INSTITUTE FOR ADVANCED

COMPUTING & SOFTWARE DEVELOPMENT

AKURDI, PUNE 411044

Documentation on

Insurance Premium Prediction

PG-DBDA AUG 2024

Submitted By-

Group No:24

Aradhya Srivastava (248512)

Shraddha Sudanrao Rananavare (24535)

Mrs Priyanka Bhor
Project Guide

Mr. Rohit Puranik
Centre Coordinator

DECLARATION

I, the undersigned hereby declare that the project report titled " Insurance Premium Prediction" written and submitted by us to the Institute for Advanced Computing & Software Development Akurdi Pune in the fulfilment of requirement for the award of degree of Post Graduate Diploma in Big Data Analytics (PG-DBDA) under the guidance of Mr.Priyanka Bhor

is original work and have not copied any code or content from any source without proper attribution ,and I have not allowed anyone else to copy our work.

The project was completed using Python libraries and ML models. This is developed as a part of our academic coursework. I also confirm the project is original, and has not been submitted previously for any academic or professional purpose.

Place: Signature

Date: Name: Aradhya Srivastava/

Shraddha Sudanrao Rananavare

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Mrs. Priyanka Bhor, Project Guide, for providing me with the guidance and support to complete this academic project. Their valuable insights, expertise, and encouragement have been instrumental in the success of this project.

I would also like to thank my fellow classmates for their support and cooperation during the project. Their feedback and suggestions were helpful in improving the quality of the project.

I would like to extend my gratitude to Mr. Rohit Puranik, Centre Goordinator, for providing me with the necessary resources and facilities to complete this project. Their support has been crucial in the timely completion of this project.

Finally, I would like to thank my family and friends for their constant encouragement and support throughout the project. Their belief in me has been a constant source of motivation and inspiration.

Thank you all for your support and guidance in completing this academic project.

ABSTRACT

The rising complexity of insurance pricing has necessitated the use of advanced machine learning techniques to predict premiums accurately. This project focuses on building a robust Insurance Premium Prediction model using timeseries forecasting techniques. The dataset for this study was sourced from Kaggle's Brazilian market insurance data, which contains extensive details about insurance premiums and policyholders.

Data preprocessing techniques, including handling missing values, outlier detection, and feature scaling, were implemented to ensure the dataset's quality and integrity. Exploratory Data Analysis (EDA) was conducted to uncover patterns, correlations, and trends in the data, providing insights into the primary drivers of premium variations.

For predictive modelling, we experimented with various time-series forecasting models, including ARIMA, SARIMA, and Facebook Prophet. Model evaluation was performed using performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared values to determine the most effective algorithm for premium prediction. After thorough analysis and model comparison, Facebook Prophet was selected as the final model due to its superior performance in capturing trends and seasonality in the data.

To enhance usability, we deployed the predictive model on Amazon EC2, allowing seamless interaction between users and the model. This deployment enables insurance companies to receive real-time premium predictions, facilitating better risk assessment and competitive pricing strategies.

The results of this study demonstrate the effectiveness of time-series forecasting in automating and optimizing insurance premiums pricing. By accurately predicting premiums, insurance companies can enhance customer satisfaction, reduce underwriting risks, and improve profitability. Future work includes incorporating additional external factors, such as economic indicators and regulatory changes, to further refine the prediction model.

This project serves as a significant step toward leveraging data-driven methodologies in the insurance sector, promoting transparency and efficiency in premium determination.

INDEX

LIST OF FIGURES

Chapter 1. Introduction

1.1) Problem Statement- The goal of this project is to develop a machine learning model that accurately predicts insurance prices based on various customer features. By analysing features, the model should be able to estimate the price of insurance premiums for new customers. The prediction will help insurance companies optimize pricing, assess risk more effectively, and provide personalized quotes to potential clients.

1.2) Aims & Objectives- Insurance pricing is a complex process influenced by multiple factors such as customer demographics, vehicle details, driving history, and claim records. Accurate price prediction can improve decision-making for insurance companies and provide tailored offers to clients. The objective of this project is to predict insurance premiums using machine learning techniques. The goal is to build a model that can predict accurate insurance premiums for new customers, ultimately aiding in risk assessment, pricing strategy optimization, and customer satisfaction.

1.3) Scope: The scope of our project focuses on leveraging data analytics and machine learning techniques to predict the premiums for various types of insurance policies, such as health, life, or auto insurance. The project involves analysing historical data, customer demographics, claims history, and other relevant features to identify patterns and factors that influence pricing decisions.

