What Drives Supplement Sales? A Predictive Modeling Approach

Introduction to the Problem and Data Set

Every industry operates in a competitive and fast-paced environment, and the supplement market is no exception. As such, the ability to accurately forecast product demand is highly advantageous for managers, enabling them to optimize inventory, maximize revenue, and minimize waste. This project aims to build predictive models to estimate the number of units a supplement product will sell in a given week using real-world sales data.

The dataset contains weekly records of various supplement products sold across different platforms (e.g., Amazon, Walmart, iHerb), locations (e.g., USA, UK, Canada), and categories (e.g., Performance, Protein, Vitamins). Key features include product name, category, discount, price, platform, location, and temporal variables such as week and year.

The target variable in this project is Units Sold, a continuous measure suitable for regression analysis. The objective is to assess how well different supervised learning models—such as Linear Regression, Decision Tree, K-Nearest Neighbors (KNN), and Random Forest—can predict weekly unit sales using these features. Performance will be evaluated using Mean Squared Error (MSE) on both training and test sets, and model interpretability will be enhanced through permutation-based feature importance.
By understanding the drivers of weekly sales volume, this project provides insights that can support more data-driven decision-making in marketing, pricing, and supply chain management.

Initial Exploratory Data Analysis

Before building any predictive models, I conducted a comprehensive exploratory data analysis (EDA) to better understand the underlying structure of the dataset and uncover preliminary relationships between features and the target variable, Units Sold. The aim of this analysis was to guide feature selection, identify potential predictors, and detect any anomalies that might affect model performance. I visualized both individual feature distributions and bivariate relationships to assess how each variable interacts with the target.

1. Distribution of Units Sold

The histogram of Units Sold reveals that weekly sales per product are approximately normally distributed, with a clear peak around 150 units. Most values fall within the 130 to 170 unit range, indicating low variance across observations. This consistency suggests a well-behaved continuous target, making the variable appropriate for regression modeling. The distribution also reflects real-world retail behavior, where most products tend to sell at stable volumes weekly, barring promotions or seasonal shifts.

2. Price vs. Units Sold

A scatterplot of Price against Units Sold shows a widely dispersed and non-linear relationship. Products priced both low and high appear throughout the entire sales range, suggesting that price alone is not a reliable predictor of units sold. However, this lack of correlation does not rule out its usefulness—price could still contribute to the model when considered in combination with other features, such as discount, category, or platform. In particular, expensive products may rely more on brand loyalty or perceived quality, while lower-priced items may compete on quantity or accessibility.

3. Discount vs. Units Sold

The boxplot of Discount versus Units Sold reveals a positive association between discount percentage and sales volume. Products offered with higher discounts tend to sell more units, which aligns with standard consumer behavior in the retail space. The presence of this relationship suggests that Discount is likely to be one of the most important predictors in the modeling process. It also reinforces the idea that promotional strategies can significantly influence short-term demand, making it a valuable feature for forecasting purposes.

4. Units Sold by Platform

To assess performance across sales channels, I examined Units Sold across different platforms using a boxplot. The platforms—iHerb, Amazon, and Walmart—display similar median sales, but iHerb exhibits a wider interquartile range and more extreme values, indicating greater variability in weekly sales. This variation could reflect differences in product assortment, customer demographics, or promotional intensity across platforms. As a result, Platform is expected to offer useful segmentation power within the model, especially when combined with Product Name or Location.

In addition to feature-specific analysis, I also reviewed broader trends in sales distribution:

- Top-Selling Products and Categories
  An analysis of average units sold by product and category revealed that items like Biotin, Ashwagandha, and Fish Oil consistently ranked among the highest-selling. Similarly, categories such as Herbal, Omega, and Performance demonstrated strong performance overall. This confirms that Product Name and Category are critical features for modeling and should be weighted heavily in feature importance analysis.

- Weekly and Yearly Sales Trends
  A line chart of average weekly sales across the year displayed modest but noticeable fluctuations, suggesting subtle seasonality in consumer demand. Although overall variation was low, this supports the inclusion of Week and Year as temporal features to help capture cyclical patterns and long-term sales behavior.

