S&P 500 prediction using Gaussian process

******

***A Project Report***

***submitted by***

**Manoah Noble**

***As part of the internship training program***

***under the supervision of***

**Shri. P. Rajashekhar, SO/E**

**Security Electronics and Cyber Technology Division**

**BHABHA ATOMIC RESEARCH CENTRE**

**Trombay, Mumbai – 400 085. INDIA**

**June 2024**

# CERTIFICATE

This is to certify that the project entitled “**S&P 500 prediction using Gaussian process**” has been carried out by **Manoah Noble** B. Tech. student of Karunya Institute of Technology and Sciences, Coimbatore from 27th May to 28th June 2024. The project training has been successfully completed at the Automation Systems Development Section (ASDS), Security Electronics and Cyber Technology Division (SECTD), Bhabha Atomic Research Centre, Mumbai, under my guidance.

**Signature of the Supervisor**

**Abstract**

S&P 500 index is a stock market index that tracks the performance of 500 large companies listed on stock exchanges in the United States. It is widely regarded as one of the best representations of the U.S. stock market and is a key benchmark for the overall health of the economy. In this project, we aim to predict the trend in the S&P 500 index from 1928 to 2018 using the Gaussian Process Regressor and gain insights into the working of the Gaussian Process Regressor while also dissecting the inference from the corresponding results.

Gaussian Process Regressor (GPR) is a kernel-based machine learning model that is specifically used for handling regression tasks. The key advantages of a GPR are its flexibility and its ability to provide probabilistic predictions with confidence intervals, making it useful in decision-making scenarios. The GPR has several parameters, of which the key parameter is the kernel function. The kernel function can be simplistically seen as the covariance matrix. It determines how the output values are correlated across different input values. Common kernels include Radial Basis Function (RBF), Matern, and others. In this project, we implement the Radial Basis Function, also known as the squared exponential kernel or the Gaussian kernel, due to its flexibility in capturing complex relationships in data. The RBF kernel, characterized by its length scale parameter, captures the covariance between data points and determines the smoothness of the learned function.

**Introduction**

Machine learning and artificial intelligence are transformative fields in modern technology, enabling systems to learn from data and make decisions or predictions without explicit programming. These technologies have revolutionized industries ranging from healthcare to finance by automating tasks, optimizing processes, and uncovering insights from vast amounts of data. At the core of machine learning and AI lies statistics, which provides the foundational principles and methodologies for understanding and manipulating data. Statistics offers crucial tools such as hypothesis testing, probability theory, and regression analysis, which are essential for modelling and interpreting patterns in data. Understanding statistics is pivotal because it enables us to assess the reliability of predictions, quantify uncertainty, and make informed decisions based on data-driven insights. Therefore, a solid grasp of statistics is indispensable for anyone working with machine learning and AI, as it forms the bedrock for developing accurate models and deriving meaningful conclusions from data-driven experiments and analyses.

In this report, we go through the statistical approach to machine learning as we delve deep into the realms of probability theory and Bayesian Statistics and finally end on the application of machine learning in performing regression tasks.

**Literature Review**

According to C. Wagh et al. (2023), controlling an induction motor drive requires accurate speed estimation, which is complex due to dependency on machine parameters. Gaussian Process Regression (GPR) offers a robust solution by handling noisy data and modelling uncertainties without the need for flux computation, thereby enhancing performance and simplifying evaluations compared to traditional methods.

In their paper, Eric S. et al. (2018) introduce Gaussian process regression as a powerful, non-parametric Bayesian tool for modelling, exploring, and exploiting unknown functions. The paper covers the basics of Gaussian processes, their application in regression problems, kernel-encoded prior assumptions, optimal design exploration, and bandit-like recommendation scenarios. Additionally, it discusses risk-averse exploration, summarizes recent psychological experiments using Gaussian processes, and provides software and literature pointers for further study.

J. Song et al. (2022) emphasize the importance of safety monitoring for dam operation, particularly deformation monitoring. They developed a high-accuracy dam deformation prediction model using Gaussian Process Regression (GPR), which is well-suited for non-stationary data. The accuracy of GPR relies heavily on the selection of hyper-parameters, and to enhance this accuracy, Bayesian Optimization (BO) was employed. Comparisons of the prediction results among BP, GPR, and BO-GPR models indicated that the BO-GPR model achieved higher prediction accuracy, making it an effective method for dam deformation analysis and prediction.

S. Bin and Y. Wenlai (2013) introduced Gaussian Process Regression (GPR) to predict thermal comfort using parameters like activity level, clothing insulation, air temperature, humidity, velocity, and radiant temperature. Their study found GPR closely matched Fanger's equation and outperformed BP neural networks and SVMs in precision and adaptability. This makes GPR ideal for real-time PMV index control in air-conditioning systems.

**Objective of the project work**

The objective of this project is to predict the trend of the S&P 500 index from 1928 to 2018 using the Gaussian Process Regressor (GPR). The project aims to gain insights into the workings of the GPR model, specifically its capability to handle regression tasks, provide probabilistic predictions with confidence intervals, and capture complex relationships in data. Additionally, the project seeks to evaluate the model's performance in both interpolation and extrapolation scenarios, thereby understanding its strengths and limitations in real-world applications such as economic forecasting.

