[Team Career Craaft]

Team ID : E63943

DECODE X 2025

Case-05

Product Demand Forecasting

1. Introduction & Business Problems

1.1 Background

In the fast-moving consumer goods (FMCG) industry, precise demand forecasting is crucial for maintaining an optimal supply chain, minimizing wastage, and ensuring product availability. **FreshMart FMCG**, a leading company in the industry, has long relied on manual forecasting methods using Microsoft Excel. While these traditional methods incorporate recent demand data and promotional schemes, they often fall short in capturing intricate demand patterns and market dynamics.
A major FMCG company's operations generate 3 months of rolling forecasts at location, C&F agent (stockist) level and SKU.

At the center of FreshMart's operations are **Mr. Rohan Malhotra**, the dynamic Chief Operations Officer (COO), and **Ms. Aditi Sharma**, the Head of Supply Chain Management. During one of their quarterly strategy meetings, Rohan voiced his concerns:

"Our current forecasting method is heavily dependent on manual inputs. While our teams use business logic and recent trends to generate forecasts, we lack a scientific approach to leverage historical data patterns effectively."

Aditi agreed, adding:

"The inaccuracies in our forecasts are leading to stockouts for some products and excess inventory for others. We need a systematic, data-driven approach to enhance forecast accuracy."

1.2 Business Problems

FreshMart's forecasting was traditionally performed using recent demand data, business intuition, and promotional schemes. However, this approach suffered from several challenges:

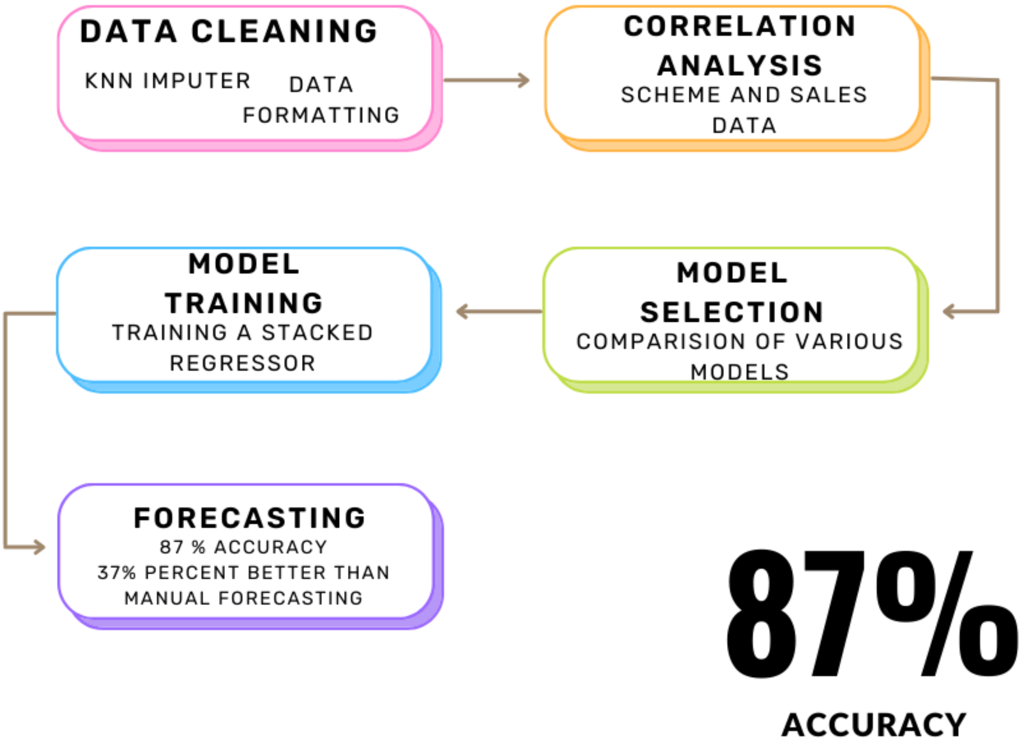
1. **Inaccuracy in Predictions** – The manual forecasts often deviated significantly from actual sales due to human bias and lack of statistical rigor.
2. **Lack of Scalability** – As FreshMart expanded into new locations and product categories, manual forecasting became increasingly unmanageable.
3. **Ignoring Seasonal Trends** – Demand fluctuations due to seasonality, festivals, and promotional schemes were not effectively captured.
4. **Inventory Mismanagement –** Inaccurate forecasts led to either **overstocking**, causing increased holding costs, or **stockouts**, resulting in lost sales opportunities.

