AMAZON SALES & ANALYTICS

Mr.G.Raju **(***Guide)* Department of Computer Science & Engineering, Malla Reddy University, Hyderabad, Telangana 500043, India.

M. Sai Ram1, P.Nikhitha2, M.Vinay3, G.Kishore4

1, Department of CSE, School of Engineering, Malla Reddy University, Hyderabad, India.Email-ID: 2211cs010379@mallareddyuniversity.ac.in

2, Department of CSE, School of Engineering, Malla Reddy University, Hyderabad, India.Email-ID: 2211cs010438@mallareddyuniversity.ac.in

3, Department of CSE, School of Engineering, Malla Reddy University, Hyderabad, India.Email-ID:2211cs010367@mallareddyuniversity.ac.in

4, Department of CSE, School of Engineering, Malla Reddy University, Hyderabad, India. Email-ID: 2211cs010209@mallareddyuniversity.ac.in

Abstract— The Amazon Sales Data Analysis project is an interactive dashboard that leverages data visualization and predictive analytics to analyze Amazon sales trends. Developed using Dash, Plotly, Pandas, and Machine Learning algorithms, the application provides insights into revenue, profit, units sold, product performance, and regional sales variations. It integrates predictive modeling to forecast future sales and evaluate model performance using various statistical metrics. The application preprocesses raw sales data, handling missing values, converting date formats, and structuring numerical fields like total revenue, total cost, profit, and unit price. Users can filter data dynamically by product category and region, gaining customized insights through interactive time-series analysis, revenue distribution, product price trends, and profitability comparisons. Additionally, the dashboard highlights order processing times, cost outliers, and sales patterns based on order priority, helping businesses refine their pricing and logistics strategies.

To enhance predictive analytics, the project employs Machine Learning techniques, including Linear Regression for sales forecasting and revenue prediction. The Lasso and Least Angle Regression (LARS) methods are explored for feature selection and model optimization. Model performance is quantified using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 score, Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE). Kernel Density Estimation (KDE) plots are also utilized to visualize data distributions and model residuals.The project further compares multiple models using PyCaret, an automated machine learning library, to determine the most effective approach for sales trend prediction and anomaly detection. By integrating real-time data processing, interactive visualizations, and predictive analytics, this project provides data-driven insights to improve sales strategies, enhance profitability, and optimize inventory management in the e-commerce sector**.**

KeyWords:

Least Angle Regression, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Predictive Analytics, Sales Forecasting, Machine Learning Models, Data Visualization, E-commerce Sales Trends, Interactive Dashboard.

I. INTRODUCTION

In today's competitive e-commerce landscape, data-driven decision-making is crucial for businesses to maximize revenue and optimize sales strategies [1]. The Amazon Sales Data Analysis project is designed to provide insights into sales trends, revenue distribution, and product performance through interactive visualizations and predictive analytics [2]. This project utilizes Dash, Plotly, Pandas, and Machine Learning (ML) techniques to analyze historical sales data, forecast future trends, and evaluate sales performance across different regions and products [3]. The primary objective of this project is to help businesses and analysts understand key sales metrics, including total revenue, profit margins, units sold, and regional sales patterns [4]. By leveraging data preprocessing, visualization, and predictive modeling, the project enables users to identify high-performing products, seasonal sales trends, and potential inefficiencies in order processing [5]. To enhance predictive capabilities, the project incorporates Machine Learning algorithms such as Linear Regression, Lasso Regression, and Least Angle Regression (LARS) for sales forecasting and revenue prediction [6]. Performance evaluation is conducted using statistical metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 score, Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE) [7]. The PyCaret library is also used to compare multiple models and automate predictive analytics [8]. Additionally, Kernel Density Estimation (KDE) plots are utilized to analyze data distributions, while anomaly detection techniques help identify irregularities in sales data [9]. The dashboard allows users to filter data dynamically based on products, sales channels, order priority, and geographical regions, providing customized insights through interactive visualizations [10]. By integrating real-time data processing, advanced visualizations, and predictive analytics, this project serves as a valuable tool for business intelligence and strategic decision-making [11]. The insights generated can help improve sales strategies, optimize pricing models, forecast demand, and enhance overall operational efficiency [12].