- Sales Distribution by Product
  A violin plot of units sold by product, with average values labeled, highlighted differences in both average performance and sales consistency. While some products showed stable weekly demand, others exhibited greater variability, reinforcing the idea that modeling should account for product-specific dynamics.

Together, these exploratory insights shaped my understanding of which features to prioritize and provided a strong foundation for selecting appropriate regression models. With this context established, I proceeded to implement and compare several machine learning models to evaluate how well they could predict weekly supplement sales using the identified patterns.

Modeling & Interpretations

To predict the number of units a supplement product will sell in a given week, I implemented and compared several regression models. The goal was to evaluate which model most accurately captures the relationships between the input features (such as product type, discount, platform, and time) and weekly sales volume.

All models were trained using an 80/20 train-test split to ensure fair evaluation on unseen data. Model performance was measured using Mean Squared Error (MSE), and feature importance was assessed using permutation importance where applicable.

The models explored include:

- Baseline model (predicting the mean)

- Multiple Linear Regression

- Decision Tree Regressor

- K-Nearest Neighbors (KNN)

- Random Forest Regressor

Each model was evaluated for both training and testing MSE, and further refined through hyperparameter tuning where necessary (e.g., using GridSearchCV for KNN and Random Forest).

Baseline Model

To establish a baseline for comparison, I calculated the mean number of units sold from the training data and used this constant value to predict all outcomes in the test set, resulting in a baseline MSE of 159.69.

Multiple Regression Model

My multiple regression model performed only slightly better than the baseline. While the training data had a lower mean squared error than the baseline, the testing data did not show significant improvement. This suggests that although the model was able to capture some variance in the training set using features like product name, discount, and platform, it struggled to generalize effectively on new, unseen data.

This outcome is likely due to the linear model's limited ability to capture complex relationships among the features. The baseline model simply predicted the mean number of units sold from the training data, resulting in a test MSE of 159.69. In comparison, the multiple regression model achieved a training MSE of 150.87 and a test MSE of 161.41, which is only marginally worse than baseline.

The most important predictors in the linear model were specific product names (e.g., Biotin, Whey Protein) and platforms (e.g., Walmart, iHerb), which showed the highest coefficients, indicating that product identity and distribution channel are meaningful drivers of weekly supplement sales, even in a simple linear framework.

Decision Tree Regression Model

To find the best tree depth, I trained decision tree models with depths ranging from 1 to 20 and plotted their training and testing mean squared errors. The optimal depth was chosen by identifying the point where testing error was lowest before it started increasing, helping avoid overfitting while still capturing meaningful patterns and as shown on the graph, this was around 4 and 6 so I went with 5.

The decision tree outperformed the multiple regression model on the training set and performed similarly on the test set, showing that it was able to learn more flexible patterns from the data.

While its test MSE was slightly above the baseline, the gap was small, indicating acceptable generalization and no major overfitting.

The visualized decision tree also revealed how the model splits based on standardized features like Price, Discount, Product Name, and Location. This made the decision-making process interpretable, especially when compared to the black-box behavior of more complex models.

Permutation importance further supported this by identifying Product Name like Vitamin C, and also Price as the top predictors. These aligned closely with what was found in the linear model, but the decision tree's structure allowed it to model non-linear splits and combinations more effectively.

In contrast, the multiple regression model relied solely on linear relationships which although provided a more interpretable summary through coefficients, it was limited in flexibility and slightly underperformed compared to the tree in training performance.
Finally, compared to the baseline model, which simply predicted the average units sold (MSE = 159.69), the decision tree demonstrated marginal improvement, reinforcing its usefulness in identifying feature interactions without significant overfitting.

K-Nearest Neighbors Regression Model
To explore how neighborhood-based models could capture localized sales behavior, I trained a K-Nearest Neighbors Regressor and used GridSearchCV to find the optimal value for k. The grid search evaluated values of k from 3 to 20, ultimately selecting k = 20 as the best-performing configuration.

