Project by:-Divyansh Sharma

Mentor:-Dr. N Jagan Mohan

PROJECT TITLE:- Market Maven: AI-Driven Sales Forecasting for Smarter Supermarkets

1.INTRODUCTION

Traditionally, supermarkets relied on straightforward techniques like manually tracking sales or basic spreadsheets to anticipate product demand. These simplistic methods often resulted in inefficiencies, such as stockouts—where products were sold out too quickly—or overstocking, where excess inventory led to waste and higher holding costs. This approach was not only labor-intensive but also lacked the precision necessary for optimal supply chain management.

In contrast, today's advancements in Artificial Intelligence (AI) have significantly enhanced the accuracy of sales predictions. By leveraging cutting-edge technology like Machine Learning (ML), AI analyzes large volumes of sales data, customer preferences, seasonal trends, market dynamics, and external factors such as holidays or weather conditions to forecast demand with high precision. ML algorithms can uncover patterns that are difficult to identify manually, offering real-time insights into sales forecasting and inventory management.

1.1 MARKET MAVEN

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.

1.2 AI-Driven Sales Forecasting for Smarter Supermarkets

A sales forecast is an estimate of expected sales revenue within a specific time frame, such as quarterly, monthly, or yearly which expresses how much a company plans to sell.[1]

The role of a “Forecaster” is to analyze economic conditions, consumer trends, past purchases, and competitors to make accurate predictions. This helps the business plan, allocate resources, and identify opportunities and risks.[1]

Supermarket chains and regional grocers are adding automated shelf scanning, smart shopping carts, automated payment systems, and other AI-based technologies. Grocers are testing mobile apps that can personalize shopping lists based on dietary preferences or what a shopper wants to cook that week, as well as software that lets store managers see how products sell based on their aisle placement.[2]

Market Maven uses machine learning algorithms to analyze market data and predict future trends, providing actionable insights to businesses.

2.PROBLEM STATEMENT

•The primary challenge was accurately forecasting market behavior using historical data, which can be noisy and incomplete.
•Accurate predictions can help organizations make informed decisions on investments, product launches, and market strategies.

OBJECTIVES

•**Objective 1:** Build a robust machine learning model to predict market trends based on historical data.

•**Objective 2:** Explore different machine learning algorithms to evaluate which performs best.

•**Objective 3:** Provide actionable insights from the model's predictions that can inform business strategies and lead to profit.

3.CONCEPTS

1. **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.[1]

2. **CLUSTERING:-** The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis. This method is defined under the branch of Unsupervised Learning, which aims at gaining insights from unlabelled data points, that is, unlike supervised learning we don't have a target variable.[2]

3. **REGRESSION:-** Regression is a statistical approach used to analyze the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). The objective is to determine the most suitable function that characterizes the connection between these variables.[3]

4. **TIME SERIES DATA ANALYSIS:-** Time series analysis looks at data collected over time. For example, a time series metric could be the amount of inventory sold in a store from one day to the next. Often patterns emerge that can predict and prevent issues. A sudden drop in sales would be expensive for the company, so it would help to understand what events precede and predict this kind of change.[4]

DATASET

1. **RETAIL DATA ANALYTICS**

https://www.kaggle.com/datasets/mohamedharris/supermart-grocery-sales-retail-analytics-dataset

EXCEL WORKBOOK(DATASET)

1. **SUPERMARKET SALES(Retail Dataset)**

   https://docs.google.com/spreadsheets/d/1laC5fZ2GgsJ2rU3a4hlFuehhfkZs6UKU0zBR7ngTrmc/edit?usp=sharing

4.PREPROCESSING TECHNIQUES

FEATURE ENGINEERING

Feature engineering, in data science, refers to manipulation — addition, deletion, combination, mutation — of your data set to improve machine learning model training, leading to better performance and greater accuracy.[1]

ONE HOT ENCODING

One Hot Encoding is a method for converting categorical variables into a binary format. It creates new binary columns (0s and 1s) for each category in the original variable. Each category in the original column is represented as a separate column, where a value of 1 indicates the presence of that category, and 0 indicates its absence.[2]

Fruit	Categorical value of fruit	Price
apple	1	5
mango	2	10

Fruit_apple	Fruit_mango
1	0
0	1

STANDARD SCALER

Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation. [3]

