Comprehensive Analysis of Multiple Linear Regression and Time Series Forecasting

Malav Naik
Master of Science in Data Analytics_A
National College of Ireland
x23271779@student.ncirl.ie

Abstract— **Multiple Linear Regression as well as Time Series forecasting has been effectively discussed within this report from time to time with the aim of showing its applications to predictive models and analysis over time. In this case, it is the target variable predicted using several independent features within an MLR section through focuses like data preprocessing, encoding categorical variables, and model evaluation. The performance of the model is highlighted through relevant metrics such as MSE and the R-squared (R^2).**

Time Series Forecasting: In this study log transformation and differencing has been performed to ensure the time series stationary before further model SARIMA fits to the data. The final process has involved exploring the autocorrelation and partial autocorrelation functions in tuning the parameters of the model. The parameters of the SARIMA are optimized using Auto ARIMA, and further, the accuracy of the model has been evaluated by performing residual analysis along with performance metrics. The techniques applied have been incorporated to show their roles in developing data-driven predictions and indications of trends and seasonal patterns to understanding practical applications of MLR and Time Series Forecasting through visual and statistical metrics that validate performance in the models used.

Keywords- Multiple Linear Regression, Time Series Forecasting, SARIMA, Auto ARIMA, Predictive Modeling, Temporal Data Analysis, Stationarity, Residual Analysis, R-squared, Mean Squared Error, Autocorrelation, Partial Autocorrelation, Data Preprocessing, Seasonal Patterns.

1 Introduction

These two approaches- MLR and Time Series Forecasting- are the more basic, fundamental statistical modelling techniques which are applied to real-world data in order to generate the parameters that would play, model the relationships, and predict accurately to assist the decision making.

MLR would then explain the relationship among one dependent variable with a few independent variables by completing the following:

- **Data Preprocessing:** It requires the management of missing values, encoding of categorical variables, and data quality maintenance.
- **Model Development:** The separation of data into training and test data, development of MLR model and consideration of its coefficients.
- **Evaluation Models:** Mean Squared Error (MSE) metrics and R-Squared (R^2) for accuracy assessment and fit.

The Time Series Forecasting Section is the investigation of time data for patterns, trends, and seasonality. The primary aims are as follows:

- **Stationarity Testing:** Test for stability of the series with the help of Augmented Dickey-Fuller (ADF) Test.
- **Data Transformation:** Apply log transformations and difference to achieve stability in variance and stationarity.
- **Model Selection and Fitting:** Determine SARIMA optimal parameters through ACF and PACF plots.
- **Forecasting and Validation:** Analysis of prediction and validation through assessing residuals and metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

MLE and Time Series Forecasting together present a two-pronged view or perspective into analytically interpretable cross-sectional and time series data for better understanding, higher accuracy in prediction, and superior basis for logical decisions.

2 Exploratory Data Analysis

Exploratory Data Analysis is quite an important part of the entire process, since through it, the measurement level, descriptive statistics, and various visualizations contain information about a thorough examination of a data set.

It's significant enough to determine relationships between variables, recognize patterns and outliers, and prepare the ground for efficient predictive modeling.

2.1 Levels of Measurement

The dataset includes variables with varying levels of measurement:

- **Nominal (categorical):** Variables such as x2 will need to be encoded for modeling.
- **Interval/Ratio (numerical):** Continuous quantities like y, x1, and the target value x2 based on the statistical analysis.

2.2 Descriptive Statistics

Escriptive statistics summarize the data: Mean, Median, Standard Deviation and Variance, Range and Interquartile Range (IQR).

The frequency tables for categorical variables like x2, ensure a balanced representation.

2.3 Visualizations

Visualizations are certainly essential tools in revealing patterns or insights from data:

Histograms and Density Plot, Box Plots, Scatter Plots, Bar Charts, Correlation Matrix.

2.4 Key EDA Insights

- **Trends:** Visible relationships guide the selection of features.
- **Outliers:** Use descriptive stats and box plots to deal with extreme values.
- **Correlations:** Identify variables that are predictive.

EDA is an approach that combines statistical summaries and visualizations to understanding the whole picture, so that subsequent modeling steps are made robust and well-grounded.[1]

3 Data Preparation

Data preparation is one of the major steps in the process of cleaning and maintaining a single coherent data set that will improve the accuracy and reliability of a model. Some of the key steps include:

3.1 Handling Missing Values

- **Numerical Variables:** Missing values (like y, x1) were replaced with mean values.
- **Categorical Variables:** x2 missing values were replaced by mode value.

Fig: Scaling Category count with Percentage

3.2 Encoding Categorical Variables

One hot encoded categorical variable such as x2 was transformed into binary columns.