Moreover, the project intends to delve into the theoretical underpinnings of the GPR, exploring its kernel functions, particularly the Radial Basis Function (RBF), and their role in determining the smoothness and covariance of the model. By implementing both library-provided and statistical formula-based versions of GPR, the project aims to compare and contrast these approaches, highlighting the practical implications of theoretical concepts. Ultimately, this comprehensive exploration will provide a robust understanding of the GPR's applicability and effectiveness in economic prediction tasks.

**Project Timeline**

The project training conducted at the Reactor Control Division (RCnD), BARC started on the 27th of May, 2024 and spanned for a period of over 30 days till the 28th of June, 2024 under the guidance of Shri. P. Rajashekar, who recommended going through the book Machine Learning: A Probabilistic Perspective by Kevin P Murphy, MIT.Press,.2012 to have a grasp of the deeper meaning of machine learning and understand the role of probability in predicting outputs. From the outset, it was emphasized that any predictions generated by the model must not only provide a specific value but also convey probabilistic significance. This includes establishing confidence intervals that indicate the model's confidence level in its predictions and the potential margin of error associated with those predictions. This approach ensures clarity on how certain the model is when predicting outputs and provides insights into the accuracy and reliability of its prediction.

The training spanned for a period of 5 weeks, where the first 3 weeks were utilized to develop a basic understanding of the probabilistic approach while following the book Machine Learning: A Probabilistic Perspective by Kevin P Murphy.

Chapter 1: Introduction

Chapter 2: Probability

Chapter 3: Generative Model for Discrete Data

Chapter 4: Gaussian Model

Chapter 5: Bayesian Statistics

For weeks 4 and 5, we aimed to implement the Gaussian Model and try to understand the working of the Gaussian Model, which was highlighted in Chapter 4 of the book and further explained in detail in Chapter 15: Gaussian Processes.

The implementation can be seen in detail in the next section

**Implementation**

***Task 1: Predicting the sine wave***

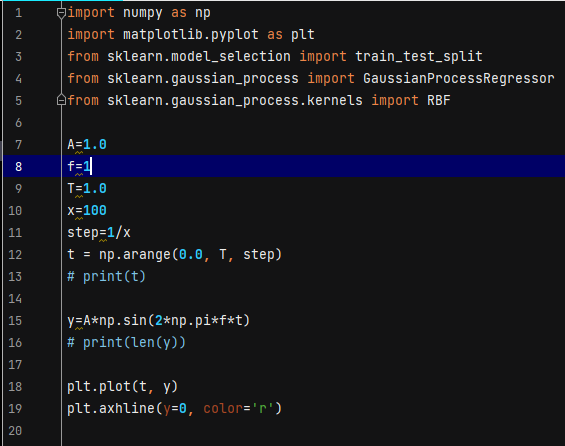
The first task at hand was to generate a sine wave by taking a random sample of 100 points in the wave and using it to predict the sine wave. The task can be identified as a regression task and the model chosen to perform the task was the Gaussian Process Regressor. The Gaussian Process Regressor (GPR) is a kernel-based machine learning model designed specifically for regression tasks. Its primary strengths lie in its flexibility and its capability to offer probabilistic predictions along with confidence intervals, which are invaluable for informed decision-making. A crucial parameter in GPR is the kernel function, akin to a covariance matrix, that dictates how output values correlate across varying input values. Common kernel options encompass the Radial Basis Function (RBF), and Matern, among others. For this project, we employ the Radial Basis Function, also known as the squared exponential or Gaussian kernel, chosen for its adeptness in capturing intricate data relationships. The RBF kernel is characterized by a length scale parameter that governs the covariance between data points and influences the smoothness of the modelled function.

***Task 1a) Implementation using the sklearn Package***

Instead of diving head first into the statical implementation, we used the Gaussian Process Regressor module provided by sklearn to understand the workings of the Gaussian Process Regressor and essentially how the results should look.

