Using Seasonal ARIMA to Improve Time series

Objective

The objective of this challenge is to generate high accuracy forecast for financial/volume variables for a telecommunication business. We are provided with a dataset of 10 time series going over 5 years. These financial variables are related to 3 products Leopard, Pather, and Hyena.
Description of products:
(i) **Falcon** – Sandesh's main broadband product (reaching maturity, available in most of the country, faster download speeds than the legacy product)

The financial/volume series are following:
(i) **Average Revenue per New Customer** – the average monthly revenue paid by new Falcon customers in the first month of the customer's contract
(ii) **Average Revenue per Existing Customer** – the average monthly revenue paid by all subscribers in the Falcon customer base per month for the service
(iii) **Gross Adds(Norm)** – the number of new subscribers to each individual product joining the brand during a month
(iv) **Net Migrations (Norm)** - The number of subscribers who remained with the brand but moved to another product. Usually and upgrade to faster broadband speed.

Gross Adds (Norm) and Net Migrations (Norm) are seasonal and discontinuous variables and vary significantly from month to month depending on competitor pressure at that point in time. ARPU New & Existing are continuous variables.

Exploratory Data Analysis

We have been provided with a anonymized dataset. Every data point in each time series is replaced by the Z-score. The formulas for the **Z-score** is
For all the variables, following is the formula used to privatize the data:

$$z_i = (x_i - \mu) / \sigma$$

where z_i = z-score of the ith value for the given variable

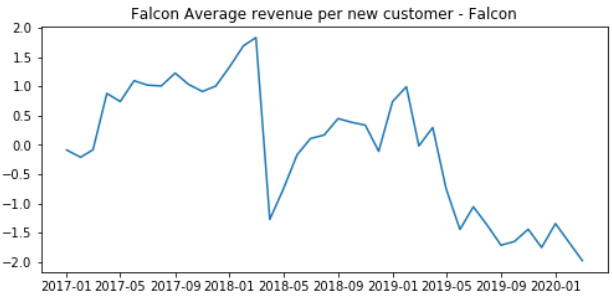
x_i = actual value

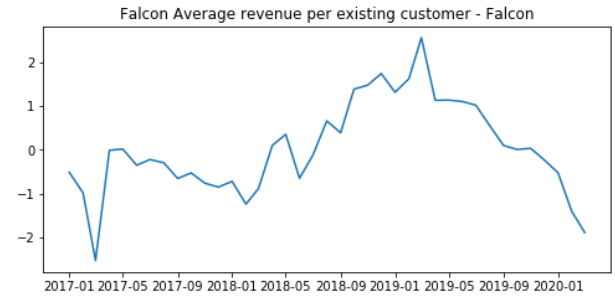
μ = mean of the given variable
σ = standard deviation for the given variable

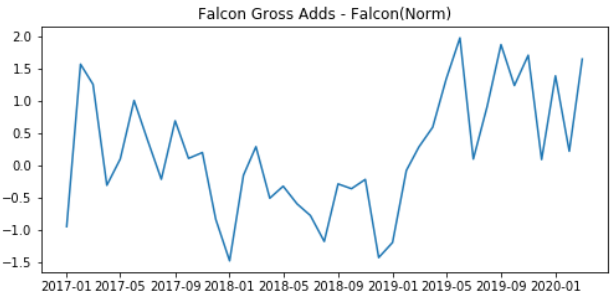
The dataset was given in excel spreadsheet form. It was processed before further data analysis.
Below is the snapshot of the dataset provided.

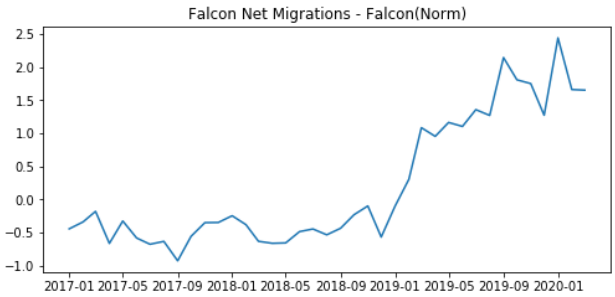
	Falcon Average revenue per new customer - Falcon	Falcon Average revenue per existing customer - Falcon	Falcon Gross Adds - Falcon(Norm)	Falcon Net Migrations - Falcon(Norm)
date				
2017-01-01	-0.088569	-0.509628	-0.950544	-0.444357
2017-02-01	-0.213121	-0.979150	1.563837	-0.339622
2017-03-01	-0.084838	-2.519182	1.251588	-0.180494
2017-04-01	0.877792	-0.011087	-0.310766	-0.664242
2017-05-01	0.738473	0.017668	0.095668	-0.325734

