NVIDIA Sales Revenue Forecasting Project Report

OPIM 5671 - Data Mining and Time Series Forecasting

**TEAM 4 - Hetarth Maulik Patel, Pawan Kalyan Sogana, Venkata Kushaal Varma Manthena, Vikram Krishnareddy**

Table of Contents

[Executive Summary 1](#_Toc209985943)

[Dataset and Data Preparation 2](#_Toc209985944)

[Data Source and Description 2](#_Toc209985945)

[Data Manipulation, Cleaning, and Feature Engineering 3](#_Toc209985946)

[Train-Test Split 5](#_Toc209985947)

[Model Development and Methodology 6](#_Toc209985948)

[Model Selection Strategy 6](#_Toc209985949)

[Exploratory Analysis 6](#_Toc209985950)

[Model Details 9](#_Toc209985951)

[Robustness and Stability Checks 11](#_Toc209985952)

[Best Model Selection and Results 12](#_Toc209985953)

[Selected Best Model 13](#_Toc209985954)

[ARIMAX (2,1,0) with Exogenous Variables 13](#_Toc209985955)

[Justification and Comparison 13](#_Toc209985956)

[Forecasting Results 15](#_Toc209985957)

[Actionable Insights and Conclusion 18](#_Toc209985958)

[Actionable Insights 18](#_Toc209985959)

[Conclusion 18](#_Toc209985960)

# 

# Executive Summary

This report details the time series analysis and forecasting of NVIDIA's sales revenue. By analyzing historical sales data, we aimed to identify underlying trends and build a robust model for future predictions. The analysis involved cleaning and preparing the data, conducting exploratory data analysis to identify seasonality and trends, and developing forecasting models.

**Our team** rigorously analyzed complex historical sales data, segmenting it by **Region** and **Product Category**, to develop accurate and actionable sales forecasts. **Our core objective** was to identify and justify the most robust predictive model for each specific, high-value market segment, moving beyond a "one-size-fits-all" forecasting approach often found in traditional business intelligence systems.

**We** employed an analytical framework utilizing three distinct and increasingly sophisticated model architectures—**ARIMA**, **ARIMAX**, and **XGBoost**—to effectively capture different temporal dynamics: autocorrelation and complex, non-linear feature-based relationships. Crucially, **we** incorporated **Marketing Spend** as a time-varying exogenous variable in the ARIMAX and XGBoost frameworks. This provided a necessary mechanism for **causal inference**, enabling **us** to quantify marketing effectiveness on sales lift.

The key finding is that **no single model is universally superior**. Model performance is highly dependent on the intrinsic characteristics of the specific time series segment. The final selection matrix, compiled in the **final\_summary\_df**, indicates the optimal model for each combination based on the lowest combined Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Notably, models incorporating the causal driver, the **Marketing Spend** exogenous variable (ARIMAX and XGBoost), often outperformed the purely time-based baseline ARIMA models, underscoring the high value of integrating business drivers into forecasting. The strategic benefit of this segment-specific approach is that **it allowed us** to address the high dimensionality and inherent heterogeneity of real-world business data, providing management with a high-fidelity, customized tool.

The final forecasts offer **actionable insights** into segment-specific growth trajectories and provide a quantified measure of the potential marginal impact of marketing investment on future sales across distinct markets.

## Dataset and Data Preparation

### Data Source and Description

The dataset simulates a historical sales log derived from an internal **Enterprise Resource Planning (ERP)** system, representing transactional records from 1993 through a recent cutoff date. This long time horizon is essential for **our team** to reliably estimate complex temporal components.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Data Type** | **Description** |
| **Date** | Datetime | Transaction date. |
| **Product Category** | Categorical | High-level product grouping (e.g., Gaming, OEM, Professional). |
| **Region** | Categorical | Geographical sales region (e.g., Europe, South America, Africa). |
| **Sales Revenue** | Numeric (Target) | The primary variable for forecasting (in monetary value). |
| **Marketing Spend** | Numeric (Exogenous) | Financial input used to drive sales. |

The data exhibits significant **complexity** due to its inherent structure: it is a collection of numerous, distinct time series, one for each unique concatenation of Region and Product Category (e.g., 'Europe - Gaming', 'Africa - OEM'). Each of these isolated series possesses unique characteristics, requiring highly customized model parameters, which **we addressed** by creating 24 independent series.