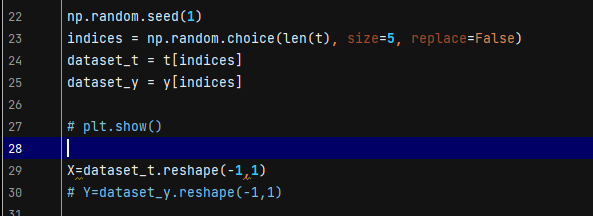
***Steps involved:***

**Step 1**: Import Libraries and Generate the Sine Wave



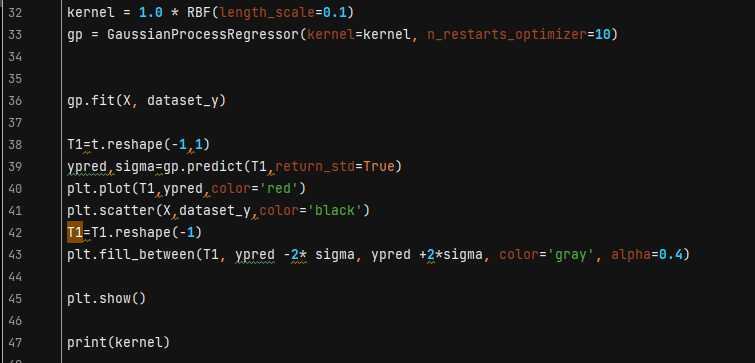
Code 4.1.1.1

**Step 2**: Selecting 5 random samples for training



Code 4.1.1.2

**Step 3**: Fitting the dataset and plotting the result

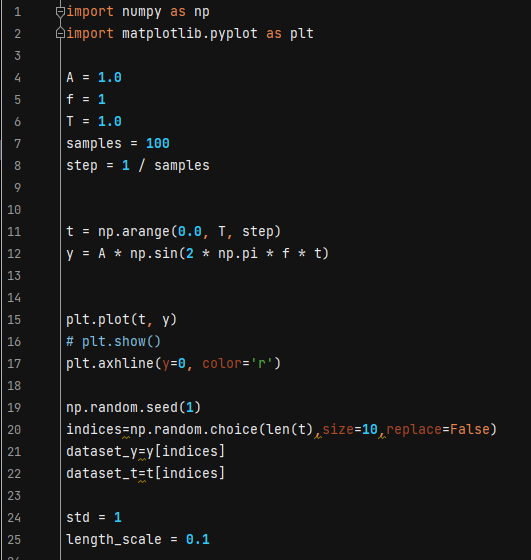


Code 4.1.1.3

***Task 1b) Implementation of the statistical formulas of the mean and its variance***

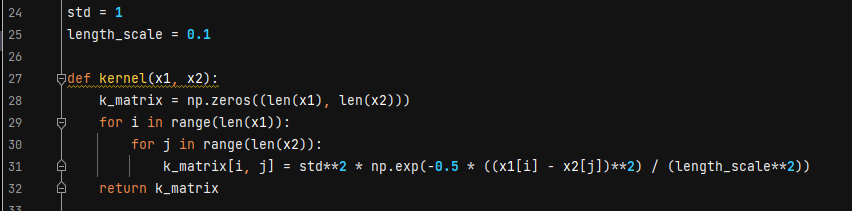
Once the general understanding of the GPR was gained and the outputs were visualized, it was time to implement the model using the statistical formulas of the mean and its variance. The formulas used for implementing the mean, variance and kernel functions can be seen in Section 15.2.1 of the book

