Project on Market Maven:

AI-Driven Sales Forecasting for Smarter Supermarkets

Created by R J N Hemalatha
Guided by Dr. N Jagan Mohan

MARKET MAVEN : (A Diffuser of Marketplace Information)

Market maven describe an investor who is "in-the-know", meaning that they are well-versed on the current state of the market and privy to information.
The term market maven usually refers to an individual who is a market participant with a great deal of knowledge and connections, thus having a trusted opinion on market events or the likelihood of success of a particular investment or speculation.
Famous market mavens include investors like Warren Buffett, John Bogle, and George Soros. [1]

- **Possession and Provision of Market Information :**
  (a) re- ported early awareness of new products across product categories
  (b) aware- ness of specific new brands within several product categories.
  (c) readership of Con- sumer Reports
  (d) use of diverse sources in acquiring market information.
  (e) enjoyment of shopping
  (f) attention. [2]

AI-Driven Sales Forecasting for Smarter Supermarkets :

➢ **Sales forecasting :**

A sales forecasting is the process of predicting sales over a defined timeframe (month or year) in the future by estimating sales and marketing efforts and a calculated predication of the revenue in future.

Elements of Sales Forecasting : Predict sales revenue, Estimate marketing efforts, Diagnose product issues, Plan product launches or releases, Streamline operations, Streamline hiring [3]

➢ **Understanding Sales Forecasting & How AI Can Improve the Process :**

- Sales forecasting is the process of predicting future sales based on various factors, including past performance, market trends, and economic conditions.
- Now, Artificial Intelligence (AI) enhances this process by providing advanced data analytics and predictive modeling.
- By integrating AI into sales forecasting, businesses like yours can achieve more precise predictions, optimize their operations, and increase their responsiveness to market changes. [4]

➢ **Guide to More Profitable Pipelines :**

- 6 Ways AI Transforms Sales Forecasting[5]
- 4 AI Technologies Revolutionizing Sales Forecasts[5]

➢ **AI Technologies :**

AI's transformative power in sales forecasting hinges on its ability to make an accurate prediction. This is achieved by leveraging various sophisticated technologies that analyze data and foresee future sales trends.

- **Regression :** Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between a dependent variable and one or more independent variables. [6]
- **Classification :** Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data. [7]
- **Clustering :** Clustering or cluster analysis is an unsupervised learning method used in machine learning and data analysis that organizes your data so that data points in the same group (or cluster) are more similar to each other than to those in other groups. Clustering helps to make sense of large and complex data sets by uncovering patterns and trends or making predictions on unlabeled data. [8]
- **Time Series Forecasting :** Time series forecasting means to forecast or to predict the future value over a period of time. It entails developing models based on previous data and applying them to make observations and guide future strategic decisions. [9]

DATASET OVERVIEW :

https://docs.google.com/spreadsheets/d/1p7HGEBg1MfwDydCAZf_SFc-rOZyTSme6RwfapKVaPhc/edit?usp=sharing

Row ID => Unique ID for each row.
Order ID => Unique Order ID for each Customer.
Order Date => Order Date of the product.
Ship Date => Shipping Date of the Product.
Ship Mode=> Shipping Mode specified by the Customer.
Customer ID => Unique ID to identify each Customer.
Customer Name => Name of the Customer.
Segment => The segment where the Customer belongs.

Country => Country of residence of the Customer.
City => City of residence of the Customer.
State => State of residence of the Customer.
Postal Code => Postal Code of every Customer.
Region => Region where the Customer belong.
Product ID => Unique ID of the Product.
Category => Category of the product ordered.
Sub-Category => Sub-Category of the product ordered.
Product Name => Name of the Product
Sales => Sales of the Product.
Quantity => Quantity of the Product.
Discount => Discount provided.
Profit => Profit/Loss incurred.

The main objective of the project are:

This project focuses on predicting supermarket profits and minimizing losses by using machine learning techniques on historical sales data. We aim to build a model that can forecast profits and losses based on a variety of input features such as sales percentage, discounts, offers, seasonality, and more.

Methodology:

Libraries and Modules Imported

The following libraries and modules were imported to facilitate data processing and machine learning tasks:

- **pandas**: For data manipulation and reading the dataset.
- **numpy**: For numerical operations.
- **matplotlib.pyplot** and **seaborn**: For data visualization.
- **sklearn.model_selection**: For splitting the dataset into training and testing sets.
- **sklearn.preprocessing**: For Label Encoding and Scaling numerical features.
- **sklearn.linear_model**: For implementing the Linear Regression model.
- **sklearn.ensemble**: For Random Forest and Gradient Boosting Regressor models.
- **sklearn.tree**: For Decision Tree Regressor.

- **sklearn.svm**: For Support Vector Regressor and LinearSVR.
- **sklearn.metrics**: For evaluating model performance (MSE, MAE, R^2).

Data Loading

The dataset is loaded from a CSV file containing sales data using pandas.read_csv():

- **Dataset Details**: The dataset contains various attributes related to the sales data, such as sales amount, discount, product category, region, etc.