With these concerns in mind, Dr. Verma and his team designed a **structured project implementation methodology** to build an advanced demand forecasting model.

The Goal of the case study is to come up with project implementation methodology with forecasting models which gives best demand forecast using the historical training and validating on the hold-out set.

To resolve these inefficiencies and solve this problem FreshMart needs a advanced forecasting model, leveraging historical data and machine learning techniques.

MODEL WORKFLOW



DATA CLEANING

KNN IMPUTER DATA FORMATTING

CORRELATION ANALYSIS

SCHEME AND SALES DATA

MODEL SELECTION

COMPARISION OF VARIOUS MODELS

MODEL TRAINING

TRAINING A STACKED REGRESSOR

FORECASTING

87 % ACCURACY

37% PERCENT BETTER THAN MANUAL FORECASTING

87%

ACCURACY

2. Data Preparation and Availability

2.1 Dataset

The Dataset we got are CPASF.csv , CPMNT.csv and SCHEME DETAILS.xls.
CPASF.csv and CPMNT.csv file contains

- Date: Sales Month
- Location: Branch location (Location of city)
- C&FAgent: Clearing and Forwarding Agent, who is responsible for handling customs clearance and forwarding logistics of products.
- Division: Category of the product
- SKU: Product name
- Sales: Units sold per month
- OpeningStock: Available stock at the beginning of the month
- Forecast: Previously predicted sales volume
- SellingPrice: Per-unit product price (INR)

The third dataset we got was SCHEME DETAILS.xls which contained 6 files in it , which are :-

- **CPASF 2004-2005**
- **CPASF 2005-2006**
- **CPASF 2006-2007**
- **CPMNT 2004-2005**
- **CPMNT 2005-2006**
- **CPMNT 2006-2007**

Each of the file contained five attributes those were :

- Primary date : Date of announcement of scheme or offers
- Product/Pack : Different SKU's
- Scheme Period : Starting and ending date of scheme.
- Scheme % details : Discounts on purchase of min. Number of products.
- Location/Cluster : It tells us where the scheme is active (Eg- ALL INDIA, or a Particular CITY)

Scheme data—Contains sales through discounts, promotional offers, and marketing campaigns.

2.2 Data Preparation

We handled the missing values in files CPASF.csv , CPMNT.csv using KNN nearest neighbour because it preserve data structure(Considers the relationships between similar data points, maintaining the original distribution of the dataset), It fills missing values based on the most similar records, leveraging the natural grouping in the data, it Handles Both Categorical & Numerical Data, it also captures Non-Linear Relationships and there is no assumption required unlike parameter methods .