Chapter 2. Overall Description

2.1) Workflow diagram-

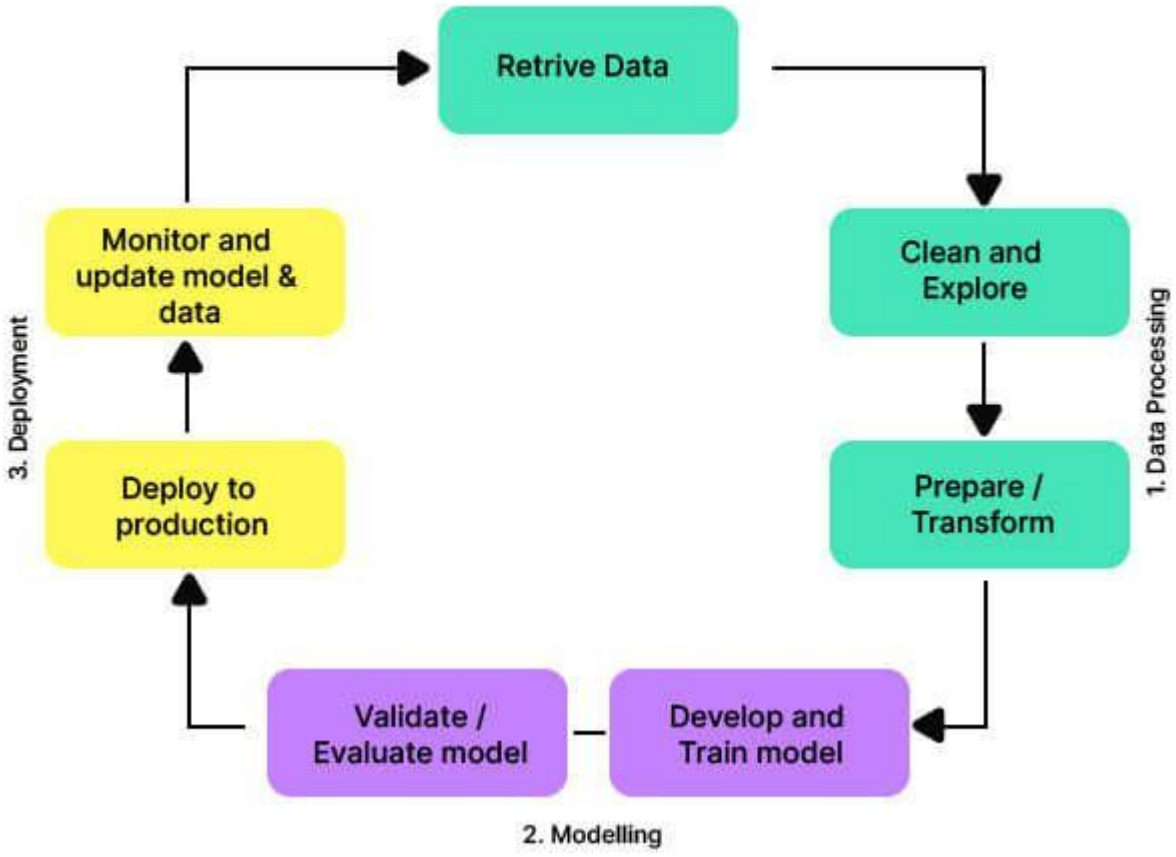

Fig2.1- Workflow

2.2) Data Cleaning and Preprocessing-

- Since our dataset is very huge of 80lakh rows of different companies their products and different state spanning from 2003 January to 2023. Hence we decided to focus on a product of a company which has the highest premiums.

- Converting year_month column to datetime

- Checking for missing values in the data if present replacing by sum

- Outlier handling of our target column ‘premiums’

