# Table of Contents

[Table of Contents 1](#_tnsdkqmv2bfp)

[Introduction 2](#_5g0ffsb2f3qj)

[Exploratory Data Analysis 2](#_3jmndz7jgi97)

[Summary of the Data 2](#_h17euegze3l8)

[Correlation 5](#_lngltv900t9k)

[Model Analysis Breakdown 7](#_4zp97g7q5jr7)

[Model Comparisons 8](#_q5s3y6g0u52s)

[Model Evaluation 8](#_2zlo9k5pwdc4)

# Introduction

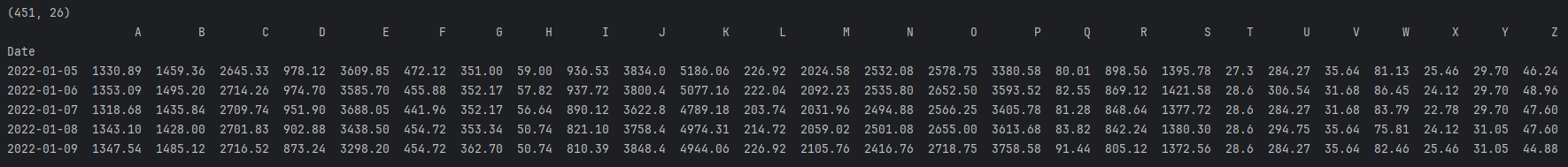
This report delves into a comprehensive analysis on an obfuscated dataset that was provided and applies predictive analytics techniques to construct a time series model. We explore various features by applying a back-shifting technique to the existing feature and aim to identify the most significant and optimal features that we can use to accurately predict the target variable, “A”, specifically forecasting its value for May 4, 2023.

# Exploratory Data Analysis

This section highlights the key features that were utilized to build the model leveraging OLS linear regression. The selection process identifies columns “A” and “F” as pivotal columns and their respective time-shifted variables were deemed significant for our predictions. We will focus on summarizing the data presented in the original dataset and further explore how these features correlate with the target variable, “A” by providing visualizations and identifying trends, cycles and seasonal movements.

## **Summary of the Data**

Figure 1 below showcases the first 5 rows of the provided dataset, comprising 451 rows indexed by their respective date and 26 columns of features denoted from “A” to “Z”. The date indexing plays a significant role in a time series model especially when making predictions or plotting the data on the graph. The data found for each row can be seen to have a wide range of values and because the dataset is obfuscated, this illustration alone can be difficult to visualize.



*Figure 1 — First 5 rows of the original dataset*

To provide a more intuitive representation, Figure 2 below depicts a time series plot for features “A” and “F”, offering deeper insights into the variation of the features over the period from January 2022 to March 2023. Feature A plot shows that the values range from approximately 600 to 3200 units indicating a fairly large range of values while feature F spans from approximately 200 to 620 units, depicting a narrower range. Both plots seem to peak very high at certain points of time such as in December 2022 for feature F, it peaks past 600 units while other values seem to be relatively within their normal range which hints at potential outliers.

|  |  |
| --- | --- |

*Figure 2 — Standard time series plot of features A and F*

Due to the data being obfuscated, we cannot tell what these values imply. However, if we assume it is some form of currency, we can calculate the percent change between the consecutive data points which can be useful to compare the rate of change over time for financial data. Figure 3 below shows the previous plots in the percent change form. Once again, feature A’s rate of change seems to be quite high as it can be seen to double its rate in some periods such as January 2023. Feature F offers a more stable rate although it can also be seen to peak its rate on a rare occasion such as June 2022. However, both plots don’t seem to show any trajectory and the rates also often return to 0 frequently even if their rate is at a very high point which shows that the values can change drastically drop or increase at any point in time.