We also did the same thing in SCHEME DETAILS.xls

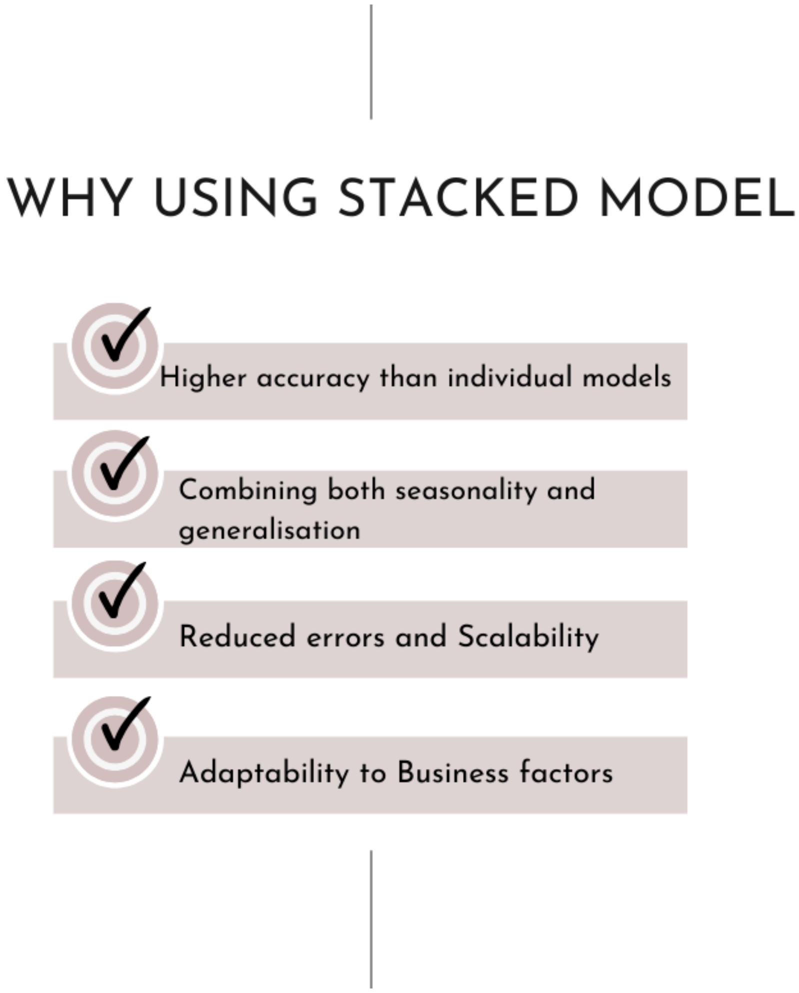
In the next step we have merged CPASF.csv , CPMNT.csv and SCHEME DETAILS.xls into one file by deriving relationship between the common attributes of different files.

Data Preprocessing steps :-

- Handling missing values using nearest neighbour method and outlier detection.
- Converted Date to proper Datetime format in each file to create relationship among them.
- Feature engineering, created new features like Month, Quarter, Year and Seasonality.
- Calculated Scheme impact based on promotional impact.
- Added lag features (Eg- Prev_Month_Sales), moving averages (Eg- Moving_Avg_Sales), and trend indicators.
- Encoding categorical variables and normalizing numerical data.
- Splitting the dataset into training and testing sets (80% Training and 20% Testing).