**Step 1**: Importing Libraries, Generating the Sine wave and splitting the dataset into train and test.



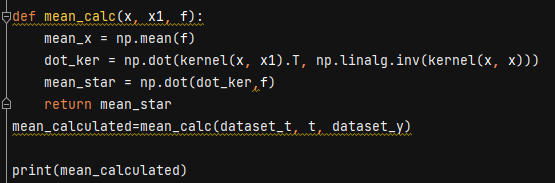
Code 4.1.2.1

**Step 2**: Initializing the parameters and the kernel function



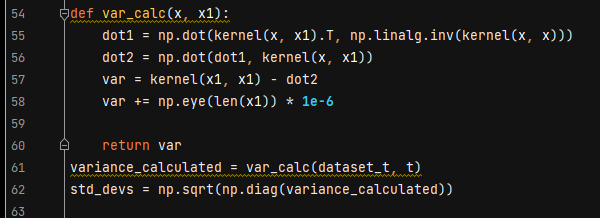
Code 4.1.2.2

**Step 3**: Calculating the mean



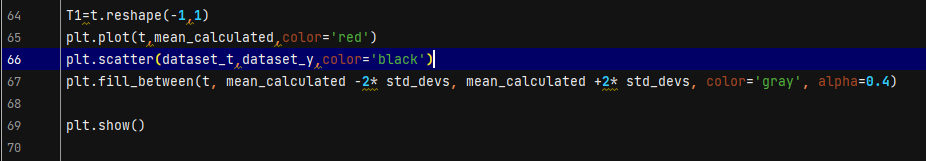
Code 4.1.2.3

**Step 4**: Calculating the variance



Code 4.1.2.4

**Step 5**: Plotting the mean and the variance

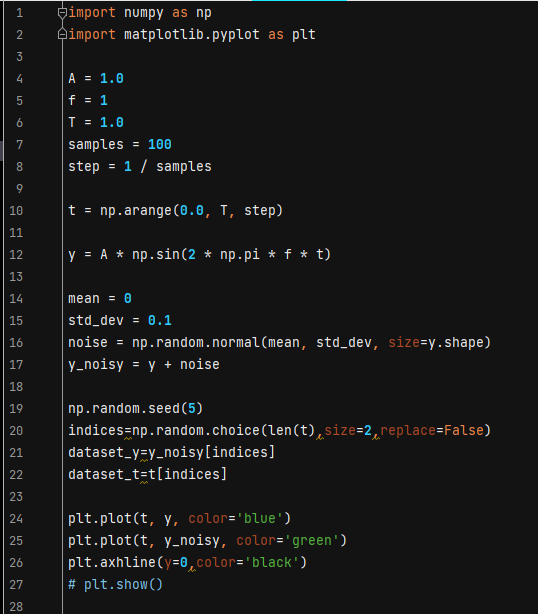


Code 4.1.2.5

***Task 1 c) Implementation of GPR for noisy observations***

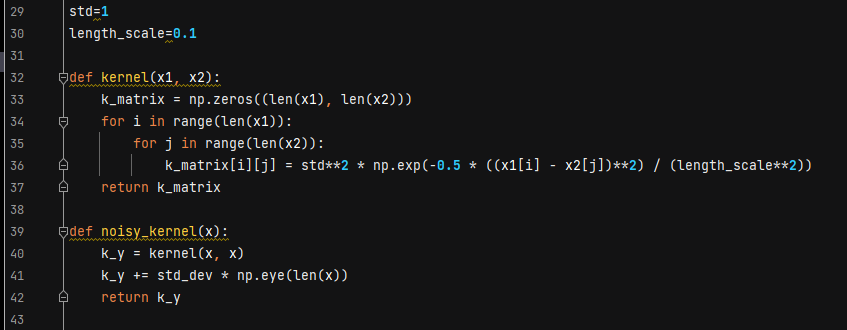
For this part of the task, the performance of the GPR was tested when it had to deal with noisy data. The formulas used in the implementation can be referred to in Section 15.2.2 of the book.

**Step 1**: Importing Libraries, Generating the Sine wave, Introducing Gaussian Noise and splitting the dataset into train and test



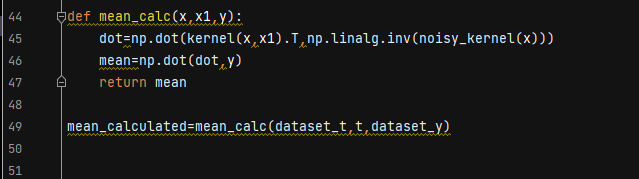
Code 4.1.3.1

**Step 2**: Introducing parameters and implementing the kernel function



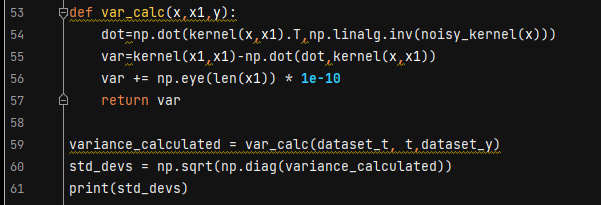
Code 4.1.3.2

**Step 3**: Mean Calculation



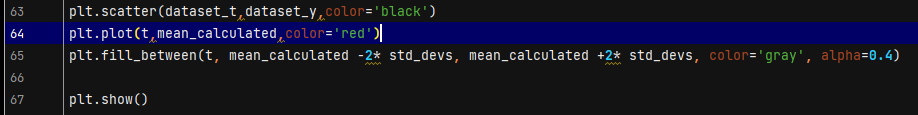
Code 4.1.3.2

**Step 4**: Variance calculation



Code 4.1.3.3

**Step 5**: Plotting the results



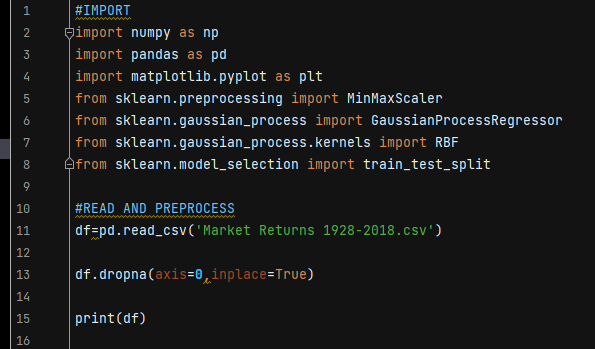
Code 4.1.3.4

***Task 2: Predicting the trend of the S&P 500 from 1928 to 2018.***

The second task was an extension of the first task where a deeper understanding of the Gaussian Process Regressor could be gained by implementing it in real-world applications. For this task a dataset from “Kaggle’ was chosen, namely “Market Returns 1928-2018.csv”. The initial aim of this task was to use the data from 1929 when the Great Depression took place, to predict the Great Recession that took place in 2009. However, this could not be achieved which will be later explained.

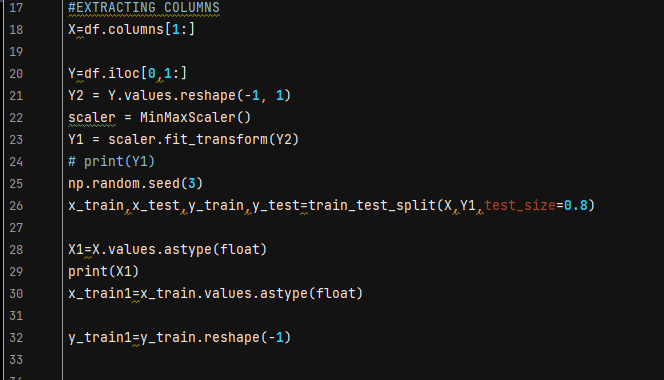
The process involved in handling this task followed the same blueprint as used in the first task. However, there still were many hurdles on the way and a lot of new insights could be gained.