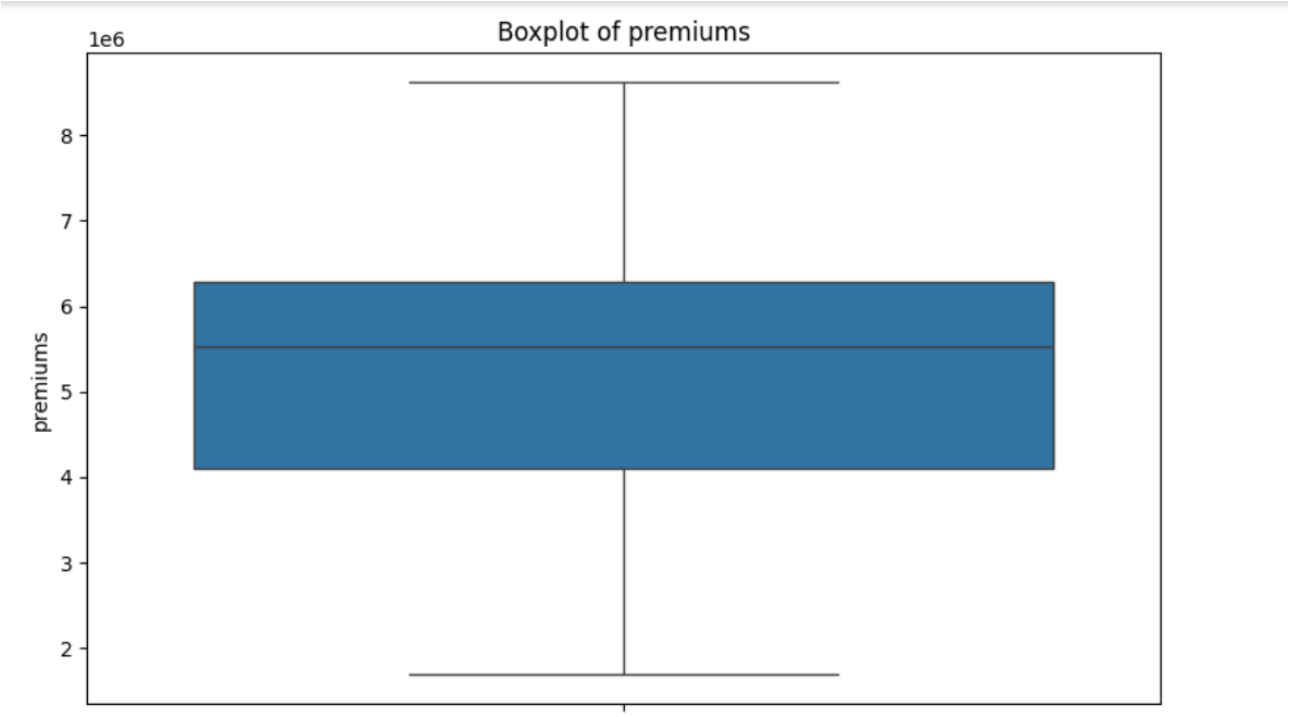

Fig 2.2.1- Boxplot to check outliers in Premium column

-Outlier Detection

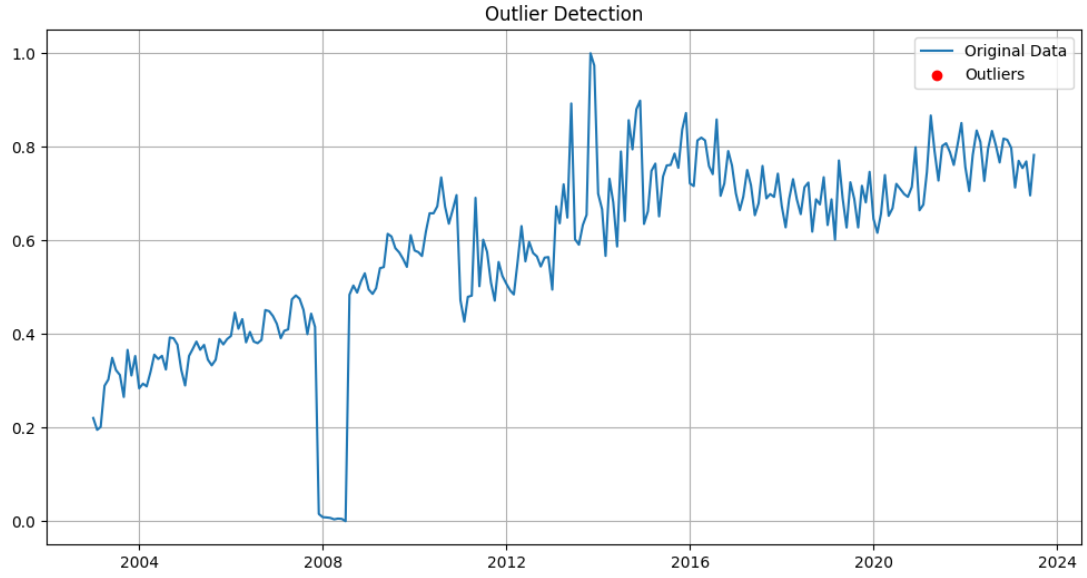

Fig 2.2.2- To check outliers

```
Outliers detected:
             company_code                    company_name  \
year_month_f
2006-09-01           6785  BRASILSEG COMPANHIA DE SEGUROS

                          product state  premiums   claims  \
year_month_f
2006-09-01    0993 - Vida em Grupo    TO       1.0 -5864.95

             claim_premium_ratio  premiums_diff    zscore
year_month_f
2006-09-01               -0.0024       0.780764  5.909851
```

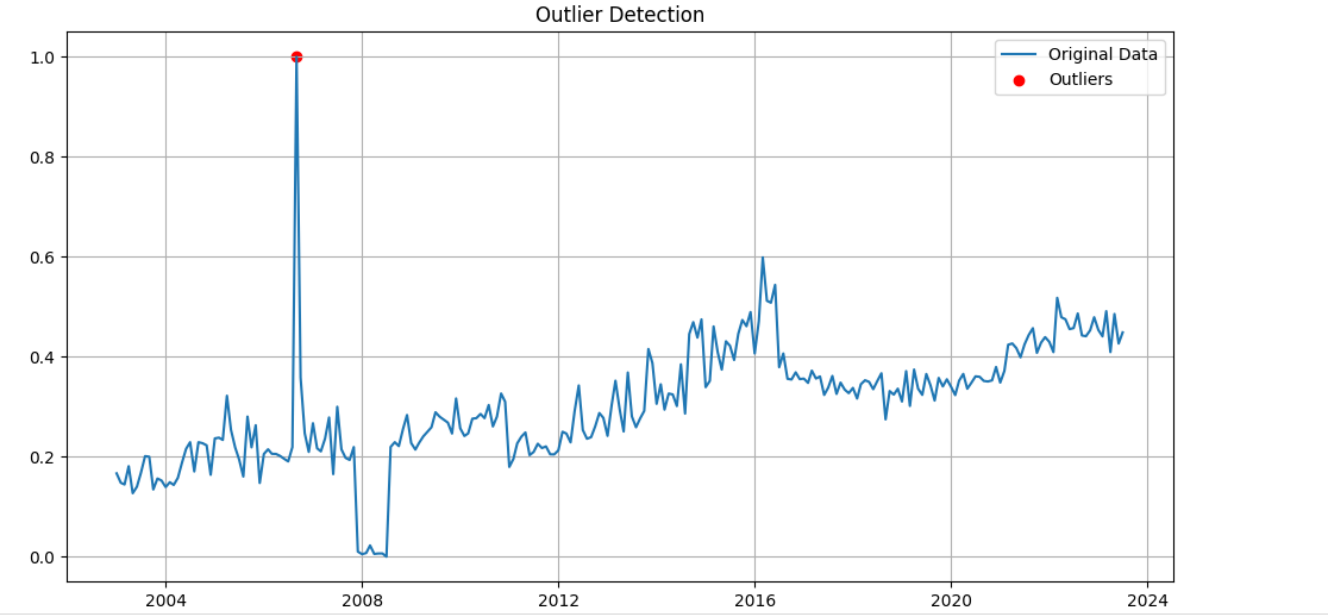

Fig 2.2.3- Significant outliers

- In 2008, the Brazilian market experienced a strong economic boom fuelled by domestic demand, but this was abruptly interrupted towards the end of the year by the global financial crisis, leading to a sharp slowdown in growth as exports declined and credit became tighter, with the industrial sector being particularly affected by the downturn.