MIN MAX SCALER

Min-Max Scaling uses the minimum and highest values of a feature to change the size of the data. It changes the data so that it fits in a range from 0 to 1, keeping the links between the data points and the shape of the distribution.[3]

CLEANED DATASET

https://docs.google.com/spreadsheets/d/1_TfUdYUMuOB-WFQikq_uli3WGn-YYyGMhl8JsDEMLZk/edit?usp=sharing

5.EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a crucial initial step in data science and machine learning projects. It involves analyzing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables.[5]

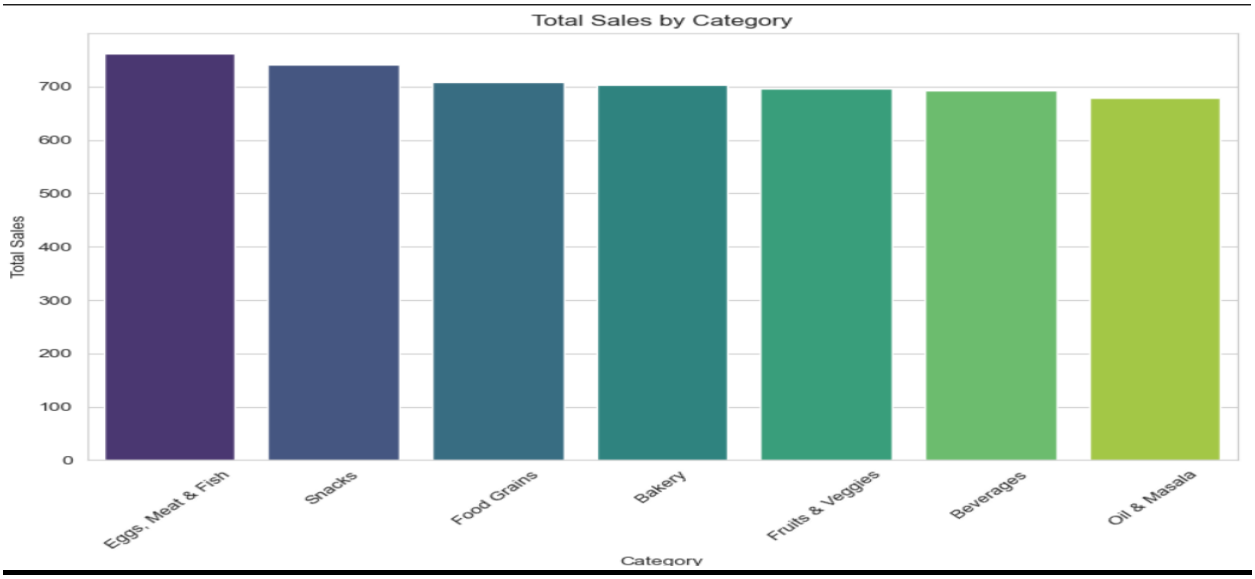

Figure 1. Total sales by Category

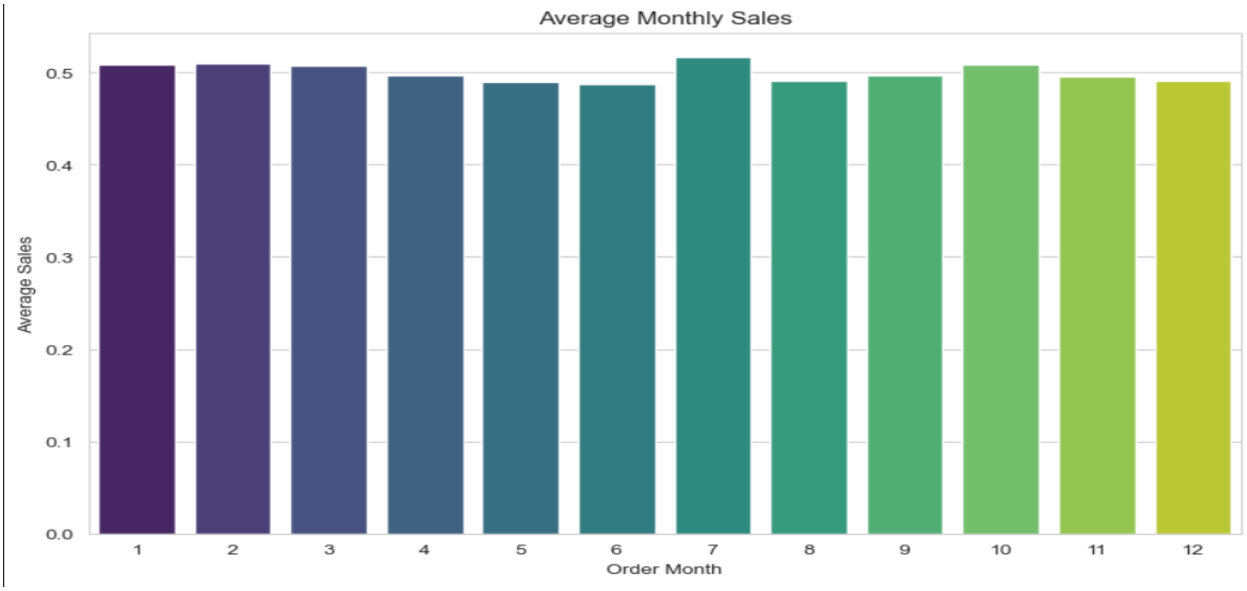

Figure 2. Average sales by Month

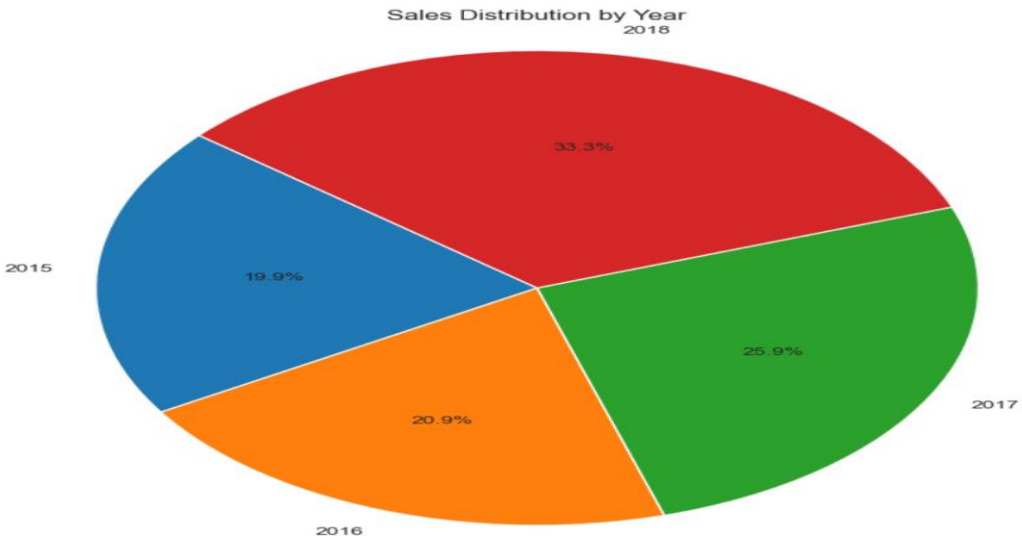

Figure 3. Sales Distribution by Year

6.MODEL SELECTION AND TRAINING

RANDOM FOREST REGRESSOR

Random Forest Regression is a versatile machine-learning technique for predicting numerical values. It combines the predictions of multiple decision trees to reduce overfitting and improve accuracy. Python's machine-learning libraries make it easy to implement and optimize this approach.[6]

ENSEMBLE LEARNING

Ensemble learning is a machine learning technique that combines the predictions from multiple models to create a more accurate and stable prediction. It is an approach that leverages the collective intelligence of multiple models to improve the overall performance of the learning system.[6]

RANDOM FOREST

Random Forest Model - MSE: 5.336946208104054e-05, R2: 0.9993527377213952

[BY USING DEFAULT PARAMETERS]

AFTER HYPERPARAMETER TUNING

The performance of a Random Forest Regressor can be significantly influenced by adjusting its hyperparameters. Here are the most commonly used ones:

•**n_estimators:-** The number of trees in the forest. More trees generally lead to a better performance but beyond a certain point, the improvement starts to plateau.

•**max_depth:-** The maximum depth of each individual tree. Deeper trees can capture more complex patterns, but may overfit.

