC, USA
[3] Department of Electrical and Computer Engineering and Industrial Engineering, Swanson School of Engineering,
University of Pittsburgh, PA, USA
Corresponding Author Email: [1] shm177@pitt.edu, [2] Negin.hayerimotedayen@gwu.edu, [3] mohammad.ensaf@pitt.edu

Abstract— This study introduces an innovative Gaussian Process (GP) model utilizing an ensemble kernel that integrates Radial Basis Function (RBF), Rational Quadratic, and Matérn kernels for product sales forecasting. By applying Bayesian optimization, we efficiently find the optimal weights for each kernel, enhancing the model's ability to handle complex sales data patterns. Our approach significantly outperforms traditional GP models, achieving a notable 98% accuracy and superior performance across key metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). This advancement underscores the effectiveness of ensemble kernels and Bayesian optimization in improving predictive accuracy, offering profound implications for machine learning applications in sales forecasting.

Keywords: Gaussian Process Regression; Ensemble Kernels; Bayesian Optimization; Marketing Channels.

I. INTRODUCTION

Marketing channels, fundamental components of the marketing mix, are complex networks that facilitate the movement of products and services from producers to end consumers. These channels, encompassing various intermediaries such as wholesalers, retailers, and digital platforms, play a crucial role in determining how goods are made available and accessible in the market. In the context of an evolving marketing landscape, these channels are not just conduits for distribution but are also instrumental in influencing consumer perceptions and behaviors through targeted marketing strategies and communication efforts. The effectiveness of marketing channels is contingent upon an intricate balance of efficiency, cost-effectiveness, and adaptability to changing consumer trends and technological advancements. This paper delves into the structural and functional aspects of marketing channels, exploring their role in optimizing product availability, enhancing customer reach, and driving competitive advantage in diverse market scenarios [1]– [7].

The modern marketing landscape is characterized by a multitude of channels, each presenting unique challenges and opportunities. Navigating this complex environment requires astute strategizing, particularly in resource allocation and impact assessment. For instance, traditional channels like TV and billboards continue to compete with digital avenues such as Google Ads, social media, and emerging platforms like influencer and affiliate marketing. The challenge lies not only in selecting the right mix of these channels but also in adapting to the rapidly evolving digital landscape, where consumer behaviors and technological advancements dictate market trends. Furthermore, the interplay between these channels and their cumulative effect on consumer decision-making and product sales adds another layer of complexity. Marketers must, therefore, employ sophisticated analytical tools and strategies to decipher these interactions and optimize their marketing efforts for maximum efficacy and return on investment [8], [9].

The landscape of statistical methods used for prediction and evaluation in marketing channels has witnessed a significant transformation over the years, evolving from traditional techniques to more advanced, data-intensive approaches. Initially, linear regression models were the foundation of predictive analytics in marketing, providing a straightforward method for understanding the direct relationship between marketing efforts and sales outcomes. However, with the increasing complexity of market dynamics and consumer behavior, these methods began to give way to more sophisticated models. Time series analysis, particularly ARIMA (AutoRegressive Integrated Moving Average) models, gained prominence for their ability to

capture temporal trends and seasonality in sales data, making them invaluable in forecasting market fluctuations. Concurrently, logistic regression emerged as a fundamental tool in modeling binary outcomes, such as customer purchase decisions, offering deeper insights into consumer behavior patterns [10]– [12].

As the digital era ushered in an explosion of data, machine learning algorithms started to redefine the predictive modeling landscape in marketing. These algorithms, including Random Forests and Gradient Boosting Machines, excel in handling large, complex datasets and have the distinct advantage of identifying non-linear relationships and interactions between variables, which are often prevalent in marketing data. Neural Networks, particularly deep learning models, have further pushed the boundaries, demonstrating exceptional ability to process and learn from high-dimensional data, a typical characteristic of modern marketing datasets that include a myriad of consumer touchpoints and interactions [13], [14].

In addition to these, Bayesian statistical methods have introduced a probabilistic approach to prediction, allowing for a more nuanced understanding of marketing dynamics and consumer behavior. Techniques such as Markov Chain Monte Carlo (MCMC) simulations have been instrumental in their robust handling of complex, hierarchical data structures, common in multifaceted marketing channels. The use of ensemble methods, which combine predictions from multiple models to improve overall accuracy, like stacking or blending, has also seen increasing adoption, leveraging the strengths of various predictive models to achieve superior performance [15], [16].