- Checking distribution of other columns in dataset

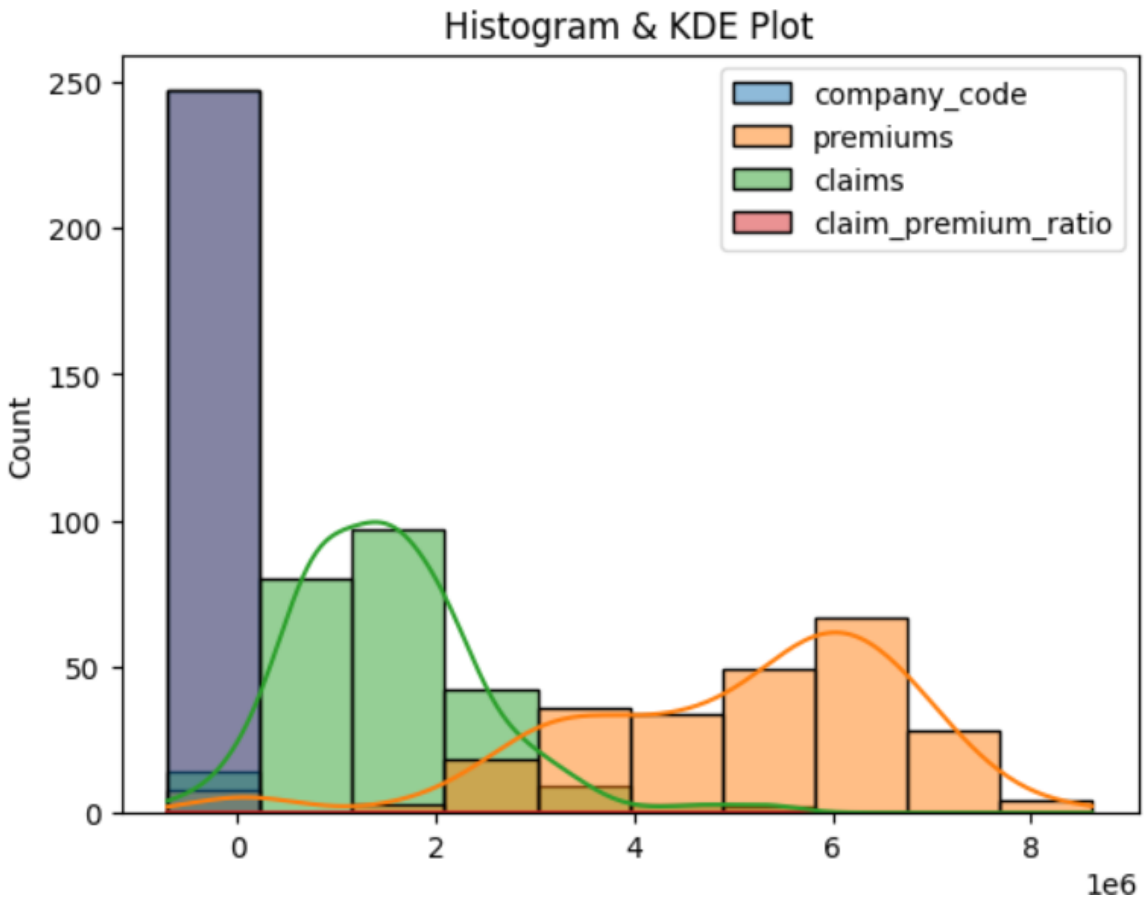

Fig 2.2.4- Plot to check distribution of various columns

-Checking the shape of data (skewness, kurtosis) and then applying min-max scaler for normalization

- One hot encoding for 3 columns (product,state and company_name) as they are categorical

- Creating plots for premium column with rolling windown , mean and std deviation

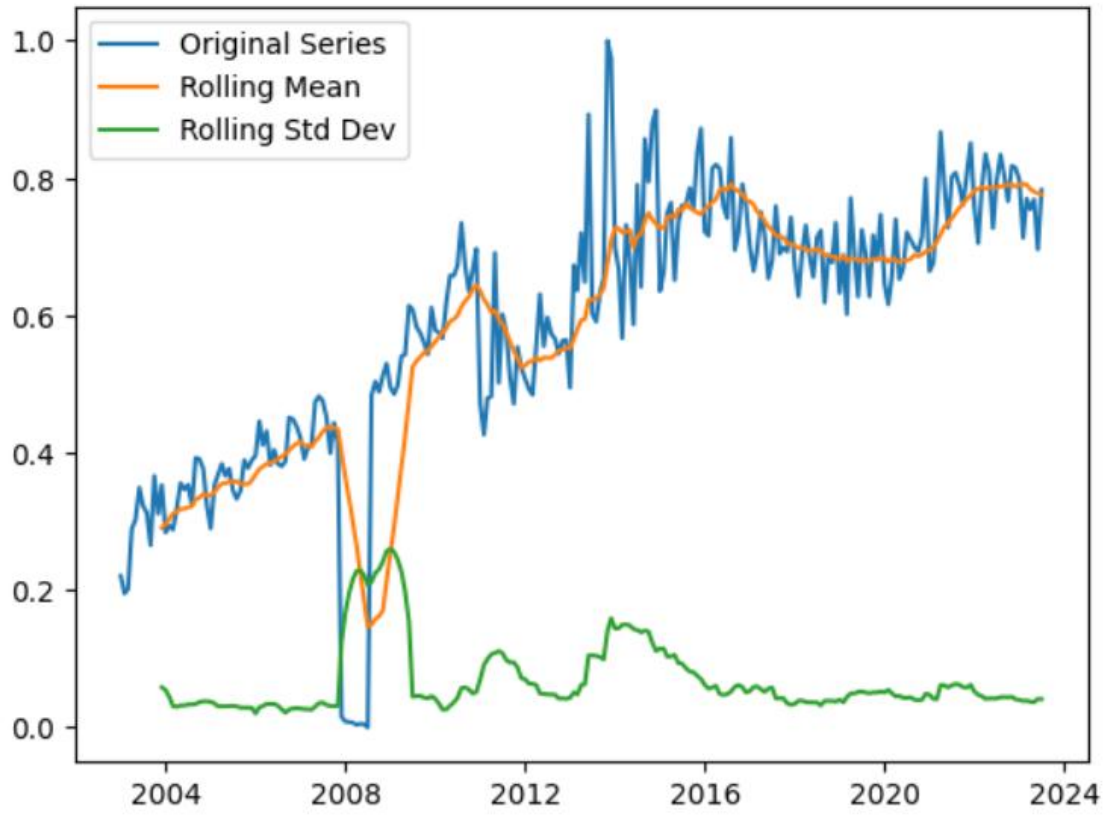

Fig 2.2.5- Line chart to check Rolling mean, Rolling Std_Dev

Key Insights:

- The period around 2008 shows a significant disruption, which reflects the Global Financial Crisis of 2008-2009
- After 2008, the rolling mean shows a smoother trend with fewer sharp fluctuations, and the standard deviation stabilizes.
- This suggests that the data becomes more predictable and less volatile after this period.

- Summary statistics of premiums-

count	247.000000
mean	0.583706
std	0.196570
min	0.000000
25%	0.447281
50%	0.632887
75%	0.725663
max	1.000000

- ADF test to check stationarity of data –

```
ADF Statistic: -1.5794087195485458
p-value: 0.4939283833112437
Critical Values: {'1%': -3.4589796764641, '5%': -2.8741347158661448, '10%': -
The data is not stationary. Differencing is required.
```

p-value clearly shows data isn't stationary

- Techniques to make data is stationary

a) Differencing – Differencing removes non-stationary components like trends and seasonality, making the data suitable for statistical modelling.

