Data Glacier Group Project: Retail Forecasting

Modeling Report

Team Member Details:

Kierra Dangerfield
kierradachelle@yahoo.com
United States of America
Freelance
Specialization: Data Science

Problem Description:

The large company which is into beverages business in Australia. They sell their products through various super-markets and also engage into heavy promotions throughout the year. Their demand is also influenced by various factors like holiday, seasonality. They needed a forecast of each of the products at item level every week in weekly buckets.

Github Repo Link:

https://github.com/KierraDangerfield/Data-Glacier/tree/main/Week_13

Modeling

Overview

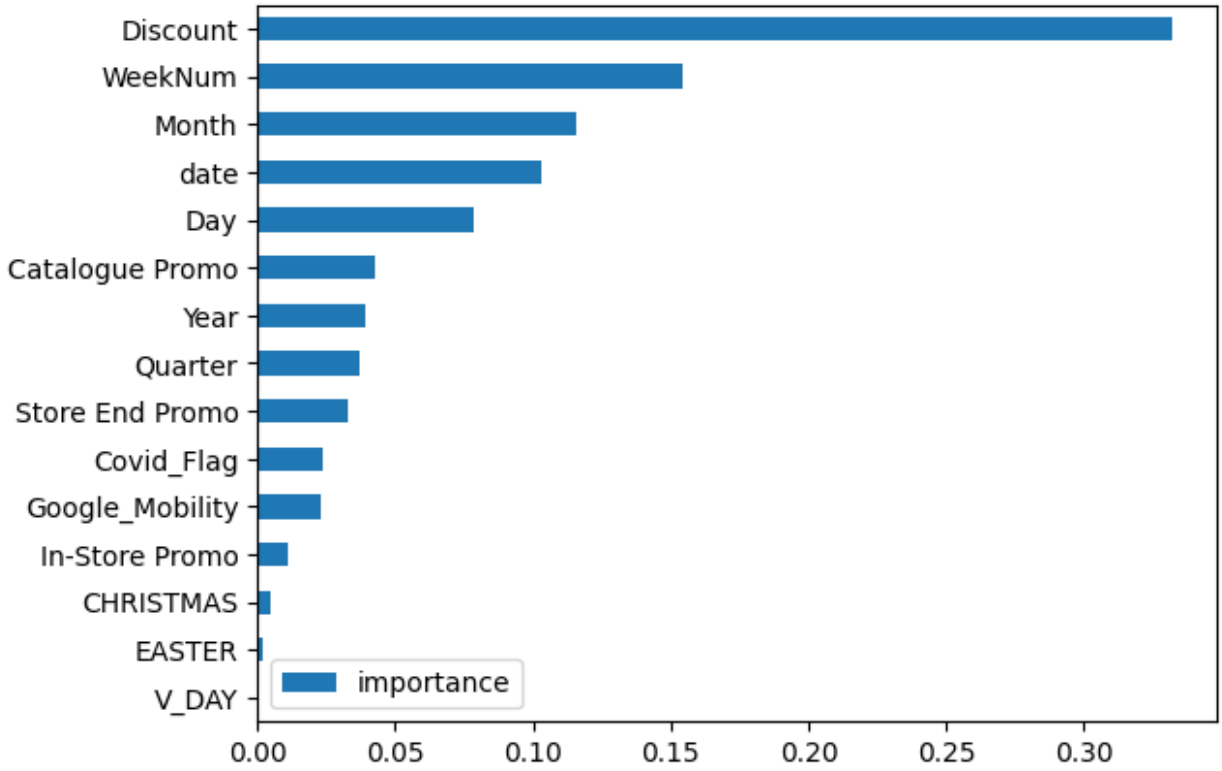
I trained 5 different models to see which one was performing the best based on mean absolute error. I used linear regression as my based model. Then I tested Random Forest, ARIMA, SARIMA, and SimpleRNN. For the model selection I only tested for product 1.

Model	Mean Absolute Error
Linear	$12,602.134
Random Forest	**$7,520.47**
ARIMA	$26,532.32
SARIMA	$26,299.80
SimpleRNN	$13,856.30

Based on the Mean Absolute Error, the Random Forest Model performed the best. Below is the Mean Absolute Error for each product.

Product	Mean Absolute Error
SKU1	$7,256.62
SKU2	**$1,067.46**
SKU3	**$13,994.45**
SKU4	$4,536.77
SKU5	$1,588.22
SKU6	$7,980.13

Product 2 performed the best using Random Forest, and Product 3 performed the worst. Below is a chart that shows the feature importance for Product 1. Customers respond well to discounts.



Key Findings

- Discount and WeekNum are the two most important features.
- The Random Forest model performed the best with a Mean Absolute error of 7,616.85 for product1.
- Product 2 and 5 performed the best with the random forest model. Product 2 had the least amount of total sales, so I wonder if that is a contributing factor.
- Product 3 had the worst performance with random forest. Product 3 also had the highest amount of total sales. I wonder if that is a contributing factor.

Final Recommendations

The ARIMA model would be great to use to predict upcoming weeks; however, it is not the best performing model.

I would suggest getting more data for each product to help build a better model. Instead of getting sales weekly, maybe collect daily sales in order to get more data.