Demand Forecasting: Certified Angus Beef Boneless Chuckeye Roast

To

Meijer Inc.

From

Eshan Sharma

Hritwik Pal

Mansha Kalra

Ranojoy Sengupta

Matthew Lanham

March 2, 2022

Table of Contents

Executive Summary

As an extrapolation to the project carried out by Gopi Krishna Mashetty, Jiacen Liu, Hui Zeng, Prachi Priyam and Prof. Lanham, we wished to reproduce and optimize results with on the foundation laid out by the team. Our primary objective was to develop models on a smaller dataset with one unique Product ID (PID) to overcome the runtime complexity challenge that the original team had faced. The goal of demand forecasting is to optimize Meijer's supply chain to estimate the stock keeping units (SKUs), which is always a challenge in the fresh produce segment of the retail industry.

Our approach to achieve our primary objective includes two parts:

1. Data preprocessing and exploratory data analysis
2. Model building and forecasting

Before our initial part, we picked one unique PID (519494.0). The data for this product, from all the PID classes, was extracted using the high-performance capabilities (HPC) of Purdue – Bell Cluster. For part one, we cleaned the dataset, performed feature engineering and used it to perform the initial exploratory data analysis. With the insights gained, we then moved towards building models to predict the units sold in the past ~5 years. Using Weight Absolute Percentage Error (WAPE) as the metric of evaluation, we derived the best model and used that to predict the future demand for the next one month. Python and Tableau were widely used in the scope of our project.

Introduction

The dataset as provided to us by the original team contained numerous features, including transformed variables, one-hot encoded variables and the market basket unit quantity (the dependent variable). The data was provided on a broad level with product hierarchies with the aggregation level on a date-wise level, with a consolidated unit quantity sold per day. Other identifier variables such as UT_ID and PID were used to organize the data into their product hierarchy and stores.

Data Preprocessing

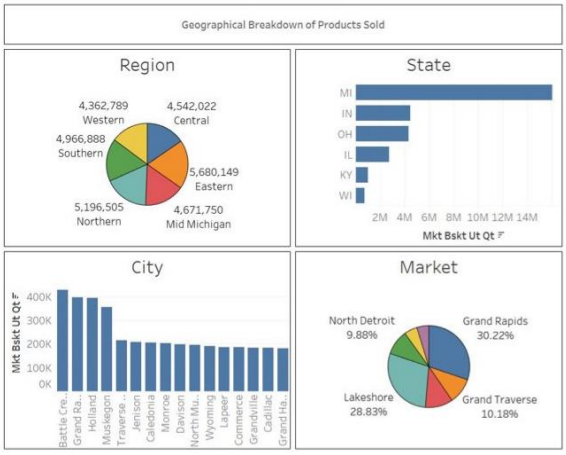
Initially, we chose a PID based on the higher ends of number of records. PID 519494.0 (Certified Angus Beef Boneless Chuckeye Roast) secured the second highest number of records and is a part of the fresh meat segment. As per Statistica, the average beef consumption in the US market has been 26.78 billion pounds from the year 2016 to 2020. With a rather consistency in the beef industry, we believe this product was a good indicator of the fresh meat segment, since we had to pick one unique product ID. The following steps were involved as a part of preprocessing the data before moving towards the exploration part:

1. *PID extraction*: Using the high-performance capabilities (HPC) of Purdue's Bell Cluster, we extracted all the information relevant to our product from all the files in the 25GB data given by the client, Meijer.
2. *Outlier Analysis*: The temperature variables had garbage values (999) which were dropped since they were less than 5% of the entire dataset.
3. *Feature Selection*: Features such as pr_drop_flg, adv_mid_wk_flg, adv_super_evnt_flg, adv_dgtl_circ_flg were completely dropped from the analysis as they contained all zero values. Furthermore, transformed variables and one-hot encoded variables were also dropped from our current analysis.0
4. *Type casting*: Relevant datatype conversions were made on features that stored dates to the datetime format.
5. *Feature Engineering*:
    a. "Holiday" is a binary variable which stores the information of whether the date was a holiday or not. This helped reduce the one-hot encoded holiday variables without losing the information.

b. “sales_price_alt” is a continuous variable that is calculated by multiplying the market basket unit quantity and actual sales price. This gives us information on the total revenue generated on a date-wise level.