Data Preprocessing

1. Selecting Relevant Columns

The dataset initially contains several columns. For the purpose of the analysis, only the relevant columns were selected:

These selected columns represent:

- Sales: The sales amount.
- Discount: The discount applied to sales.
- Quantity: The number of items sold.
- Segment: Customer segment (e.g., Consumer, Corporate, Home Office).
- Region: Geographical region where the sale occurred.
- Category: Product category.
- Profit: Profit from the sale.

2. Handling Missing Values

Checks for any missing values in the dataset and handles them accordingly by dropping rows with missing values

- If missing values are detected, rows containing missing values are removed using df.dropna().

Feature Engineering

1. Encoding Categorical Variables

To prepare categorical variables for machine learning models, **Label Encoding** is applied. Label Encoding converts each category into a numeric value.

- **Segment**: Encodes customer segments (e.g., Consumer, Corporate, Home Office).
- **Region**: Encodes different geographical regions.
- **Category**: Encodes product categories.

2. Scaling Numerical Features

Numerical features such as Sales, Discount, and Quantity are scaled to ensure they have similar ranges for better model performance. **Standard Scaling** is applied to normalize these features

- StandardScaler scales each feature to have a mean of 0 and a standard deviation of 1.

3. Creating New Features

A new feature, **Profit Margin**, is created by calculating the ratio of Profit to Sales

- **Profit Margin** provides insights into the profitability of each sale relative to the sales amount.

Exploratory Data Analysis (EDA)

The process of visualizing and analyzing data to understand its distribution, relationships, and key trends.

1. Univariate Analysis

Univariate analysis involves analyzing individual columns to understand their distribution and characteristics.

- **Distribution of Sales:** A histogram with a Kernel Density Estimate (KDE) is plotted for the Sales column to visualize its distribution
- Countplot for Categories: A count plot is created for the Category column to visualize the distribution of product categories

2. Bivariate Analysis

Bivariate analysis explores the relationships between two variables to uncover any patterns or correlations.

- **Sales vs. Profit:** A scatter plot is used to show the relationship between Sales and Profit, with points colored by Category

3. Correlation Analysis

Correlation analysis is performed to study the relationships between numerical features in the dataset.

- **Correlation Heatmap:** A heatmap of the correlation matrix is plotted to identify strong or weak correlations between numerical variables

The heatmap displays the correlation values, helping to identify any variables that are strongly related to each other, which can be useful for feature selection**.**

4. Category-Level Insights

- **Profit by Category:** The total profit is calculated for each product category and visualized using a bar chart

5. Geographic Insights

- **Sales Share by Region:** A pie chart is used to visualize the share of total sales by different regions

6. Anomaly Detection

- **Outlier Analysis (Sales by Category):** A box plot is generated for Sales by Category to identify any outliers.

Model Training and Evaluation

1. Splitting the Data

The dataset is split into features (X) and the target variable (y, which is Profit). The data is then further divided into training and testing sets.

2. Model Training

Multiple machine learning models are initialized and trained using the training data.

3. Model Evaluation

Each model is evaluated using three metrics:

- **Mean Squared Error (MSE)**
- **Mean Absolute Error (MAE)**
- **R^2 Score**

4. Grid Search for Hyperparameter Tuning

For the Gradient Boosting Regressor model, a grid search is performed to find the best hyperparameters

5. Model Evaluation (Post-Tuning)

After tuning the Gradient Boosting model, the performance is evaluated on the test set

6. Predictions and Feature Importance

The actual vs predicted values are plotted to visualize the performance of the model

7. Saving Results

Finally, after optimizing the model, the updated dataset is saved:

python

```
df.to_csv('optimized_sales_data.csv', index=False)
```

Conclusion

In this section, After preprocessing and feature engineering, several machine learning models were trained to predict sales profits, with the Gradient Boosting Regressor being fine-tuned using Grid Search for better performance. The model's results were evaluated using key metrics, and predictions were visualized for further analysis.

REFERENCES :

[1]https://www.investopedia.com/terms/m/marketmaven.asp
[2]https://www.researchgate.net/publication/235361395_The_Market_Maven_A_Diffuser_of_Marketplace_Information
[3] https://goodmeetings.ai/blog/sales-forecasting/
[4]https://goodmeetings.ai/blog/ai-in-sales-forecasting-uses-benefits-best-practices/
[5] https://www.scratchpad.com/blog/ai-sales-forecasting
[6]https://www.investopedia.com/terms/r/regression.asp,https://corporatefinanceinstitute.com/resources/data-science/regression
[7] https://www.datacamp.com/blog/classification-machine-learning
[8]https://www.mathworks.com/discovery/clustering.html,https://developers.google.com/machine-learning/clustering/overview,https://www.geeksforgeeks.org/clustering-in-machine-learning/
[9] https://github.com/MRYingLEE/Time-series-Preprocessing-Studio-in-Jupyter, https://www.influxdata.com/time-series-forecasting-methods/
[10]**https://support.google.com/a/users/answer/9604139?hl=en#zippy=%2Clearn-how (data cleaning techniques)**