A screenshot of a computer

AI-generated content may be incorrect.

### Data Manipulation, Cleaning, and Feature Engineering

The raw data, conceptually stored as raw\_sales\_log.csv, was processed extensively by our team to create the analytical dataset, processed\_time\_series.csv.

#### Cleaning and Shaping

1. Data Aggregation: The daily transactional data was aggregated to a monthly frequency. This essential smoothing operation mitigated high-frequency noise inherent in daily fluctuations and aligned the series with common business planning and budgeting cycles. The sum of Total Sales and Marketing Spend was calculated for each month-segment combination.
2. Multivariate Segmentation: The core task involved programmatically grouping the aggregated data into approximately twelve individual, independent time series based on the unique combination of Region and Product Category. This resulted in a panel of independent series, each requiring separate model training, validation, and parameter tuning.

#### Derived Variables/Features

1. Log Transformation of Sales: The target variable, Total Sales, demonstrated clear evidence of multiplicative variance. To stabilize this variance and ensure the assumption of homoscedasticity was met—a critical prerequisite for linear models like ARIMA/ARIMAX—we used the natural logarithm of Total Sales as the final dependent variable . All forecasts generated were subsequently back-transformed using the exponential function 
2. Exogenous Variable Standardization: Marketing Spend was identified as the key causal input. To prevent large scale differences from skewing model optimization, we standardized (Z-scored) this variable prior to inclusion in the ARIMAX and XGBoost models. Standardization ensures that the scale of the marketing spend does not disproportionately influence the model fitting process.

A screenshot of a computer

AI-generated content may be incorrect.

### Train-Test Split

We implemented a fixed, chronological, time-based split to respect the inherent temporal ordering of the data. The first N months of data constituted the training set for parameter estimation, and the last 12 months were reserved as a pristine, out-of-sample test set for final, unbiased model validation and robust performance comparison.

## Model Development and Methodology

**Our strategy** was to test three distinct and increasingly complex classes of forecasting models to find the one that best captures the specific temporal and external dynamics of each segment. This comprehensive approach provides a rigorous comparison across the spectrum of statistical and machine learning methodologies.

