Market Sales Prediction Using Machine Learning

Presented By : Man Desai

May 2, 2024

1 Introduction

In today's digital world, the rise of e-commerce has transformed the way businesses operate and consumers shop. E-commerce refers to the buying and selling of goods and services over the internet. With many online platforms and many sellers available for a single product, the e-commerce market has become fiercely competitive.

In this highly competitive landscape, businesses must constantly adapt and innovate to stay ahead of the competition. Customer expectations are continuously evolving and the demand for seamless shopping experience, personalized recommendations and speedy delivery.

Among the many online e-commerce platforms, Amazon stands out as the world's largest online retailer. Amazon's e-commerce platform enables third party retailers to showcase and sell their products alongside Amazon's own items. With millions of products available under numerous categories, ranging from electronics and apparels to home goods and beyond, Amazon offers the customers with an unmatched shopping experience. However, this also leads to increase in competition among the sellers and brands of different products. This is where market sales prediction helps the sellers or retailers make the right business decisions.

Market Sales Prediction refers to the process of forecasting future sales of a product or service within a particular market or industry. This involves analyzing historical sales data, market trends or consumer behavior, economic indicators and other relevant factors to estimate the potential demand for the product or service over a specific period of time.

This report explores the various analysis methodologies aimed at constructing a dashboard to assist retailers or brands in analyzing market sales and product demand across various categories. Additionally, the objective is to forecast or predict the monthly sales volume for individual products.

2 Dataset Overview

The dataset consists of categorical and numerical data which are listed below:

- **URL:** URL of the product being sold online.
- **ASIN:** Amazon Standard Identification Number
- **Title:** Product name
- **Brand:** Brand of the product
- **Fulfillment:** Defines where the product is stored. It has two values-
  - FBM (Fulfillment by Merchant)
  - FBA (Fulfillment by Amazon)
- **Category:** Category of the product being sold.
- **BSR:** The best seller rank. It is a metric that appears on an item's product page and indicates how it's selling compared to other items within the same product category in the Amazon store.
- **Subcategory:** The subcategory of the product within the main category.
- **Price:** Price of product in INR.
- **Price Trends:** Percentage of price increase or decrease in the last 90 days.
- **Monthly Sales:** Average monthly sales for the product.
- **Sales Trend:** A sales trend appears when a minimum number of customers have recently purchased a particular ASIN.
- **Monthly Revenue:** How much revenue the product generates in a month.
- **Review Count:** Number of total reviews of the product.
- **Reviews Rating:** The average rating of the product.
- **Seller:** The seller of the product.
- **Number of Active Sellers:** Number of sellers who are selling this product.
- **Last Year Sales:** Count of the sales for a particular product in the last year.
- **Size Tier:** The category of size tier in which the product falls.
- **Best Sales Period:** The month in which the sales were highest.
- **Age (Month):** The age of the product in months since its initial listing or release.

- **Number of Images:** Number of images of the product.
- **Variation Count:** Product variations of a particular product available for sale.
- **Sales to Reviews:** Ratio between the number of product sales and the number of customer reviews received for that product.

3 Methodology

First, We go to the Amazon website and collect data. Then, We clean up the data to make sure it's accurate. After that, We analyze the data to find important information. We create a dashboard to see the results clearly.

Next, We choose which data to use to predict sales. We build a model using this data. We train and test the model to make sure it works well. Finally, We use the model to predict future sales.

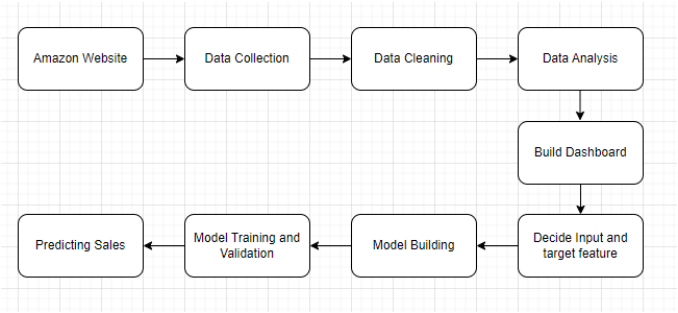

Figure 1: Pipeline of market sales prediction

3.1 Data Collection

During the data collection phase, information from Amazon was acquired through web scraping techniques. This encompassed details such as product names, category, subcategory, brands, and prices etc. The gathered data was comprehensive, including various attributes crucial for analysis.