**Step 1**: Importing the necessary libraries and reading the data



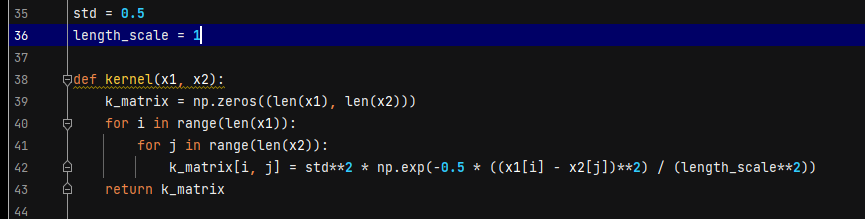
Code 4.2.1

**Step 2**: Normalizing the data and splitting it into train and test set



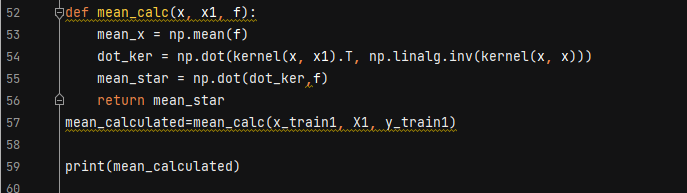
Code 4.2.2

**Step 3**: Initializing parameters and defining the kernel function



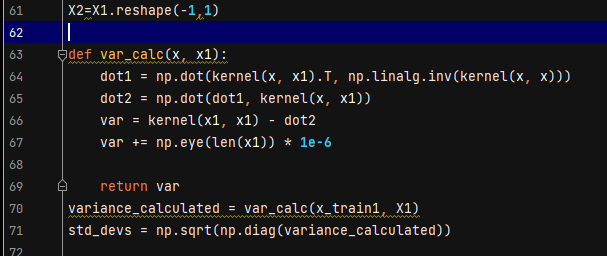
Code 4.2.3

**Step 4**: Calculating the mean



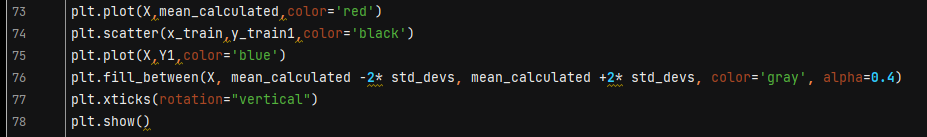
Code 4.2.4

**Step 5**: Calculating the variance



Code 4.2.5

**Step 6**: Plotting the results



Code 4.2.6

**Result**

***Task 1a) Visualizing the Results of the implementation of the sine wave using the sklearn package***

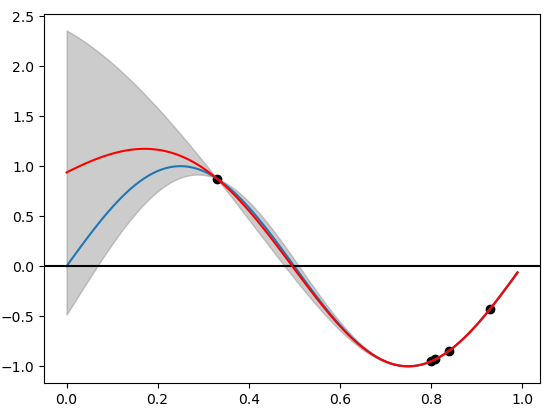


Fig 5.1

Fig 5.1 depicts the behaviour of the GPR while trying to emulate the sine wave. The Blue curve represents the original sine wave whereas the red curve depicts the prediction made by the GPR.

***Task 1b) Implementation of the statistical formulas of the mean and its variance***

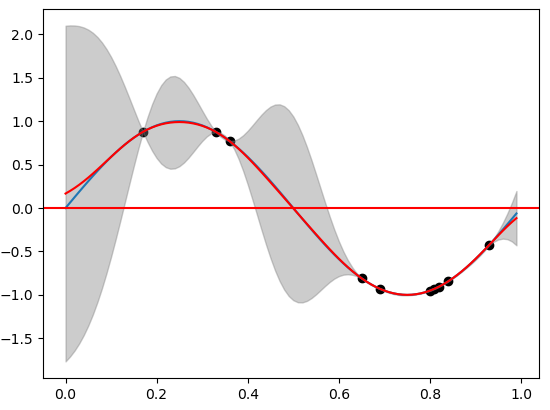


Fig 5.2

Figure 5.2 represents the curve that is being generated by the statistical formulas of the mean and variance. A common trend while we correlate the two figures is that the variance becomes very low near the data that is being used for training whereas at points where there are no trained data, even though the model can predict the curve, the variance is high.

***Task 1 c) Implementation of GPR for noisy observations***

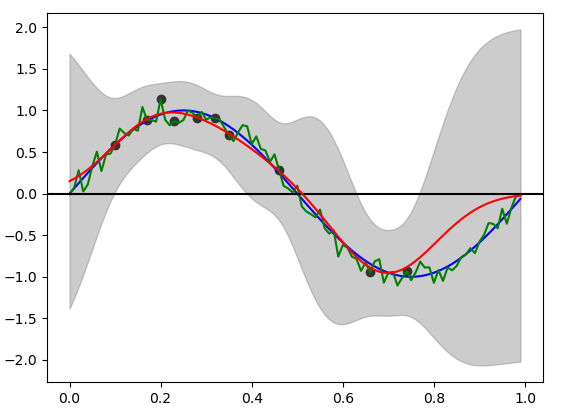


Fig 5.3

Fig 5.3 shows the performance of the model when it is being trained using noisy data. As we can see, the inference we can infer is that once the noise increases, the variance also increases. But still, the advantage of the GPR can be seen here as it is still able to predict the curve.