3.3 Outlier Detection and Handling

- Box plot analysis has detected outlier values in variables like y and x1.
- Capping and flooring such outliers helped in minimizing their effects.

3.4 Scaling Numerical Variables

The scaling process-Min-Max that was applied normalized features such as y and x1 into [0,1] without shifting the distributions.[1]

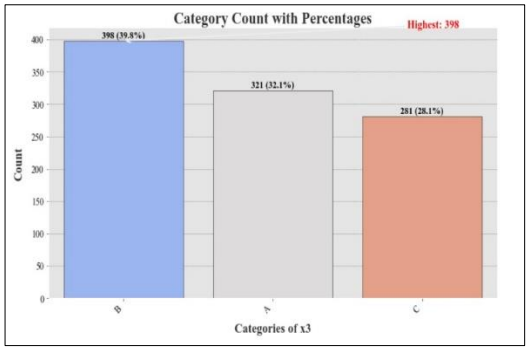


Fig: Scaling Category count with Percentage

This Count Plot, created using Seaborn, visualizes the frequency distribution of the categorical variable x2 from the dataset df. Each bar represents a unique category, with its height showing the count of occurrences, and exact values labeled above.

3.5 Splitting the Dataset

- **Training Set (80%):** Train Set-for developing the model.
- **Testing Set (20%):** Test Set-Alone used to validate the model on unseen data.

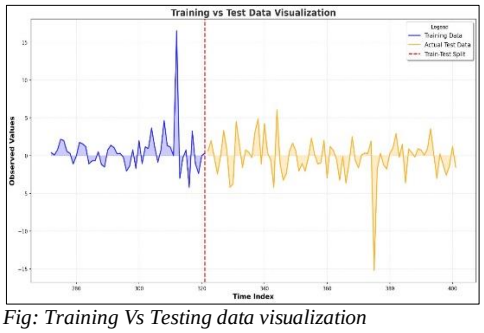


Fig: Training Vs Testing data visualization

3.6 Treatment of Multicollinearity

- **Correlation Matrix**: Identified that those variables which have correlation value very high.
- **Variance Inflation Factor (VIF):** High VIF values cause dropping or transformation of variables.

4 Modelling

The phases of modelling encompass developing, evaluating and improving predictive models to deliver precise, accurate results; and this section describes everything that has gone into developing the final Multiple Linear Regression and Time-Series Forecasting models, including justifying the selection, and reasons why interim models were jettisoned.[2]

4.1 Initial Model Building

First, those models fitted with the cleansed datasets served as baselines against which to measure the relative applicability of final approaches.

- **Multiple Linear Regression (MLR):**

The regression model takes the form of the variables: y, x1, and the encoded x2. The model's performance was measures using Mean Squared Error (MSE) and R-squared (R^2).

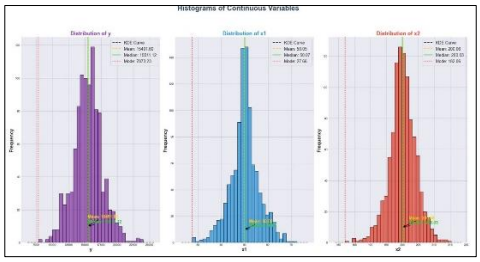


Fig: Histograms of MLR continuous variable

- **Time Series Forecasting:**

An initial SARIMA model was fitted with manually selected parameters based on observations from ACF and PACF plots. The residuals were analyzed for evaluations of autoregression and stationary existence with positive and negative regions.

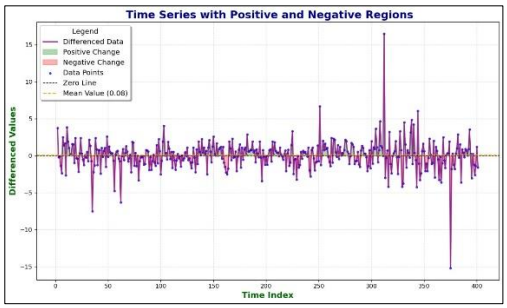


Fig: Time series with Positive and Negative Regions

4.2 Missing Data and Outlier Handling

At model building, missing and outlier observations have been managed to avoid bias or skewed predictions:

- **Missing Values:** Numerical features used mean imputation to preserve distributions. The mode was used for consistent imputation on the categorical variable x3.
- **Outliers:** Upper and Lower Limits were applied at the 95th and 5th percentiles, respectively, for extreme numerical outliers.[2]

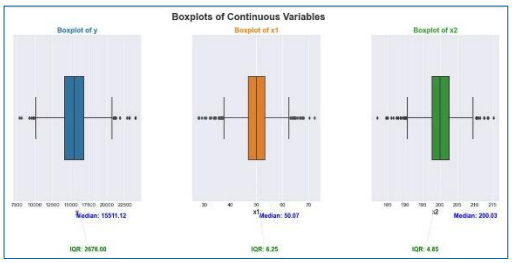


Fig: Plotting of Time series with Gradient Moving

Missing data and outlier management provided a clean and consistent dataset without deleting too much information. Hence these models are more robust.