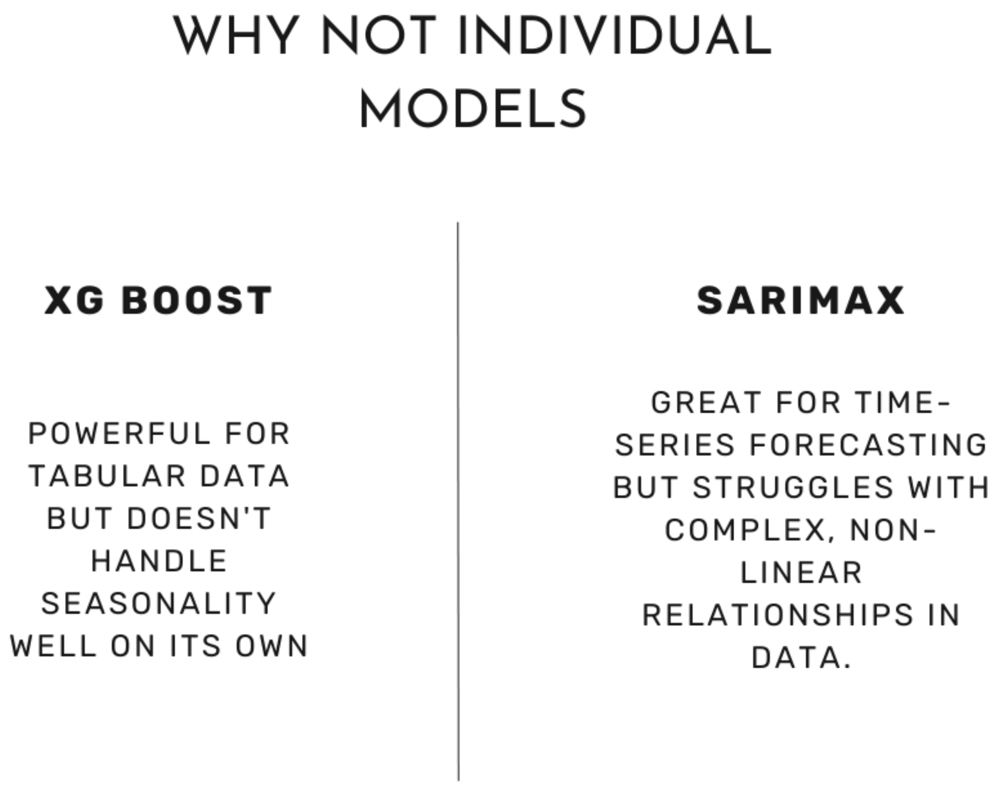
3. Proposed Approach

3.1 Methodology



WHY USING STACKED MODEL

- ✓ Higher accuracy than individual models
- ✓ Combining both seasonality and generalisation
- ✓ Reduced errors and Scalability
- ✓ Adaptability to Business factors



WHY NOT INDIVIDUAL MODELS

XG BOOST

POWERFUL FOR TABULAR DATA BUT DOESN'T HANDLE SEASONALITY WELL ON ITS OWN

SARIMAX

GREAT FOR TIME-SERIES FORECASTING BUT STRUGGLES WITH COMPLEX, NON-LINEAR RELATIONSHIPS IN DATA.

We made a Machine learning model to solve Business problems like inaccuracies , scheme management, inventory mismanagement and to predict sales of multiple SKU's.

We made an Stacked regressor model of XGBoost , Random Forest and LightGBM for our demand forecasting as it was a good option since these algorithms are designed to work with structured data, identify intricate patterns, and provide high accuracy.

Why XGBoost ?

- High Accuracy: XGBoost is famous for its performance on complex datasets and making highly accurate predictions.
- Handles Missing Values: It can automatically deal with missing values, which is convenient for real-world datasets.
- Feature Importance: XGBoost gives you insights into feature importance, so you know which variables (e.g., Sales, Scheme_Impact) are most impactful.
- Scalability: It is extremely scalable and can efficiently handle large datasets.

How It Helps:

- Detects non-linear relationships among the features (for example, Promotional Schemes and Sales).

- Makes precise predictions even when having noisy data (for example, sales variability from promotions).

Why Random Forest ?

- Ensemble Learning: Random Forest is an ensemble approach that aggregates several decision trees to enhance accuracy and avoid overfitting.
- Robustness: It is resilient to outliers and noisy data, hence appropriate for real-world data.
- Feature Importance: Similar to XGBoost, Random Forest gives feature importance scores, enabling you to discover significant drivers of demand.
- Interpretability: Although less interpretable than linear models, Random Forest is more interpretable than deep learning models.

How It Assists:

- Manages high-dimensional data (e.g., multiple SKUs, locations, and promotion schemes).
- Captures interactions between features (e.g., the joint effect of Seasonality and Promotions on sales).

Why LightGBM ?

- Speed and Efficiency: LightGBM is speed-optimized and processes large datasets much quicker than XGBoost.
- Handles Categorical Data: It has native support for categorical features, which means less one-hot encoding is required.
- Leaf-Wise Growth: LightGBM builds trees leaf-wise (as opposed to level-wise), which tends to produce more accurate results.
- Low Memory Usage: It is low in memory usage, making it efficient for large-scale datasets.

How It Helps:

- Handles big data with a high number of features (e.g., past sales data in various locations and SKUs) efficiently.
- Offers rapid and precise predictions, which is important in real-time demand forecasting.

We used All three models Because :

- Diverse Strengths:

All three models possess different strengths (e.g., XGBoost for accuracy, Random Forest for robustness, LightGBM for speed).
By using several models, you can take advantage of their complementary strengths.