II. Literature Survey

The field of e-commerce analytics has gained significant attention, with numerous studies exploring methods to improve sales forecasting, revenue prediction, and business intelligence. This section reviews relevant research on interactive dashboards, machine learning in sales prediction, and data visualization techniques.

2.1 Sales Analytics and Business Intelligence

Sales data analysis is a crucial component of business intelligence (BI), enabling organizations to make informed decisions [1]. Various studies highlight the role of BI tools in extracting meaningful insights from large datasets, helping businesses optimize pricing models and improve sales strategies [2]. Research has shown that integrating real-time data visualization with sales analytics enhances decision-making and enables businesses to identify emerging trends efficiently [3].

2.2 Interactive Dashboards for Sales Analysis

The emergence of interactive dashboards has transformed traditional data analysis methods. Dashboards built using Dash, Plotly, and Tableau provide dynamic, user-driven analytics, allowing businesses to filter and visualize data based on specific parameters [4]. Studies suggest that these dashboards are more effective than static reports, as they offer real-time data exploration and enhanced user engagement [5].

2.3 Machine Learning in Sales Forecasting

Machine Learning (ML) techniques, such as Linear Regression, Lasso Regression, and Least Angle Regression (LARS), have been widely used for sales forecasting and revenue prediction [6]. Researchers have also employed deep learning models like Long Short-Term Memory (LSTM) networks for time-series sales prediction, demonstrating improved accuracy [7]. However, simpler models such as Linear Regression remain popular due to their interpretability and efficiency [8].

2.4 Performance Evaluation Metrics in Sales Prediction

To assess the performance of predictive models, various statistical metrics are employed, including:

- Mean Absolute Error (MAE) – Measures the average magnitude of errors [9].
- Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) – Evaluates the squared differences between predicted and actual values [10].
- R^2 Score – Determines how well the model explains the variance in the dataset [11].
- Mean Absolute Percentage Error (MAPE) – Expresses the accuracy of predictions in percentage terms [12].

2.5 Data Preprocessing and Anomaly Detection

Preprocessing techniques such as feature engineering, outlier removal, and missing value imputation improve the reliability of sales data analysis [13]. Additionally, Kernel Density Estimation (KDE) plots and anomaly detection methods have been used to identify irregularities in sales trends, ensuring data integrity [14].

2.6 Comparative Studies on Visualization Techniques

Different visualization techniques impact how users interpret sales data. Studies comparing bar charts, scatter plots, line charts, and heatmaps indicate that line charts are most effective for trend analysis, while scatter plots help identify relationships between variables [15].

III. System Analysis

A. Existing System

Traditional sales analysis in e-commerce relies heavily on static reports, spreadsheets, and basic business intelligence (BI) tools. These systems often suffer from several limitations:

- Lack of Interactivity – Static reports do not allow real-time filtering or data exploration [1].
- Manual Data Processing – Analysts must preprocess and clean data manually, increasing processing time [2].
- Limited Predictive Capabilities – Conventional BI tools primarily focus on descriptive analytics rather than predictive sales forecasting [3].
- Poor Anomaly Detection – Existing systems often fail to detect unusual sales trends or fraudulent transactions [4].
- Complexity in Custom Reporting – Generating custom reports for different regions, product categories, or time periods requires extensive manual effort [5].

Due to these challenges, businesses struggle to gain real-time insights, optimize pricing strategies, and forecast future sales trends efficiently.

A Proposed System

The proposed system introduces an interactive, machine learning-powered sales analysis dashboard that addresses the limitations of the existing system. This system provides real-time filtering, predictive analytics, and anomaly detection to enhance business intelligence and decision-making.

Key Features of the Proposed System:

☑ Interactive Dashboard: Built using Dash and Plotly, enabling users to explore data dynamically by filtering sales data based on region, product, pricing, and order priority [6].