4.3 Variable Transformation

Transformations were applied to make the variance stable and make those relationships between variables more interpretable.

- **Log Transformation:** Applied to the dependent variable in MLR model and to time series data to handle skewness and stabilize variance.[2]

Fig: Log Transformation Time series

- **Differencing:** Various transformations to achieve stationarity, which is a requirement for SARIMA modelling.[3]

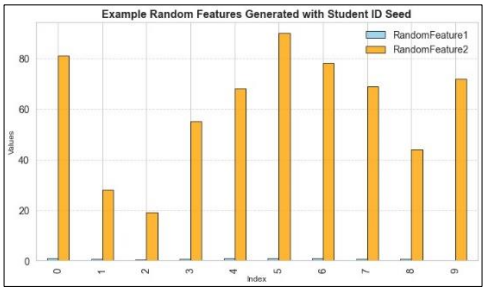


Fig: Rando, feature to generated Student ID seeds

These transformations helped the model to identify patterns and relationships between them so that all possible improvements in predictive accuracy could be achieved.

4.4 Evaluation and Rejection of Intermediate Models

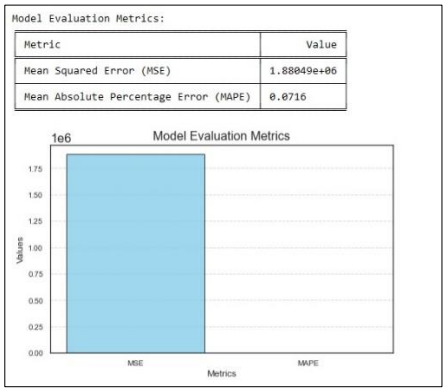
All intermediate models were evaluated on the performance metrics revamping them as:

- **Multiple Linear Regression:** High variance inflation factor (VIF) values indicated multicollinearity among the predictors. Variables with high VIF were eliminated iteratively for Interpretability and Models to improve without statistically significant predictors (p-value > 0.05) were rejected.
- **Time Series Forecasting:** Manually selected SARIMA parameters resulted in poor residual patterns, indicating a bad fit. Auto ARIMA was used to optimize parameter selection for the SARIMA model.

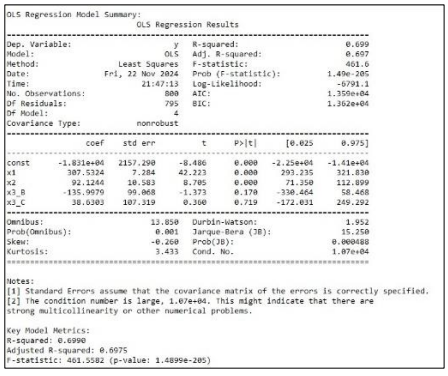
Fig: Correlation Matrix for evaluating models

4.5 Final Model Selection

- **Multiple Linear Regression:** The model was final on predictors that were statistically significant (p-value < 0.05) and uncorrelated, which ensures reliable coefficients and interpretability Metrics:



Mean Squared Error (MSE): 1,880,489.72



R-squared (R^2): 0.699

- **Time Series Forecasting:** Auto ARIMA selected optimal SARIMA parameters according to Akaike Information Criterion (AIC). The final SARIMA model showed quite low residual autocorrelation with high forecast accuracy Metrics:

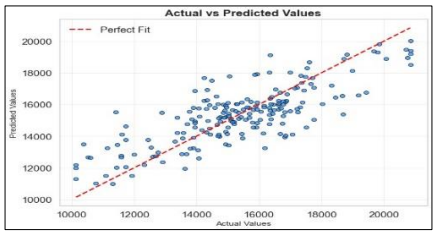


Fig: Actual Vs Predicated values of RMSE.

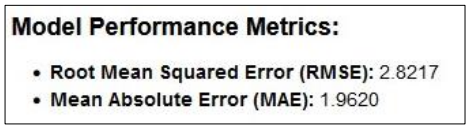


Fig: Model performance of Time series forecasting.

4.6 Model Validation

- **Residual Analysis:** Residuals of both models were analysed to ascertain if they had followed a normal distribution and although showed significant pattern.
- **Test Set Predictions:** Predictions on the test have been compared against actual values with error metrics confirming the reliability of the models.