3.2 Data Cleaning

Data cleaning, also called data preprocessing, is a very crucial step in the machine learning pipeline. It involves identifying and correcting errors, inconsistencies, inaccuracies in the raw data before it is further analyzed. For the process of data preprocessing, the following steps were done:

- Handling null values

  - For the rows with very few null values, the rows were dropped. But in cases where the brand was null, the brand names were found manually using the product url.

- Renaming the columns
  - Renaming the columns as per the naming convention.
- Dropping unwanted columns
  - The columns such as length, width, height, storage information were not relevant for the market sales prediction, so they were dropped.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) refers to the process of performing initial investigations on data to discover hidden patterns, trends, identifying anomalies or to test hypotheses or assumptions. We did the EDA, to identify useful charts that can be used in making predictions for the market sales.

We have collected the data for the following categories:

- Beauty
- Car & Motorbike
- Clothing & Accessories
- Computers & Accessories
- Electronics
- Health & Personal Care
- Home & Kitchen
- Industrial & Scientific
- Musical Instruments
- Office Products
- Shoes & Handbags
- Sports, Fitness & Outdoors
- Toys & Games
- Watches

Overall Analysis:

- Top Categories by Sales:

– Bar chart depicting the top 10 categories that had the highest selling. Beauty is top selling category followed by Clothing and Accessories.

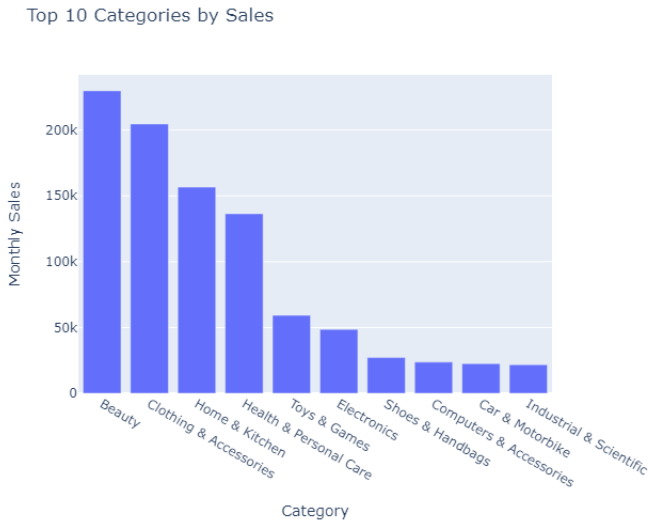

Figure 2: Top Categories by Sales

- Distribution of top categories:
  - Pie chart depicting the overall distribution for the top 10 categories.

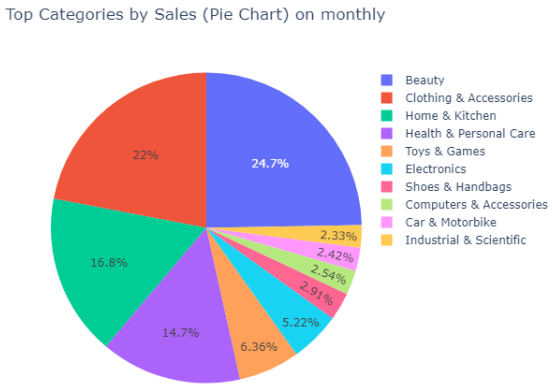

Figure 3: Distribution of top categories

- Top Selling products:
  - A table describing the highest selling products in each category.

	Category	Title	Last Year Sale
1	Beauty	Maybelline New York Matte Lipstick, Intense Colour, Moisturised Lips, Color Sensatior	307,867
2	Car & Motorbike	Autofy 100% Waterproof (Tested) RE Size Bike Cover Dustproof UV Protection Bike Bo	31,748
3	Clothing & Accessories	minicult Cotton Pajama Pants with All Over Print for Boys and Girls (Multicolor Desigr	60,826
4	Computers & Accessories	TP-Link Archer C50 AC1200 Dual Band Wireless Cable Router, Wi-Fi Speed Up to 867 M	35,097
5	Electronics	boAt Airdopes 121 PRO TWS Earbuds Signature Sound, Quad Mic ENx™, Low Latency	83,250
6	Health & Personal Care	MANGALAM Bhimseni Camphor Chunk 50G Jar - Pack Of 1	85,901
7	Home & Kitchen	Wakefit 100% Waterproof Premium Cotton Mattress Protector | Breathable and Hypoa	176,618
8	Industrial & Scientific	3M 1110 Ear Plugs Corded, Extra Soft, Reusable Earbuds Noise Cancellation, Soundpr	32,501
9	Musical Instruments	JNKC Mic Lapel Collar Mic Voice Recording Filter Microphone for Singing YouTube Sm	3,759
10	Office Products	Reynolds AEROSLIM BP 5 CT POUCH - BLACK | Ball Point Pen Set With Comfortable Gr	25,849
11	Shoes & Handbags	DOCTOR EXTRA SOFT Care Diabetic Orthopedic Pregnancy Flat Super Comfort Dr Flip	6,646
12	Sports, Fitness & Outdoors	Boldfit Heavy Resistance Band for Exercise & Stretching (Black)(Material: Natural Rub	5,062
13	Toys & Games	Pikipo Froggy Face Rattle Soft Toy with Squeeze Handle for Squeaky Sound (Green)	55,355
14	Watches	Fire-Boltt Phoenix Pro 1.39" Bluetooth Calling Smartwatch, AI Voice Assistant, Metal E	34,578