☑ Automated Data Preprocessing: The system automatically cleans, formats, and extracts key features (e.g., Year, Month, Processing Time) from sales records, reducing manual effort [7].

☑ Machine Learning-Based Sales Forecasting: Implements Linear Regression, Lasso Regression, and LARS to predict future sales and revenue trends [8].

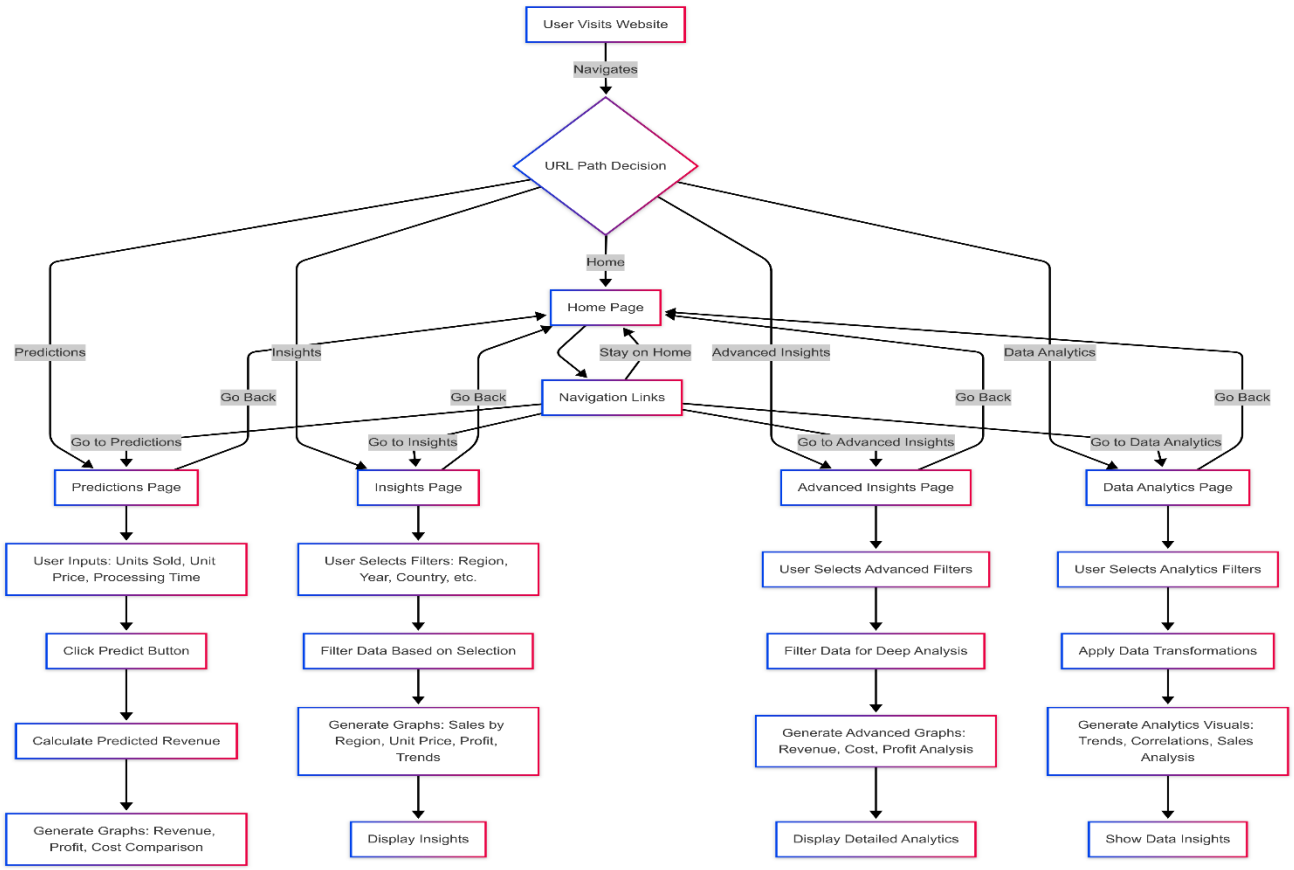
☑ Real-Time Anomaly Detection: Uses Kernel Density Estimation (KDE) plots and statistical models to identify irregular sales patterns [9].