5 Interpretation & Diagnostics of MLR

Interpretation is an essential step in gaining insights from the results generated by the models. It involves analyzing coefficients, statistical significance, and confidence intervals for MLR.

Diagnostics are important to ensure the reliability and robustness of models in Multiple Linear Regression (MLR).

By using visualizations and statistical tests, the models are refined to meet theoretical assumptions, resulting in accurate and meaningful outputs.

5.1 Model Coefficients

The coefficients of the regression model express the effect of a one-unit change in predictor variable on the average of the target variable (y), assuming that all other variables remain constant:

Coefficient of y: A positive value suggests that as y increases, the predicted value of y also increases.

Coefficient of x1: A negative value indicates that as x1 increases, the predicted value of y decreases.

Categorical Variable (x2): Binary-encoded variables for x2 indicate the difference in the predicted value of y relative to the reference category.

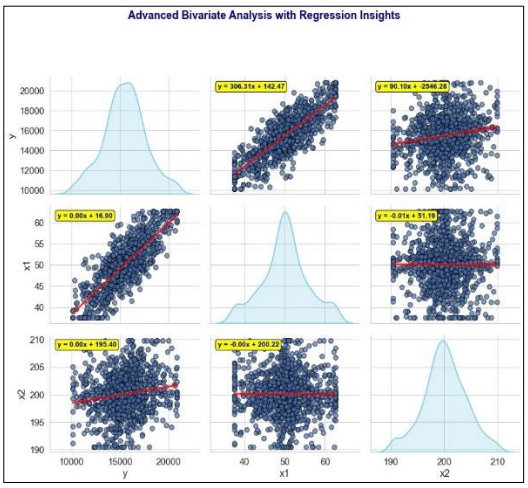


Fig: Advance Bivariate Analysis with Regression Insights.

5.2 p-Values for Coefficients

p-values test whether the coefficient of a predictor differs significantly from zero (meaning it affects target variables):

p < 0.05: The predictor is statistically significant and make meaningful contributions in the model.

p > 0.05: The predictor is not statistically significant, and may not be critical for the model.

5.3 Confidence Intervals

Confidence Interval: defines the range in which the actual coefficient values are likely to fall at defined confidence level (example: 95%).

Interpretation: For instance, if x1's confidence interval is [2.1, 3.5], there is a 95% chance that the actual effect of x1 on y lies inside that interval.

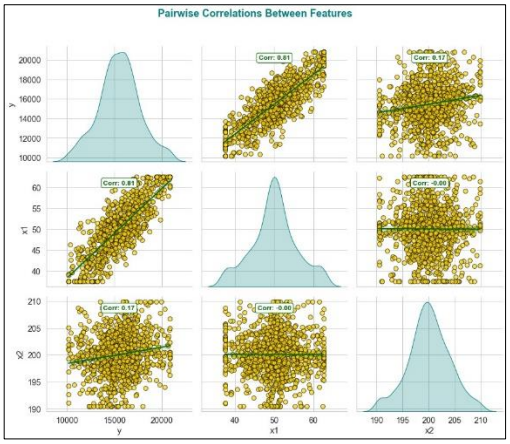


Fig: Pairwise Correlation between features.

5.4 Analyzing Variable Relationships

The relationships between the target variable (y) and the predictors were explored. Scatter Plots reflected that generally linear trends exist between y and the numerical predictors (y, x1).

Box Plots illustrated differences in the dependent variable (y) across categories of x2. Correlation Matrix found evidence for multicollinearity among the predictors.

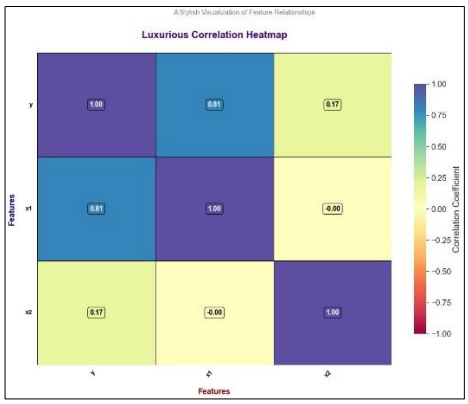


Fig: Luxurious Correlation heatmap between relationship

5.5 Encoding Categorical Variables

The categoric variable x3 was transformed using One-Hot-Encoding:

Preserving its non-ordinal character, enhancing the model, treating each category as independently.

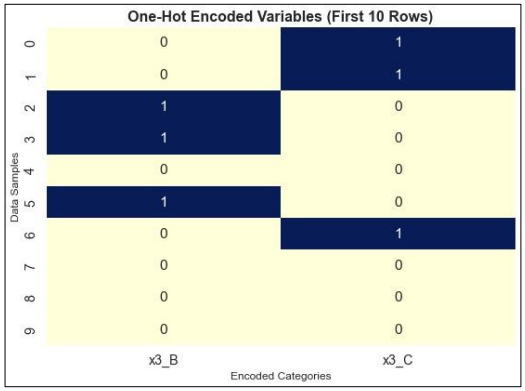


Fig: One-hot encoding variables

5.6 Splitting the Data

Data were split for model evaluation - **Training Set, 80%** for building the MLR model and **Testing Set, 20%** for testing the model on observations that were not used during the training.

5.7 Fitting the MLR Model

The model was trained using Ordinary Least Squares (OLS) in order to minimize the lesser residual error. The proper fit of the model would be evaluated using residual plots and also by using metrics like R2 and MSE.[2]

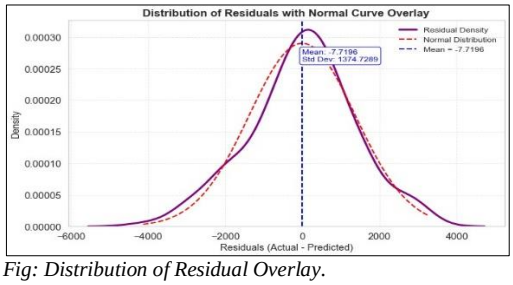


Fig: Distribution of Residual Overlay.

5.8 Model Summary

Coefficients: As per the predictors impact on target variables.

p-Values: Hypothesis testing about the statistical significance of predictors; retention was only for variables where $p < 0.05$.

Confidence Intervals show reliability in the estimates of coefficients. F-Statistic: the overall significance of the model is assessed.

6 Interpretation & Diagnostics of Time Series Forecasting

Interpretation gives insight to the results obtained from the various models. Understanding the parameters of the Time Series Model will interpret insights into approaches of applicability. Diagnostics are essential for ensuring reliability and robustness for Time Series Analysis methods. Models-purposed diagnostics using visualizations and statistical tests refine the model to meet theoretical assumptions and, thus generate accurate and meaningful outputs.

6.1 SARIMA Model Parameters

Those are also the parameters of the SARIMA model. Such as, non-seasonal components:

(p, d, q): p (Autoregressive term): Number of lagged observations used. A positive coefficient indicates a direct relationship between past and current values. **d** (Differencing term): Number of steps taken to make the series stationary. **q** (Moving Average/MA term): Number of lagged forecast errors used. A positive coefficient suggests past forecast errors contribute to current values.

(P, D, Q, m): Represent the seasonal components: **P** (Seasonal AR): Number of seasonal lagged observations. **D** (Seasonal Differencing): Number of seasonal differencing steps. **Q** (Seasonal MA): Number of seasonal lagged forecast errors. m (Seasonal Period): Length of the seasonal cycle (e.g., 30 for monthly data).

```
"""
Trains an Auto ARIMA model based on the given parameters.

Args:
    data: The training data (time series) as a pandas Series or array.
    start_p (int): Initial AR order.
    start_q (int): Initial MA order.
    max_p (int): Maximum AR order.
    max_q (int): Maximum MA order.
    seasonal (bool): Whether to include seasonality.
    m (int): Seasonal period.
    d (int): Non-seasonal differencing order.
    D (int): Seasonal differencing order.
    stepwise (bool): Use stepwise approach for optimization.
    trace (bool): Enable tracing of progress.
    error_action (str): Action to take on errors.
    suppress_warnings (bool): Suppress warnings during fitting.

Returns:
    A fitted auto_arima model.
"""
```

Fig: SARIMA model parameters

6.2 Interpretation of Coefficients

The coefficients in a SARIMA model show how strongly each component affects predictions:

AR Coefficients: Positive values indicate a strong influence of past observations on current predictions.

MA Coefficients: These tend to reflect the influence of past forecast errors on current observations.

Seasonal Coefficients: indicate repeating patterns and have effects on predictions.

The most appropriate SARIMA model should have a low residual autocorrelation, which means that it has captured most of the predictable patterns in data effectively.[6]

6.3 Initial Plotting

Raw time series- They exhibited inputs indicating trends, seasonality, and irregularities. They were indicative of a need for variance stabilization.

6.4 Testing for Stationarity

The test employed was the Augmented Dickey-Fuller (ADF) test: Non-stationarity required transformations like logarithmic adjustments and differencing.

6.5 Analyzing Correlations

ACF Plot: The different lag correlations identified for the selection of moving average (MA) component.