Figure 4: Top Selling products

- Number of Products with Desirable BSR:
  - BSR refers to the Best Seller Rank on Amazon and a good rank falls between the range 1-10,000. Watches have the most desirable BSR.

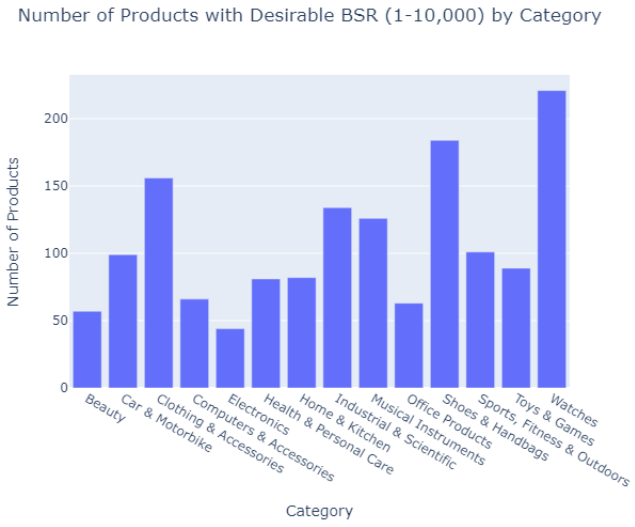

Figure 5: Number of Products with Desirable BSR

- Top Selling Brands:

– Bar chart that shows the Top Selling brands across all the categories. Maybelline is the top selling brand.

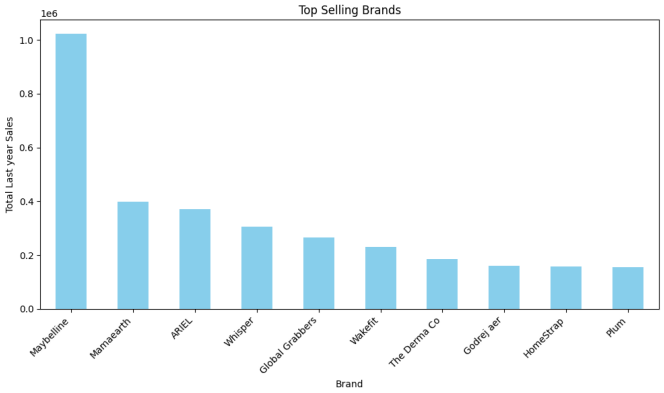

Figure 6: Top Selling Brands

- Top Reviewed Brands:
  - It shows the brands that have the highest reviews across all categories. boAt has the highest reviews indicating a good engagement with the customers.

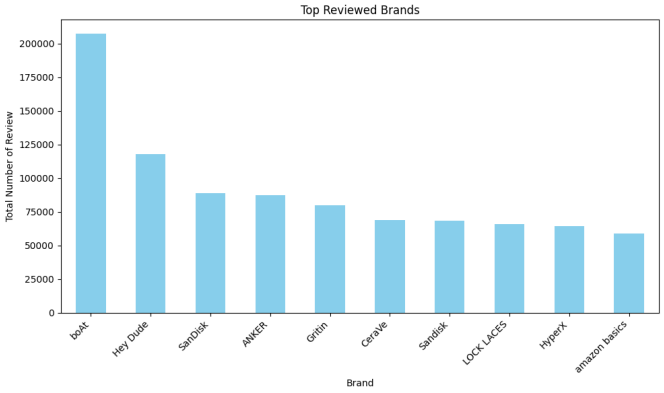

Figure 7: Top Reviewed Brands

- Average last year sales:
  - Table showing the average last year sales for each category.

	Category	Last Year Sale
1	Beauty	8343
2	Car & Motorbike	717
3	Clothing & Accessories	4812
4	Computers & Accessories	866
5	Electronics	1541
6	Health & Personal Care	3786
7	Home & Kitchen	4997
8	Industrial & Scientific	489
9	Musical Instruments	124
10	Office Products	610
11	Shoes & Handbags	548
12	Sports, Fitness & Outdoors	253
13	Toys & Games	1685
14	Watches	291

Figure 8: Average last year sales

- Best Sales month:
  - Chart describing the month in which the particular category is sold the most.