|  |  |
| --- | --- |

*Figure 3 — Percent change time series plot of features A and F*

Another form of visualizing the dataset is performing a seasonal decomposition for both features. Figure 4 below displays a multiplicative decomposed plot for feature A that visualizes trends, seasonality, and cyclic behaviours which are often exhibited by time series data. The trend in the time series model is defined as a constant slope increase or decrease. Feature A doesn’t seem to exhibit a trend as we can see the slope often changes its trajectory and instead seems to reveal a seasonal behaviour as we can see a more distinct pattern in the seasonal decomposed plot occurring every month. Feature A also doesn’t seem to show a cyclic behavior which is essentially a constant rise and fall pattern on fixed time intervals that generally lasts longer than a year. As for the residual plot, we can see it multiplicative decomposition does a decent job of decomposing the error component since the residuals range from 0.4 to 1.6 units while additive decomposition does poorly as it ranges from approximately -1100 to 1200 units.

|  |  |
| --- | --- |
|  |  |

*Figure 4 — Seasonal decomposition plots of Trend, Seasonal and Residual for feature A*

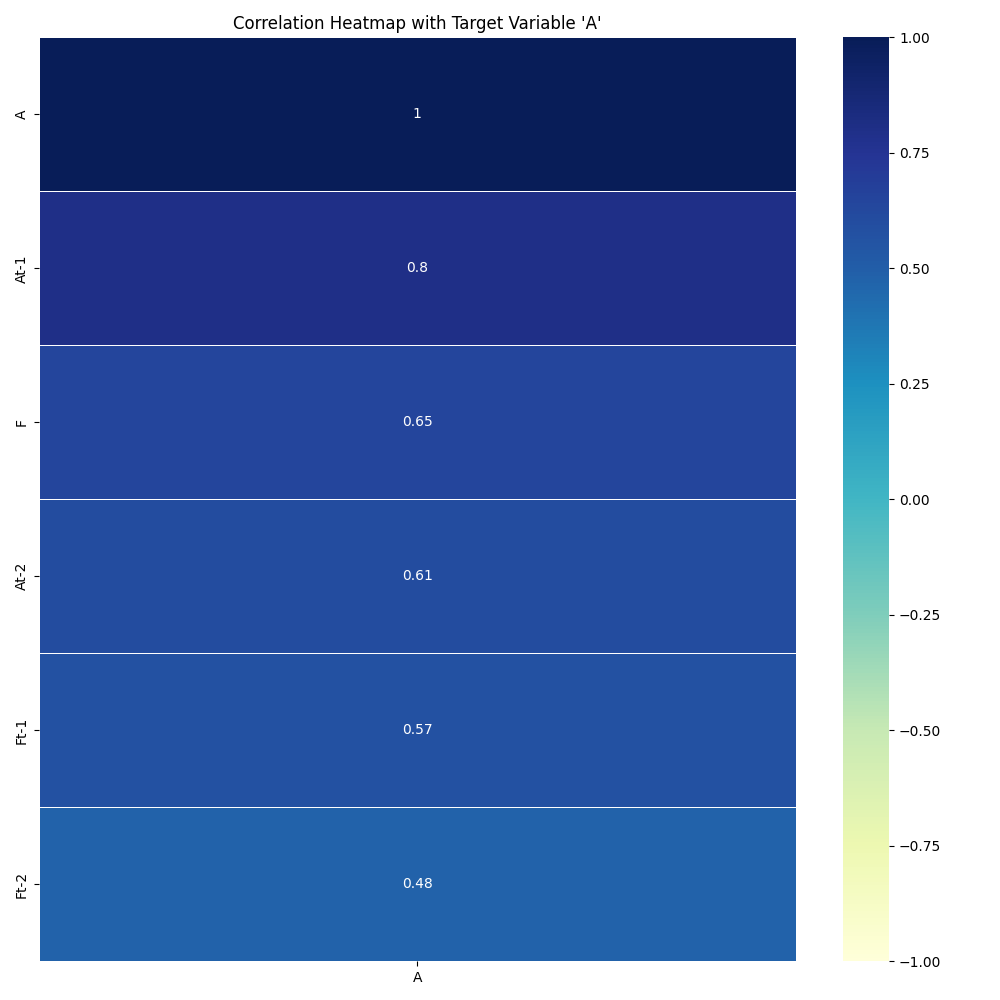
Feature F presents very similar behaviours as illustrated in Figure 5 below. It depicts a similar pattern as feature A of exhibiting only a seasonal behaviour. It also does a better job at decomposing the residual plot in multiplicative decomposition since its residuals range from 0.6 to 1.4 units while additive decomposition ranges from -150 to 150 units offering a larger range.