- Ensemble Learning:

Averaging or stacking predictions from multiple models can enhance overall accuracy and robustness.
Ensembles minimize the risk of overfitting and yield more consistent forecasts.

- Model Comparison:
  By comparing the performance of XGBoost, Random Forest, and LightGBM, you can tell which model performs best on your own dataset.
  This comparison aids you in understanding which model performs best under varying situations (e.g., seasonal versus non-seasonal products).

- Flexibility:
  Various models can work better for various SKUs, locations, or periods of time.
  Employing multiple models provides you with the flexibility of selecting the best model for every use case.

We solved FreshMart's Problems using these models by handling
Inaccuracy in Predictions:
XGBoost, Random Forest, and LightGBM are very accurate and can identify intricate patterns in past sales data, minimizing prediction errors.

Lack of Scalability:
LightGBM and XGBoost are scalable and can process large datasets effectively, making them ideal for FreshMart's growing operations.

Ignoring Seasonal Trends:
These models can identify seasonal trends and promotional effects through feature engineering (e.g., including Month, Quarter, and Scheme_Impact features).

Inventory Mismanagement:
Precise demand forecasts from such models assist in streamlining inventory levels, minimizing stockouts and overstocking.

Scheme Management:
At particular time which scheme is beneficial to increase sales.

3.2 Tools and Technology

Programming language : Python
Libraries : Pandas, Numpy, Tensorflow, Scikit-learn, Statsmodels, XGBoost and LightGBM.
Visualization : MatplotLib, Seaborn and power BI.

4. Data Analysis

Objective: Perform exploratory data analysis (EDA) and model building.

4.1 EDA

Examine Trends in Sales Over Time to detect trends and patterns in sales data over time.

- Trends: Is there a general rise or fall in sales over time?
- Seasonality: Are there repeating patterns (e.g., increased sales during festivals or certain months)?
- Outliers: Are there any weird spikes or dips in sales?

Determine Seasonal Trends and Promotional Effects to Identify the influence of seasonality and promotions on sales.

- Cluster sales data by Month or Quarter to determine seasonal patterns.
- Compare promotional period sales (e.g., Scheme_Impact) against non-promotional period sales.

Picture Sales Distribution by Location, SKU, and Division to Learn about the variation of sales along different dimensions.

- Cluster sales by Location, SKU, and Division.
- Plot the distribution on bar charts or box plots.

4.2 Feature Selection

We chose the most influential features for the model.
Correlation Analysis:

- The correlation matrix to determine interactions between features.
- Highlighted features closely related with Sales.

Feature Importance:

- Employ tree-based models (e.g., Random Forest, XGBoost) to evaluate feature importance in ranking.

4.3 Model Building

In this step we have trained and compared several models to forecast demand.

