**Predictive Analysis for Amazon Sales Data**

**Data Source and Acquisition:**

For this analysis, we intend to use second-hand data obtained from Kaggle, specifically the Amazon sales report dataset. The dataset will be acquired by downloading it directly from Kaggle <https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data>

The following file was used for analysis. Various machine learning algorithms were applied to test their accuracy for the depth analysis.

A computer screen shot of a computer

Description automatically generated

**Data Description:**

This data collection offers in-depth insights into Amazon sales, such as the SKU Code, Design Number, Stock, Category, Size, and Color, to assist in optimizing product profitability.

**Predictive Variables:**

We plan to make predictions about the following variables:

* **Sales Amount Distribution Visualization:** To visualize the distribution of sales amounts, we can use a histogram with Kernel Density Estimation (KDE) to provide insights into the frequency and distribution of the sales amounts across orders.
* **Order Quantity Distribution Visualization:** To visualize the distribution of order quantities, we can use a histogram with Kernel Density Estimation (KDE) to illustrate the frequency and distribution of the quantities ordered across the dataset.
* **Exploratory data analysis (EDA)**:
* **Date Distribution Visualization:** To analyze the distribution of orders over time, we use a histogram to visualize the frequency of orders across different dates. This helps in understanding how order volume varies over time.
* **Order Status Distribution:** The bar plot shows the frequency of each order status category. This helps in identifying the most common order statuses and the proportion of each status in the dataset.
* **Fulfilment Type Distribution:** By plotting the fulfillment types, we can visualize the distribution of different fulfillment methods used for the orders. This provides insights into the preferred or most common fulfillment types.
* **Sales Channel Distribution:** This bar plot displays the distribution of sales channels used for the orders. It helps in understanding which sales channels are most frequently used and their relative frequency.
* **Shipping Service Level Distribution:** The bar plot illustrates the distribution of different shipping service levels. This visualization shows the frequency of various shipping service levels across orders.
* **Product Category Distribution:** The bar plot reveals the frequency of each product category in the dataset. This helps in understanding the popularity and distribution of different product categories.
* **Courier Status Distribution:** By plotting the courier statuses, we can see the distribution of different courier statuses. This provides insights into the performance and distribution of courier services.
* **Quantity Distribution:** The histogram visualizes the distribution of quantities ordered. It helps in understanding how frequently different quantities are ordered and any patterns in the quantity distribution.
* **Currency Distribution:** This bar plot shows the distribution of currencies used in the sales. It helps in identifying the most common currencies and understanding their relative frequency.
* **Sales Amount Distribution:** The histogram with KDE visualizes the distribution of sales amounts. It provides insights into how sales amounts are distributed and helps in identifying patterns or outliers in sales data.
* **Top 10 Shipping Cities:** By plotting the top 10 shipping cities, we can see which cities have the highest frequency of shipments. This helps in identifying key shipping locations.
* **Top 10 Promotion IDs:** This bar plot shows the top 10 promotion IDs with the highest frequency. It helps in understanding which promotions are most frequently used.
* **B2B Orders Distribution:** The bar plot visualizes the distribution of B2B orders. It helps in understanding the proportion of B2B orders compared to other order types.
* **Bivariate analyses:**
* **Amount vs. Quantity Analysis:** To explore the relationship between order quantity and sales amount, we use a scatter plot. This visualization helps in understanding how variations in quantity ordered impact the total amount of sales, revealing any potential correlation or trends**.**
* **Sales Channel vs. Amount Analysis:** The box plot compares the distribution of sales amounts across different sales channels. This analysis helps in understanding how sales amounts vary with different sales channels, identifying any significant differences or outliers associated with each channel.
* **Fulfillment vs. Amount Analysis:** The box plot visualizes the distribution of sales amounts based on different fulfillment types. This helps in assessing how different fulfillment methods impact the sales amount, revealing any notable differences in sales performance across fulfillment types.
* **Regression Analysis for Sales Amount Prediction:** This Linear Regression model predicts sales amounts based on features like quantity, category, size, and B2B status. It evaluates model performance using metrics such as Mean Squared Error (MSE) and R-squared (R²), focusing on how well these features explain the variation in sales amounts.
* **Classification for Order Fulfillment Status:** This Decision Tree Classifier model predicts the order fulfillment status based on features like quantity, category, size, amount, and B2B status. The model’s performance is assessed using accuracy and detailed classification metrics, which help determine how effectively the model predicts different fulfillment statuses.
* **Order Fulfillment Status Prediction:** We use a Decision Tree Classifier to predict order fulfillment status, trained on features like quantity, category, size, amount, and B2B status. The model's performance is evaluated using accuracy and classification metrics.
* **Clustering for Segmenting Orders:** K-Means clusters orders into 3 groups based on quantity, amount, and B2B status. It helps identify and analyses distinct order segments.
* **Time Series Forecasting:** This ARIMA model forecasts future sales amounts based on historical sales data. It fits an ARIMA model to monthly aggregated sales data and provides a forecast for the next 12 months, helping in predicting future sales trends.
* **Association Rule Mining:** This code uses the Apriorism algorithm to find frequent item sets and generate association rules for product categories. It helps identify product associations based on purchase patterns.
* **Sales Amount Prediction:** The code uses Linear Regression to predict sales amount from quantity and B2B status, handling missing values, and evaluates the model's performance with MSE and R² scores.
* **Courier Service Prediction:** The code uses an SVC model to predict courier status based on quantity and B2B features, processing the data, training the model, and evaluating it with accuracy and classification metrics.

**Business Value:**

The predictive analyses and visualizations provide substantial business value by:

* **Optimizing Inventory Management:** Visualizing and predicting sales amounts and quantities helps manage inventory levels efficiently, reducing risks of overstock or stockouts.
* **Enhancing Order Fulfillment**: Understanding the distribution of order statuses and predicting fulfillment types ensures timely deliveries, boosting customer satisfaction.
* **Improving Marketing and Promotions**: Insights from sales amount and distribution data enable targeted marketing strategies and effective promotions, enhancing customer engagement and conversion rates.
* **Increasing Operational Efficiency**: Clustering and segmenting orders help in identifying key customer segments and optimizing order processing and fulfillment strategies.
* **Supporting Strategic Decisions**: Time series forecasting provides foresight into future sales trends, assisting in strategic planning and resource allocation.
* **Boosting Revenue and Profitability**: Accurate predictions and analyses of sales amounts, order quantities, and other key metrics lead to improved revenue management and profitability.
* **Enhancing Customer Insights:** Visualizing order patterns, shipping locations, and product associations helps understand customer preferences and optimize product offerings.
* **Improving Shipping and Courier Management**: Predicting courier service performance and understanding shipping trends help in optimizing shipping strategies and improving service levels.