***Task 2: Predicting the trend of the S&P 500 from 1928 to 2018.***

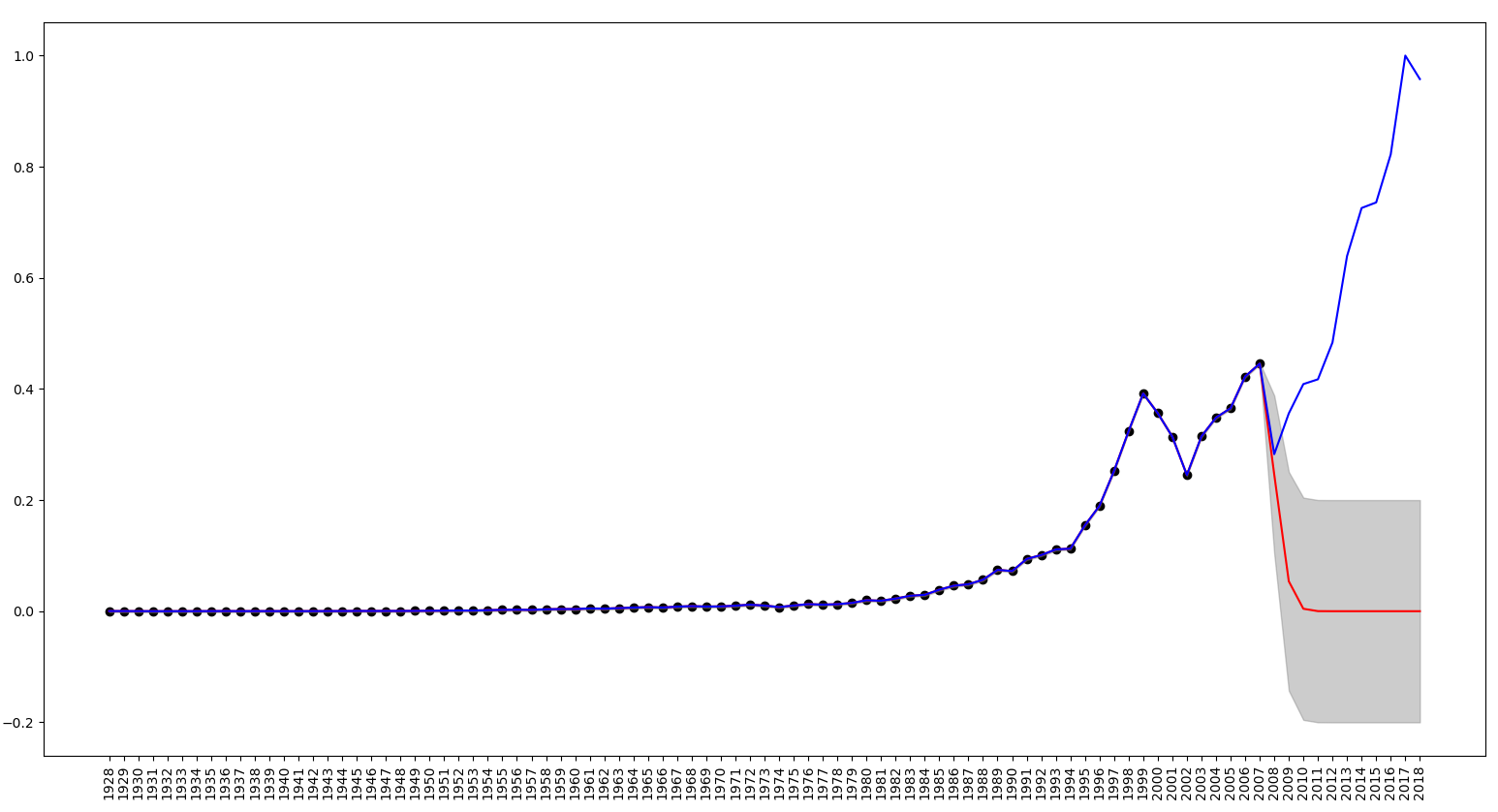


Fig 5.4

As mentioned previously, the initial aim was to predict the Great Recession using the trends of the Great Depression in 1929. This was not possible to predict using GPR as we can see in Fig 5.4. This brings us to our second important conclusion that Gaussian Process Regressors are regression models that extremely well in tasks that require interpolating data rather than extrapolating them. That is, it specializes in predicting the missing data rather than predicting the unknown data.

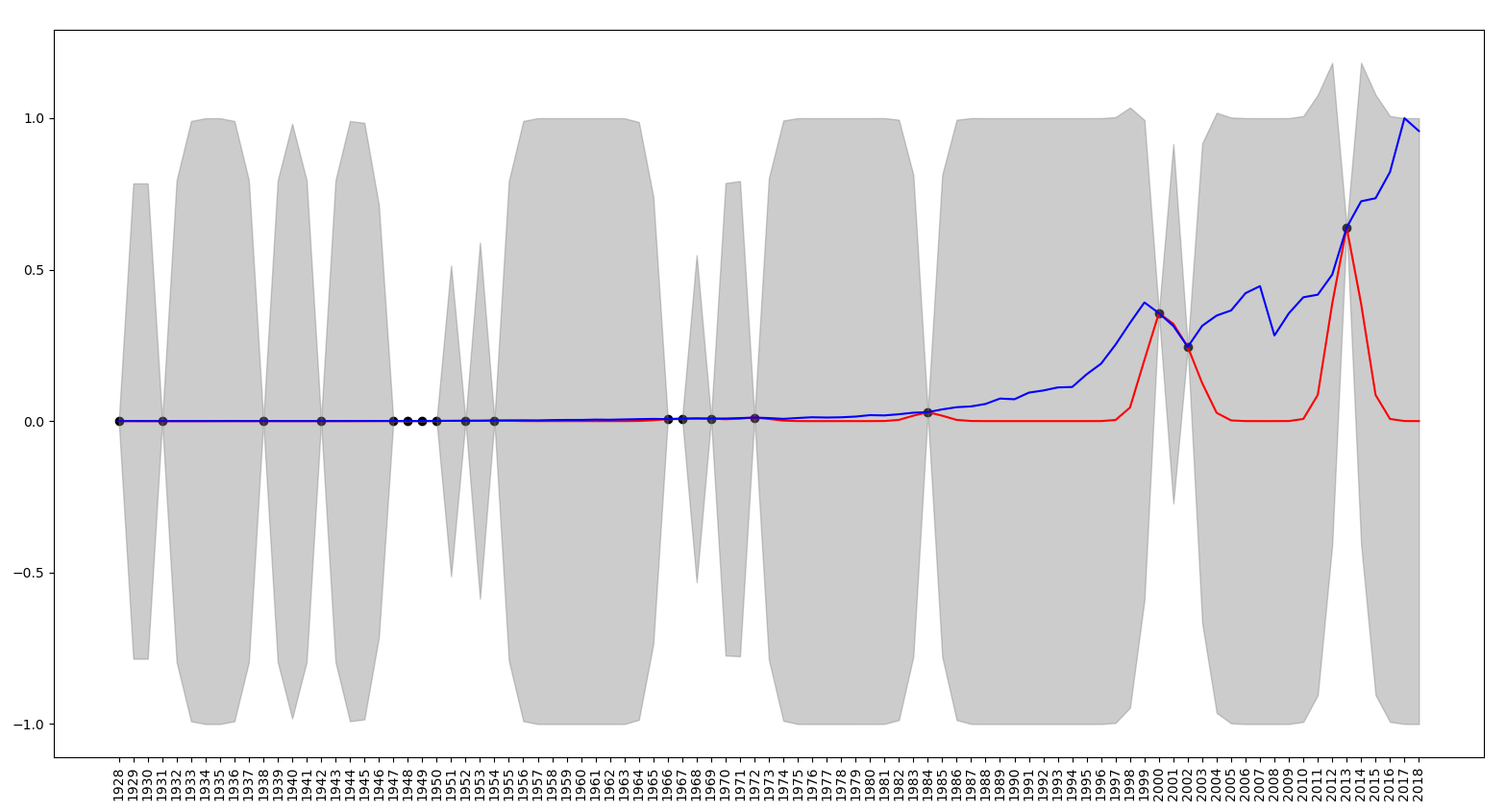


Fig 5.5

Hence due to the failure shown in Fig 5.4, we proceeded with interpolating missing data and split the data randomly into train and test sets. Fig 5.5 shows the behavior of the GPR when passed with 20% of the data

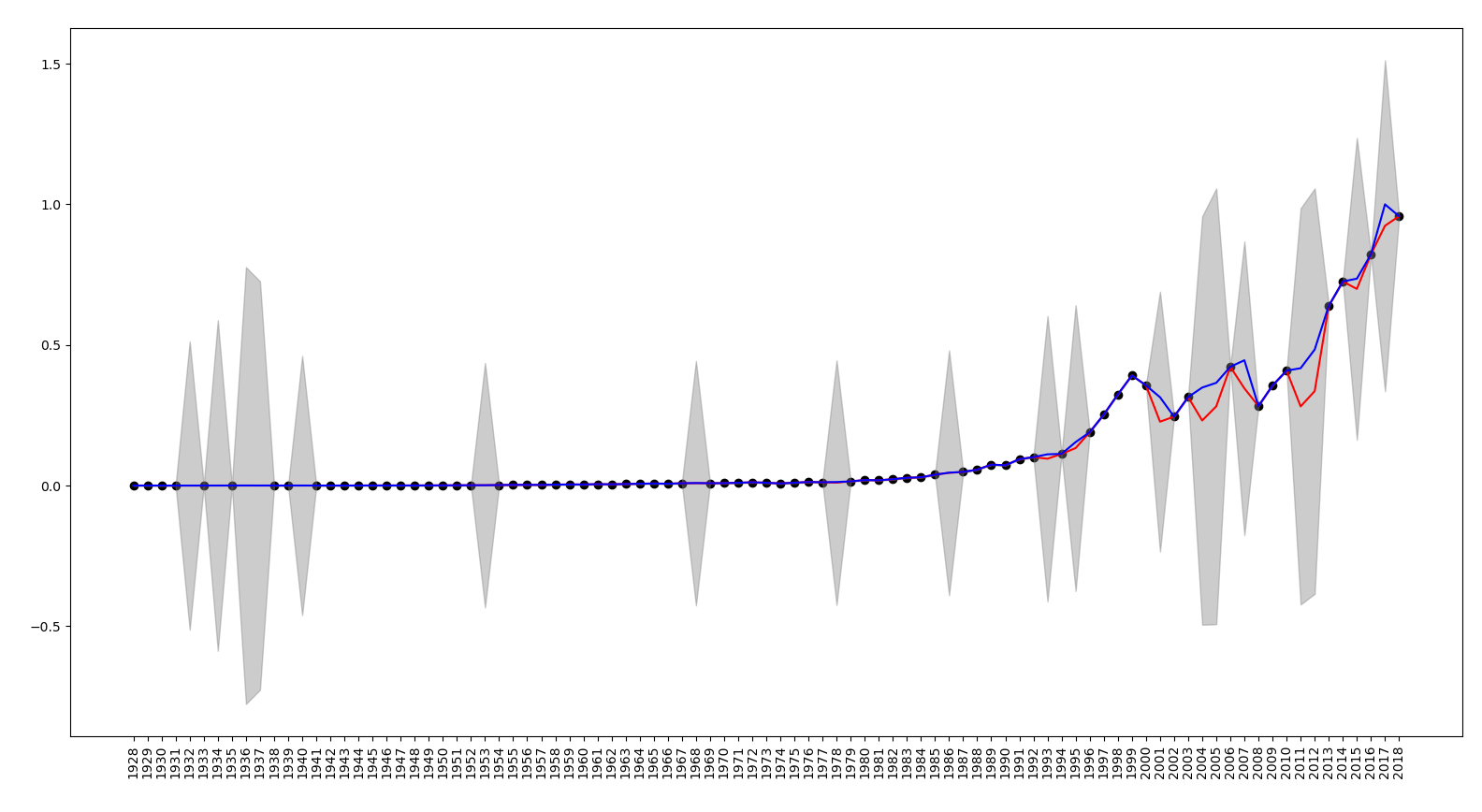


Fig 5.6

Fig 5.6 depicts the behaviour of the GPR when trained with more data. In this case, 80% of the data is being used as the training set and we can see the variance significantly decreasing. This shows that the model is much more confident while predicting the data as compared to Fig 5.5. This brings us to our third important conclusion that the model becomes more confident as it receives significantly more data.

**Conclusion**

Based on the detailed exploration and implementation of the Gaussian Process Regressor (GPR) in this project, several key insights have been gleaned regarding its application and performance.

Firstly, the GPR model, leveraging the Radial Basis Function (RBF) kernel, demonstrated its flexibility and capability in predicting complex relationships in data, such as emulating sine waves and handling noisy observations. The model's ability to provide probabilistic predictions with accompanying confidence intervals proved invaluable, offering insights into the uncertainty associated with predictions.

In the context of real-world applications to predicting the S&P 500 index, the GPR excelled in interpolating data within the observed range but faced limitations in extrapolating beyond it. Specifically, attempts to predict economic downturns like the Great Recession using patterns from the Great Depression era (1929) were unsuccessful. This underscores a critical aspect of GPR: its proficiency in interpolating existing data but its caution in extrapolating to unknown scenarios.

Furthermore, the experiment highlighted the significance of data quantity in model performance. As observed, increasing the size of the training dataset led to reduced prediction variance, indicating higher confidence in predictions. This reinforces the importance of sufficient and representative data for enhancing model accuracy and reliability.

In conclusion, while Gaussian Process Regressors are robust tools for tasks involving data interpolation and predictive modelling within known ranges, they require careful consideration and supplementary techniques when applied to scenarios requiring extrapolation.

**Scope for future work**

Future advancements could explore hybrid approaches combining GPR with complementary models to enhance its predictive capabilities across broader ranges of data. Thus, this project not only deepened our understanding of GPR but also illustrated possibilities for further research and application in predictive analytics and economic forecasting.

**References**

[1] C. Wagh, C. Shivam, G. Revati, S. Shadab and S. R. Wagh, "Speed Estimation of Induction Motor Using Gaussian Process Regression," 2023 9th International Conference on Control, Decision and Information Technologies (CoDIT), Rome, Italy, 2023, pp. 01-06,

<https://doi/10.1109/CoDIT58514.2023.10284307>.

[2] Eric S, Maarten S, Andreas K, A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions, Journal of Mathematical Psychology, 85, 2018, Pages 1-16, ISSN 0022-2496,

<https://doi.org/10.1016/j.jmp.2018.03.001>**.**

[3] J. Song, Y. Chen, M. Duan and Z. Zeng, "Dam deformation monitoring data prediction using Bayesian optimization Gaussian process regression," 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2022, pp. 980-983,

<https://doi.org//10.1109/ICSP54964.2022.9778799>.

[4] S. Bin and Y. Wenlai, "Application of Gaussian Process Regression to prediction of thermal comfort index," 2013 IEEE 11th International Conference on Electronic Measurement & Instruments, Harbin, China, 2013, pp. 958-961,

<https://doi.org/10.1109/ICEMI.2013.6743191>.

.