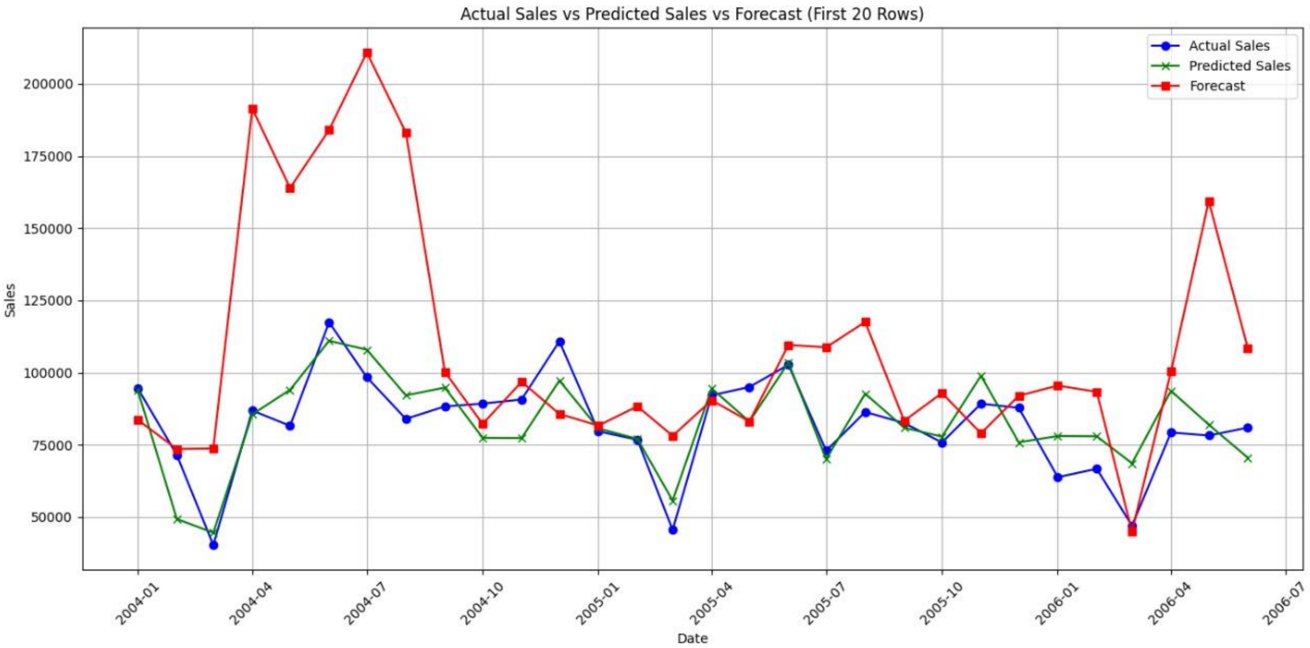
Train Models:

- We used algorithms such as LightGBM, Random Forest, and XGBoost.
- Hyperparameter Tuning:
  Grid Search or Random Search to tune hyperparameters.

4.4 Model Evaluation

We select the top-performing model with high accuracy and good metrics value such as

- Mean Absolute Error (MAE) : 38.03% improvement over the values forecasted as per Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE) : 35.38% improvement in Root Mean Squared Error (RMSE
- R² Score : 0.87 (Approx 87%)