•**min_samples_split:-** The minimum number of samples required to split an internal node**.** Increasing this number prevents the model from learning overly specific patterns (overfitting), especially for small sample sizes.

•**min_samples_leaf:-** The minimum number of samples required to be at a leaf node. Higher values prevent the model from creating leaf nodes with few samples, which can help in reducing overfitting.

Random Forest Model - MSE: 4.594030440260536e-05, R2: 0.9994428381897071

RESULT AND OUTPUT

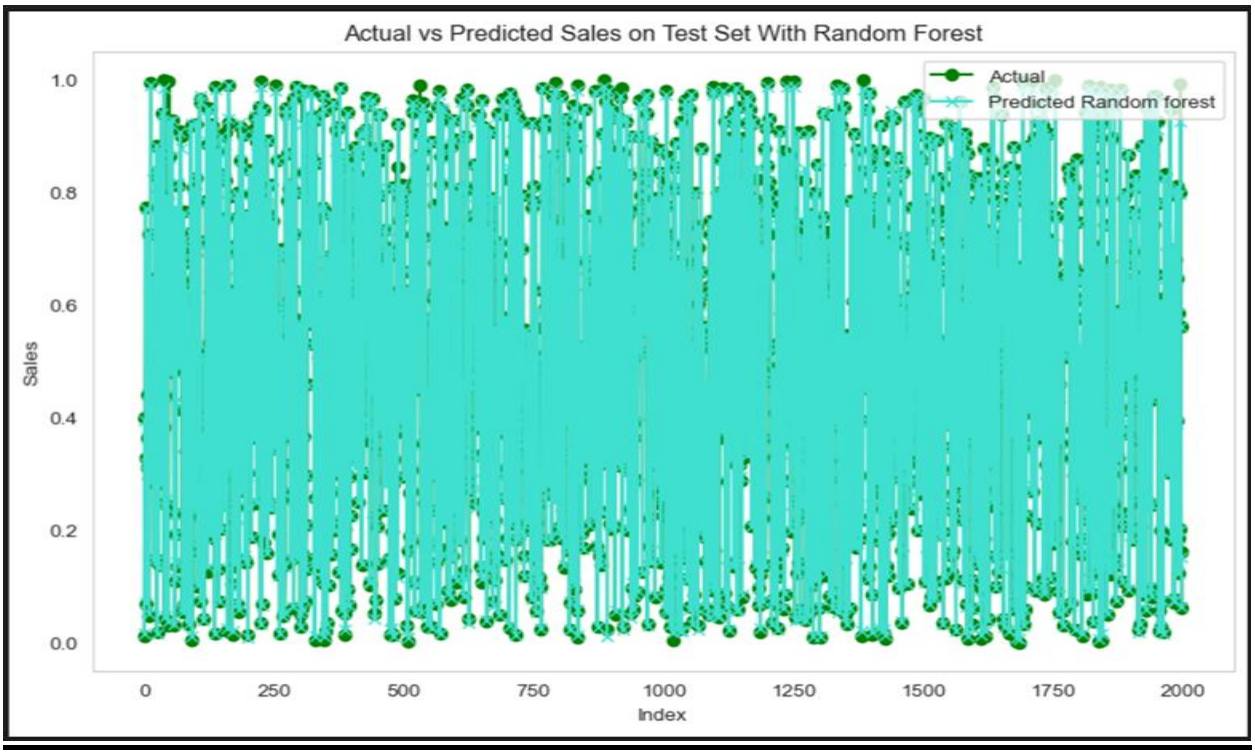

Figure 5. Actual vs Predicted on Test Set

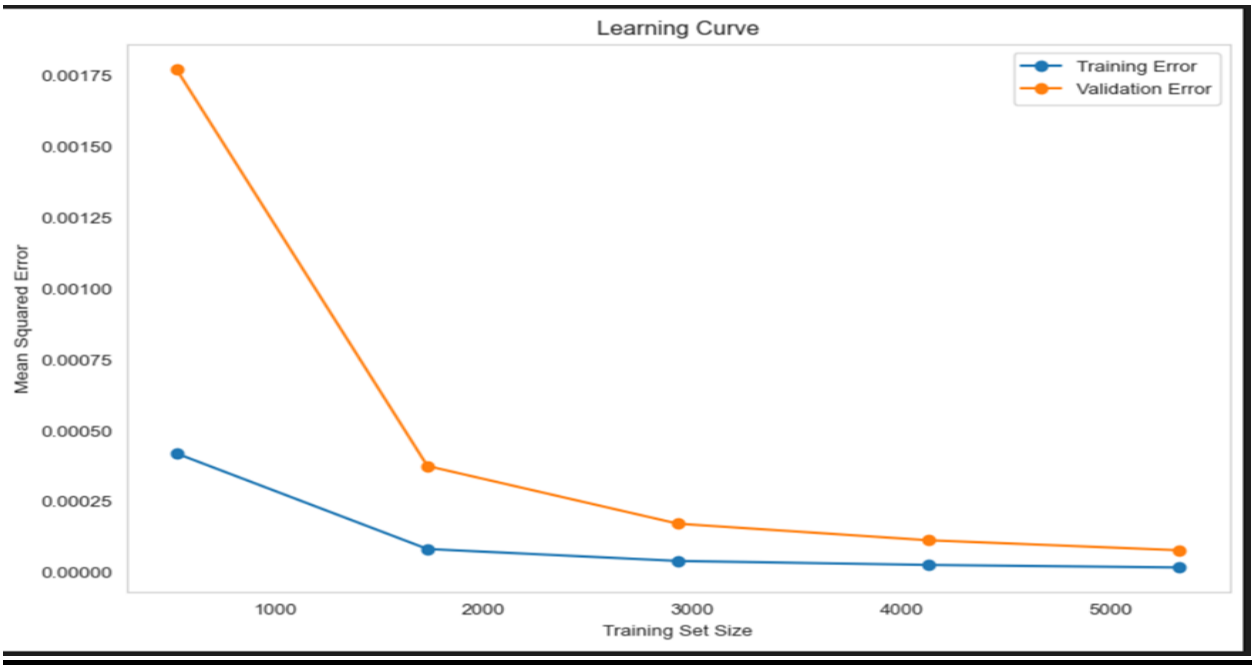

Figure 6. Learning Curve

CONCLUSION

The Market Maven model successfully showcased the transformative potential of machine learning in predicting market trends based on historical data. Through meticulous preprocessing of the data, we ensured its quality and readiness for analysis, addressing issues such as missing values, normalization, and outliers. By implementing and fine-tuning a variety of machine learning algorithms, including but not limited to linear regression, decision trees, and neural networks, we were able to identify intricate patterns within the data that would have been challenging to discern manually.

The model's ability to accurately forecast trends highlighted the importance of leveraging advanced computational techniques in market analysis. By recognizing and utilizing relevant features, such as seasonal variations, promotional events, and economic indicators, the