TEAM 133

INDEX:

INTRODUCTION

This report outlines the analysis done on the sales data of fans and their relationship with various factors like Location, Shop ID, Seasonality etc. Our Task is to analyze any trends or correlations in the data across all regions and months and predict the sales of June 2021 for all shop IDs using Time Series Forecasting and other Machine Learning Methods.

DATA DESCRIPTION -

The data available to us is divided into 4 regions - particularly 'NORTH', 'EAST', 'WEST', 'SOUTH', mapping to Wh1, Wh2, Wh3, and Wh4 respectively and we have monthly sales of different Shops from April 2018 to May 2021.

The number of distinct Regions = 4
The number of distinct shop ids = 408
The number of distinct months = 39
Due to the repetition of shop id across regions, the total number of data items is 1039.

Few Simple Observations:

1. The data is quite small
2. The year 2020 will behave differently than others due to covid
3. Very few features are available to us, thus feature engineering is necessary
4. There are various outliers in months (maybe due to festivals) and shop ids (maybe due to it being located in a metropolitan city)
5. Outliers across different regions occur at different months/shop IDs

FEATURES IN THE DATASET PROVIDED

S.No.	Feature	Data type	Details
1	Region	String	Mapping to (‘NORTH’,’ EAST’,’ WEST’,’ SOUTH’)
2	WarehouseID	String	Mapping to (‘Wh1’,’ Wh2’,’ Wh3’,’ Wh4’)
3	SKU id	int	Mapping shop to unique ID
4	Monthly Sales	int	Mapping month to its sales for each shopID

DATA CLEANING AND PROCESSING

Corrupted or missing data imbues a huge amount of error in the prediction of an analysis model which in turn drastically reduces the accuracy of the prediction, therefore data cleaning is an essential part of data preprocessing.

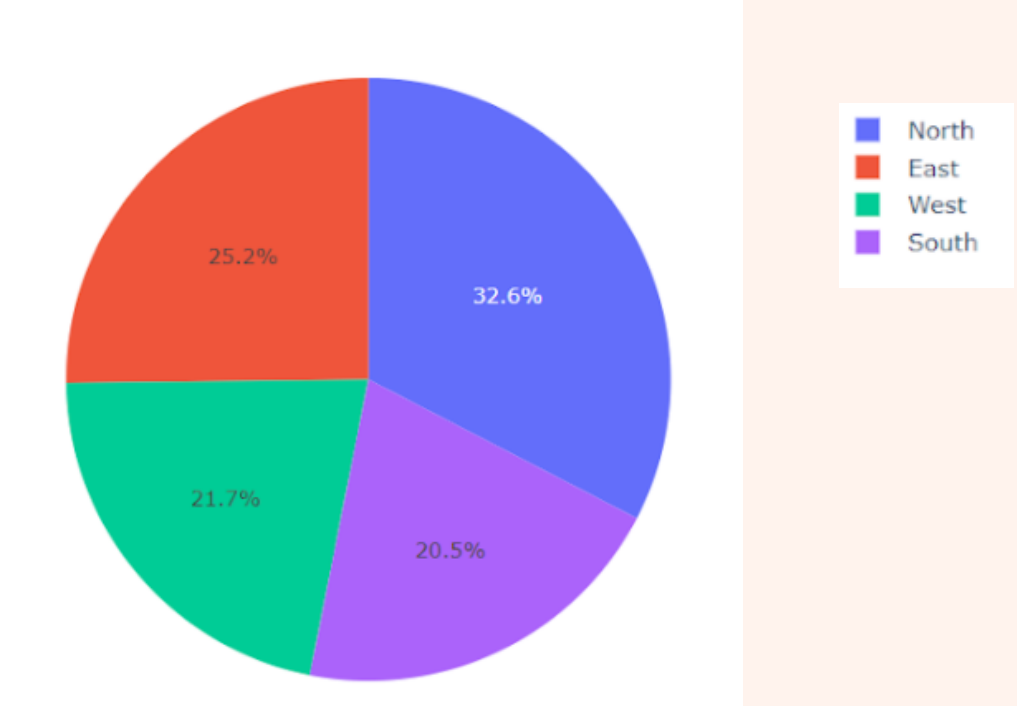
The dataset was looked at for null and NaN values. The dataset provided does not contain any null value, but extra features generated through feature engineering resulted in null values, which were replaced by the mean of their respective columns.

