CHAPTER I

INTRODUCTION

The importance of mathematical and statistical models in forecasting cannot be overstated, as they empower us to uncover patterns within historical data and make predictions about the future. The utilization of these models brings forth a multitude of benefits, including:

- Unveiling Hidden Trends: These models are capable of recognizing trends in historical data that may not be visible to the unaided eye.
- Forecasting a Various Set of Variables: By utilizing these models, we are able to forecast a wide range of variables, like those that present difficulties for conventional modeling methods.
- Accurate Future Predictions: Through mathematical and statistical models, we are able to predict the future with some degree of precision.
- Evaluation of Forecast Accuracy and Identification of Risks: These models can be employed to evaluate forecast accuracy and find potential risks related to anticipated outcomes.

There exist various types of mathematical and statistical models that can be utilized for forecasting, encompassing time series models, econometric models, and machine learning models, among others. It is important to note, however, that these models are not flawless and may occasionally produce inaccurate predictions. Nevertheless, they remain a valuable tool for forecasting future trends and facilitating well-informed decision-making processes (De Jong and Heller, 2008).

Generalized linear models (GLMs) are statistical models specifically designed to handle data that deviates from a normal distribution. This characteristic renders them highly valuable in the field forecasting since numerous real-world variables exhibit non-normal distributions. Take, for instance, sales data, which often displays a skew towards higher values. GLMs effectively capture such patterns and enable accurate forecasts for this type of data. Although GLMs are effective tools for modeling non-normal data, it is important to keep in mind that they may have limitations when dealing with data that does not follow a particular distribution or exhibit non-linear relationships. GLMs might not be able to fully capture the complexity of such data patterns, despite their advantages. Generalized additive models (GAMs) serve as an extension of GLMs, offering increased flexibility in capturing the intricate relationship between predictors and the response variable. Consequently, they prove to be an invaluable tool for forecasting since many real-world variables defy specific distribution assumptions and exhibit non-linear relationships (Zanin and Marra, 2012). The following are some advantages of extending GLMs to GAMs:

- Flexibility: GAMs allow for greater adaptability in modeling the complex relationships between predictors and the response variable. This is particularly advantageous when dealing with variables that do not adhere to specific distribution assumptions and showcase non-linear patterns.

- Accuracy: GAMs often yield more accurate forecasts compared to GLMs, especially when handling data that does not follow specific distributions or exhibits non-linear relationships. The flexibility of GAMs enables them to capture more nuanced patterns and variations, resulting in improved forecasting accuracy.

- Interpretability: GAMs can provide better interpretability compared to GLMs, as they can uncover and depict non-linear relationships between predictors and the response variable. This offers valuable insights into the underlying factors influencing the response variable, aiding in informed decision-making for future planning.

Here are some instances of applying GAMs to forecasting:

- Sales Forecasting: GAMs can be employed to forecast sales of products or services, enabling businesses to plan production, inventory management, and targeted marketing strategies more effectively.

- Demand Forecasting: GAMs can predict the demand for products or services, assisting businesses in optimizing resource allocation, staffing levels, pricing strategies, and inventory management.

- Risk Forecasting: GAMs can be utilized to forecast the likelihood of events such as defaults, bankruptcies, or natural disasters. This empowers businesses to assess potential financial exposures and make well-informed risk management decisions.

- Healthcare Forecasting: GAMs can forecast healthcare service demand, including hospital admissions and doctor visits. Healthcare providers can utilize these forecasts to optimize resource allocation, staffing, and financial projections.

Extending generalized linear models (GLMs) to double generalized linear models (DGLMs) can be important for forecasting in cases where the variance of the response variable is not constant. In GLMs, the variance of the response variable is assumed to be constant, but this assumption may not hold true in all cases. For example, the variance of sales data may be higher for products that are more popular. DGLM is a model that extends the GLM by allowing the variance of the response variable to be modeled as a function of the mean. This is in contrast to GLMs, which assume that the variance of the response variable is constant. In this case, extending GLMs to DGLMs can be a valuable tool for forecasting. DGLMs are more flexible, accurate, and interpretable than GLMs, making them a better choice for modeling data that does not follow a specific distribution and has non-linear relationships.

Croux, Gijbels, and Prosdocimi (2012) developed robust estimation methods for mean and dispersion functions within the framework of extended generalized additive models. Their approach utilized P-splines and quasi-likelihood methods to model both the mean and dispersion components, demonstrating improved performance in the presence of outliers. Gijbels and Prosdocimi (2011) further illustrated these concepts through applications to Italian abortion data, showing how flexible modeling of both mean and dispersion can reveal important patterns that would be missed by standard GLMs or GAMs with fixed variance structures. In 2022, Chang, Caruana, and Goldenberg extended this research by employing several techniques, namely Explainable Boosting Machine (EBM), splines, nodes, eXtreme Gradient Boosting (XGBoost), and Random Forest (RF), to find the parameters of GAMs models. The results of their study demonstrated that these models were capable of identifying interesting patterns within the data.While GAMs provide excellent interpretability of nonlinear relationships, their performance can be enhanced through integration with other machine learning techniques. Support Vector Regression (SVR), based on Vapnik's statistical learning theory (1995), has emerged as a complementary approach that excels in handling high-volatility data regions where GAMs may struggle. Hybrid GAM-SVR frameworks capitalize on the interpretability of GAMs for elucidating fundamental variable relationships while leveraging SVR's robust error minimization capabilities in regions characterized by extreme fluctuations (López et al., 2025). This synergistic combination enables predictive models that maintain both transparency and superior accuracy across diverse data conditions.

In this thesis, we develop and evaluate advanced parameter estimation techniques for Double Generalized Additive Models (DGAMs) that integrate traditional statistical approaches with modern machine learning methods, including Smoothing Splines, Support Vector Regression, Explainable Boosting Machine, eXtreme Gradient Boosting, and Random Forest. The research systematically compares multiple estimation frameworks to identify optimal strategies for different forecasting contexts and data characteristics. Comprehensive validation will be conducted using standard benchmark datasets and real-world health data encompassing critical life-stage outcomes including birth, aging, morbidity, and mortality patterns. Model evaluation will assess predictive accuracy, interpretability, computational efficiency, and robustness through rigorous cross-validation, sensitivity analyses, and practical case studies in healthcare forecasting applications.

1.1 Research Objectives

1. To develop techniques for determining model parameters in Double Generalized Additive Models (DGAMs) using Smoothing Splines, Explainable Boosting Machine, eXtreme Gradient Boosting, Random Forest, and Support Vector Regression. This will be accomplished by employing all proposed techniques to create forecasting models across multiple datasets, each containing multiple variables examined at various proportion levels.

2. To evaluate and compare the performance of all proposed methods for the obtained forecasts based on predictive accuracy, interpretability, computational efficiency, and robustness across different dataset configurations, variable selections, and proportion specifications.

1.2 Scope and Limitations

In this study, a range of techniques has been utilized to address the challenge of solving the double generalized additive models (DGAMs) problem. These techniques encompass spline, Explainable Boosting Machine (EBM), eXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Regression (SVR). Each technique brings its unique strengths and capabilities to the table, providing diverse approaches for effectively tackling the problem at hand. In addition, the proposed theoretical developments are empirically evaluated using four publicly available healthcare datasets obtained from Kaggle:

1. Fetal Health Monitoring Dataset;

2. Life Style Data;

3. Health Checkup Result Dataset;

4. Life Expectancy (WHO) Fixed Dataset.

1.3 Research Procedure

The research work will proceed as follows:

1. Study background techniques for estimating the mean and dispersion/variance in relation to covariates is critical for focusing on the flexibility of Generalized Additive Models.

2. Study and prepare the knowledge of mathematics essential to comprehend the techniques that the relationship uses to estimate model parameters.

3. Modify the source code of the statistical computing program to improve the prediction of the Double Generalized Additive Models.

4. Apply the improved code in conjunction with real-world data to generate relevant predictive or problem-solving abilities models.

5. Evaluate the predictive performance of the model by measuring its accuracy and analyze the experimental results to identify areas for potential future research advancement.

1.4 Expected Result

A performance analysis of DGAMs obtained using various techniques in parameter estimation will be presented.