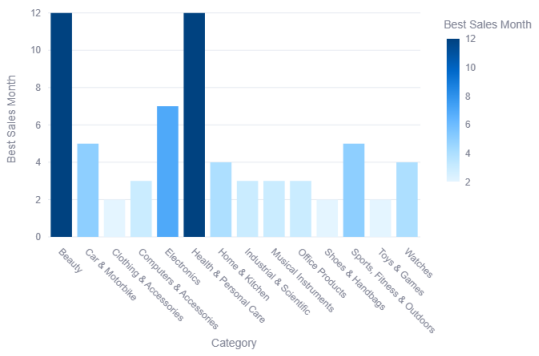

Figure 9: Best Sales month

- Reviews Count:
  - Bar chart that shows which categories have the highest number of reviews. Shoes and handbags are the most reviewed products followed by electronics, indicating a better engagement with audience.

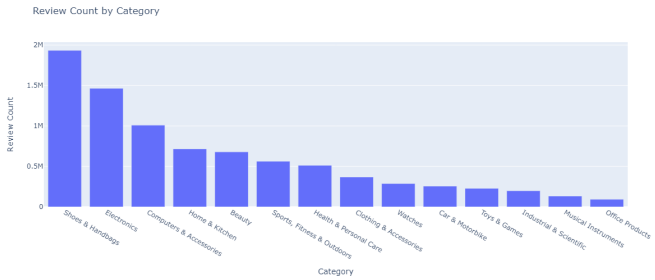

Figure 10: Reviews Count

- Fulfillment type distribution:
  - Pie chart indicating the ratio of Fulfillment type by amazon vs fulfillment by merchant.

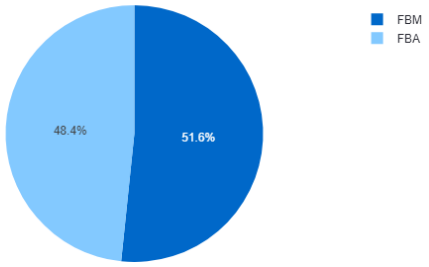

Figure 11: Fulfillment type distribution

Category-wise Analysis:

- Top subcategories by sales:
  - The top 10 subcategories within each category that have the highest sales.

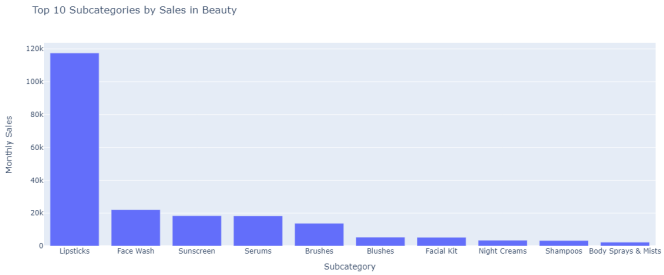

Figure 12: Top subcategories by sales

- Top selling brands:
  - Bar chart indicating the top selling brands in the particular category.

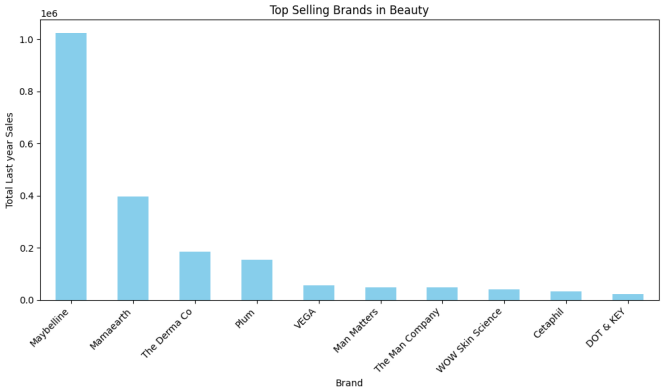

Figure 13: Top selling brands

3.4 Dashboard

Based on the exploratory data analysis, a dashboard was created using Streamlit to visualize the various charts for each category and do an overall analysis in an efficient way. The dashboard provides a centralized platform where multiple data visualizations, metrics and performance indicators are presented in a clear and concise manner.