After applying differencing -

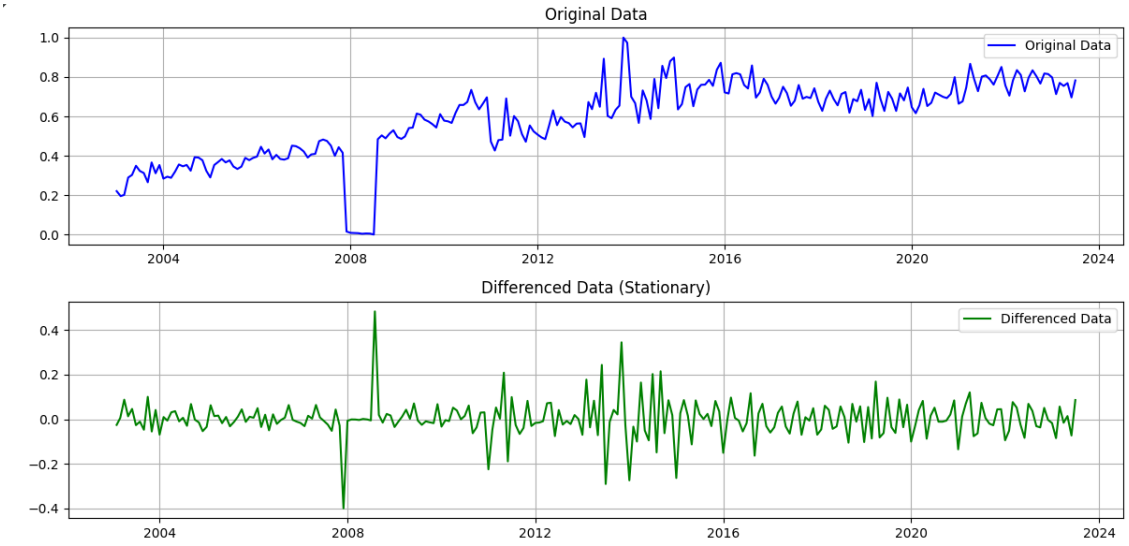

Fig 2.2.6- Original and Differenced data plots

b) Seasonal decompose- Breaks the series into components (trend, seasonality, residual) for analysis.

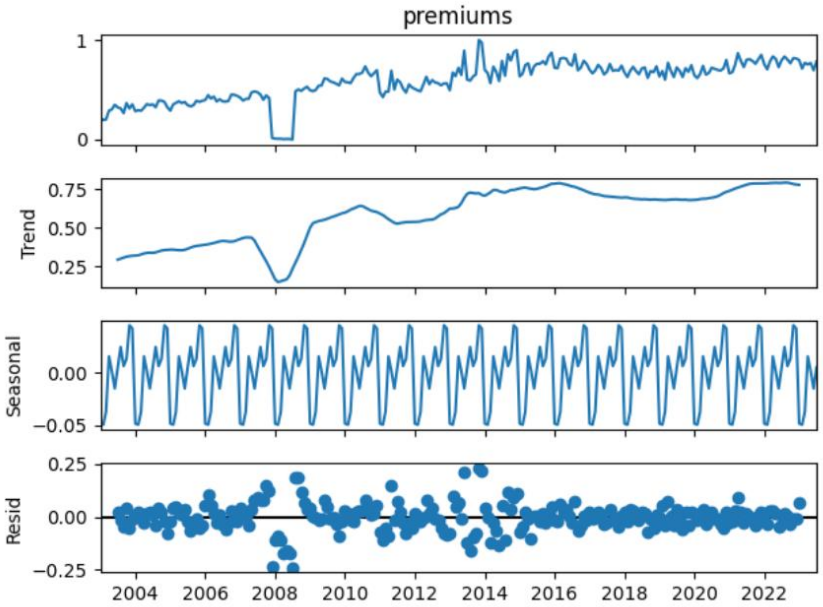

Fig 2.2.7-Breakdown into plots showing Trend, Seasonality, Residuals

Chapter 3. Model Training

1) SARIMA-

SARIMA (Seasonal Autoregressive Integrated Moving Average) extends ARIMA by incorporating a **seasonal component**, making it suitable for time series data that exhibits repeating patterns over fixed intervals (e.g., monthly sales, quarterly revenue). It adds seasonal terms to ARIMA, denoted as **(p, d, q) × (P, D, Q, s)**, where:

(P, D, Q, s) represent the **seasonal** counterparts:

- **P** (Seasonal AR terms)
- **D** (Seasonal differencing)
- **Q** (Seasonal MA terms)
- **s** (Seasonal period, e.g., 12 for yearly seasonality in monthly data)

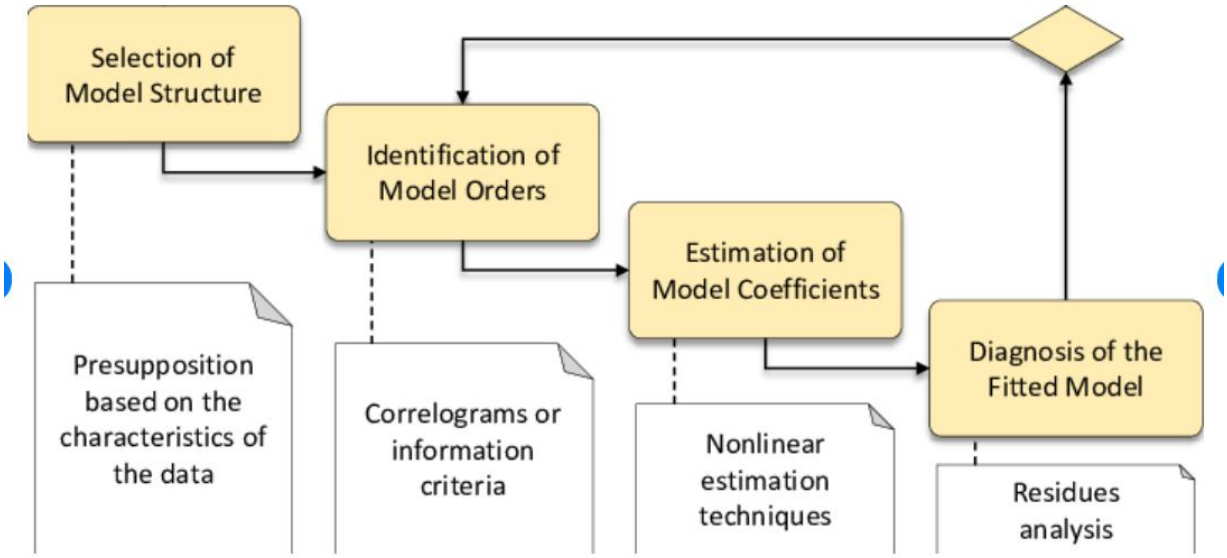

Fig 3.1- Working of SARIMA

Results-

```
Best model:  ARIMA(0,1,1)(1,0,0)[12]
Total fit time: 43.926 seconds
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  247
Model:             SARIMAX(0, 1, 1)x(1, 0, [], 12)   Log Likelihood                410.721
Date:                             Mon, 03 Feb 2025   AIC                            -815.441
Time:                                     17:28:17   BIC                            -804.925
Sample:                                 01-01-2003   HQIC                           -811.207
                                      - 07-01-2023
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.4359      0.046     -9.564      0.000      -0.525      -0.347
ar.S.L12       0.1467      0.057      2.563      0.010       0.035       0.259
sigma2         0.0021      0.000     18.652      0.000       0.002       0.002
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):               141.66
Prob(Q):                              0.94   Prob(JB):                         0.00
Heteroskedasticity (H):               0.24   Skew:                             0.24
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.69
===================================================================================
```

Fig 3.2- Results of SARIMA

Model Fit (AIC = -815.441, BIC = -804.925)

- A **lower AIC/BIC** indicates a better model fit. These values suggest the model is relatively good.
- **Heteroskedasticity Test (H = 0.24, p < 0.01)**: Variance is **not constant**, indicating some periods may be more volatile.
- The low variance of residuals indicates **a good fit**.