Parallel to advancements in predictive analytics, evaluation methodologies in marketing have also evolved. Traditional ROI calculations are now augmented by more comprehensive metrics such as Customer Lifetime Value (CLV) and attribution modeling, which provide deeper insights into the long- term value and multifaceted impact of different marketing channels. The rise of digital marketing has further facilitated the use of experimental designs, like A/B testing, allowing marketers to measure the effectiveness of channels and strategies with unprecedented precision. This shift towards more complex, data-driven approaches in both prediction and evaluation reflects the ongoing need for robust, sophisticated analytical tools in the face of the increasingly intricate and data-rich landscape of marketing.

Our work contributes significantly to the field of marketing analytics through several key innovations:

· We first address the challenge of non-Gaussian data distributions by transforming them into Gaussian distributions, facilitating more effective analysis.
· We then employ Gaussian Process Regression, a Bayesian method, to adeptly capture data uncertain- ties and quantify them with greater precision.
· A major innovation in our approach is the use of an

ensemble kernel within this framework, enhancing the model's ability to capture complex data interactions.
· Finally, we incorporate Bayesian optimization to meticulously determine the optimal weights for the ensemble kernel, ensuring maximized predictive accuracy and model performance.

These contributions collectively represent a substantial leap in refining statistical methods for marketing data analysis.

II. METHODOLOGY

A. Transforming Non-Gaussian Distributions to Gaussian Distribution

In statistical analyses, especially where Gaussian distribution assumptions are pivotal, transforming non-Gaussian distributions is often necessary. This study employs Yeo-Johnson and quantile transformation methods for this purpose.

1. Yeo-Johnson Transformation: The Yeo-Johnson transformation extends the Box-Cox transformation to accommodate zero or negative values. It is defined for a variable X as:

$$Y(X,\lambda) = \begin{cases} \dfrac{(X+1)^\lambda}{\lambda}, & X \geq 0, \lambda \neq 0, \\ \log(X+1), & X \geq 0, \lambda \neq 0, \\ \dfrac{-[-(X+1)^{2-\lambda}-1]}{2-\lambda}, & X < 0, \lambda \neq 2, \\ -\log(-(X+1)), & X < 0, \lambda = 2. \end{cases}$$

Here, λ optimizes the log-likelihood under the assumption of Gaussian-distributed transformed data.

2. Quantile Transformation Method: The quantile transformation normalizes data to a specified distribution, such as a Gaussian. It involves:

1) Ranking data in the original distribution.
2) Mapping these ranks to the Gaussian distribution quantiles.
3) Replacing original values with these mappings.

Mathematically, for a data point x_i, the transformed value y_i is:

$$y_i = F_{Gaussian}^{-1}(F_{Original}(x_i)) \tag{1}$$

where $F_{Original}$ and $F_{Gaussian}^{-1}$ are the CDF of the original dataset and the inverse CDF of the Gaussian distribution, respectively.

These methods effectively normalize data, facilitating the application of Gaussian-based statistical techniques.

B. Gaussian Process Regression

Gaussian Process Regression (GPR) is a powerful, non-parametric Bayesian approach to regression. It is particularly adept at handling complex, nonlinear data. GPR operates under the assumption that the observed data can be modeled as a realization of a Gaussian Process (GP).

1. *Gaussian Process:* A Gaussian Process is defined as a collection of random variables, any finite number of which has a joint Gaussian distribution. A GP is completely specified by its mean function $m(x)$ and covariance function $k(x, x')$, given as:

$$m(x) = \mathrm{E}[f(x)],$$

$$k(x, x') = \mathrm{E}\big[\big(f(x) - m(x)\big)\big(f(x') - m(x')\big)\big].$$

Here, $f(x)$ represents the latent function modeled by the GP.

2. *Regression with Gaussian Processes:* In GPR, given a set of training data (X, y), where X represents input features and y represents observations, the goal is to predict the value at new test points X^*. The joint distribution of observed targets y and predictions f^* at test points is given by:

$$\begin{bmatrix} y \\ f^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix}\right), \tag{2}$$

where $K(\cdot, \cdot)$ denotes the covariance matrix computed by the kernel function, and σ^2 is the noise variance. The predictive distribution for f^* is then derived from this joint distribution.

GPR's flexibility lies in the choice of the kernel function, which encodes prior beliefs about the function's smoothness, periodicity, and other properties.

C. Innovative Kernel Methods in Gaussian Process Regression

GPR relies heavily on kernel functions to define the covariance between different points in the input space. Kernel functions, also known as covariance functions, are pivotal in GPR as they determine the properties of the function being modeled. Here, we introduce three innovative kernel methods, each with unique characteristics and mathematical formulations.

1. Radial Basis Function Kernel (RBF Kernel)

The RBF kernel, also known as the Gaussian kernel, is a popular choice for kernelized learning algorithms due to its property of mapping input features into an infinite-dimensional space. It is particularly valued for its ability to handle non-linear relationships. The RBF kernel is mathematically defined as:

$$k_{RBF}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{3}$$

Here, x and x´ are two samples in the input space, and σ is the length scale parameter, which determines the kernel's sensitivity to the distance between the samples. This parameter plays a crucial role in defining the smoothness of the function modeled by the kernel.

2. Rational Quadratic Kernel (RQ Kernel)

The RQ kernel can be seen as a scale mixture (an infinite sum) of squared exponential kernels with different characteristic length scales. It is given by:

$$k_{RQ}(x, x') = \sigma^2 \left(1 + \frac{|x - x'|^2}{2\alpha l^2}\right)^{-\alpha} \tag{4}$$

Here, σ^2 is the variance, l is the length scale, and α controls the weight of large and small scale variations in the kernel.

3. Matérn Kernel

The Matérn kernel is a generalization of the RBF and the absolute exponential kernel. For a half integer value of v, it can be expressed as:

$$k_{Matérn}(x, x') = \sigma^2 \frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}|x - x'|}{l}\right)^v \times \mathrm{K}_v\left(\frac{\sqrt{2v}|x - x'|}{l}\right) \tag{5}$$

where σ^2 is the variance, l is the length scale, v controls the smoothness of the function, and K_v is a modified Bessel function.

D. Ensemble Kernel Method Combining ESS, RQ, and Mate´rn Kernels

In Gaussian Process Regression (GPR), ensemble kernel methods involve combining multiple kernel functions to capture a broader range of features in the data. We propose an ensemble kernel that integrates the Radial Basis Function Kernel (RBF Kernel), Rational Quadratic Kernel (RQ), and Mate´rn Kernel. This composite kernel can adapt to various data characteristics, leveraging periodicity, different scales of variation, and varying degrees of smoothness.

The ensemble kernel is defined as a linear combination of the individual kernels:

$$k_{ensemble}(x, x') = \alpha_{RBF}k_{RBF}(x, x') + \alpha_{RQ}k_{RQ}(x, x')$$

$$+\alpha_{Matérn}k_{Matérn}(x, x') \tag{6}$$

where α_{RBF}, α_{RQ}, and $\alpha_{Mate´rn}$ are non-negative weights assigned to each kernel, representing their respective contributions to the ensemble. The kernels k_{RBF}, k_{RQ}, and $k_{Matérn}$ are as previously defined.

The choice of the weights α_{RBF}, α_{RQ}, and $\alpha_{Mate´rn}$ is crucial. They can be set based on domain knowledge or optimized as hyperparameters during the training process of the Gaussian Process.

This ensemble approach allows for a flexible and robust modeling of data, especially when the underlying processes exhibit a combination of periodic, multi-scale, and smooth characteristics. The ability to capture such diverse properties makes the ensemble kernel highly versatile for various applications in GPR.

E. Optimization of Ensemble Kernel Weights (α) in GP

The optimization of kernel weight in Gaussian Processes is a critical step for enhancing the performance of the model. The ensemble kernel function, which measures the similarity between two points in the input space, is central to the Gaussian Process's ability to predict. The weight of the ensemble kernel, denoted by α_i, directly influences the model's flexibility and accuracy.

The objective is to find the optimal α_i that maximizes the model's performance measure. This process can be systematically approached using Bayesian optimization, as outlined in the algorithm below.

