Product Demand Prediction with Machine Learning

Objective:

To build a model that assigns a score to each given sales data and then based on a threshold value, classifies the 'Product in Demand' from the overall sale data of a store, which can then be used to increase the overall sale rate of the store.

Abstract:

To improve a sales team's gains by defining a clear classification within a given set of demand so that the store focuses on the supply chain for the demanded product**.**

To help them focus on the demand with the most priority, thus increasing the overall sales rate and reducing the risk of the sales for non-demand products

Problem Definition and Design Thinking:

To understand the problem statement and create a document on what have we understood and how will we proceed ahead with solving the problem.

Problem Definition:

The problem is to create a machine learning model that forecasts product demand based on historical sales data and external factors. The goal is to help businesses optimize inventory management and production planning to efficiently meet customer needs. This project involves data collection, data pre-processing, feature engineering, model selection, training, and evaluation.

Design Thinking:

1. Data Collection: Collect historical sales data and external factors that influence demand, such as marketing campaigns, holidays, economic indicators, etc.
2. Data Pre-processing: Clean and pre-process the data, handle missing values, and convert categorical features into numerical representations
3. Feature Engineering: Create additional features that capture seasonal patterns, trends, and external influences on product demand.
4. Model Selection: Choose suitable regression algorithms (e.g., Linear Regression, Random Forest, XG Boost) for demand forecasting.
5. Model Training: Train the selected model using the preprocessed data.

6. Evaluation: Evaluate the model's performance using appropriate regression metrics (e.g., Mean Absolute Error, Root Mean Squared Error).

Dataset
Link: https://www.kaggle.com/datasets/chakradharmattapalli/product-demand-prediction-with-machine-learning

Sale Demand score prediction

The model uses a Logistic regression library as its core to fit in the given dataset. This library is referred by the model to observe the underlying pattern within the given dataset. In a brief it tries to apply and fit o the logistic regression formula,

$$Y = b0 + b1x1 + b2x2 +b2x2 +...+bnxn$$

Where,

b0...bn denotes coefficients of the input features and

x1...xn denotes the input features.

But this approach uses a transformed version of this formula to specifically find a

probabilistic value for a positive outcome such as,

$$P(Y = 1/x) = 1 / (1 + exp(-(b0 + b1x1 + b2x2 +b2x2 +...+bnxn)))$$

The above formula can be described as the probability of the outcome being 1 given the x

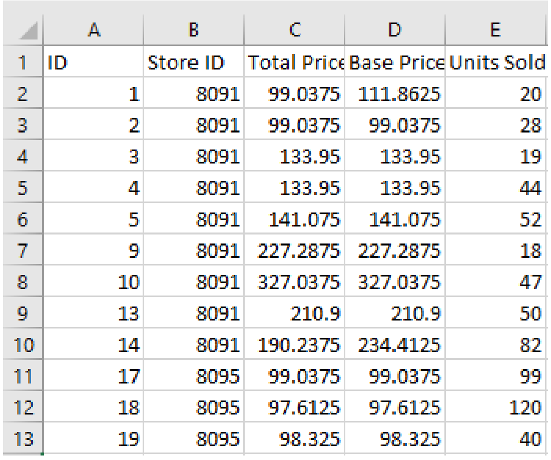
features.

Where,

b0...bn denotes coefficients of the input features and

x1...xn denotes the input features

Insights from the dataset:



ID	Store ID	Total Price	Base Price	Units Sold
1	8091	99.0375	111.8625	20
2	8091	99.0375	99.0375	28
3	8091	133.95	133.95	19
4	8091	133.95	133.95	44
5	8091	141.075	141.075	52
9	8091	227.2875	227.2875	18
10	8091	327.0375	327.0375	47
13	8091	210.9	210.9	50
14	8091	190.2375	234.4125	82
17	8095	99.0375	99.0375	99
18	8095	97.6125	97.6125	120
19	8095	98.325	98.325	40

Each column represents different aspects of the data:

1. ID: This column typically represents a unique identifier for each data entry or record. It's often used to distinguish one data point from another.

2. Store ID: This column likely represents the identifier of the store where the sales occurred. Each store may have a unique ID, which can be used to analyze sales performance across different stores.

3. Total Price: This column likely represents the total revenue generated from the sales of a product. It's calculated by multiplying the "Base Price" by the "Units Sold" for each transaction.

4. Base Price: This column likely represents the original or base price of the product before any discounts or promotions. It's used to calculate the "Total Price."

5. Units Sold: This column represents the number of units of the product sold in each transaction. It's a key factor in understanding product demand and sales.

THE PROCESS INVOLVED TO BUILD AND USE THE MODEL

CREATION:

Step 1: The '.csv' files are loaded into the model.

Step 2: The dataset is pre-processed by handling any missing values and removing the least

relevant features

Step 3: The dataset is classified as Y for the prediction and X for the features.

Step 4: The dataset is then split into training and testing sections.

Step 5: The parameters for randomness are defined.

Step 6: The dataset is observed for underlying patterns.

Step 7: The dataset is then fit into the model.

TUNING:

Step 8: A test metric is used to evaluate the model's accuracy.

Step 9: Based on the accuracy the model's hyper-parameters are tuned.

Step 10: Step 9 is repeated until the model reaches the desired accuracy.

TESTING:

Step 11: A new dataset is loaded.

Step 12: The dataset is then pre-processed and cleaned.

Step 13: The dataset is then loaded into the model for prediction.

LIBARIES/MODULES USED:

- PANDAS for dataset manipulation,
- NUMPY for encoded variables manipulation,
- SEABORN for visualization,

- SCI-KIT LEARN for the core

MODEL SELECTION

- train, test, split for the division within the dataset

PRE-PROCESSING

- Label encoder for encoding strings and categorical variables

LINEAR MODEL

- Logistic Regression for the core

METRICS

- confusion matrix,

- accuracy score and

- classification report for the evaluation.

In conclusion, businesses must recognize the pivotal role demand generation plays in their success. These strategies to estimate demand score can ensure a steady understanding and relationship between sales, greater brand recognition, and long-term customer loyalty.