2) ARIMA

- Auto ARIMA is an advanced time series forecasting technique that automates the process of selecting the optimal ARIMA model parameters. By analysing historical data, it identifies the best combination of autoregressive (AR), differencing (I), and moving average (MA) terms to minimize forecasting error. This eliminates the need for manual parameter tuning, making the modelling process more efficient and accurate.

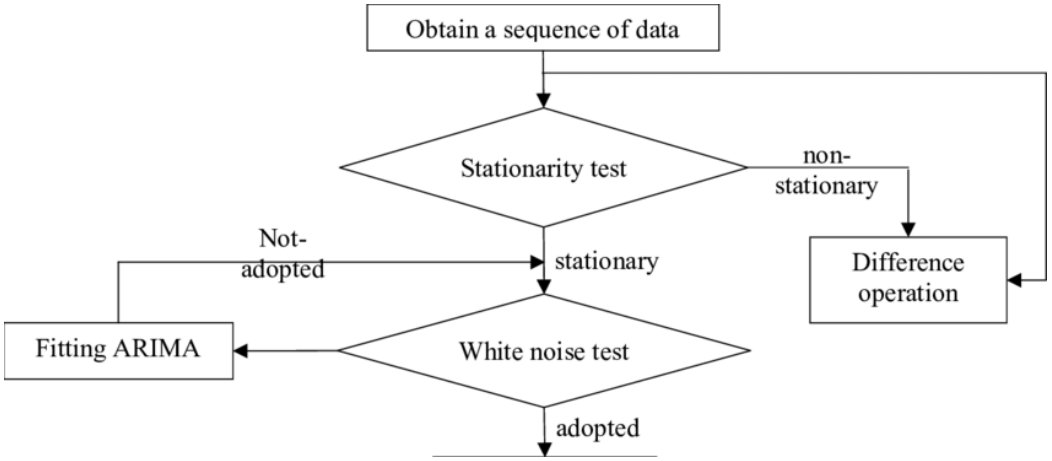

Fig 3.3 ARIMA Working

- **ARIMA Evaluation-**

SARIMAX Results

Dep. Variable:	premiums	No. Observations:	247
Model:	ARIMA(1, 1, 1)	Log Likelihood	408.359
Date:	Mon, 03 Feb 2025	AIC	-810.718
Time:	17:28:17	BIC	-800.202
Sample:	01-01-2003 - 07-01-2023	HQIC	-806.484
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1141	0.132	0.863	0.388	-0.145	0.373
ma.L1	-0.5424	0.119	-4.557	0.000	-0.776	-0.309
sigma2	0.0021	0.000	18.619	0.000	0.002	0.002

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	141.33
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	0.26	Skew:	0.14
Prob(H) (two-sided):	0.00	Kurtosis:	6.70

Fig 3.4 ARIMA Results

Results-

AIC = -810.718, BIC = -800.202:

- Lower values suggest a good fit, but these should be compared with other models for confirmation.
- J**arque-Bera Test (JB = 141.33, p < 0.01)**:Residuals **are not normally distributed**, which may affect prediction intervals.
- **Heteroskedasticity (H = 0.26, p < 0.01)**:The variance of residuals is **not constant**, indicating potential volatility in premiums.

3) Facebook Prophet –

Prophet is a powerful time series forecasting model developed by Meta, designed for handling **trend, seasonality, and holiday effects** in data. Prophet is particularly useful because it automatically detects patterns in time series data and provides robust predictions with minimal parameter tuning. It follows an **additive model structure**, where the overall forecast is a combination of a **trend component**, **seasonal effects**, and **holiday adjustments**. In out project, Prophet is applied to predict **insurance premiums**, leveraging its ability to handle missing data, detect yearly/weekly seasonality, and incorporate external factors if needed. Unlike traditional ARIMA models, Prophet can effectively **capture nonlinear trends** and **adjust dynamically** to changes in data patterns, making it well-suited for real-world financial forecasting. Its interpretability and automated hyperparameter selection make it a valuable tool for improving the accuracy of my premium predictions.

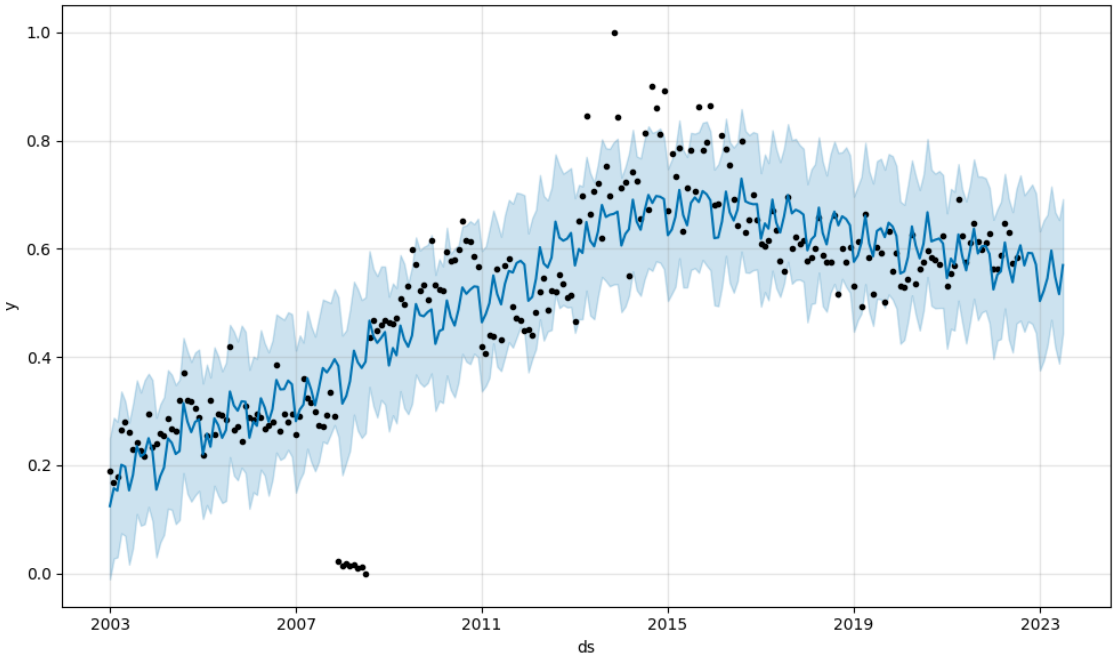

Fig 3.5-Checking outliers with Prophet

Overall Trend:

- The model captures a clear upward trend in the data until around 2015, followed by a gradual downward trend.
- This suggests a peak in the observed variable around 2015, followed by a decline.