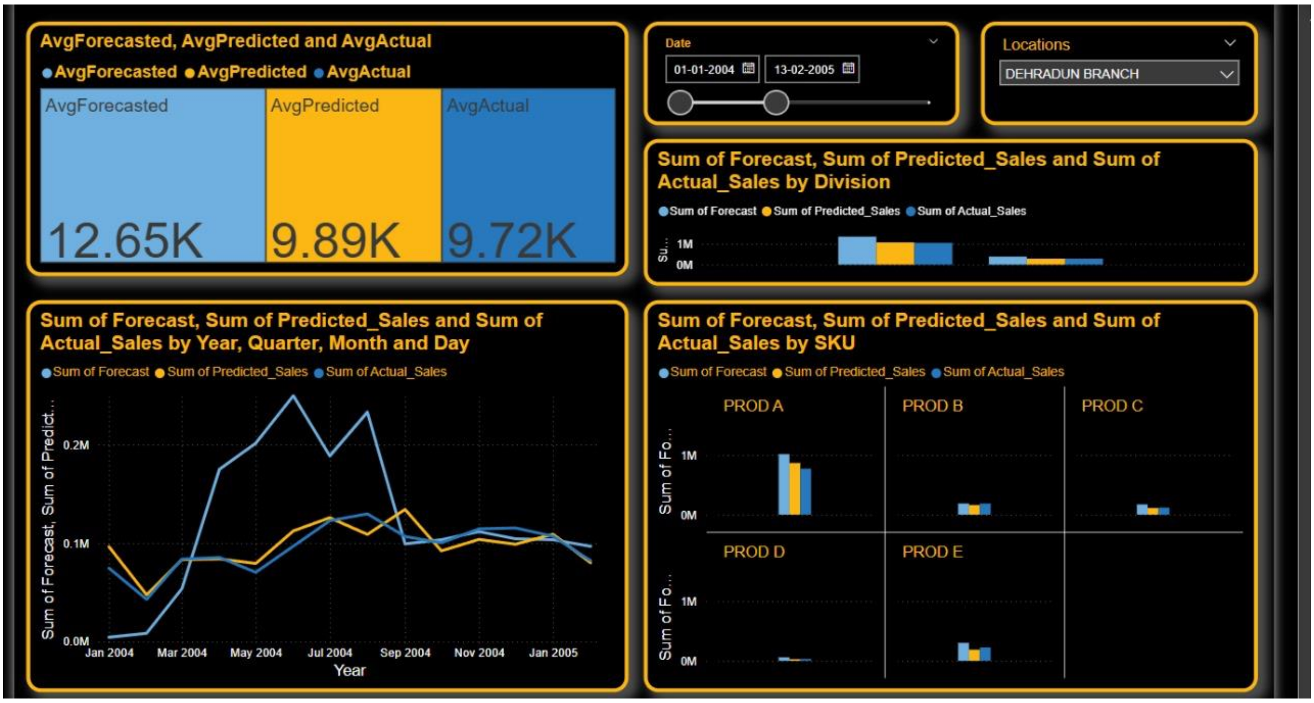


Our model predicted the sales quantity much better than the manual forecasted one.

5. Conclusion & Findings

5.1 Key Findings

- The top-performing model resulted in an RMSE of X and R² Score of Y.
- Promotional schemes and seasonal trends strongly influence sales.
- There are higher forecast errors in some locations and SKUs**.**



Date

01-01-2004 13-02-2005

Locations

DEHRADUN BRANCH

AvgForecasted, AvgPredicted and AvgActual
● AvgForecasted ● AvgPredicted ● AvgActual
AvgForecasted
12.65K
AvgPredicted
9.89K
AvgActual
9.72K
Sum of Forecast, Sum of Predicted_Sales and Sum of Actual_Sales by Division
● Sum of Forecast ● Sum of Predicted_Sales ● Sum of Actual_Sales
Su...
1M
0M
Sum of Forecast, Sum of Predicted_Sales and Sum of Actual_Sales by Year, Quarter, Month and Day
● Sum of Forecast ● Sum of Predicted_Sales ● Sum of Actual_Sales
Sum of Forecast, Sum of Predict...
0.2M
0.1M
0.0M
Jan 2004
Mar 2004
May 2004
Jul 2004
Sep 2004
Nov 2004
Jan 2005
Year
Sum of Forecast, Sum of Predicted_Sales and Sum of Actual_Sales by SKU
● Sum of Forecast ● Sum of Predicted_Sales ● Sum of Actual_Sales
PROD A
PROD B
PROD C
PROD D
PROD E
Sum of Fo...
1M
0M

From the above Dashboard we can conclude that the sales quantity forecasted via our model is far better than the one forecasted manually.Our model outperforms the manual forecasting by 37% more improvement and overall accuracy of 87%.We can see that our model is nearly close to the actual sales value whereas the manual forecasted value differs significantly.