PACF Plot: Significant lags highlighted for autoregressive (AR) components.

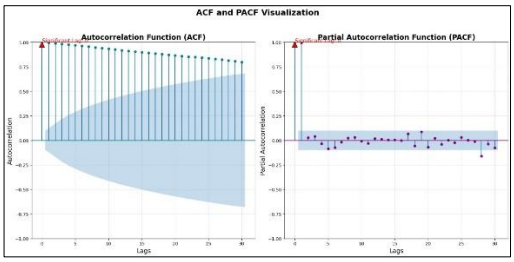


Fig: ACF and PACF Visualization.

6.6 Decomposing the Series

Almost all of the series were decomposed into: Trend Long-term movement, Seasonality Recurring patterns, Residuals: Noise after removing trends and seasonality. Seasonal patterns were included in the SARIMA model.[5]

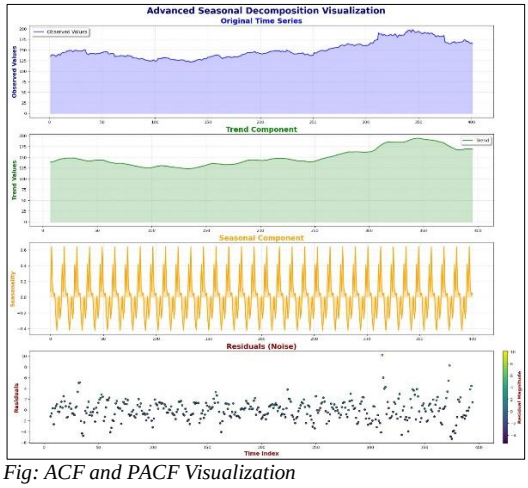


Fig: ACF and PACF Visualization

6.7 Feature Engineering

Rolling mean and standard deviation averages and their derivatives clarify trends while reducing the noise associated with the series.

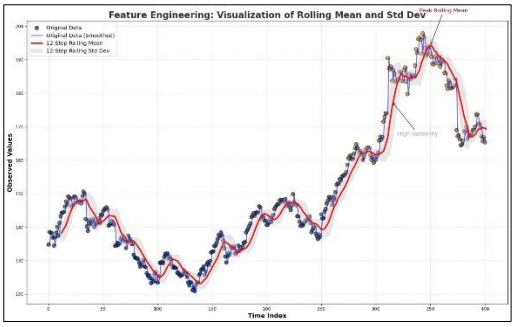


Fig: Feature Engineering of Rolling Mean and Std Dev.