Seasonality:

- The oscillating pattern in the blue line indicates seasonal effects, meaning the observed variable fluctuates periodically (e.g., annually or quarterly).

Confidence Intervals:

- The light blue shaded area represents the model's uncertainty intervals. Wider intervals suggest higher uncertainty in the predictions, especially at the edges of the forecast range.

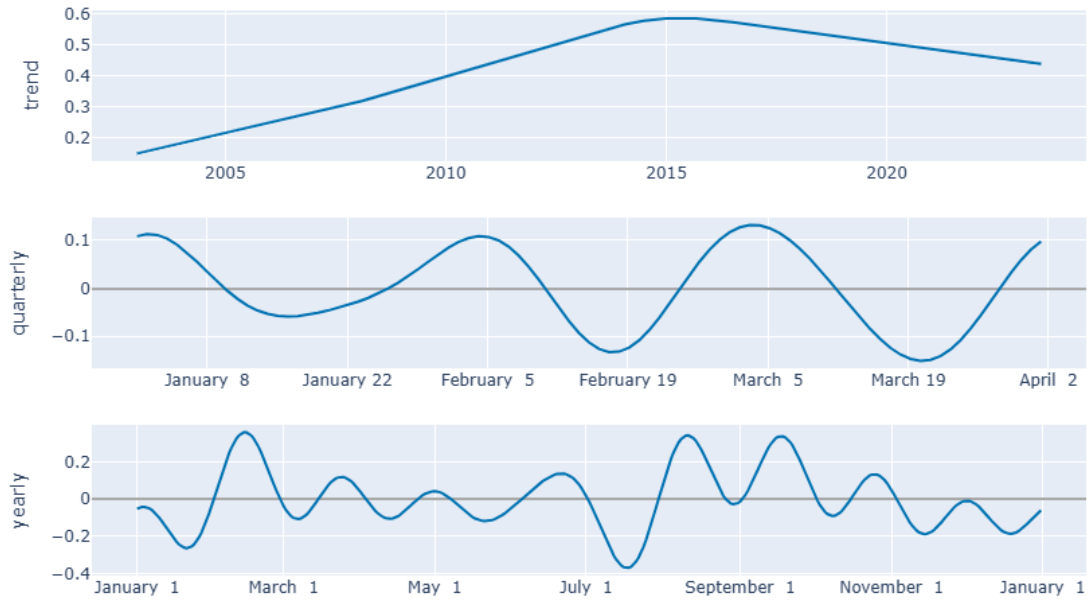

Fig 3.7- Plots of yearly, quarterly trends

Evaluation metrics-

MAE: 28.10

RMSE: 1476.45

MAPE: 4.74%

Performance:

- The low MAPE (4.74%) indicates the model performs well in capturing relative changes in the data.
- However, the high RMSE (1476.45) suggests that while most predictions are accurate, there are some significant errors or outliers affecting the overall performance.

Chapter 4. User Interface

After training the models and determining that Facebook Prophet performed best for our forecasting tasks, we proceeded with building a user interface for seamless interaction with our models. To achieve this, we utilized the Flask web framework, known for its simplicity and robustness. Additionally, we deployed the application on AWS EC2 using Flask, ensuring efficient and scalable access to our models for making predictions.

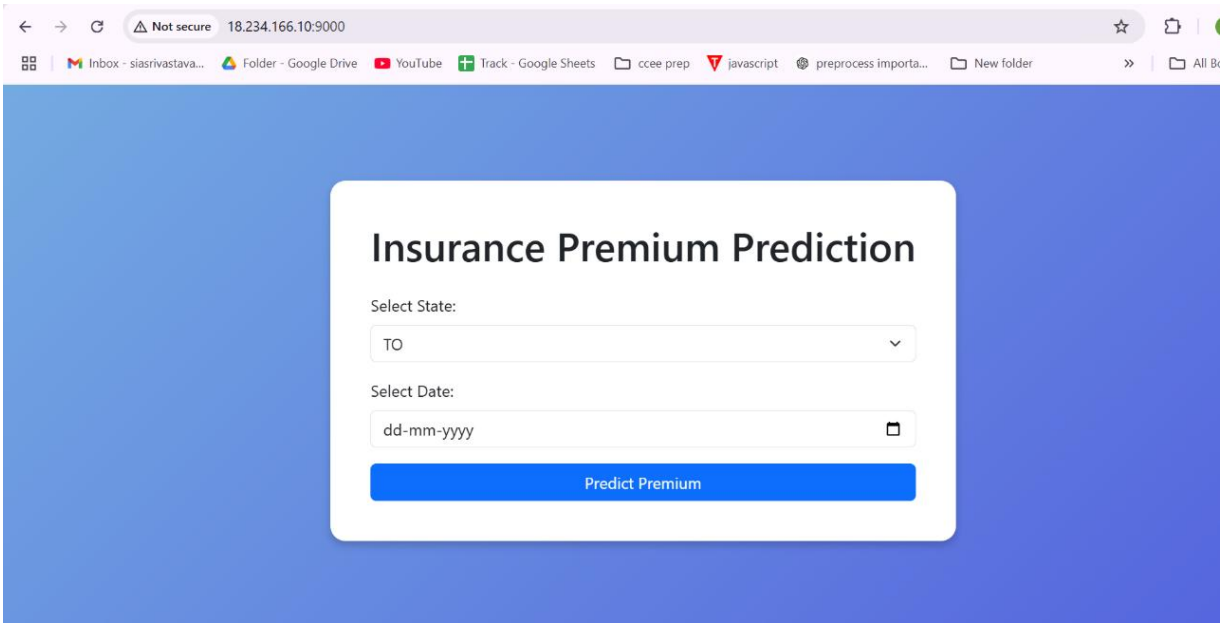

Fig 4.1- Main Landing page

- The user will select state in dropdown and date for which they want to predict.
- On selecting predict button, it will generate results.