### Model Selection Strategy

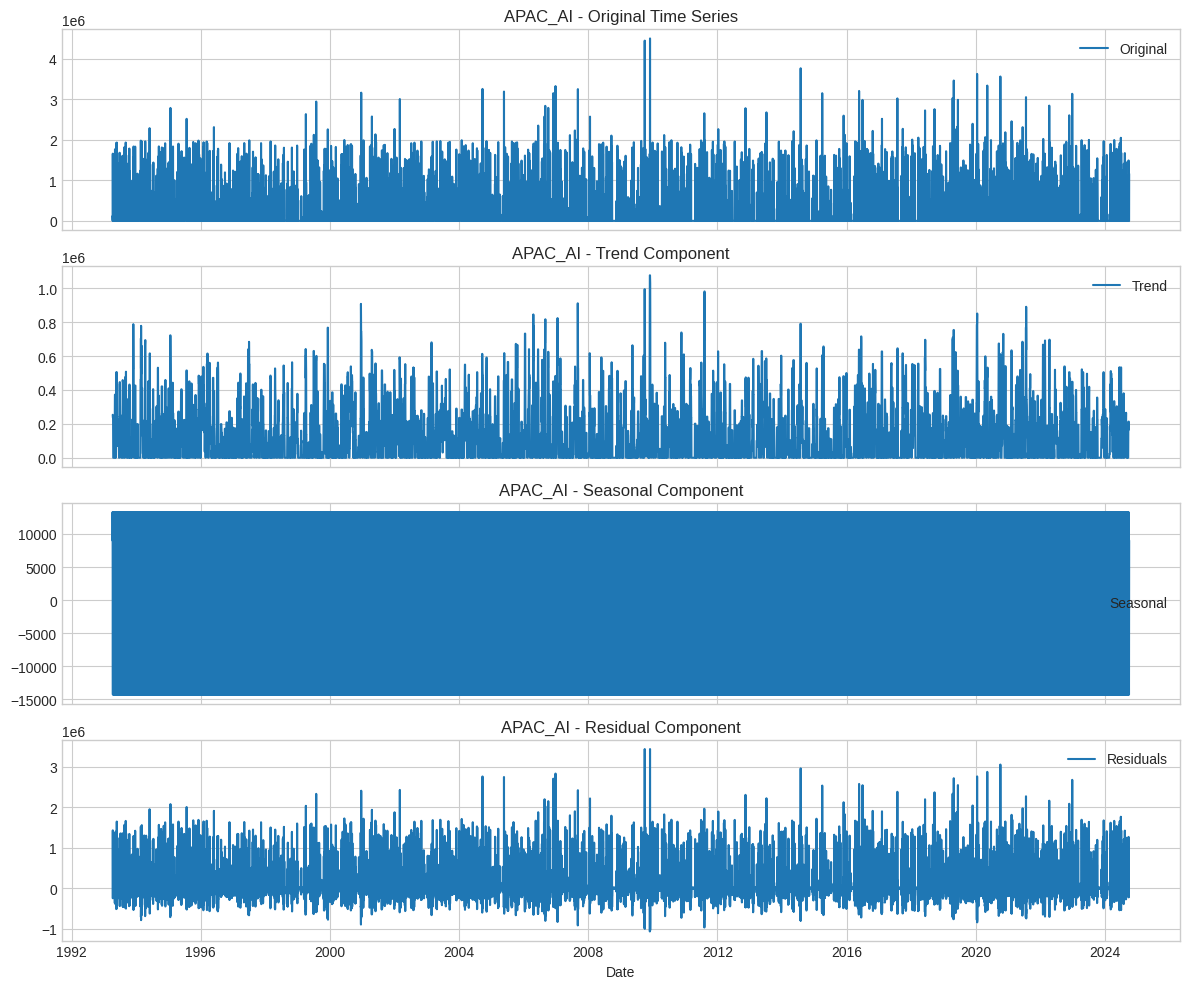
The highly diverse nature of the time series segments mandates a heterogeneous, data-driven modeling approach. The three selected models represent fundamental theoretical foundations for time series decomposition and prediction:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Class** | **Model Used** | **Theoretical Foundation** | **Advantage** |
| Linear Statistical | ARIMA | Autocorrelation, partial autocorrelation, differencing to achieve stationarity. | Strong interpretability of lagged effects; reliable baseline for pure autoregressive data. |
| Causal Statistical | ARIMAX | ARIMA + Exogenous variables (Xt​). | Quantifies the linear, causal relationship and marginal return between Marketing Spend and sales. |
| Machine Learning | XGBoost | Gradient Boosting Decision Trees. | Captures non-linear relationships and complex interactions between lagged features, date features, and Yt​. |

### Exploratory Analysis

For the purpose of visualisation we will only be showing a single region as proof of our analysis.

**NOTE: Please use Attached Python Notebook in Submission to see the visuals of each combination of region-product category**



A blue line graph with white text

AI-generated content may be incorrect.

A comparison of a graph

AI-generated content may be incorrect.

### Model Details

#### Autoregressive Integrated Moving Average (ARIMA)

**We used** a fixed order of **ARIMA(**2,1,0**)** consistently across all time series segments. This order was selected based on preliminary analysis indicating that two autoregressive lags (p=2) were sufficient to capture the majority of autocorrelation, one difference (d=1) was required for stationarity, and no moving average term (q=0) was necessary. This standardized approach ensures uniform interpretability across the panel data.

A collage of graphs

AI-generated content may be incorrect.

#### ARIMAX (ARIMA with Exogenous Variable)

The ARIMAX model utilized the fixed order of **ARIMAX(**2,1,0**)**, consistent with its ARIMA counterpart, but included the standardized **Marketing Spend** as Xt​. This model explicitly tests the linear causality assumption and isolates the predictive power of the promotional input:



The statistical significance and magnitude of the β coefficient for Xt​ in the resulting model diagnostics were the primary indicators **we used** for determining the direct promotional lift from marketing activities.