|  |  |
| --- | --- |
|  |  |

*Figure 5 — Seasonal decomposition plots of Trend, Seasonal and Residual for feature F*

## **Correlation**

Figure 6 below presents a heatmap that reveals the correlation between the key features and the target variable A, displaying a correlation coefficient value that depicts the correlation strength. Observing the heatmap, we can see that At-1 has a fairly strong positive correlation with the target variable A as it inherits a correlation coefficient of 0.8. The rest of the features show a coefficient ranging from 0.4 to 0.6 which presents a decently positive correlation, indicating that they are all potential predictor variables that can be used for our model.



*Figure 6 —Correlation heatmap of time-shifted and base features with target variable A*

Another way to depict the correlation for time-shifted variables is by observing an autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. ACF is used to observe the correlation of the current observation with its previous time steps while PACF is used to represent the correlation amount specifically between each lag and the current time step which is located at time step 0.

Figure 7 below illustrates the ACF and PACF plots for feature A for 50 lags. In the ACF plot on the left, we can see that the current time step on the far left is correlated to its previous time steps arguably until lag 8 as most of the points sit outside of the 95% confidence interval shaded in the blue region. However, some of those lags sit inside the region and also exhibit a low correlation coefficient, so, we can confidently claim that the first 2 to 3 lags present the strongest correlation as they all have a correlation coefficient roughly above 0.4 indicating again a strong potential for features to use in our model. The PACF plot on the right side of Figure 7 depicts that its initial lag is correlated with the first time step only with a coefficient of around 0.8 indicating a very strong positive correlation.

|  |  |
| --- | --- |

*Figure 7 — Partial autocorrelation and autocorrelation plots for feature A*

Figure 8 below now presents the ACF and PACF plots for feature F for 50 lags. Similar results can be observed in comparison to feature A’s plots since the ACF for feature F shows that the first 3 lags present a strong correlation with its current time step all exhibiting a correlation coefficient above 0.4. PACF also shows the first lag being highly correlated positively with its initial lag.

|  |  |
| --- | --- |

*Figure 8 — Partial autocorrelation and autocorrelation plots for feature F*

In conclusion, we can see that the first time step for both features presents a strong correlation with its initial lag leading to potential features that will be used to construct our OLS model.

# Model Analysis Breakdown

This section evaluates the performance of three predictive models including OLS regression, Triple Exponential Smoothing Multiplicative Holt-Winters, and SARIMAX, for predicting our target variable A. We will focus on comparing their metrics and further evaluate which of the three models presents the most optimal features and accurately predicts our target variable.