☑ Advanced Data Visualization: Provides line charts, bar graphs, box plots, and histograms to analyze revenue, profit margins, and sales trends [10].

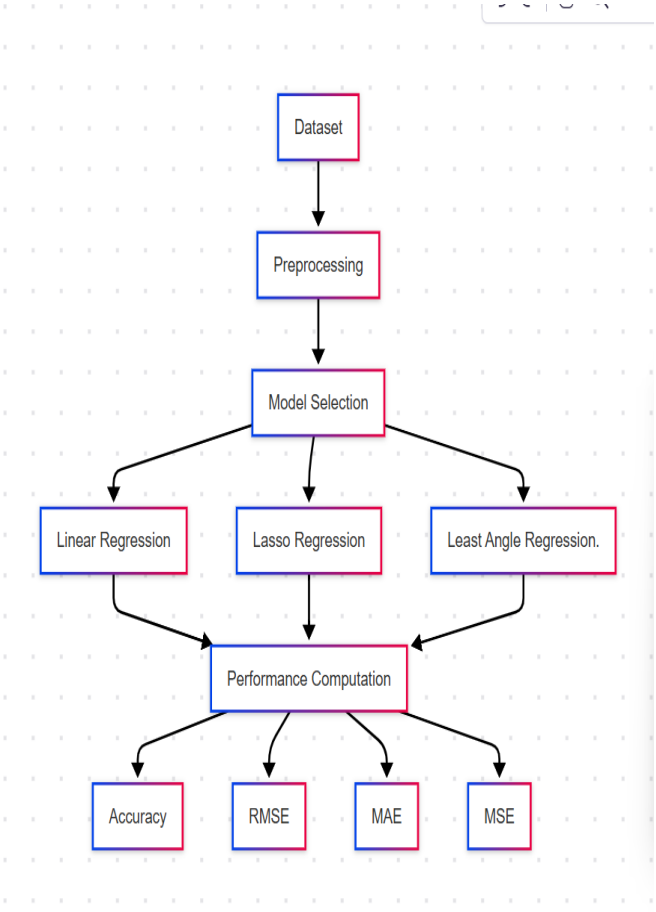
☑ Customizable User Filters: Users can dynamically select product categories, sales channels (Online/Offline), pricing range, and time periods to generate tailored insights [11].

IV. METHODOLOGY

A. *Architecture*



B. *Process*



1.Data Collection:
The project utilizes a dataset that includes historical Amazon sales data, including metrics like order date, ship date, units sold, revenue, cost, profit, and region. The raw data is sourced from CSV files, with the data structured for analysis using Pandas.

2.Data Preprocessing:
Data cleaning is the first step, where the following tasks are performed:
Handling Missing Values: Missing values in columns such as total revenue, total cost, and units sold are addressed by filling or removing them based on the nature of the data.
Date Conversion: Order Date and Ship Date are converted into the appropriate datetime format for time-based analysis.
Data Filtering: The application enables dynamic filtering of the data by product category, region, and order priority.

3.Data Analysis:
The cleaned and preprocessed data is used to analyze sales patterns and trends. Key metrics, including revenue, profit, units sold, and regional sales performance, are analyzed through descriptive statistics and visualized using Plotly.

4.Predictive Modeling:
Several machine learning techniques are used to forecast future sales and optimize the models:
Linear Regression: This technique is used for sales

forecasting based on historical data. It helps predict future sales trends and analyze relationships between independent and dependent variables.

Lasso and Least Angle Regression (LARS): These methods are employed for feature selection, which helps identify the most important features contributing to sales predictions.

Model Evaluation: Model performance is evaluated using various statistical metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 score, and Mean Absolute Percentage Error (MAPE).