The outliers were kept as the dataset is already too small to reduce further.

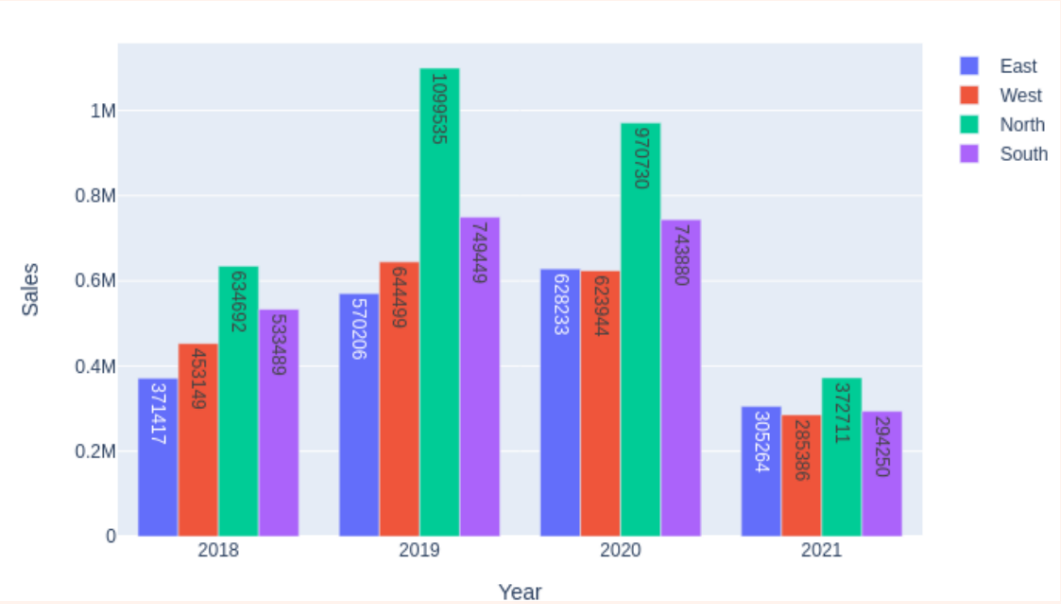
The data was split into training and test, the split was made in Dec-2020. Also, we also trained models after splitting them into Regions, i.e **‘North’,’ South’,’ East’,’ West’**.

DATA VISUALISATION & EXPLORATORY DATA ANALYSIS

REGION-WISE AGGREGATE DATA



YEAR-WISE CUMULATIVE SALES



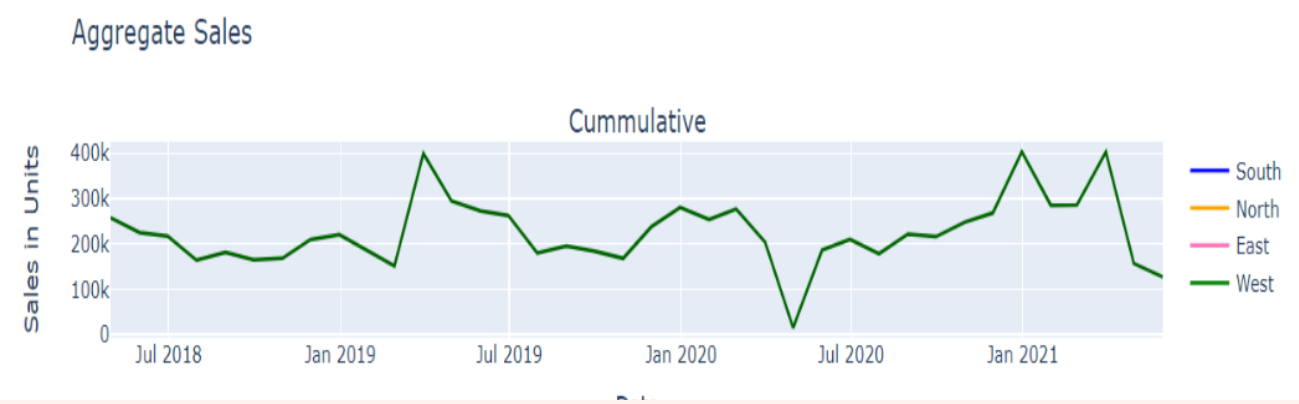
FEATURE ENGINEERING

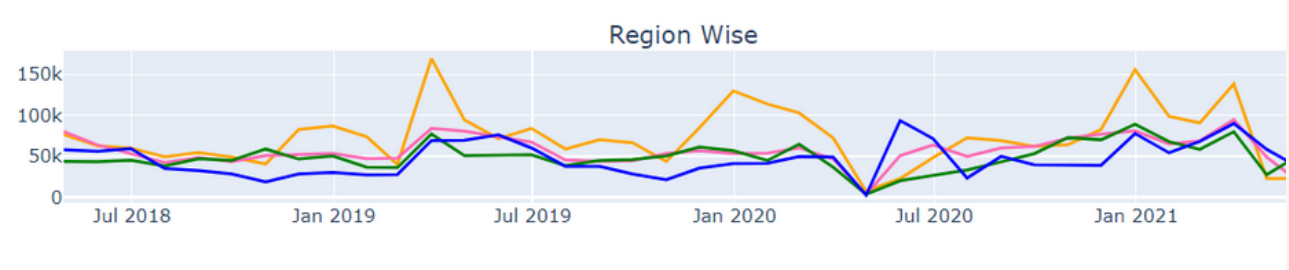
List of features used for training:-

S. No.	Feature	Data-type	Information
1	sales prev month	int	sales of the previous month.
2	isSankranti	int	One hot encoding for the month of January(because that's when Sankranti occurs)
3	isHoli	String	One hot encoding for March
4	isDiwali	String	One hot encoding for October
5	isChristmas	String	One hot encoding for December
6	isCovid	int	One hot encoding for Aug, Sep, Oct 2020 and Mar, Apr, May 2021
7	Month	int	corresponding month
8	time_idx	int	Set beginning month as 0 and accordingly number each month with the last one being 37
9	isSum	int	summer/non-summer month(Apr, May, June)
10	isWin	int	winter/non-winter month(Nov, Dec, Jan, Feb)

As we had only one training feature(timestamp) and to be predicted feature(sales), using statistical methods to gain features for the time series data like lag, differences, ewm, ra and categories like summer, winter to follow the trend.

PLOTS

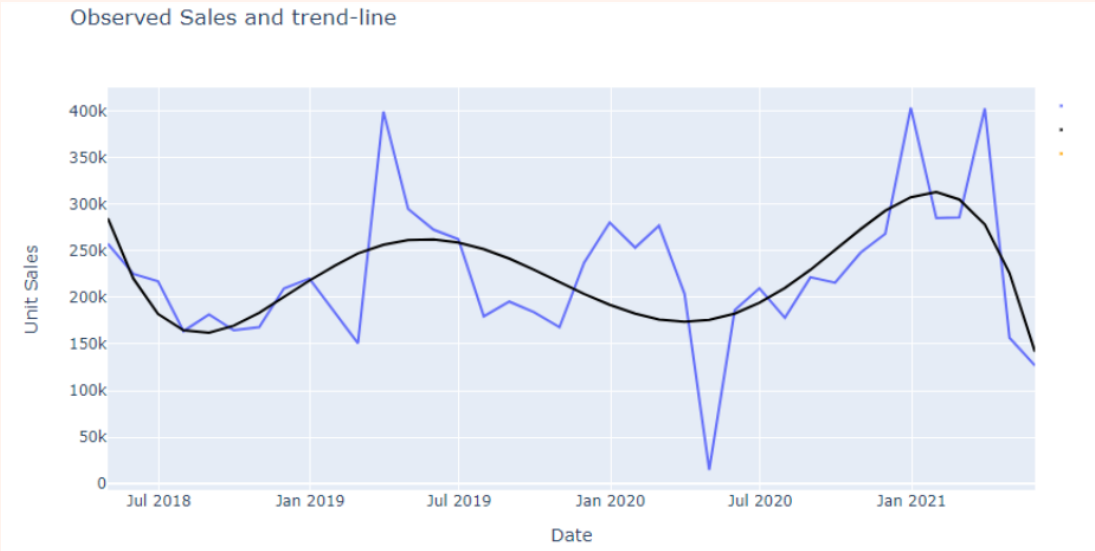


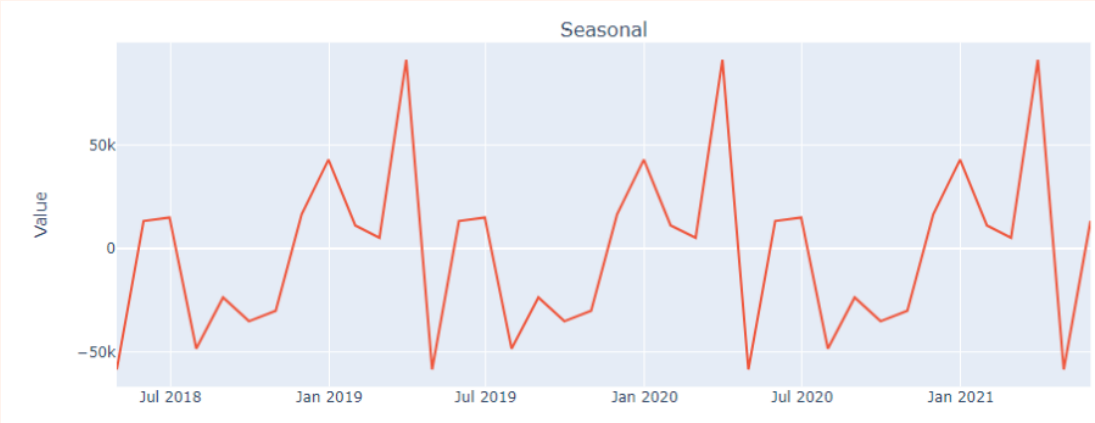


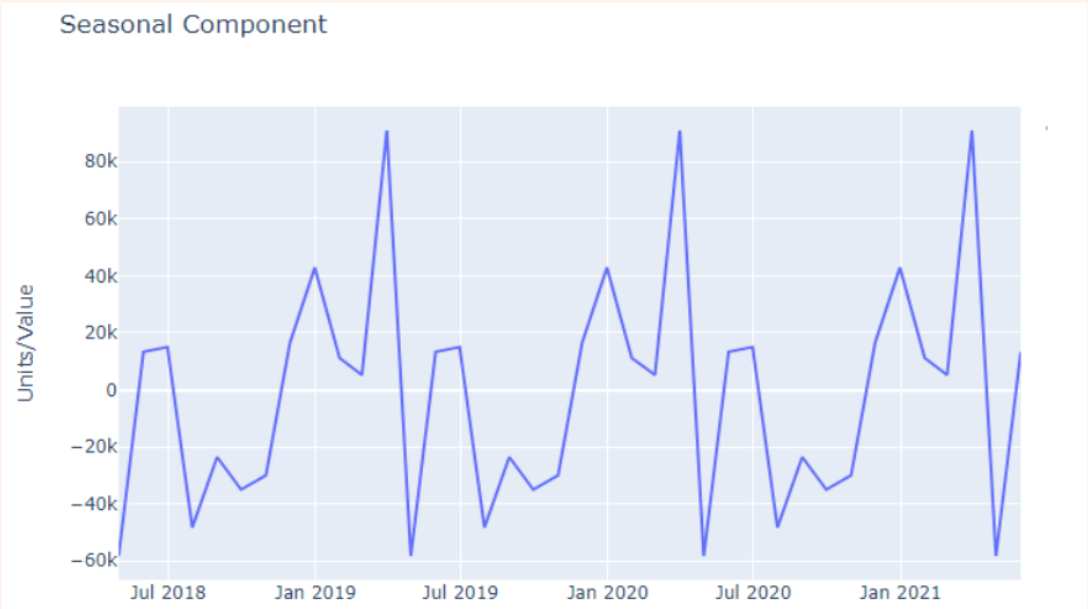
The plot which shows that during the time of festivals and summer, winter the amount of sales increases and aggregate plots of whole data and region-wise

From the above-average plots of every month for every region we can conclude that there is a rise and dip in sales during summer and winter months respectively. Now, from the above plots, we can also infer that there is a specific increase in sales due to the presence of some festivals in those months like Holi, Christmas, Diwali, etc.

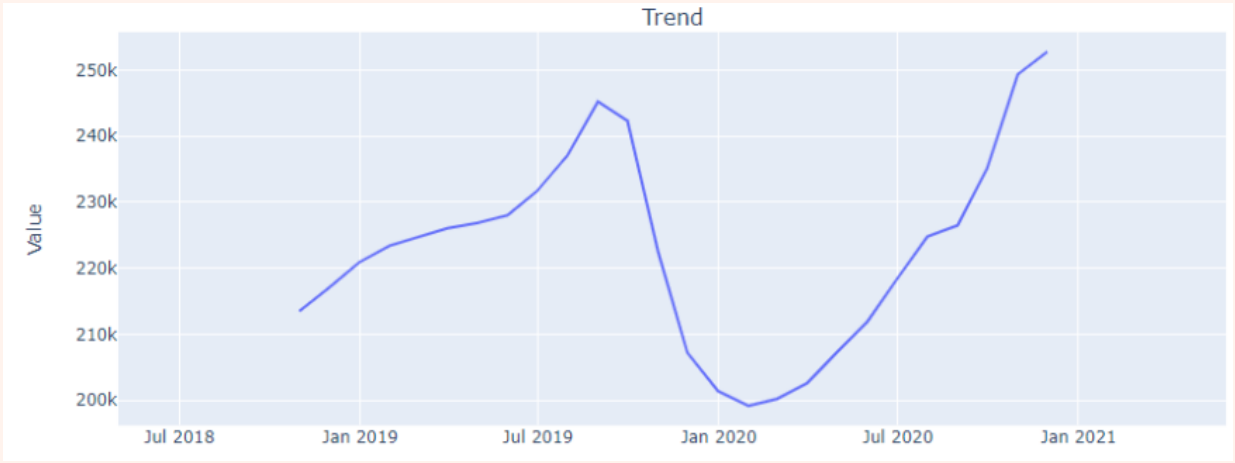
PLOTS- TRENDS AND SEASONAL







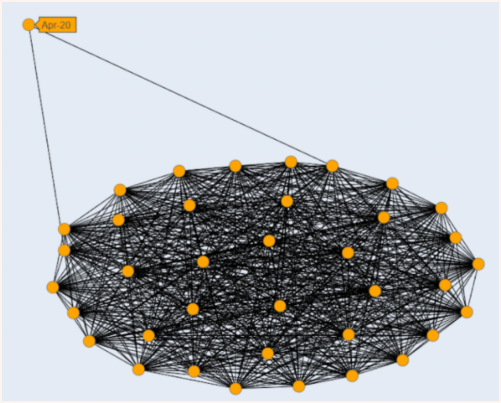
PLOTS- TRENDS AND SEASONAL



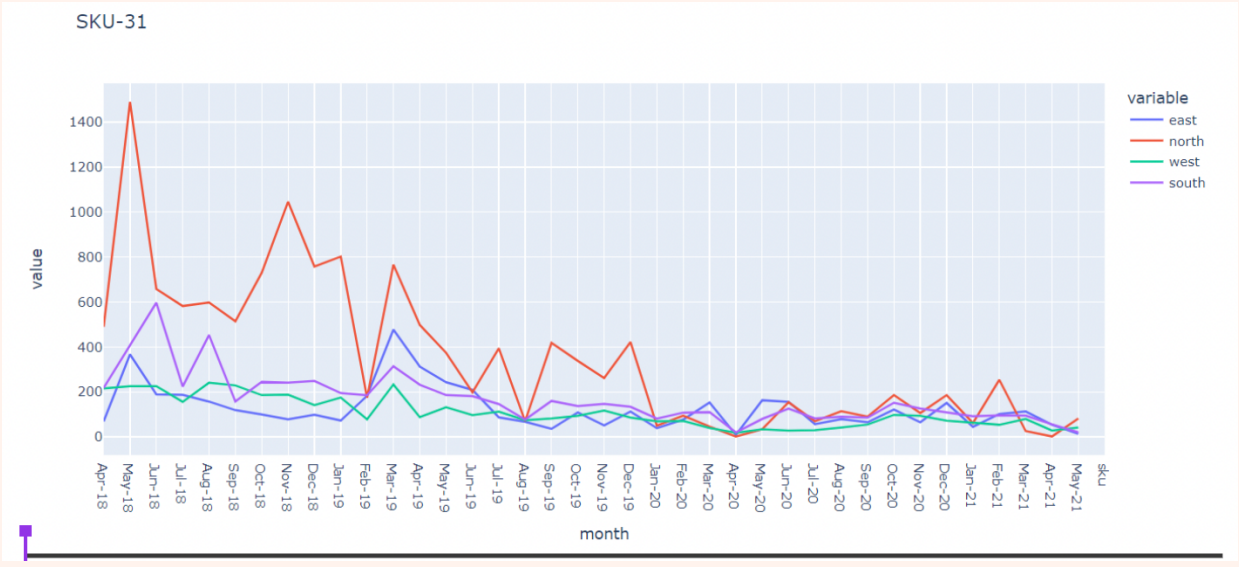
1. From the trend graph, We can observe that there is a little increase in sales from 2018 to, 2019 but there's a steep decrease in 2020 because of COVID- 19 and there's a steep increase due to recovery from COVID. The trend component is supposed to capture the slowly-moving overall level of the series.
2. From the seasonal component, We see the cyclicity of the data. The seasonal component captures patterns that repeat every season.
3. The residual is what is left. It may or may not be autocorrelated. For example, there can be some autocorrelated pattern evolving quickly around the slowly moving trend plus the seasonal fluctuations. This kind of pattern cannot be ascribed to the trend component (the former moves too fast) or the seasonal component (the former does not obey seasonal timing). So it is left in the remainder.

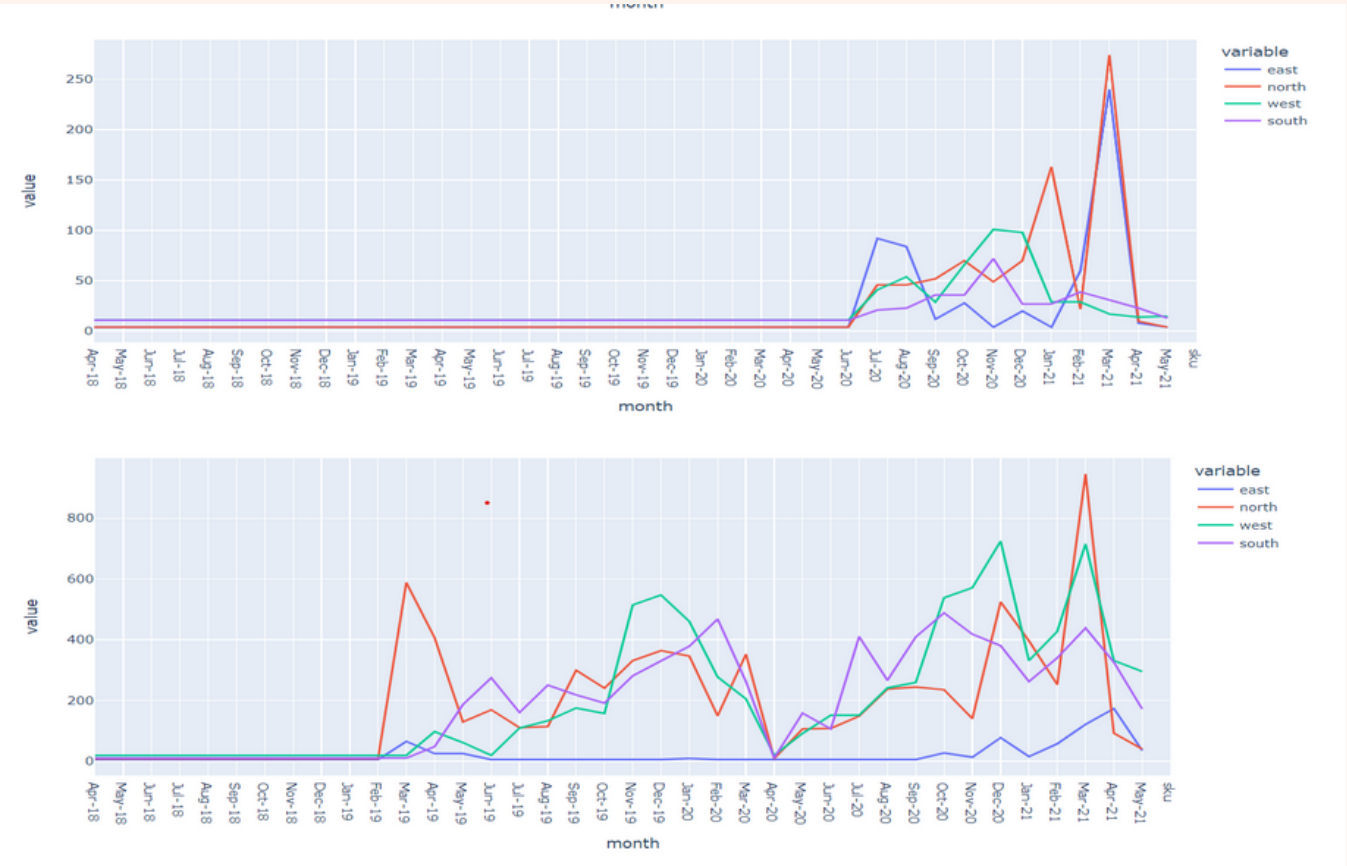
correlation of graph of data together where threshold shows Apr 20 as an outlier

correlation graph



From the above plot, We can conclude that there is a dip in sales of Apr 2020 dataset and we can associate this effect with the rise of the global pandemic i.e, COVID-19 in India





-> plot-1 is the plot of under-performing(decreasing sales wrt time) SKU units in the pre-covid period in all regions, after the covid lockdown the sales decreased drastically. (failed/Outdated stock)

-> plot-2 is the plot of under-performing(constant sales at low frequency wrt time) SKU units in the pre-covid covid period in all regions, after the covid hit the SKU, the sales were increased exponentially. (Trending Stock).

->plot-3 is the plot of SKUs where the sales experienced a short-term dip in the covid lockdown period and regained the sales frequency. (basic/popular/established stocks)

APPROACH AND MODELS

APPROACH

Several models were trained on the training set and tested on a validation set. The model that performed better on the final task was finally chosen for the prediction on the test set. It has been described below and the remaining approaches have been compiled under other approaches. The evaluation metric that we finally used was based on the minimum mape score on the predictions of the validation set.

Our problem can be categorized as a short-term sales forecasting problem. Classical forecasting methods, such as autoregressive integrated moving average (ARIMA) or exponential smoothing (ETS), fit a single model to each individual time series, and then use that model to extrapolate the time series into the future. In our problem, we have much similar time series across a set of cross-sectional units. Thus, it is beneficial to train a single model jointly over all-time series.
we divided the features into two categories to be passed in the model.

1. Time-varying known categoricals: Time-varying features with known values. eg festivals, seasons, month
2. Time-varying unknown reals: Time-varying features with unknown values. eg sales, sales next month, rolling mean

Other approaches:
Several ensembles and solo models were used for forecasting.
For We kept the same features as the DeepAR model and fine-tuned learning rate, tree-specific, and regularisation parameters.

- XGBoost Regressor
- Random Forest Regressor
- Light Gradient Boosting Machine

For classical forecasting methods, we passed individual time series for each row in the data frame.

- Auto-Arima
- TBATS
- GARCH

Evaluation results:
Mape scores for different models

MODEL	MAPE
XGBOOST	3.557705
DeepAR	1.945911
Random Forest Regressor	3.788959
Auto ARIMA	8.299338
Prophet	5.795035
DeepAR + AWS forecast	1.290658

FINAL MODEL DESIGN:

Ensembled the top-2 best models DeepAR and Amazon Forecast to get the predicted forecast of June 2021.

Formula:-

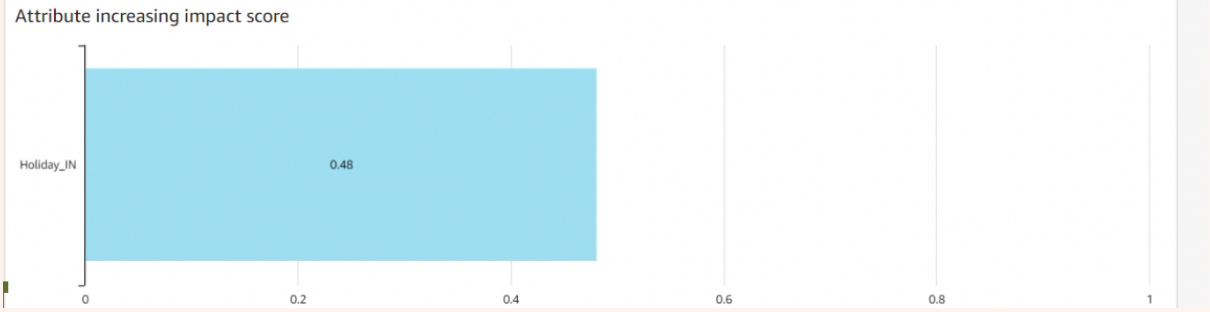
$$Ensemble\ model = \frac{\frac{Loss\ of\ AF}{Pred.\ of\ AF} + \frac{Loss\ of\ DeepAR}{Pred.\ of\ DeepAR}}{\frac{1}{Loss\ of\ AF} + \frac{1}{Loss\ of\ DeepAR}}$$

DeepAR model design:

We first apply the 'melt' function to the entire dataset to transfer the columns of individual dates into rows. Thus we obtain 39482 rows. Certain categorical features are added such as the important festivals of India(Holi, Diwali, Christmas, Sankranti), summer and winter months, a column for month values of each date, and previous month sales for each product.
The training dataset uses 32 months of history and the validation dataset contains the last 6 months. The 6 MAPE scores are the cumulative values for the months from Dec 20 to May 20. Early stopping has also been implemented which allowed the training to be complete in 2 mins as opposed to 8 mins without early stopping. The model tackles categorical variables internally and it doesn't need one-hot encoding to be done externally. A relative time index scale was also
added which assigns 0 to 2018-04-01 and 37 to 2020-05-01. The batch size for gradient descent was set to be 128

Amazon Forecast model design:

Created a dataset group in AWS forecast console with features timestamp,sku_ids, region, and sales. trained the predictor model with autopredictor feature(autoML) and trained for 1 hour adding additional feature festivals to the dataset taken from April,18 to December 20 with the forecasted frequency of 1 month to predict the forecast of Jan,21 to June,21. The mape score of the trained model is 1.248 with an average weighted quantile loss value of 0.64.since the cost of being understock is often higher than the cost of being overstocked we calculated the WQL at 0.1 (P10)=0.24, 0.5 (P50)=0.764, 0.7 (P70)=0.856 and 0.9 (P90)=0.678.
(p[i] means the true value is expected to be lower than the predicted value i% of the time.).
->The impact score of the additional feature holidays on the forecasting is 0.4% which implies that to keep high inventory stock on festivals.



BUSINESS STRATEGY:

1) We have observed mainly 3 types of trends in the most popular SKUs in the given point of time as mentioned on page B.

- In retail, the cost of being understocked is often higher than the cost of being overstocked, and so maintaining more inventory for the percentage of stocks that are in great demand after covid. For maintaining stocks, we can use the metrics Weighted Quantile Loss (wQL) to avoid understock.
- Selling the low demand stocks at discounted price and marketing for the average demanded stocks to increase the sales of SKUs that are performing low may increase the sales and sell out the remaining stocks and maintaining a low inventory helps

2) Maintaining more inventories for the region specified

3) Maintaining the inventories of economical SKUs.

4) The impact score of the holidays on the forecasting is 0.4% which implies that to keep high inventory stock on festivals.