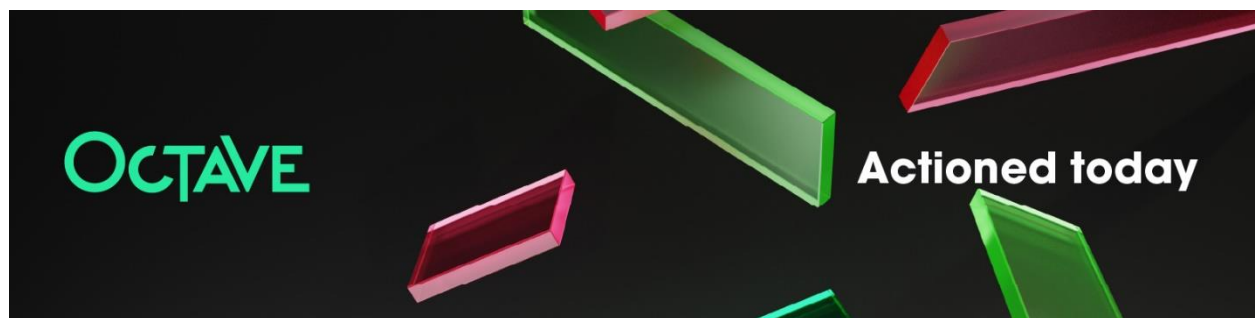


Round 01 – Kaggle Competition Case Study

DataStorm 3.0

OCTAVE &
Rotaract Club of University of Moratuwa.
March 2022



Sales Demand Forecasting

Business Problem

Over the years, the retail landscape underwent remarkable transformation through innovations on different verticals within the industry. During uncertain times, retail businesses are challenged to take key decisions within a short amount of time. Leading retail giants are turning towards advanced analytics for better data-driven decision making.

Forecasting sales is one of the most fundamental problems most retail chains have. Solving sales forecasting doesn't only help project revenue but also prepare for the necessary supply, reduce wastage, and better manage storage warehousing. Thus, it lies as one of the key building blocks in improving efficiency across several verticals within the retail landscape.

Use-case Definition

Retail Chain A has used traditional forecasting methods to estimate projected sales for each item across stores. However, in recent times, these approaches were found to be inaccurate, and they want to improve forecasting through advanced analytics and machine learning techniques. They have been collecting various data for the past 6 months at a store level.

They seek to understand how we can derive insights from the data and better estimate the sales 4 weeks in advance.

In this round of Data Storm 3.0, you are required to provide an analytics solution to Retail Chain A in order to solve their sales demand estimation problem.

Data Sources

You are provided with the following data source files to develop a data analytics solution.

1. Training dataset – Historical item sales dataset consists of 194 items from 4 different categories with daily sales in quantity units collected through transactions for 5 months. (train-data.csv)
 - a. CategoryCode: The items are classified into 4 different categories; this column refers to its code.
 - b. ItemCode: An identifier for each item

- c. DateID: The date for the respective sales value. The time period starts from 1st October 2021 to 13th February 2022.
 - d. DailySales: Sales value in quantity units for the respective item on the given date.
- 2. Validation dataset – Similar to training dataset, for specific items with weekly sales in quantity units for the 4 weeks from 14th February to 13th March 2022 (validation_data.csv)
 - a. CategoryCode: The items are classified into 4 different categories; this column refers to its code.
 - b. ItemCode: An identifier for each item (only selected items are present in the validation_data.csv)
 - c. Week: The time period starts from 14th February to 13th March (Monday to Sunday). The weeks are as follows:
 - i. w1: 14th February 2022 – 20th February 2022
 - ii. w2: 21st February 2022 – 27th February 2022
 - iii. w3: 28th February 2022 – 6th March 2022
 - iv. w4: 7th March 2022 – 13th March 2022
 - d. WeeklySales: Sales value in quantity units for the respective item for the given week.
- 3. Testing dataset – You are required to test your solution using test_data.csv. Final weekly sales for the next 4 weeks should be submitted in a CSV format with columns.
 - a. CategoryCode: The items are classified into 4 different categories; this column refers to its code.
 - b. ItemCode: An identifier for each item (only selected items are present in the validation_data.csv)
 - c. Week: The time period starts from 14th February to 13th March. The weeks are as follows:
 - i. w1: 14th February 2022 – 20th February 2022
 - ii. w2: 21st February 2022 – 27th February 2022
 - iii. w3: 28th February 2022 – 6th March 2022
 - iv. w4: 7th March 2022 – 13th March 2022
 - d. PredictedSales: You are required to predict the sales value in this column, in quantity units.

Deliverables & Evaluation Metrics

In this competition, you are required to submit the following.

1. Analytical Solution – Analytical solution to predict the demand for each item for the next 4 weeks separately. You must submit a CSV file, which contains the item code (ItemCode), and the predicted sales quantity for each week in the below format.

CategoryCode	ItemCode	Week	PredictedSales
category_2	4379	w1	XXX
category_2	4379	w2	XXX
category_2	4379	w3	XXX
category_2	4379	w4	XXX
category_4	111459	w1	XXX
...

You can use the Total Mean Absolute Percentage Error (Total MAPE) for evaluation. Total MAPE is given by:

$$\text{Total MAPE} = \frac{\text{Sum}(\text{Abs}(\text{Predicted Sales (I,W)} - \text{Actual Sales (I,W)}))}{\text{Sum}(\text{Actual Sales(I,W)})}$$

Where I is for each item, and W is for each week. – 50 Points

2. Technical Report - Report of your solution with clearly defined steps, features, feature engineering steps, modelling approaches, evaluation metrics, all the necessary plots/figures and interesting business findings that you can derive from this analysis – 50 Points.
 - a. Discuss any additional attributes that the retail management should start collecting
 - b. Based on the results you achieved, list interventions that the management team can take as next steps?