## **Model Comparisons**

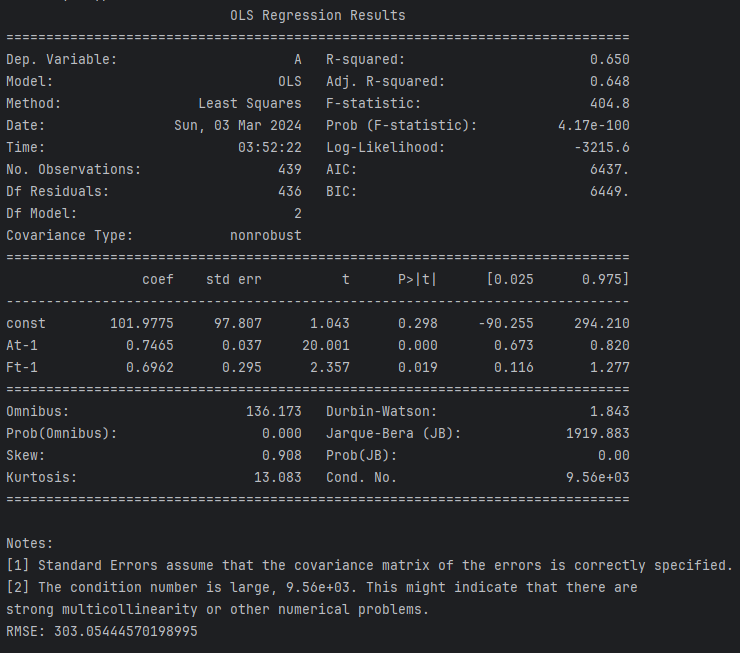
After delving into the correlations between the features and target variable in the previous section, it's evident that both features A and F and their respective time step variables played a significant role in the model as we can see in Table 1 below that the OLS model presents the most optimal results when compared with the Triple Exponential Smoothing Multiplicative Holt-Winters and SARIMAX models. OLS presents the fewest features as it only requires one previous time step for features A and F while Holt-Winters requires all time-shifted variables since it's a univariate model and SARMIAX requires three autoregressors and moving average components along with the sigma2 component. While retaining the fewest variables for OLS, it also presented the lowest Root Mean Square Error (RMSE) score out of the three models when predicting the most recent ten days of the original dataset. Although the RMSE scores are all within the low 300 range which is unfortunate, the other two models require far more variables that may be an indicator for overfitting issues which we would want to avoid.

| **Model** | **OLS Regression** | **Holt-Winters (Triple ES Multiplicative)** | **SARIMAX(3, 0, 3)** |
| --- | --- | --- | --- |
| **Features** | At-1, Ft-1 | At-n | ar.L1, ar.L2, ar.L3, ma.L1, ma.L2, ma.L3, sigma2 |
| **# of Features** | 2 | N-1 | 7 |
| **10-Day RMSE** | 303.05444570198995 | 327.41289304415693 | 323.4765114421665 |

*Table 1 — Comparison table for the 3 constructed models and its respective features and RMSE scores*

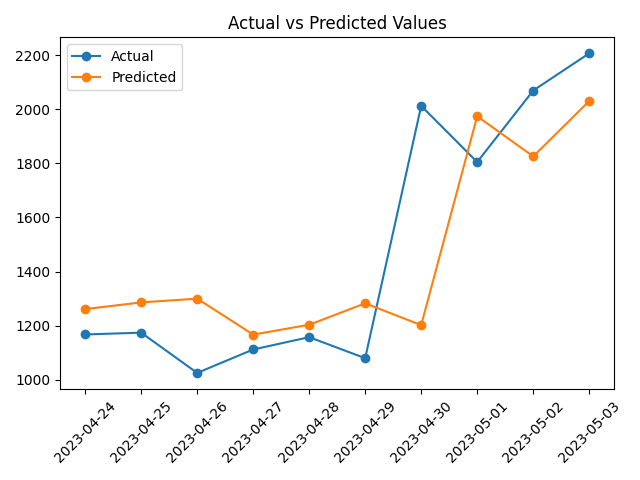
## **Model Evaluation**

To reiterate, the OLS regression model was deemed the most optimal model presenting the best set of features that also showed high correlation coefficients and performed the best with the lowest RMSE score out of the three constructed models. Further evaluating the model, Figure 9 below illustrates a model summary that presents its feature coefficients and more. The R-squared value shows a value of 0.65 which indicates a decent fit and does a fairly good job in explaining the variance. We can additionally observe the p-value scores presented for the features At-1 and Ft-1, both exhibiting an ideal low value indicating they are both significant features for the model.



*Figure 9 — Statistical OLS regression model summary and metrics*

Further analyzing the RMSE metric for the OLS, we have plotted a graph for the most recent ten days of the prediction for the target variable A value versus the actual values from the original dataset as presented in Figure 10 below. We can see that the RMSE of 303 is apparent since the prediction falls behind by a lag after the actual results which can be critical for certain scenarios.



*Figure 10 — Plots of Predicted vs. Actual Results*