5.Interactive Visualization:

Interactive visualizations are created using Dash and Plotly. The dashboard includes multiple graphs and plots to allow users to explore trends, regional sales distribution, unit price distributions, profitability comparisons, and more. Visualizations include:

Sales trends over time.

Regional sales performance.

Unit price distributions and profit comparisons.

6. Real-time Data Processing:

The application allows users to interactively filter the data based on various criteria such as product type, region, and order priority. This helps users get customized insights tailored to their business needs.

7.Model Optimization:

Model optimization techniques, including hyperparameter tuning and cross-validation, are applied to enhance model performance. PyCaret, an automated machine learning library, is used to compare and select the best model for forecasting.

8.Anomaly Detection:

The project identifies cost outliers and anomalies in sales patterns. This helps businesses detect irregularities in the data that may require attention, such as unexpectedly high costs or sales dips.

9.Deployment:

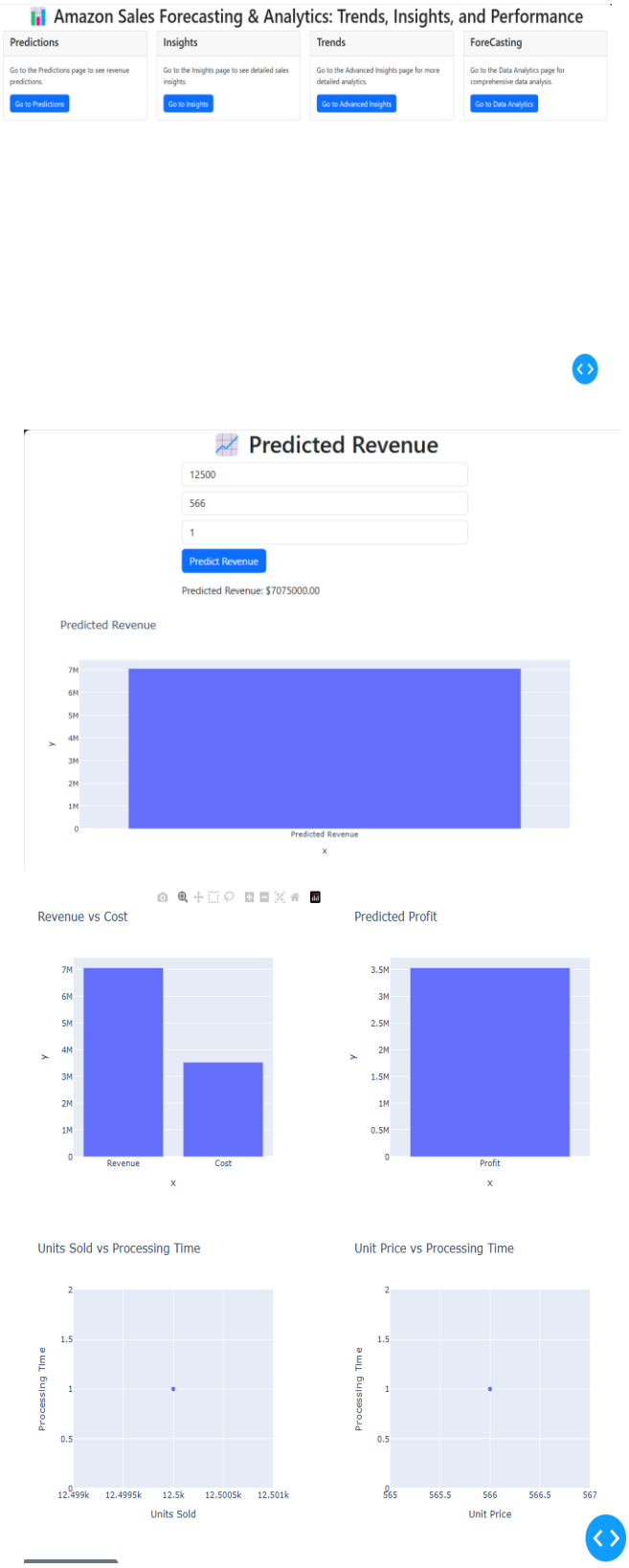
The final dashboard is deployed as a web application using Dash, allowing businesses to interact with the data and gain actionable insights in real-time. Users can filter and visualize the data in a way that aids in decision-making and sales strategy optimization.