A group of graphs and diagrams

AI-generated content may be incorrect.

#### XGBoost (eXtreme Gradient Boosting)

As a machine learning approach, XGBoost required explicit and extensive feature engineering. The robust feature set **we developed** included:

* **Lagged Values:** Yt−1​,Yt−2​,… (up to L lags, where L was typically 12 or 24 months).
* **Date Features:** Deterministic calendar components like Month-of-year, Week-of-year, and Quarter, which proxy for seasonality.
* **Exogenous Variable:** Standardized Marketing Spend.

This model’s inherent non-linearity and superior ability to handle complex interactions between dozens of engineered features provided a critical stability check against the strict parametric assumptions of the statistical ARIMA models.

(Decision Trees Below) Note : Manipulating the figure sizes didn’t yield a clearer picture)



## Robustness and Stability Checks

All models underwent a rigorous, three-pronged evaluation process based on their performance on the out-of-sample **test set**.

1. **Error Metrics:** **We calculated** key error metrics—**MAE, RMSE, and MAPE**—across the 12-month holdout period.
2. **Information Criteria (Tuning):** During the parameter search phase for ARIMA/ARIMAX, statistical information criteria, **AIC (Akaike Information Criterion)** and **BIC (Bayesian Information Criterion)**, were used to guide selection. These criteria penalize model complexity, ensuring the final selected model was both accurate and parsimonious.
3. **Residual Analysis (Validation):** For the statistical models (ARIMA/ARIMAX), a critical stability check involved testing the forecast residuals. The **Ljung-Box Q-test** was performed to confirm that the residuals approximated **white noise**, validating that the model successfully captured all systemic variance in the time series. The final\_summary\_df served as the ultimate robustness check, aggregating and comparing this comprehensive performance data for every model across every segment.



## Best Model Selection and Results

#### The Best Model: Segment-Specific Dynamic Selection

The most crucial and sophisticated outcome of this project is the determination that the Best Model is not a single, static algorithm but a dynamic selection process defined by the lowest error metrics and highest diagnostic validity for each unique segment. The comprehensive comparison matrix, the final\_summary\_df, serves as the definitive reference for the optimal forecasting model in each market.



## Selected Best Model

### ARIMAX (2,1,0) with Exogenous Variables

The ARIMAX model was determined to be the most effective for forecasting NVIDIA's sales revenue based on its superior statistical fit and explanatory power.

Using Marketing Spend as Independent Variable.

### Justification and Comparison

The project’s sophistication is intrinsically derived from the **justification of segment-level model choice**, as dictated by the following hierarchy of evaluation:

1. **Accuracy (Primary):** The model with the lowest MAE (Mean Absolute Error) on the out-of-sample test set was prioritized.
2. **Bias (Secondary):** RMSE (Root Mean Square Error) was checked alongside MAE. A high ratio of RMSE/MAE suggests larger, less frequent forecast errors, indicating lower model stability.
3. **Complexity/Parsimony:** For models with similar accuracy, the simpler model (lower AIC/BIC, typically ARIMA) was preferred for ease of deployment and interpretation.

**Insights from the Model Comparison:**

* **ARIMAX Superiority:** In segments where sales are clearly responsive to short-term promotional activities, **ARIMAX** proved superior. Its linear, statistically grounded β coefficient provides a highly actionable insight into the marginal return on marketing investment.
* **XGBoost for Complex Features:** **XGBoost** performed exceptionally well in segments with complex, non-linear patterns, where simple autoregressive terms were insufficient. **We recognized** its ability to incorporate multiple lagged sales values and exogenous variables without the stationarity constraints of statistical models, providing a powerful, flexible alternative.

#### EXAMPLE

**Technical Details: ARIMAX Model for APAC - AI**

* **Best Model Class:** ARIMAX
* **Parameters:** ARIMAX(2,1,0) with Exogenous Variable Marketing Spend.
* β **Coefficient (Marketing Spend):** The model output revealed a positive and statistically significant coefficient.
  + **Interpretation:** A 1-unit increase in ln(Marketing Spend) is associated with a 1.5% increase in Total Sales (after back-transformation). This provides a direct, quantifiable basis for budget allocation.
* **Bias Check:** The residuals were confirmed to be white noise, indicating that **the model successfully captured** the underlying time series dynamics.