3.5 Model Building

The target variable for market sales prediction is "monthly_sales" which is a numerical feature. Hence, this is considered as a regression problem. We have used the following models:

- Linear Regression:
  - Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes that this relationship is approximately linear, meaning that changes in the independent variables are associated with proportional changes in the dependent variable.
- RandomForest Regressor:
  - Random Forest Regressor is a powerful machine learning algorithm that belongs to the ensemble learning family. It works by constructing multiple decision trees during training and outputs the average prediction of the individual trees.Each decision tree in the Random Forest is trained on a random subset of the training data and a random subset of the features. This randomness helps to prevent overfitting and improves the generalization performance of the model.
- XGBoost Regressor:
  - XGBoost stands for Extreme Gradient Boosting and it is an implementation of gradient boosting algorithms. It works by building an ensemble of weak predictive models sequentially with each model focusing on the errors made by previous ones. It optimizes a specific objective function by minimizing the residuals of the predictions, which leads to the construction of strong predictive models. In summary, the XGBoost Regressor is well-suited for market sales prediction projects due to its ability to handle complex relationships in data, its robustness to overfitting, and its high predictive accuracy.

3.6 Model Training and Validation

During model training, the dataset was divided into training and test sets using an 80:20 split ratio. Categorical features were encoded using label encoding. A pipeline was created for the regression models, which included scaling the numeric features and applying preprocessing within the pipeline. The model was then evaluated on the test set, and the results were recorded for analysis.

3.7 Model Prediction

During the prediction phase of the model, we input the product name and category to generate sales predictions. This process involves computing the cosine similarity of the product name to identify the most similar products to the query, matching their categories, and then forecasting monthly sales for the selected product.

4 Results

Deriving observations from the data.

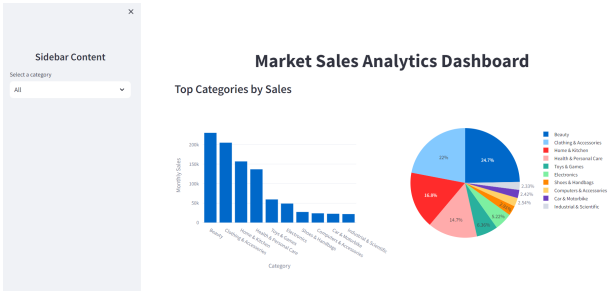

Figure 14: Sales Dashboard for all category

Figure 15: Sales Dashboard for all Beauty

Below R^2 and MAE value suggesting that the model provides a good fit to the data.

Model	R squared	Mean Absolute Error
XGBoost Regressor	0.8655692	25.0474
RandomForest Regressor	0.899869	27.0711
Linear Regression	0.768033	104.96847

Figure 16: R^2 and MAE value for different models

5 Discussion

In this section, We discuss the following things.

- By aggregating sales data over time, it identifies patterns of seasonality or specific months where sales tend to peak for different product categories. For example, certain products may experience higher demand during holiday seasons, while others may be more popular during specific weather conditions or cultural events.
- Knowing the best sales month for each category enables businesses to optimize their inventory management and marketing strategies. They can allocate resources more effectively, plan promotional campaigns, and adjust pricing strategies to capitalize on peak demand periods.
- Beauty products and Health & Personal care have December as the best sales month, indicating a higher demand for healthcare and beauty products in the winter, whereas Car and Motorbike are more sold in summer months during May due to the vacation.

6 Conclusion

We conclude that By analyzing sales data over time, we've uncovered patterns of seasonality, revealing peak sales months for different product categories. For instance, beauty and health products are in high demand during December, while car and motorbike sales peak in May, coinciding with summer vacations.

Understanding these trends allows businesses to optimize inventory and marketing strategies. By focusing resources during peak demand periods, they can enhance sales and profitability. Among all models, XGBoost performs best in predicting market sales, providing valuable insights for sellers.

7 Future Scope

In the future, We Integrating additional data sources such as social media trends, competitor pricing provide a more comprehensive understanding of market dynamics and further improve sales predictions. Utilizing customer data and behavior analysis to provide personalized product recommendations can improve customer satisfaction and increase sales. Additionally, we aim to implement automatic report generation for brand performance analysis. Moreover, applying additional techniques such as neural networks, deep learning, and reinforcement learning can help achieve better accuracy in sales predictions.

References

[1] M. Sajawal, S. Usman, H. Alshaikh, A. Hayat, and M. U. Ashraf, "Predictive analysis of retail sales forecasting using machine learning techniques," vol. 6, pp. 23–33, 02 2023.

[2] "Documentation - Streamlit." https://docs.streamlit.io/.

[3] "Web Scraping - Python." https://www.geeksforgeeks.org/python-web-scraping-tutorial/.

[4] "XGBoost." https://www.geeksforgeeks.org/xgboost/.

[1], [2], [3], [4]