Exploratory Data Analysis and Insights

After preprocessing, we got a cleaner version of the data. We moved towards exploring the data through visualizations and gaining a better understanding and insights into our chosen PID.

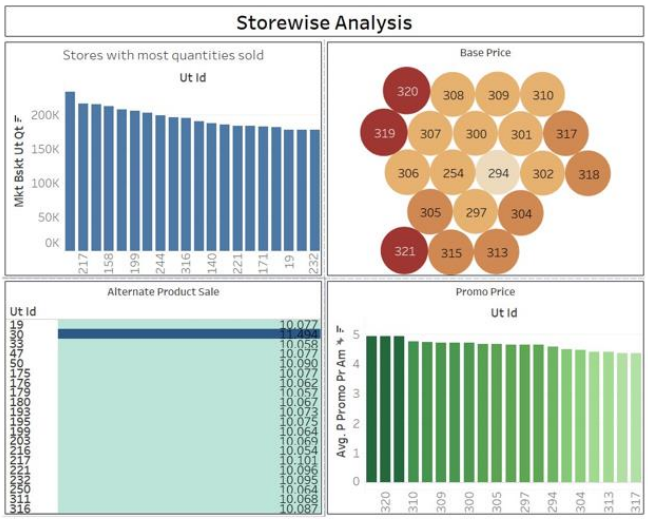


The Eastern Region contributes to over 19.3% of the sales, followed by Northern – 17.6% and Southern regions– 16.8%

Battle Creek, Grand Rapids and Holland are three cities with highest product quantities sold.

The state of Michigan has over 16M quantities sold- 4 times than second placed state Indiana.

In the market level, Great Rapids and Lakeshore dominate in terms of units sold over the years.

