CFO Forecasting - Mobile Market - Sandesh Brand 2 - Final Challenge

By: Rishabh Singhal

Introduction:

My model is capable of predicting a wide range of values. I have submitted the **.ipynb** file that contains the code written in the language python 3.

For running this notebook, library requirements are - **Pandas**, **NumPy**, **Matplotlib**, **Sklearn**, **Statmodels**(for installation use **conda install -c conda-forge statsmodels**).

I have written separate codes for different variables starting from forecasting for - 'Equipment Revenue - Leopard', 'Average Revenue per existing customer (Equipment) - Leopard', 'Mobile Data Revenue - Leopard', 'Closing Base - Panther' in a serialized manner.

Let us label them as **variable1**, **variable2**, **variable3**, and **variable4**.

So before deciding the model for training them I have followed some basic steps for each variable step :

- Created a new dataframe d having columns time and sales
- For each variable, extract the horizontal row and store them in the sales column in the dataframe d.
- Store time corresponding to each value in the time column in the dataset d.
- Merging the Iphone factors launch date and affordability factor with the dataframe which in future will be used as an exogenous variable.
- Before merging them I have also eliminated multiple roles having the same launch date in terms of the affordability factor.

- After doing the previous step, I have converted the above column into an index by using set_index, after which time values will be the index of the dataframe d.
- After this, using the describe function I have analyzed the dataset.
- After this, I have plotted the dataframe d.

For variable1 :

- After plotting, we observe that the curve is following a seasonal pattern, so I used a SARIMAX model with exogenous variables to make the predictions.
- Each feature was analyzed using the Augment Dickey-Fuller test to determine if the data had a trend.
- I have also plotted prediction-actual value which we found out is very less, which validates our approach.
- I finally predicted the next 6-time stamp and stored it in x1,x2,x3,x4,x5 and x6 respectively.
- I have plotted the RMSE curve for the 6-time stamp.
- After this using sliding window for h=6, I got the result better than the previous one so, my final model is calculated using h=6 and I have not used september to february data for calculation of validation data (i.e. prediction). In my model we can calculate the new value and then use it for next prediction in case if the validation size is greater than 6, in our case we can predict using training data only.

For variable4 :

- After plotting the curve, I observed that the curve does not follow any pattern(from visual inspection).
- So, I used adfuller test to check if the curve is stationary.

- I want to check if the curve is stationary in order to check the validity of the ARIMA(**Autoregressive Integrated Moving Averages**) algorithm and hence obtain the predictions.
- I wanted to know if our data if following any seasonal relationship. Still, after using adfuller test on seasonal data, I found that this data was again not stationary and therefore the seasonal relationship is not present. I did this seasonal test for all, but I have shown only for this as for all data, this was found not useful because either data can be found stationary using other approaches or not found stationary in any of them.
- So for most of the data in different variables, I have used the First Shift differencing method to obtain the data stationary so that it can be used in further prediction.
- In this method rows of sales column is shifted by one and after this differencing between the rows takes place.
- After performing adfuller test on the above-obtained data, I obtained the result as data is stationary, so now we can use arima mode to obtain better results.
- After this, I plotted the autocorrelation correlation and partial autocorrelation figure. From which I find (p,d,q) values (order).
- From the above, I predicted using ARIMA for a wide range in which I found for a known dataset our model fitted relatively good as compared to other models.
- I also predicted for the next 6-time stamp and labeled it as sub2.
- For robustness calculation, similar steps were followed as previously.

For Variable2, Variable3:

- SARIMAX model used for the same prediction steps with a little variation in parameters were followed as they were followed for Variable1 but the sliding window approach is not used for prediction in case of Variable3.

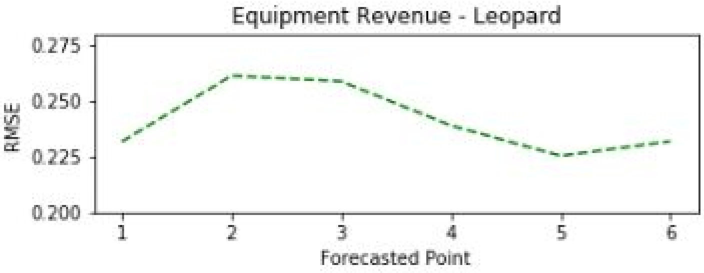
All of the prediction values are stored in the dataframe name **'Prediction'** in the format as asked in the problem statement and stored the values of tt in the **required format** and then downloading it as **'Prediction.xlsx'** file.

I have also submitted the required 6 sliding windows file.

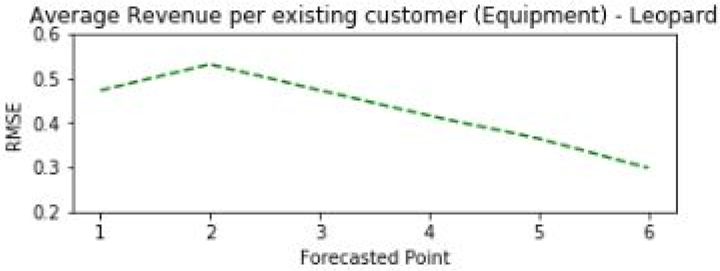
- Since there were two missing values in **variable2 validation data** I **extrapolated** them using the differencing method.

Results/Plots of some variables :

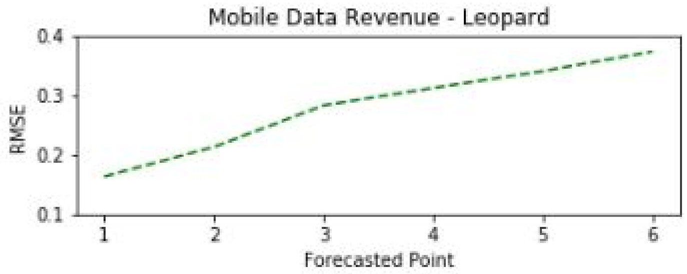
Variable1: (Robustness)



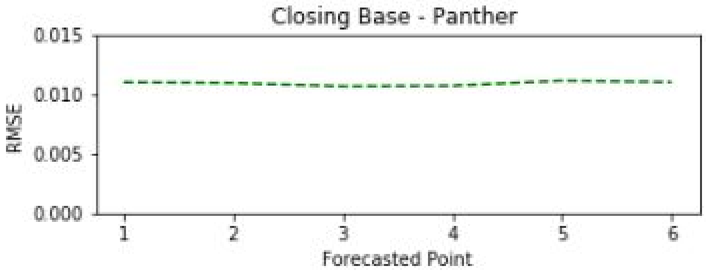
Variable2: (Robustness)



Variable3: (Robustness)



Variable4: (Robustness)


Closing Base - Panther
RMSE
Forecasted Point