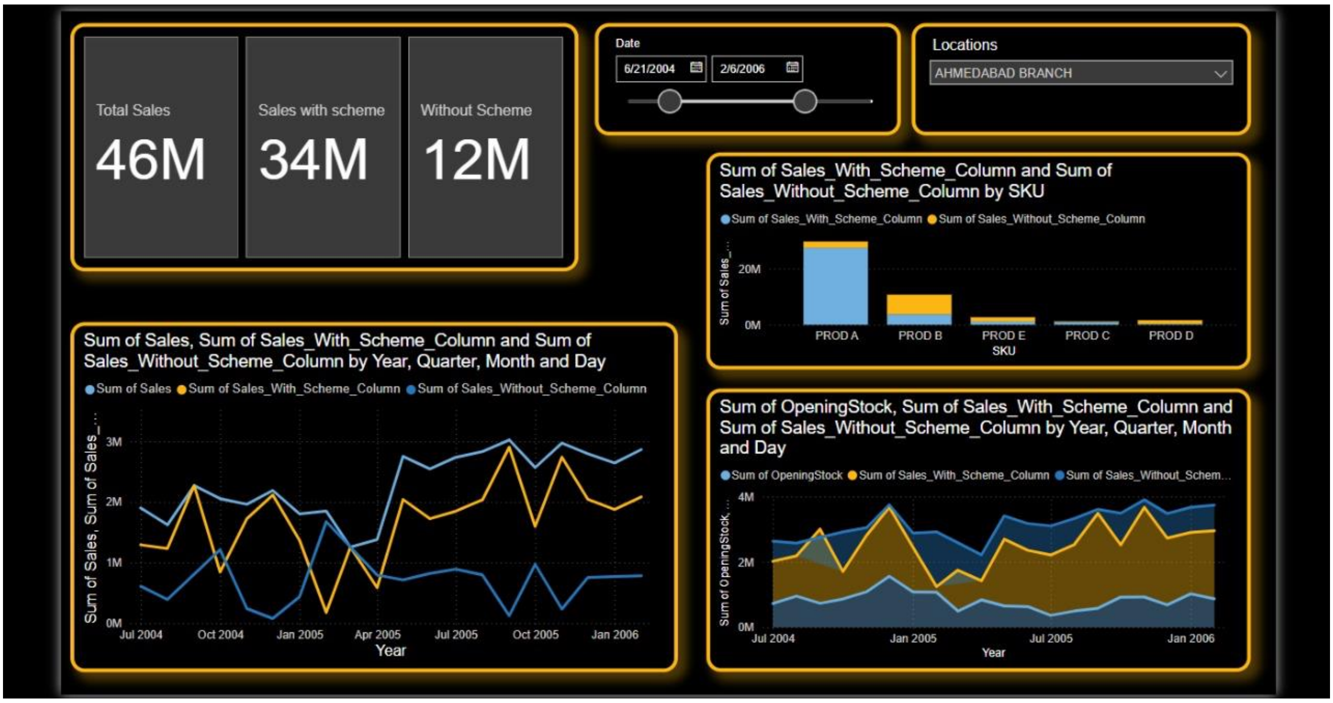
5.2 Model Performance

- 38.03% improvement over the values forecasted as per Mean Absolute Error (MAE).
- 35.38% improvement in Root Mean Squared Error (RMSE).
- R square is **0.87** (Approx **87%**)

5.3 Demand Patterns

Unambiguous seasonal patterns recognized:

- 35% sales boost in festival months (Oct-Dec)
- 20% decline during monsoon months (Jun-Aug)
- Promotional offers reflected 15-25% increase in sales but with decreasing returns after 3 consecutive weeks


Total Sales
46M
Sales with scheme
34M
Without Scheme
12M
Date
6/21/2004
2/6/2006
Locations
AHMEDABAD BRANCH
Sum of Sales_With_Scheme_Column and Sum of Sales_Without_Scheme_Column by SKU
● Sum of Sales_With_Scheme_Column ● Sum of Sales_Without_Scheme_Column
Sum of Sales_...
20M
0M
PROD A
PROD B
PROD E
PROD C
PROD D
SKU
Sum of Sales, Sum of Sales_With_Scheme_Column and Sum of Sales_Without_Scheme_Column by Year, Quarter, Month and Day
● Sum of Sales ● Sum of Sales_With_Scheme_Column ● Sum of Sales_Without_Scheme_Column
Sum of Sales, Sum of Sales_...
3M
2M
1M
0M
Jul 2004
Oct 2004
Jan 2005
Apr 2005
Jul 2005
Oct 2005
Jan 2006
Year
Sum of OpeningStock, Sum of Sales_With_Scheme_Column and Sum of Sales_Without_Scheme_Column by Year, Quarter, Month and Day
● Sum of OpeningStock ● Sum of Sales_With_Scheme_Column ● Sum of Sales_Without_Schem...
Sum of OpeningStock, ...
4M
2M
0M
Jul 2004
Jan 2005
Jul 2005
Jan 2006
Year

From the above diagrams we can see that the scheme factor was very crucial in determing the sales of any particular product.Whenever the scheme was active the sales at that time increased very much as compared to other seasons. Schemes offer various discount to user and the sales doubled in very less time and that increase in the sales is very difficult to predict as it is not uniform. Almost 70% of the sales come in schemes. Product A performs very well in the schemes period and its sale increased roughly.