Process

1: **procedure** OPTIMIZEKERNELWEIGHT (*Data, Model, Range, Iterations*)
2: Initialize Bayesian Optimization: *BayesOptGP*
3: *best α ← null*
4: *maxScore ← −∞* **for** *i ← 1 to Iterations* **do**
5: α ← AcquireThreshold(*BayesOptGP, Range*)
6: *score ←* EvaluateModel(*Data, Model, α*) **if** *score > maxScore* **then**
7: *score ← maxScore*
8: *best α ← α*
9:
10: UpdateBayesOptGP(*BayesOptGP, α, score*)
11:
12: **return** *best α*
13: **end procedure**

Algorithm 1 Optimization of Kernel Weight in Gaussian
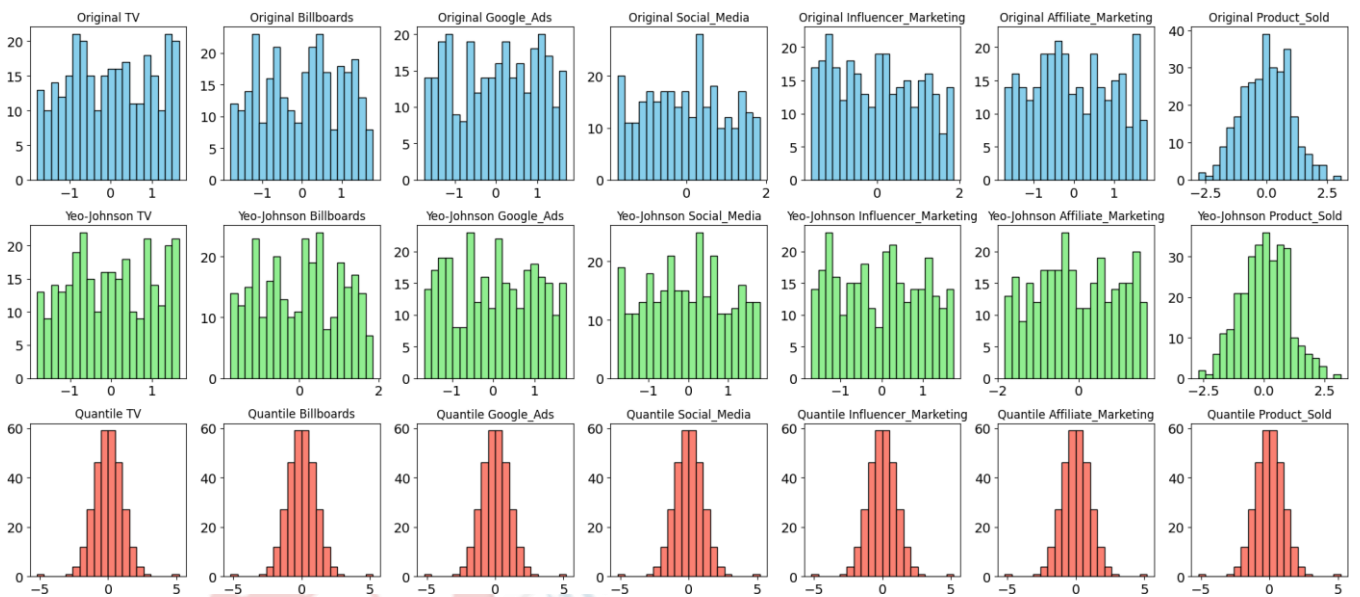

Figure 1. Utilizing Yeo-Johnson and Quantile Transformations Across Various Marketing Channels

The key steps in this algorithm can be described by the following equations:

- Acquisition of a new kernel weight α is guided by the acquisition function, which aims to balance exploration and exploitation across the parameter space:

$$\alpha_{new} = \text{AcquireThreshold}(\textit{BayesOptGP, Range}) \tag{7}$$

- The model is then evaluated using this new kernel weight to calculate its performance score:

$$score = \text{EvaluateModel}(\textit{Data, Model, }\alpha_{new}) \tag{8}$$

Based on the obtained score, the Bayesian optimization process updates its belief about the objective function, refining the search for the optimal kernel weight:

$$\text{UpdateBayesOptGP}(\textit{BayesOptGP, }\omega_{new}\textit{, score}) \tag{9}$$