6.8 Fitting the SARIMA Model

The SARIMA model was optimized with Auto ARIMA in accordance with the minimization of Akaike Information Criterion (AIC) parameters selected. The final model fits well with both seasonal and non-seasonal components.[4]

```
Performing stepwise search to minimize aic
 ARIMA(0,1,0)(1,1,1)[30]             : AIC=inf, Time=1.12 sec
 ARIMA(0,1,0)(0,1,0)[30]             : AIC=1505.420, Time=0.09 sec
 ARIMA(1,1,0)(1,1,0)[30]             : AIC=1370.268, Time=0.42 sec
 ARIMA(0,1,1)(0,1,1)[30]             : AIC=inf, Time=1.67 sec
 ARIMA(1,1,0)(0,1,0)[30]             : AIC=1403.601, Time=0.07 sec
 ARIMA(1,1,0)(2,1,0)[30]             : AIC=1354.129, Time=1.49 sec
 ARIMA(1,1,0)(2,1,1)[30]             : AIC=inf, Time=5.06 sec
 ARIMA(1,1,0)(1,1,1)[30]             : AIC=inf, Time=1.93 sec
 ARIMA(0,1,0)(2,1,0)[30]             : AIC=1450.269, Time=1.20 sec
 ARIMA(2,1,0)(2,1,0)[30]             : AIC=1323.191, Time=1.98 sec
 ARIMA(2,1,0)(1,1,0)[30]             : AIC=1341.215, Time=0.48 sec
 ARIMA(2,1,0)(2,1,1)[30]             : AIC=inf, Time=6.69 sec
 ARIMA(2,1,0)(1,1,1)[30]             : AIC=inf, Time=2.58 sec
 ARIMA(3,1,0)(2,1,0)[30]             : AIC=1309.824, Time=2.51 sec
 ARIMA(3,1,0)(1,1,0)[30]             : AIC=1327.630, Time=0.88 sec
 ARIMA(3,1,0)(2,1,1)[30]             : AIC=1296.598, Time=7.00 sec
 ARIMA(3,1,0)(1,1,1)[30]             : AIC=inf, Time=3.21 sec
 ARIMA(3,1,0)(2,1,2)[30]             : AIC=1298.556, Time=12.62 sec
 ARIMA(3,1,0)(1,1,2)[30]             : AIC=inf, Time=9.01 sec
 ARIMA(4,1,0)(2,1,1)[30]             : AIC=inf, Time=7.63 sec
 ARIMA(3,1,1)(2,1,1)[30]             : AIC=inf, Time=17.40 sec
 ARIMA(2,1,1)(2,1,1)[30]             : AIC=inf, Time=19.01 sec
 ARIMA(4,1,1)(2,1,1)[30]             : AIC=inf, Time=32.17 sec
 ARIMA(3,1,0)(2,1,1)[30] intercept   : AIC=1298.591, Time=8.14 sec

Best model:  ARIMA(3,1,0)(2,1,1)[30]
Total fit time: 144.386 seconds
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  320
Model:             SARIMAX(3, 1, 0)x(2, 1, [1], 30)   Log Likelihood                -641.299
Date:                              Fri, 29 Nov 2024   AIC                           1296.598
Time:                                      12:42:01   BIC                           1322.263
Sample:                                           0   HQIC                          1306.882
                                                - 320
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7836      0.036    -21.851      0.000      -0.854      -0.713
ar.L2         -0.4886      0.050     -9.808      0.000      -0.586      -0.391
ar.L3         -0.2352      0.038     -6.244      0.000      -0.309      -0.161
ar.S.L30       0.0202      0.137      0.148      0.883      -0.248       0.288
ar.S.L60      -0.1568      0.104     -1.508      0.132      -0.361       0.047
ma.S.L30      -0.7264      0.175     -4.146      0.000      -1.070      -0.383
sigma2         4.4866      0.243     18.438      0.000       4.010       4.964
===================================================================================
Ljung-Box (L1) (Q):                   0.95   Jarque-Bera (JB):               852.61
Prob(Q):                              0.33   Prob(JB):                         0.00
Heteroskedasticity (H):               1.41   Skew:                             0.94
Prob(H) (two-sided):                  0.09   Kurtosis:                        11.20
===================================================================================
```

Fig: SARIMA model results.

6.9 Validating Forecasts

Forecast accuracy was measured using Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE).

Residual analysis established little evidence of autocorrelation, indicating a good fit.[3]

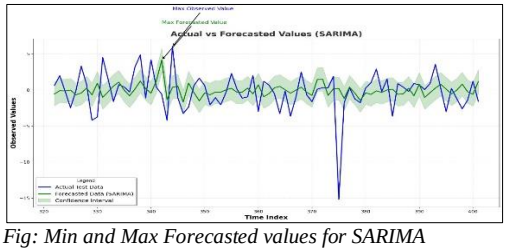


Fig: Min and Max Forecasted values for SARIMA

6.10 Model Summary

This SARIMAX model summary table information consists of SARIMAX model results, an important point-based evaluation in the model. It includes information about dependent variable (y), model specification (SARIMAX (3, 1, 0) x (2, 1, [1], 30)), as well as other performance indicators: Log Likelihood (-641.299), AIC (1296.598). To explain the fit of the model through residual pattern testing: Ljung-Box (Q=0.95, p=0.33) and heteroskedasticity (H=1.41, p=0.09). the covariance type "opg" and all warnings would be included.[4]

This helps evaluate suitability and reliability of a SARIMAX model.

	Metric	Value
0	Dep. Variable	y No. Observations
1	Model	SARIMAX(3, 1, 0)x(2, 1, [1], 30) Log Likelihood -641.299
2	Date	Fri, 29 Nov 2024 AIC 1296.598
3	Time	12
4	Sample	0 HQIC 1306.882
5	Covariance Type	opg
6	Ljung-Box (L1) (Q)	0.95 Jarque-Bera (JB)
7	Prob(Q)	0.33 Prob(JB)
8	Heteroskedasticity (H)	1.41 Skew
9	Prob(H) (two-sided)	0.09 Kurtosis
10	Warnings	

Fig: SARIMA Model Summary.

7 Evaluation

Such a model has evaluation phases wherein the trained model is tested with unseen data to validate predictive accuracy and reliability of the model making it with robustness over real-world applications.

7.1 Multiple Linear Regression -MLR

The performance evaluation metrics for MLR correspond to the test set as follows:

Mean Squared Error (MSE): The MSE is measured as the mean of all the squared differences between the actual and predicted values. The lesser this figure is, the better. MSE for this MLR model is 1,880,489.72, which shows it does quite well in representing the variance of the target variable.