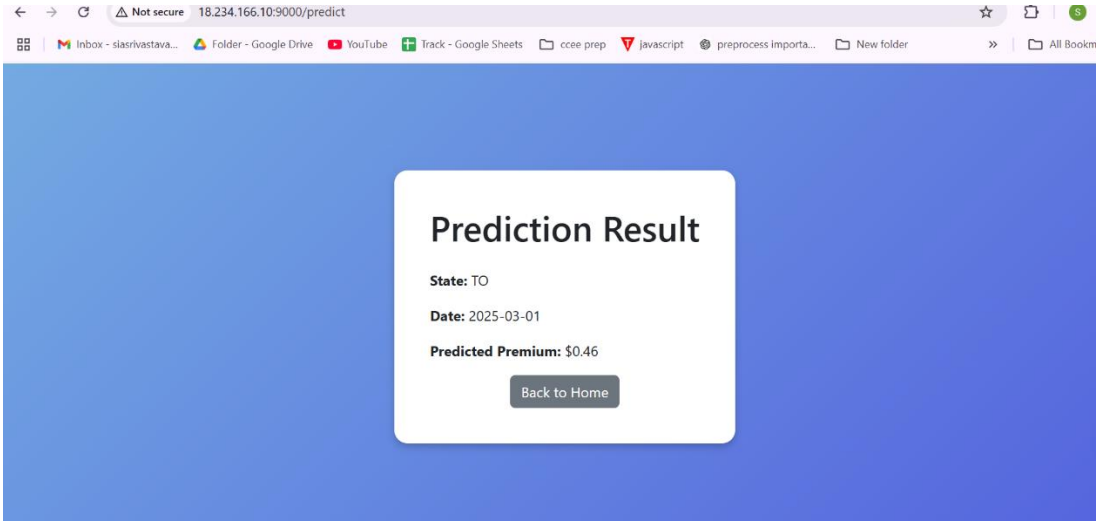

Fig 4.2- Result page

- The results are displayed for the particular user inputs.

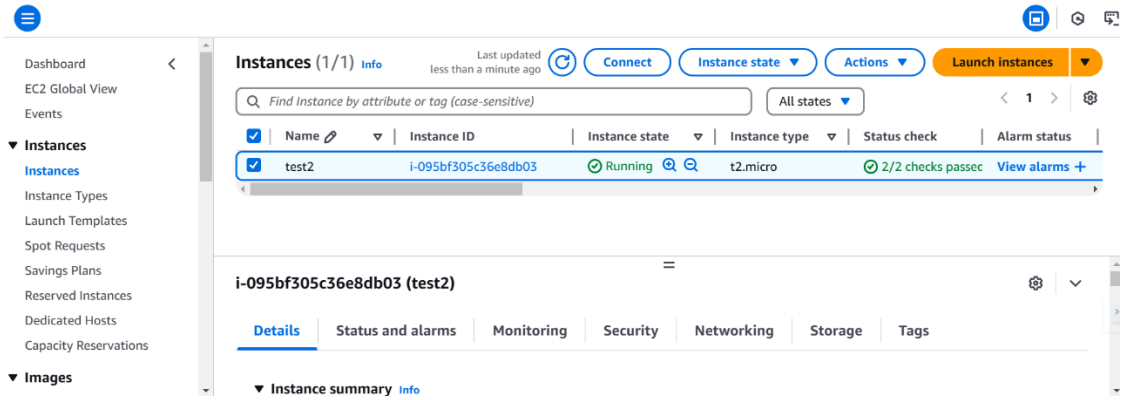

Fig 4.3- Deployment using EC2

Chapter 5. Requirements & Specifications

5.1 Hardware Requirement:

- 500 GB hard drive (Minimum requirement)
- 8 GB RAM (Minimum requirement)
- PC x64-bit CPU

5.2 Software Requirement:

- Windows/Mac/Linux
- Python-3.9.10
- VS Code/Anaconda/Google Colab/Jupyter
- Python Extension for VS Code
- **Libraries:**
  - Flask=1.1.1
  - prophet==1.1.5
  - numpy=1.9.2
  - scipy>=0.15.1
  - scikit-learn=0.18
  - matplotlib=1.4.3
  - pandas=0.19
  - Any Modern Web Browser like Google Chrome
    To access the web application written in Flask
  - AWS Cloud Platform ==> EC2 service

Chapter 6. Future Scope & Enhancements

To enhance the project, multiple forecasting models can be integrated using an ensemble approach for better accuracy. Real-time data streaming with Apache Kafka will enable dynamic adaptation, while interactive visualizations with Plotly or D3.js will improve insights.

Deployment can be optimized with AWS Lambda or containerization using Docker and Kubernetes for scalability. Automated model retraining will ensure updates with new data patterns. Incorporating external data sources like economic indicators or social media sentiment can further refine predictions. Lastly, integrating deep learning-based forecasting models will enhance precision, making the project more robust and adaptable.

Chapter 7. Conclusion

This project successfully implemented a predictive modelling system using Facebook Prophet to forecast trends and make data-driven predictions. By leveraging time series forecasting techniques, we identified key patterns and insights, enabling more informed decision-making. The integration of Flask allowed us to build an interactive and user-friendly interface, making it easy to access and utilize the predictive models. Furthermore, deploying the application on AWS EC2 ensured scalability and efficient model serving, allowing for real-time predictions with minimal latency.

The project demonstrated the effectiveness of machine learning and cloud deployment in solving real-world forecasting challenges. Additionally, extensive evaluation and fine-tuning of the model ensured optimal accuracy, making it a reliable tool for predictive analytics. The seamless combination of data preprocessing, model selection, and deployment highlights the importance of a well-structured end-to-end machine learning pipeline.

This project serves as a foundation for further exploration in predictive analytics and showcases the potential of cloud-based AI solutions in various industries.

Chapter 9. References

- https://www.kaggle.com/datasets/ex0ticone/brazilian-insurance-market-data
- https://pandas.pydata.org/docs/
- https://facebook.github.io/prophet/
- https://www.sciencedirect.com/science/article/pii/S2666827023000695
- https://www.ibm.com/think/topics/arima-model
- https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide