Invent Analytics II

Assignment B)

While creating a predictor model, I sliced the data for each product-store pair. (For example, Fast products and Slow Stores) Then, I trained a model for each pair, meaning that there are **9 different models** contributing to the final prediction.

Considering prior time series analyses I made on the total sales, I used 1,4,5,7, and 30 interval lags as features in the model which helped the model perform much better.

Also, I created the following features using the date data:

week = The week number of the current week.

day = The number of the current day. Some days may have higher sales than others.

is_weekend = A boolean column labeled as 1 if the row's date falls on a weekend. Weekends often create a distinctive shopping environment, as people have more leisure time and are more likely to engage in shopping activities.

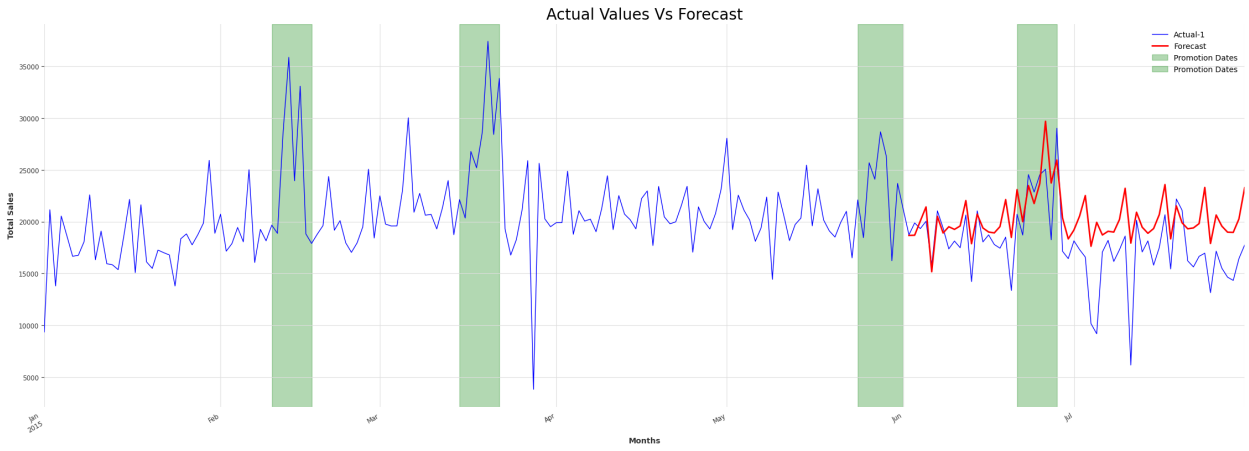
is_holiday = A boolean column labeled as 1 if the row's date corresponds to a holiday. Holidays often lead to unique shopping patterns, with increased sales due to special promotions, celebrations, or extended shopping hours. As a result, this column helps capture the distinct impact of holidays on sales.

is_promo_date = A boolean column labeled as 1 if the row's date corresponds to a day with an ongoing promotion. On promotion days, customers are more inclined to make purchases due to the perceived value and savings. This column allows the model to analyze how promotions influence sales on the specific days they are active.

is_first_week_after_promotion = A boolean column labeled as 1 if the date of the row falls within the first week after a promotion. During the initial week after a promotion, the amount of returns coming from unhappy customers is higher than the average. The return drops the total sales as they are entered as negative. The data shows big downward spikes, especially on the first weekend after promotions.

I split the data into train and test parts with a **70-30 ratio**. This allowed me to test the model's ability to capture the effect of the last promotion(Promotion4) in the initial dataset. The forecast proved that the model is sensitive to promotions which was my primary goal.

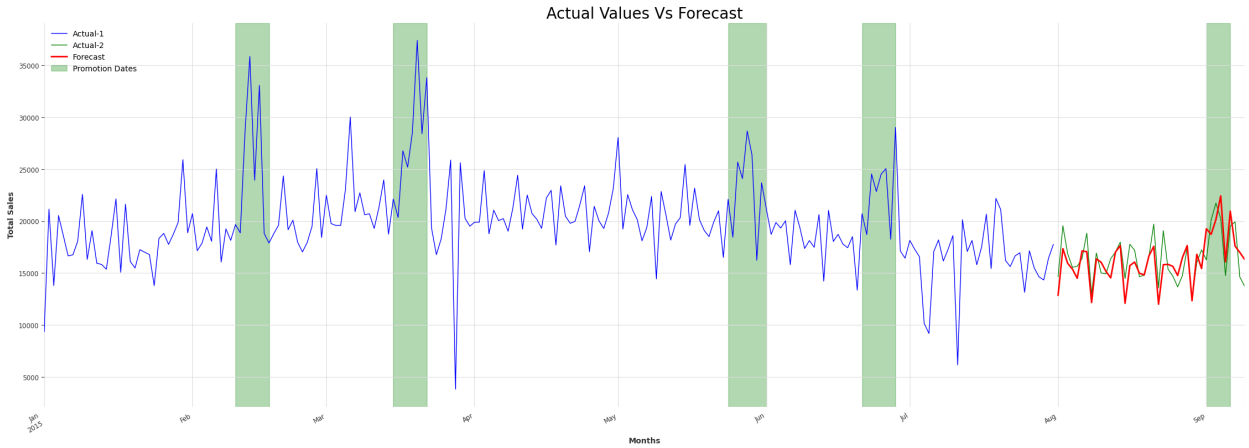
Since the overall trends of the training and test datasets are slightly different and there are big spikes in the test data, the predictions failed to fit the trend in the very last part of the data. The mean absolute percentage error was **14.23%.**



MAPE:14.23%

To see the effect of promotion 5, I trained the model using the entire dataset with the same parameters and made a forecast. The model was able to showcase the promotion effect in its predictions. Upon seeing that, I compared the forecast with the actual values

coming from the second dataset. The results were even better, achieving **8.04**% mean absolute percentage error (MAPE) value.



MAPE: 8.046

According to predictions generated by 9 different models, I calculated the mean sales of each product-store group pair for both the week before the promotion5 and during the promotion 5. The models predicted that "**Fast Store x Fast Product**" pair is the one that yields the most from the promotion5 with 44% increase in sales. The performance of other store-product pairs are shown as follows:

	Before	During	Promo_effect
Fast store x Fast product	5773.683594	8362.573242	44.839478
Medium store x Slow product	25.106565	31.725767	26.364422
Slow store x Slow product	7.191667	8.869360	23.328293
Slow store x Fast product	1380.594238	1662.380371	20.410501
Medium store x Medium product	1534.118652	1816.035156	18.376446
Medium store x Fast product	4765.367676	5407.257324	13.469887
Slow store x Medium product	513.900696	573.436707	11.585116
Fast store x Medium product	1599.580688	1726.469604	7.932639
Fast store x Slow product	19.678225	21.161581	7.538056

Questions

1- What measure would you use for goodness of fit?

I used **mean absolute percentage error(**MAPE) as my primary measure to understand the goodness of fit since it is easy to interpret. It represents the average percentage deviation of forecasted values from actual values. Also, it does not get affected by outliers as much compared to other metrics like Mean Squared Error. Considering the data includes many spikes, it is a good option for model evaluation.

2 - How good is your model developed in step 1?

The model initially showed relatively a poor performance. (MAPE = 14.23%) It did not bother me because my initial goal was to create a model that predicts the impact of promotions on sales.

3 - What are the main problem points causing bad fits?

The initial challenge faced in the analysis was the shifting trend in the sales data. At the beginning of the dataset, there was a noticeable upward trend, but towards the end, it changed to a downward trend. This inconsistency posed a significant issue, leading to inaccuracies in the predictions. However, incorporating the data with the downward trend into the training phase helped mitigate this problem.

Furthermore, having large spikes in the data was another problem. These spikes caused considerable disruptions in the model's performance, affecting its ability to accurately forecast sales.

Additionally, the data does not include enough data for an entire year. This may prevent the model from learning a yearly seasonality pattern in the sales. Having a full year’s worth of data would help the model detect and leverage the recurring patterns and seasonal trends in sales.

4 - What would you change in step 1?

I would be more careful with where I am splitting the data, paying more attention to the overall trend. Also, I would try to remove the entire data with the upward trend.

Recommendations:

The clients can take managerial decisions by looking at the model predictions on store-product pairs. Here are some points to take into considerations:

- **Supply Chain Optimization:** The sales predictions reveal a substantial 45% increase in sales for the Fast store and Fast product combination during Promotion 5. This surge in demand poses a significant challenge for the supply chain operations. To capitalize on these sales opportunities, it becomes critical to ensure a seamless and uninterrupted supply of the desired products to the Fast stores. Failing to meet the demand could result in missed sales opportunities, underscoring the necessity of maintaining a well-organized and agile supply chain for the high-performing product-store pairs.
- **Product-Store Synergy:** Notably, the model predictions demonstrate that Slow stores paired with Slow products unexpectedly perform remarkably well, generating 23.3% more sales compared to selling Slow products at Fast stores, which only yields a 7.5% increase. This observation highlights an opportunity for resource optimization within the company. Prioritizing the sales of Slow products at Slow stores, instead of Fast stores, proves to be more profitable. Managers may consider reallocating inventory by reducing or transferring Slow products from Fast stores to other locations, optimizing sales potential and resource allocation.
- **Strategic Promotional Decisions:** The forecasted performance of various store-product pairs should guide promotional strategies. To maximize efficiency and resource utilization, it is logical to select the best-performing store-product pairs for

promotions. Based on the results, pairs like Medium product - Fast store are not predicted to perform well during promotions, leading decision-makers to reconsider or cancel promotions for such combinations to enhance overall promotional effectiveness.

- **Identifying Operational Inefficiencies:** A comparison between the predictions and actual sales values can be indicative of operational inefficiencies. Consistently underperforming store-product pairs may indicate internal issues such as ineffective sales processes, inefficient inventory management, or equipment malfunctions. Addressing such challenges can bring sales performance in line with expectations and improve overall operational efficiency.
- **Customer Segmentation:** The analysis allows for the creation of buyer personas for store-product pairs based on characteristics like store location or product prices. This customer segmentation approach provides valuable insights for tailoring future promotions and marketing strategies to specific target audiences, leading to more effective and personalized customer engagement.

External Datasets:

I incorporated an external dataset into the analysis, specifically the holidays data for Turkey. This dataset enabled the model to capture patterns related to holidays during the training process. The holidays data was sourced from the "holidays" library in Python.

Additionally, I sought to explore potential patterns that could influence sales by using a weather temperature dataset for Istanbul, Turkey, corresponding to the given dates. However, despite the effort, I did not observe significant benefits from this weather dataset. The weather data was obtained through an API service provided by **https://www.timeanddate.com**.

Bonus Question:

In Turkey, the allowable time frame for item returns is **14 days**, which applies to items purchased during promotions as well. To analyze the return rates, I considered this 14-day window and calculate the return rates for each day. Subsequently, I performed an

independent Student's t-test to assess the significance of any differences in return rates when promotion days and regular days are compared.

The test returned a p-value of **0.066,** which is just above the critical value or it is in the danger area if we lower our significance tolerance by raising the critical value. Anyways, traditionally speaking, there is no significant difference between the regular days and promotion days in terms of item returns. However, still there is practical importance in returns on promotion days. Since the p-value is close to the critical level, it is worth investigating the underlying factors in item returns. **It is also important to consider that the chart shows deep downward spikes after the promotions which are likely to result of big number of returns.**