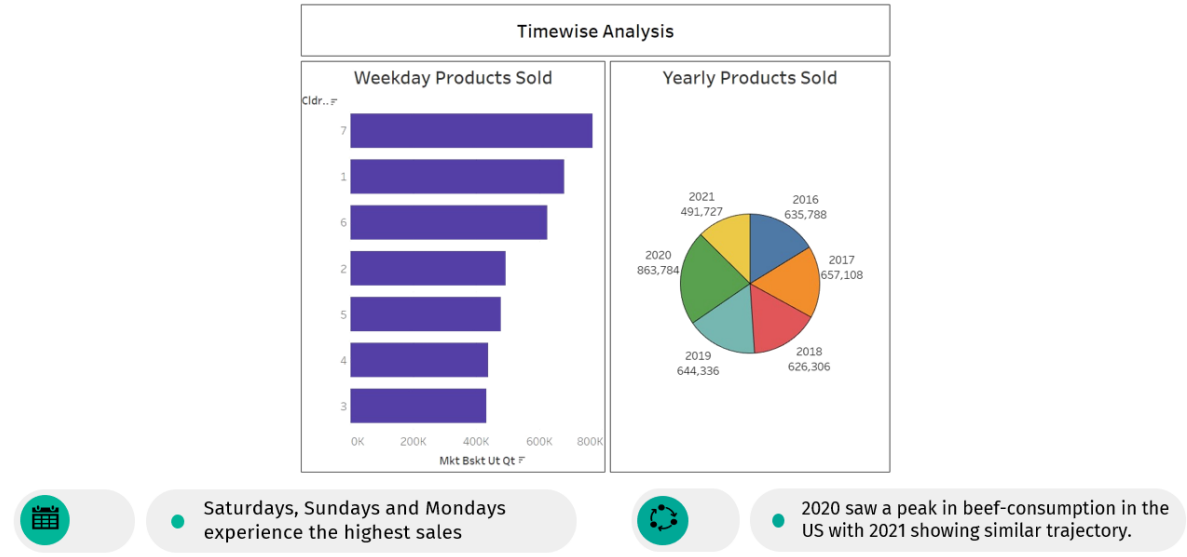


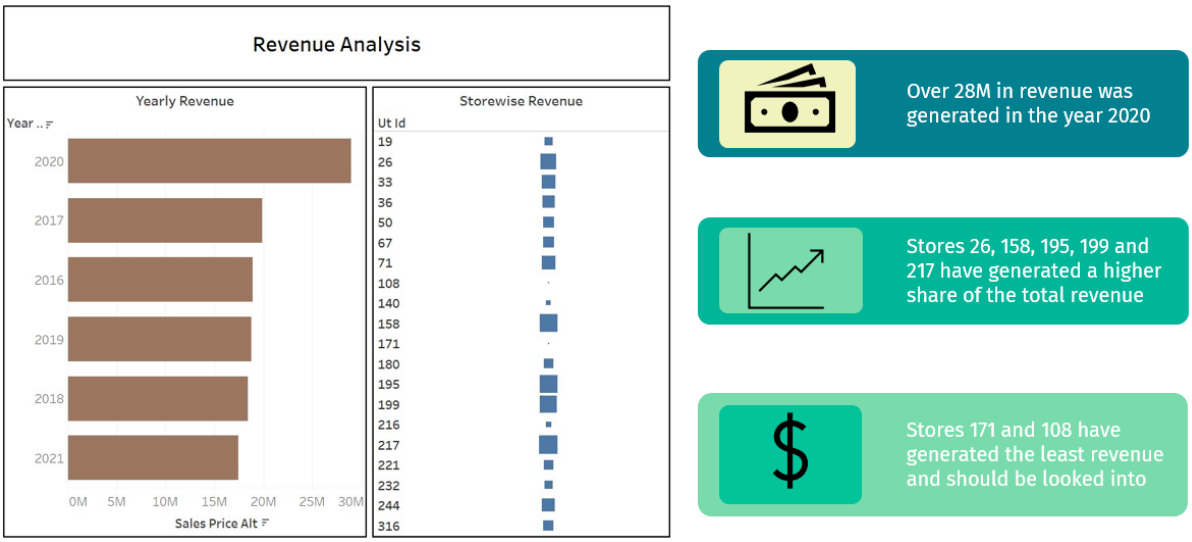
Stores 195, 217 and 33 are the top 3 stores in terms of units sold

Store 294 has the lowest average base price & stores 319,320, 321 have the highest average base price

Store number 30 has sold the most number of alternate product.

Stores 319, 320, 321 have the highest average promotional base price





Model Building

After preprocessing the data and gaining some deeper insights, we started with the model building process to achieve our primary goal of forecasting the demand for our product. In order to predict the demand, the market basket unit quantity was used as the dependent variable. We used the following models to choose the best model based on the lowest average WAPE:

Deep Neural Network: WAPE – 0.34

After aggregating the data on a store level, a deep neural network was implemented with the following simple architecture:

1. Three deep neural layers
2. Pyramidical neurons' structure: 32,8,1
3. Activation function: ReLU
4. Optimizer: Adam
5. Evaluation metrics: Mean Squared Error (MSE)

Based on the above architecture, the WAPE that was achieved with this model is: 0.34

Recurrent Neural Network: WAPE – 0.35

After aggregating the data on a store level, a recurrent neural network was implemented with the following architecture:

1. Three recurrent neural layers
2. Pyramidical neurons' structure: 32,16,1
3. Activation function: ReLU
4. Optimizer: Adam
5. Evaluation metrics: Mean Squared Error (MSE)

Based on the above architecture, the WAPE that was achieved with this model is: 0.35

Random Forest Regressor: WAPE – 0.34

After aggregating the data on a store level, a random forest regressor was implemented with the following architecture:

1. N-estimators: 100
2. Random State: 0

Based on the above architecture, the WAPE that was achieved with this model is: 0.34