Mean Absolute Percentage Error: It is considered as the average absolute difference between the predicted value and the actual value. MAPE gives intuitive prediction accuracy in the raw scale of the data. MAPE is low, about 0.0716, indicating the average error made in the forecast.

R-squared (R^2): R^2 stands out to quantify how many portions of variance in the dependent variable is being explained by the model. The higher R2 value, the better performance from the model.

R^2 was found to be 0.638; independent variables were shown to explain about 70 percent of the variance in a target variable.

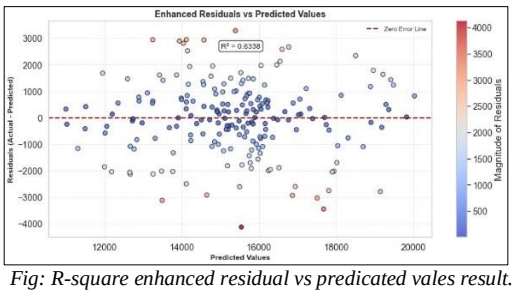


Fig: R-square enhanced residual vs predicated vales result.

7.2 Time Series Forecasting

The real prove of a predictive model, as a SARIMA compose, is depicted in the assay of test cases correlated actual test data with forecasted result values. The following metrics were used in this evaluation:[6]

Root Mean Squared Error (RMSE): RMSE is a square root of MSE indicating an error measure in the same unit as that of original data. Value of **2.8217** indicates that forecasted values are very close to the actual values.

Mean Absolute Error (MAE): Average of all absolute differences calculated from the forecast and actual value was MAE which was found to be **1.9620**. The value indicated the model was at minimum variation in prediction.

Residual Analysis: The residuals shown are checked for randomness-i.e., are there any signs of autocorrelations. It was found that residual analysis proved justifiable by the SARIMA model in showing that all those important patterns in data have been captured.

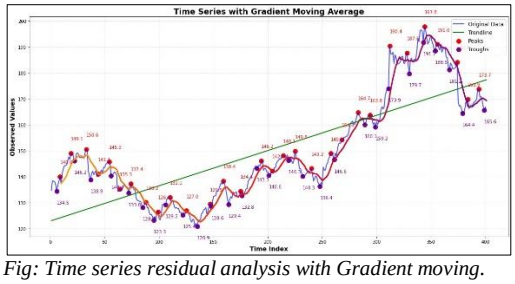


Fig: Time series residual analysis with Gradient moving.

Summary

However, both the MLR Model and SARIMA Model did present a good performance on the test data set. There is low error rates and high R^2 value in MLR that indicates good explanation from different aspects toward the variances of target variable. SARIMA also exhibited strength in dealings with time series data with its ability of predicting future values accurately.

However, whatever the above reasons are, in simple terms the two models are powerful and reliable, making them powerful tools for data-driven decision-making.

References

1 L. T. Skovgaard, "Applied regression analysis. 3rd edn. N. R. Draper and H. Smith, Wiley, New York, 1998. No. of pages: xvii+706. ISBN 0-471-17082-8," Statistics in Medicine, vol. 19, no. 22, pp. 3136–3139, Nov. 2000, doi: 10.1002/10970258(20001130)19:22<3136::AID-SIM607>3.0.CO;2-Q.

2 H. Oh, "Introduction to Linear Regression Analysis, 5th edition by MONTGOMERY, DOUGLAS C., PECK, ELIZABETH A., and VINING, G. GEOFFREY," Biometrics, vol. 69, no. 4, p. 1087, Dec. 2013, doi: 10.1111/biom.12129.

3 G. M. Ljung, J. Ledolter, and B. Abraham, "George Box's contributions to time series analysis and forecasting," Applied Stochastic Models in Business & Industry, vol. 30, no. 1, pp. 25–35, Jan. 2014, doi: 10.1002/asmb.2016.

4 Y. Funde and A. Damani, "Comparison of Arima and Exponential Smoothing Models in Prediction of Stock Prices," Journal of Prediction Markets, vol. 17, no. 1, pp. 21–38, 2023, doi: 10.5750/jpm.v17i1.2017.

5 S. Lee, J. Choi, and Y. Son, "Efficient visibility algorithm for high-frequency time-series: application to fault diagnosis with graph convolutional network," *Annals of Operations Research*, vol. 339, no. 1/2, pp. 813–833, 2024, doi: 10.1007/s10479-022-05071-x.

6 X. Li, F. Petropoulos, and Y. Kang, "Improving forecasting by subsampling seasonal time series," *International Journal of Production Research*, vol. 61, no. 3, pp. 976–992, 2023, doi: 10.1080/00207543.2021.2022800.