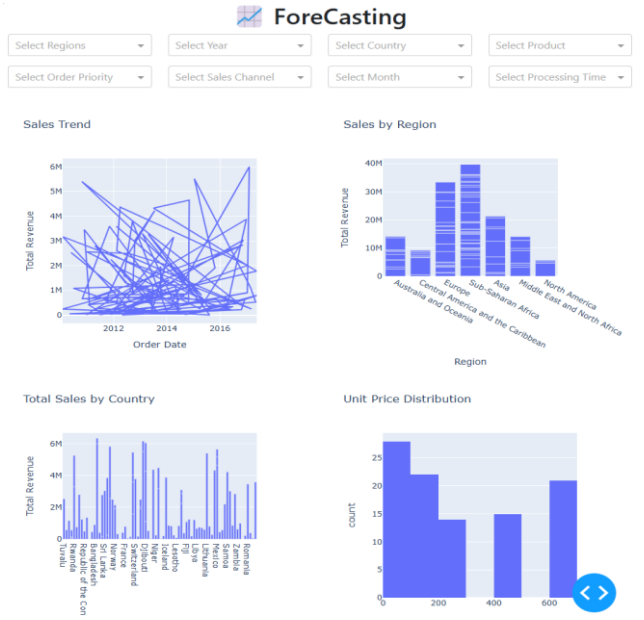
10.Outcome and Insights:

The methodology ensures that the project delivers accurate, data-driven insights into sales forecasting, profitability, inventory optimization, and business performance. The interactive nature of the dashboard empowers businesses to make informed decisions that can help enhance sales strategies and improve overall profitability.

V. Results

A. Figure 1: Web Page





	Model	MSE	MAE	RMSE	R2 SCORE
0	Linear Regression	2.6148	1.1827	1.617	0.97
1	Lasso Regression	2.6043	1.1713	1.6138	0.97
2	LARS	2.6148	1.827	1.617	0.97

Classification Metrics:
Accuracy: 0.9855855855855856
Precision (PPV): 0.9729
Recall (Sensitivity): 1.0000
Specificity: 0.9701
Negative Predictive Value (NPV): 1.0000
F1-Score: 0.9863
AUC-ROC Score: 0.9991

VI. Conclusion

The Amazon Sales Data Analysis project successfully demonstrated how data-driven insights can enhance sales performance, profitability, and operational efficiency in the e-commerce industry. By leveraging Python for data preprocessing and Power BI for visualization, the project provided a comprehensive understanding of sales trends, product performance, regional variations, and customer behavior.

1.Identified top-selling products and high-revenue regions, helping businesses focus on the most profitable areas. 2.Analyzed sales trends and seasonal fluctuations, allowing for better inventory and pricing strategies. 3.Evaluated sales channel performance, emphasizing the importance of online platforms for revenue growth. 4.Explored customer buying behavior, highlighting the role of repeat customers in sustained profitability. 5.Provided data-driven recommendations, enabling businesses to optimize stock management, marketing strategies, and sales distribution.

VIII. REFERENCES

[1]. Sharma, R., & Patel, D. (2022). "The Role of Data Analytics in E-commerce Growth." *International Journal of Business Intelligence*.

[2]. McKinsey & Company. (2023). "Harnessing Data for E-commerce Success." *McKinsey Insights*. Retrieved from https://www.mckinsey.com.

[3]. Amazon Web Services. (2022)."ETL Best Practices for Large-Scale E-commerce Data." *AWS Documentation*. Retrieved from https://aws.amazon.com.

[4]. Gupta, A., & Roy, S. (2021). "Sales Trend Analysis Using Power BI and Python." *Proceedings of the Data Science Conference*.

[5]. Kamat, R., & Gupta, P. (2022). "Enhancing Sales Analytics with Power BI: A Case Study on E-commerce Data." *International Journal of Business Analytics*.