XGBoost: WAPE – 0.32

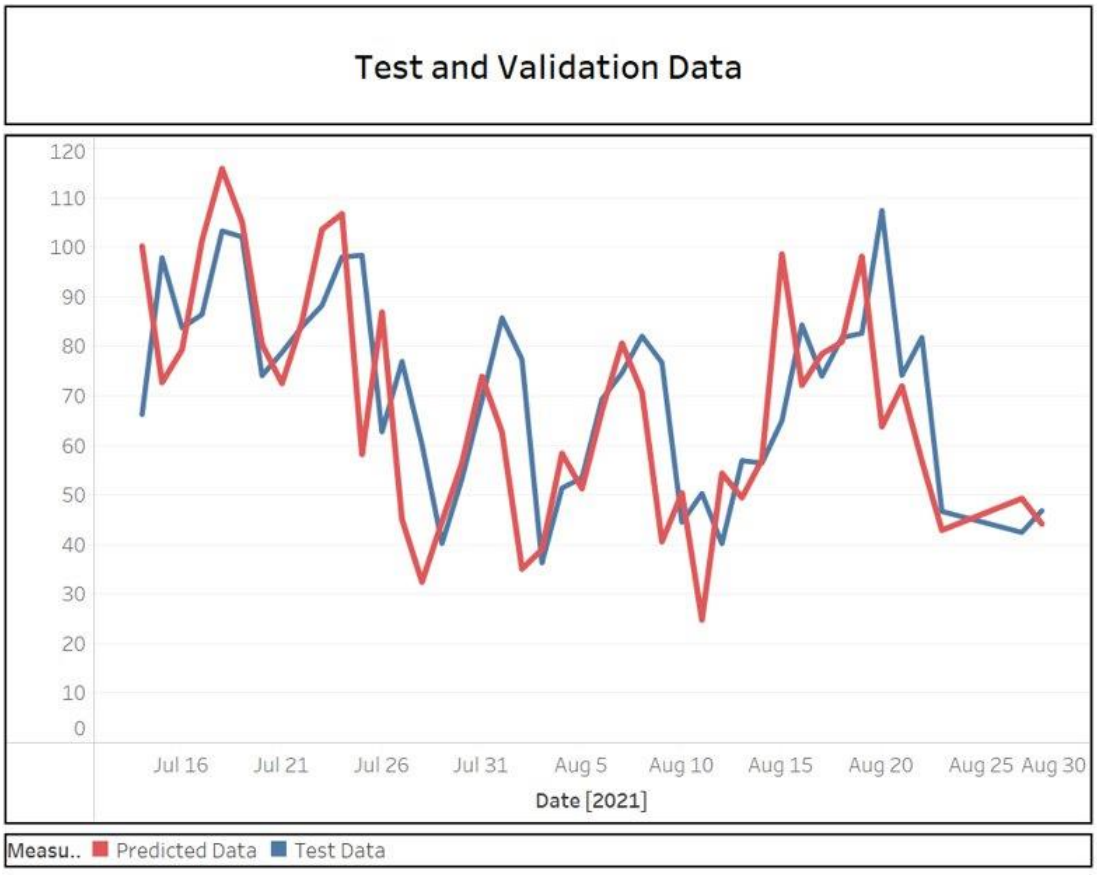
After aggregating the data on a store level, an XGBoost model was implemented using a grid search. The parameters set for optimizing the grid search are listed as follows:

1. Learning Rates: 0.01, 0.001, 0.0001
2. Max Depths: 3, 4, 5
3. N-estimators: 50, 100, 200, 500

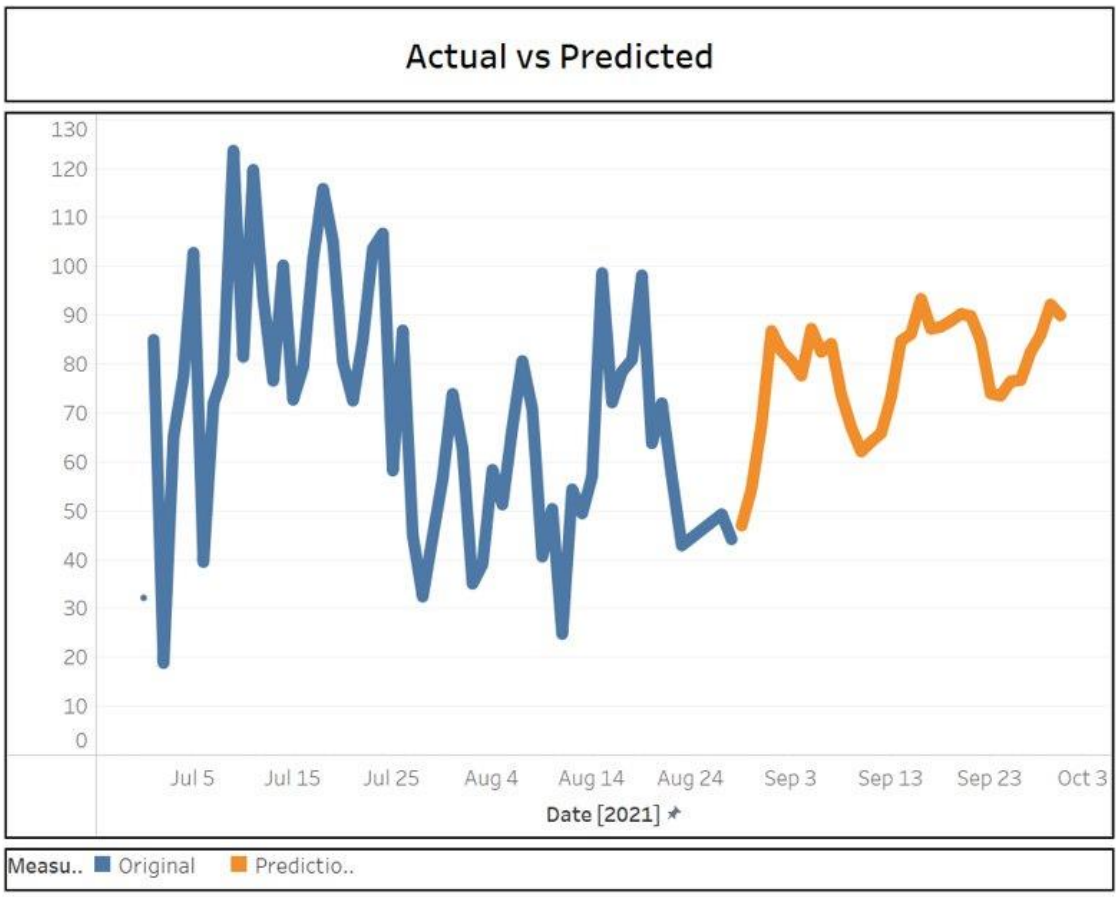
Based on the above architecture, the WAPE that was achieved with this model is: 0.32

Evaluation and Results

Basis the evaluation, we can conclude that the best output is given by the XGBoost Model with an average WAPE of 0.32 at the store level. This may be attributed to the fact that they are a more advanced form of decision trees which are used sequentially. Weights of the independent variables used also play a very important role in predictions as inputs into the decision tree. Incorrect weights are readjusted before being fed as inputs into the next tree, thus, giving a strong ensemble model. The graph for the model which was implemented on the validation test set is attached below:



The overall predictions are displayed in the graph below:



As per our predictions, we can conclude that store with the UT_ID of 71 had the best WAPE of 0.21.

Recommendation

Our recommendations to Meijer Inc. would be the following:

1. Stock up on beef roast for Sept 16th, 29th, 1st, to avoid losing sales.
2. Make combos with complementary products for an increase in revenue.
3. Consistently track products sales, even on days with low sales within regular time intervals.
4. Provide a new set of offers and discounts on weekdays to boost sales.

References

1. Statista. (2022, March 2). Per capita consumption of beef in the U.S. 2000–2031. https://www.statista.com/statistics/183539/per-capita-consumption-of-beef-in-the-us/