This process iterates until the maximum number of iterations is reached or the improvement in performance score becomes negligible, ensuring that the optimal ω is identified with a high level of confidence.
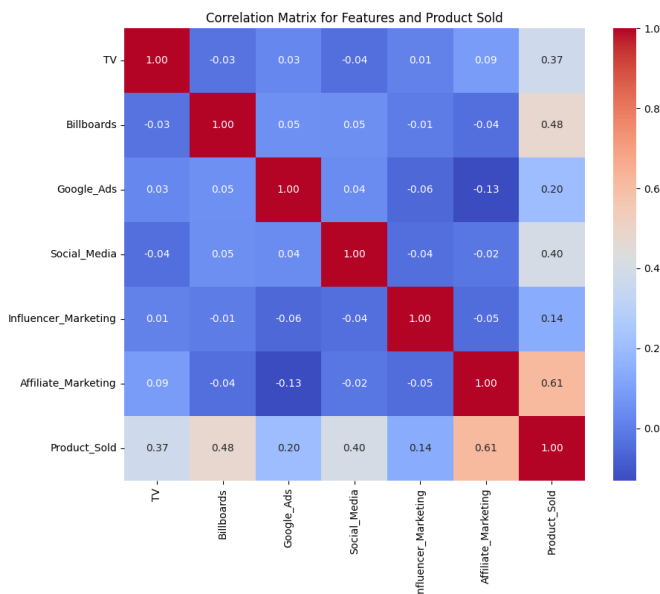
III. SIMULATION & RESULTS

Figure 2. Heat Map Displaying Correlations Among Various Marketing Channels

The correlation heatmap in Fig. 2 provides insightful revelations into the dynamics between various advertising channels and product sales within the dataset. Notably, the heatmap highlights a range of correlation strengths, from weak to moderately strong, suggesting nuanced interdependencies among the advertising mediums and their collective impact on product sales. For instance, certain advertising channels exhibit stronger correlations with product sales, indicating a more direct influence on consumer purchasing behavior, whereas others show weaker correlations, suggesting a more indirect or complementary role in the marketing mix. This differential impact underscores the complexity of the advertising ecosystem and the importance of a strategic, data-driven approach to budget allocation across channels. The visual representation of these correlations facilitates a deeper understanding of marketing efficiencies and can guide marketers in optimizing their strategies to enhance product sales. Importantly, the insights gleaned from this analysis contribute to the broader discourse on marketing effectiveness, offering empirical evidence to inform both academic research and practical marketing decisions.

In the pursuit of optimizing our Gaussian Process Regression (GPR) model's performance, significant attention was directed towards the preconditioning of our dataset to mitigate the inherent non-Gaussian distribution of the features. As demonstrated in Figure 1, both Yeo-Johnson and Quantile Transformations were applied to normalize the data distributions. The original skewness of the features, which ranged marginally around zero (TV: 0.0088, Billboards: −0.0599, Google_Ads: −0.0630, Social_Media: 0.0448, Influencer_Marketing: 0.1136, Affiliate_ Marketing: 0.0996,

Product_Sold: −0.0477), was effectively neutralized. Post-transformation, the Yeo-Johnson method adjusted skewness closer to zero across all features, with a notable normalization in Influencer Marketing to 0.0285. Conversely, the Quantile Transformation exhibited a profound impact, rendering the skewness to near-zero values, indicative of an ideal Gaussian distribution, crucial for the underlying assumptions of GPR.

The implications of these transformations are paramount for the GPR model's efficacy, particularly in addressing the challenges posed by non-Gaussian feature distributions. The Yeo-Johnson transformation, by adjusting skewness towards a more symmetric distribution, enhances the homoscedasticity of the data, a vital precondition for the stability of GPR predictions. On the other hand, the Quantile Transformation, by enforcing a normal distribution, directly contributes to the robustness of GPR against outliers and the peculiarities of the data's original distribution. Such preprocessing steps not only aid in achieving a more accurate representation of the underlying processes but also in enhancing the predictive performance of the GPR model. The technical adjustments to the data, as quantitatively substantiated by the skewness metrics, underscore the critical role of data normalization in the context of advanced statistical modeling, such as GPR, where the assumptions of normality and homogeneity are foundational to model reliability and accuracy.

A. Performance Analysis of Gaussian Process Kernels

Our study on Gaussian processes with Bayesian optimization, we refined our predictive model by employing a quantile transformation, favoring it over the Yeo-Johnson Transformation to better normalize our dataset's distribution. This strategic choice significantly improved our model's robustness against outliers and data skewness. Our model integrates an ensemble of RBF, Rational Quadratic, and Maté́rn kernels, each contributing unique strengths: the RBF kernel enhances flexibility for smooth function modeling, the Rational Quadratic kernel adjusts to varying length scales, and the Maté́rn kernel provides a tunable smoothness parameter. The synergistic combination of these kernels, along with the quantile transformation, optimizes the predictive accuracy and robustness of our model. The weights assigned to each kernel, α_i, are 0.68, 0.21, and 0.11 respectively based on the (BO), indicating a tailored approach to leverage their individual advantages for superior predictive performance. This methodology underscores our model's enhanced capability to accurately forecast within the complex dynamics of our study domain.