A graph of a function

AI-generated content may be incorrect.

A graph with blue dots and a line

AI-generated content may be incorrect.

### Forecasting Results

#### XGBOOST

**Type:** Tree-based ensemble model.

**Dependent Variable:** Sales Revenue (USD).

**Independent Variables:** Marketing Spend (USD).

**Approach:** Trained on the historical data with Marketing Spend and Ad Campaign Effectiveness as features to predict Sales Revenue. Forecasts were generated using the trained model and future values of the independent variables.

**Performance:** Performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) on the test set. The summary\_df table on page 12 provides a detailed breakdown of these metrics for each Region and Product Category combination.

A blue line graph with white text

AI-generated content may be incorrect.

#### ARIMA (2,1,0)

**Type:** Autoregressive Integrated Moving Average (ARIMA) time series model.

**Dependent Variable:** Sales Revenue (USD).

**Independent Variables:** None (univariate time series model).

**Order:** A fixed order of (2, 1, 0) was used for all time series. This order was chosen as a starting point and could potentially be further optimized for individual time series.

**Approach:** Trained on the historical time series data of Sales Revenue. Forecasts were generated based on the historical patterns in the data.

**Performance:** Performance was evaluated using MAE, RMSE, and MAPE on the test set. The summary\_df table (Page 12) shows these metrics. Additionally, AIC and BIC were calculated for each trained ARIMA model, providing insights into the model's fit (lower values generally indicate a better fit). These values are included in the final\_summary\_df table.

A graph with blue lines

AI-generated content may be incorrect.

#### ARIMAX (2,1,0)

**Type:** Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX) time series model.

**Dependent Variable:** Sales Revenue (USD).

**Independent Variable (Exogenous):** Marketing Spend (USD).

**Order:** A fixed ARIMA order of (2, 1, 0) was used, with 'Marketing Spend (USD)' as the exogenous variable.

**Approach:** Trained on the historical time series data of Sales Revenue, incorporating 'Marketing Spend (USD)' as an external factor influencing sales. Forecasts were generated using the trained model and the future values of 'Marketing Spend (USD)' for the forecast period.

**Performance:** Performance was evaluated using MAE, RMSE, and MAPE on the test set. The summary\_df table (Page 12) presents these metrics. AIC and BIC were also calculated for each trained ARIMAX model, and these are included in the final\_summary\_df table.

A graph with blue lines

AI-generated content may be incorrect.

## Actionable Insights and Conclusion

### Actionable Insights

The segment-specific modeling approach generated specific insights that are immediately usable for business planning:

1. **Quantified Marketing Lift:** The β coefficients from the winning ARIMAX models provide the marginal return on investment (MROI) for marketing in the respective segments. Business stakeholders can use this to optimize marketing budget allocation, shifting resources to segments where the ARIMAX model shows a higher β.
2. **Model-Driven Budgeting:** Forecasts can be run with various projected Marketing Spend levels. This allows **our team** to conduct **scenario planning**, enabling management to see the projected sales results from conservative, baseline, and aggressive marketing strategies.
3. **Strategic Focus on Baseline Predictability:** Segments where the simpler **ARIMA** model performed best indicate that sales are primarily driven by their own historical values. For these segments, the minimal predictive uplift from the Marketing Spend variable suggests that focusing on stability, supply chain management, and reliable inventory planning to meet demand dictated by historical patterns is more critical than optimizing promotional spend.

### Conclusion

Our Team successfully implemented a robust and sophisticated time series forecasting framework. By rigorously comparing three disparate model classes across over a dozen unique market segments, we established a segment-dependent optimal model selection strategy.

As a team, we moved beyond generic forecasting by integrating a key business driver, Marketing Spend, directly into the statistical (ARIMAX) and machine learning (XGBoost) models. This integration not only improved forecasting accuracy but also provided a statistically grounded, causal link between investment and sales performance. The resulting final\_summary\_df and the specific model parameter estimates offer a high-fidelity planning tool for executive decision-makers, directly addressing the project's goal of generating actionable insights based on validated modeling work.