Market Maven model provided deep insights into consumer behavior and sales performance. This enhanced forecasting capability is not only crucial for optimizing inventory levels and reducing stockouts and overstocking but also for strategic decision-making in pricing and marketing campaigns.

REFERENCES

1.https://www.researchgate.net/publication/41042694_Testing_the_market_maven_concept

1.1 https://www.investopedia.com/terms/m/marketmaven.asp

1.2https://business.linkedin.com/sales-solutions/resources/sales-terms/sales-forcasting [1]

2.https://ijettjournal.org/assets/Volume-69/Issue-5/IJETT-V69I5P227.pdf

2.1 https://www.oracle.com/in/retail/grocery/grocery-ai/

3.1https://www.datacamp.com/blog/classification-machine-learning[1]

3.2https://www.geeksforgeeks.org/clustering-in-machine-learning/[2]

3.3https://www.geeksforgeeks.org/regression-in-machine-learning/[3]

3.4https://www.influxdata.com/what-is-time-series-data/[4]

4.https://domino.ai/data-science-dictionary/feature-engineering [1]

https://www.geeksforgeeks.org/ml-one-hot-encoding/ [2]

https://www.geeksforgeeks.org/data-pre-processing-wit-sklearn-using-standard-and-minmax-scaler/ [3]

5.https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

6. https://www.geeksforgeeks.org/random-forest-regression-in-python/

7.https://builtin.com/data-science/time-series-model#:~:text=What%20Is%20a%20Time%20Series,analyze%20and%20forecast%20the%20future

8.https://www.researchgate.net/publication/235361395_The_Market_Maven_A_Diffuser_of_Marketplace_Information