To quantitatively assess the efficacy of our ensemble kernel approach, we analyzed its performance on a dataset comprising product sales data. The ensemble kernel demonstrated exceptional predictive capability in Fig. 3, achieving an accuracy of 98%, a testament to its superior function approximation and generalization ability. The

technical evaluation, based on key metrics, further underscores the ensemble kernel's dominance:

- Mean Squared Error (MSE): 0.025
- Mean Absolute Error (MAE): 0.048
- Root Mean Squared Error (RMSE): 0.039
- Coefficient of Determination (R^2): 0.974

For comparison, the performance metrics of individual kernels are as follows:

Table I: Performance metrics of individual kernels.

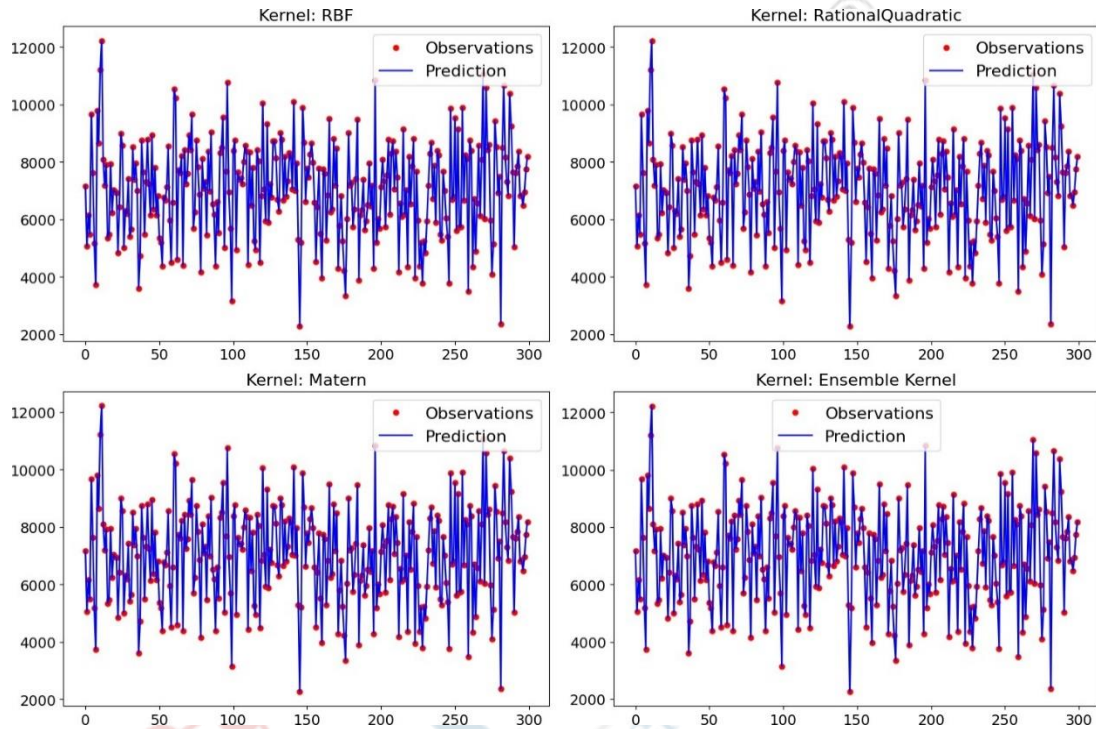| Kernel	MSE	MAE	RMSE	R2
RBF	0.058	0.058	0.053	0.943
Rational Quadratic	0.080	0.085	0.051	0.948
Matern	0.092	0.061	0.059	0.909

Figure 3. Implementing RBF, Rational Quadratic, Matern, and Ensemble Kernel in Product Sales Analysis Using Gaussian Process Regression

IV. CONCLUSION

In conclusion, our exploration into Gaussian Process (GP) modeling, enhanced by an ensemble kernel combining Radial Basis Function (RBF), Rational Quadratic, and Mate´rn kernels with respective weights of 0.68, 0.21, and 0.11, has demonstrated a significant improvement in predictive performance for product sales forecasting. By employing quantile transformation for data normalization, we have further increased the model's robustness against outliers and skewed distributions. The ensemble model achieved remarkable performance metrics, with a Mean Squared Error (MSE) of 0.025, Mean Absolute Error (MAE) of 0.048, Root Mean Squared Error (RMSE) of 0.039, and a Coefficient of Determination (R^2) of 0.974, markedly outperforming individual kernel models (RBF, Rational Quadratic, Matern) across all metrics. These results underscore the efficacy of our ensemble approach and Bayesian optimization in capturing complex, non- linear trends in sales data, showcasing a robust framework for advanced predictive modeling in various domains.