Training MSE: 144.99

Testing MSE: 164.88

Baseline MSE: 159.69

This model slightly outperformed the baseline on training data but only marginally improved on the test set. While KNN is a non-parametric model capable of adapting to local trends in the data, the relatively high test error suggests that the model may be too sensitive to noise, particularly in datasets where the majority of products sell within a tight range each week.

Permutation importance revealed that the most influential features included Discount, Category_Performance, and products like Iron Supplement, Creatine, and Whey Protein. This is consistent with business intuition — product type and discount strategy significantly impact weekly sales.

Compared to the multiple regression model, KNN had a lower training error but similar test error, indicating better training fit but slightly more overfitting. It also provided more flexibility in capturing nonlinear relationships between features and sales, but at the cost of slightly worse generalization.

Random Forest Regression Model

To build a more robust model capable of capturing complex feature interactions, I implemented a Random Forest Regressor and optimized it using GridSearchCV. The search evaluated combinations of n_estimators and max_depth, ultimately selecting 200 trees and a maximum depth of 3 as the best configuration.

Training MSE: 149.87

Testing MSE: 160.31

Baseline MSE: 159.69

The Random Forest model slightly outperformed the multiple regression, Decision Tree and KNN models on the test set, while maintaining a relatively low training MSE. This suggests it was able to capture meaningful patterns in the data without overfitting. The shallow depth (3) likely contributed to this by enforcing model simplicity and reducing variance.

Permutation importance revealed that the top contributing features included Year, Product Name like Pre-Workout, Whey Protein, and Category_Fat Burner. These findings align closely with earlier models and also reinforced the importance of product identity, temporal variables, and category in predicting weekly unit sales from our initial exploratory data analyses.

Conclusion & Next Steps:

This project investigated the use of supervised regression models to predict weekly supplement sales, with the objective of identifying the most influential predictors of demand and evaluating which algorithms perform best on unseen data. After training and comparing a baseline model, multiple linear regression, decision tree, K-nearest neighbors (KNN), and a tuned Random Forest, the Random Forest Regressor emerged as the most effective model overall.

Summary of Findings:

Several key insights emerged from this analysis:

- Ensemble Models Generalize Best
  The Random Forest Regressor slightly outperformed all other models on the test set (MSE = 160.31), indicating its strength in capturing nonlinear feature interactions while minimizing overfitting. Its consistent performance across training and testing data made it the most reliable model for this task.

- Product and Category Features Were Most Predictive
  Across all models, specific product names (e.g., Biotin, Whey Protein) and product categories (e.g., Performance, Omega) consistently appeared as top predictors. This finding emphasizes the importance of product-level segmentation and targeted forecasting strategies.

- Temporal and Promotional Variables Played a Supporting Role
  Although Week, Year, and Discount were not dominant features, they still contributed meaningful signal—especially in ensemble models—by capturing seasonal variation and price sensitivity.

- Model Performance Trade-offs

  - Linear Regression provided interpretability but limited modeling flexibility.

  - Decision Trees and KNN captured more nuanced relationships but showed signs of overfitting and inconsistent generalization.

  - Random Forest balanced complexity and generalization, producing the lowest test error while maintaining interpretability through feature importance.

Next Steps & Improvements:

To further improve prediction accuracy and enhance model applicability, the following steps are recommended:

- Incorporate External Variables
  Integrating data on promotional campaigns, seasonal trends (e.g., New Year's resolutions), or competitor pricing could provide additional context to improve the model's predictive power.

- Add Lag Features and Rolling Averages
  Including time-dependent features such as prior week's units sold or rolling averages could help capture recent momentum and short-term trends.

- Explore Time Series Models
  While this project focused on standard regression algorithms, future work could apply time series-specific models (e.g., ARIMA, Prophet, LSTM) to more effectively model autocorrelation and long-term seasonality.

- Enable Business Integration
  Deploying the final model in a real-time forecasting dashboard would provide operational value to inventory and sales managers, enabling data-driven planning and replenishment decisions.

In summary, this project provided a clear understanding of the primary drivers behind supplement sales and demonstrated how machine learning can be leveraged to build accurate, interpretable, and actionable forecasting models in a retail environment.