REFERENCES

[1] S. R. Dutta, S. Das, P. Chatterjee," Smart Sales Prediction of Pharmaceutical Products," in 2022 8th International Conference on Smart Structures and Systems (ICSSS), pp. 1-6, Apr. 2022, IEEE.

[2] S. Punia, S. Shankar," Predictive analytics for demand forecasting: A deep learning-based decision support system," in Knowledge-Based Systems, vol. 258, p. 109956, 2022.

[3] J. F. Torres, D. Hadjout, A. Sebaa, F. Mart´ınez-A´lvarez, A. Troncoso," Deep Learning for Time Series Forecasting: A Survey," in Big Data, vol. 9, no. 1, Mary Ann Liebert Inc., pp. 3–21, Feb. 01, 2021, doi: 10.1089/big.2020.0159.

[4] Pelton, Lou E., David Strutton, and James R. Lumpkin. "Marketing channels." A Relationship (1997).

[5] Watson IV, George F., et al. "The evolution of marketing channels: Trends and research directions." Journal of Retailing 91.4 (2015): 546-568.

[6] Kozlenkova, Irina V., et al. "The role of marketing channels in supply chain management." Journal of Retailing 91.4 (2015): 586-609.

[7] Alderson, Wroe. "Factors Governing the Development of Marketing Channels, in." (1954).

[8] P. K. Mianaei, M. Aliahmadi, S. Faghri, M. Ensaf, A. Ghasemi, and A. A. Abdoos, "Chance-constrained programming for optimal scheduling of combined cooling, heating, and power-based microgrid coupled with flexible technologies," Sustainable Cities and Society, vol. 77, p. 103502, 2022. [Online]. Available: https://doi.org/10.1016/j.scs.2021.103502.

[9] R. Rathipriya, A. A. Abdul Rahman, S. Dhamodharavadhani, A. Meero, G. Yoganandan," Demand forecasting model for time-series pharmaceutical data using shallow and deep neural network model," in Neural Computing and Applications, vol. 35, no. 2, pp. 1945-1957, 2023.

[10] Y. Han," A forecasting method of pharmaceutical sales based on ARIMA-LSTM model," in 2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT), pp. 336-339, Nov. 2020, IEEE.

[11] L. P. E. Yani, A. Aamer," Demand forecasting accuracy in the pharmaceutical supply chain: a machine learning approach," in International Journal of Pharmaceutical and Healthcare Marketing, vol. 17, no. 1, pp. 1-23, 2023.

[12] M. R. Moradi, S. R. N. Kalhori, M. G. Saeedi, M. R. Zarkesh, A. Habibelahi, et al., "Designing a Remote Closed-Loop Automatic Oxygen Control in Preterm Infants," Iran J Pediatr., vol. 30, no. 4, p. e101715, 2020. [Online]. Available: https://doi.org/10.5812/ijp.101715.

[13] S. Ratre, J. Jayaraj," Sales Prediction Using ARIMA, Facebook's Prophet and XGBoost Model of Machine Learning," in Machine Learning, Image Processing, Network Security and Data Sciences: Select Proceedings of 3rd International Conference on MIND 2021,

[14] pp. 101-111, Jan. 2023, Springer Nature Singapore, Singapore.

[15] A. E. Jery, M. Aldrdery, N. Ghoudi, M. Moradi, I. H. Ali, H. H. Tizkam, S. S. Sammen," Experimental Investigation and Proposal of Artificial Neural Network Models of Lead and Cadmium Heavy Metal Ion Removal from Water Using Porous Nanomaterials," in Sustainability, vol. 15, no. 19, p. 14183, 2023.

[16] R. Gustriansyah, E. Ermatita, D. P. Rini," An approach for sales forecasting," in Expert Systems with Applications, vol. 207, p. 118043, 2022.