Let's draw a time series plot of each of the variable separately. Some of the variables have visibly high variance while some don't. We also see big anomalous data points in some of the time series. We see a big drawdown in mar-2018 for APNC and it's again normal by the next month. This definitely looks like an anomaly. We also see some small anomalies in Gross Adds. Only Net-Migrations show a clear cut trend which is upward.



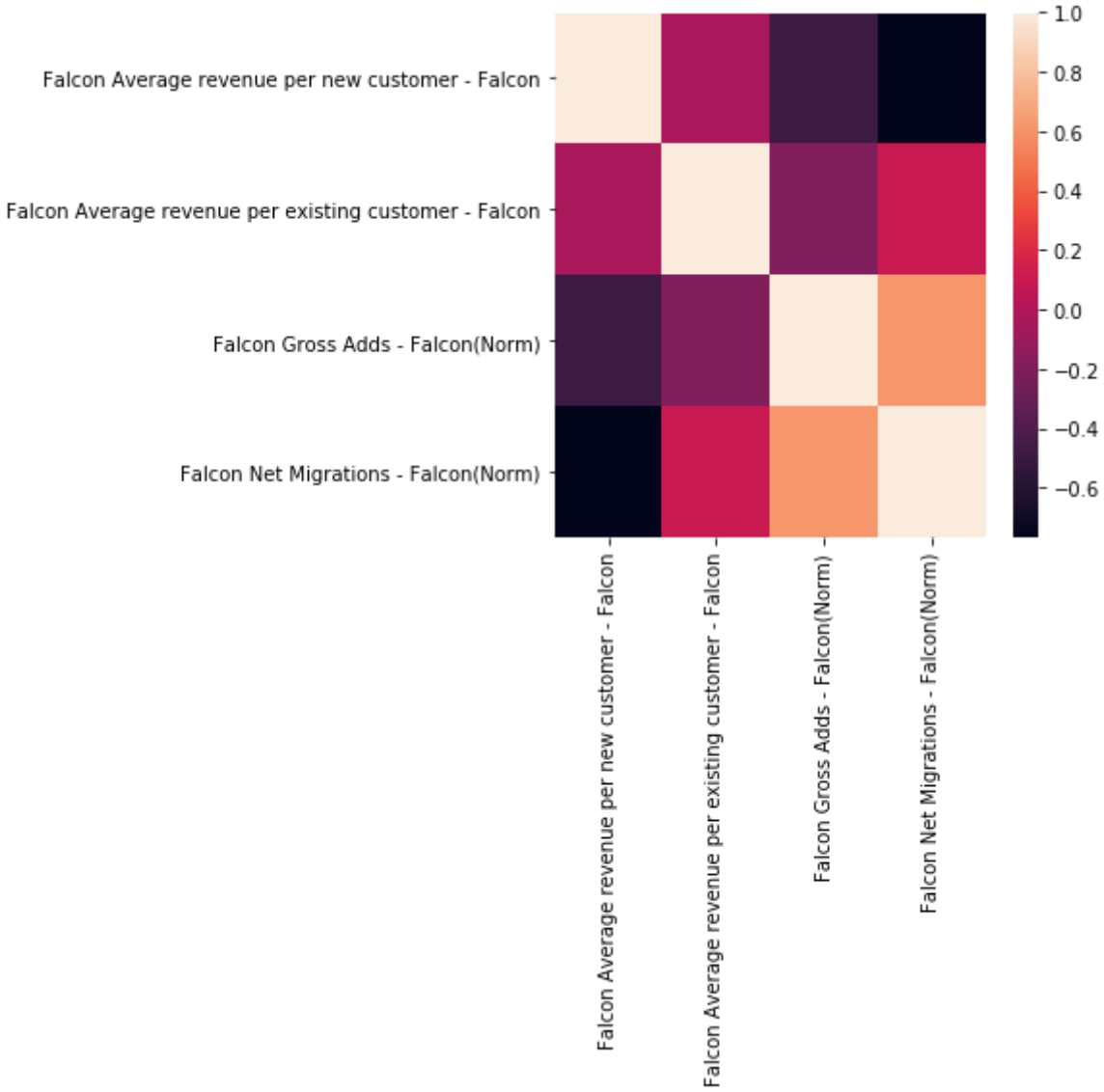






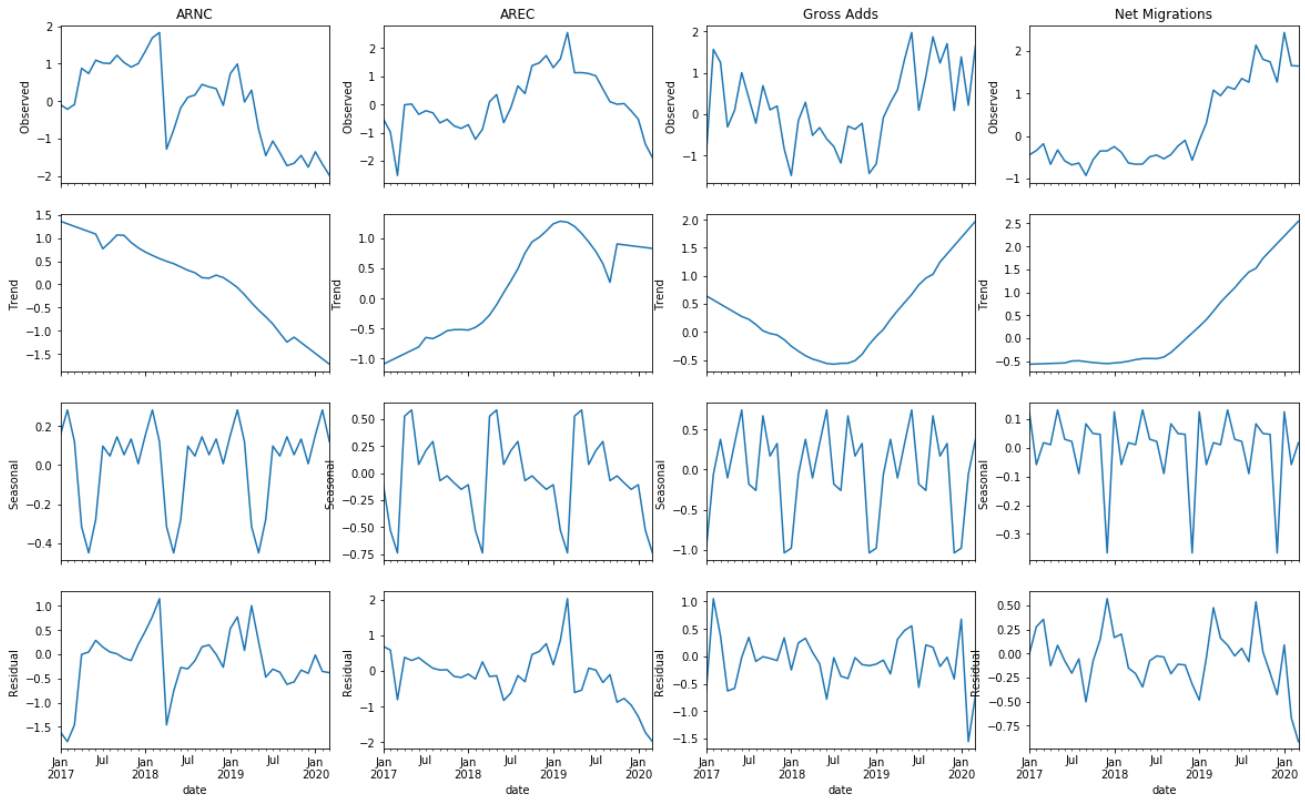
Correlation among the different time series

Moving further, Let's check the correlation among the different time series. We see that many time-series are heavily correlated with each other. For ex. APNC is heavily uncorrelated with Net Migrations.



Let's check for seasonal decompose

Analysis of Time series is incomplete without seasonal_decompose and analysis of seasonality, trend and residual values separately. There are two types of decomposition addtive and multiplicaative. since, our data contains negative and zero values, it is best suited for additive model. Except of ARNC, others time series don't exhibit a clear trend.. We see that yearly seasonality is clearly visible in all the time series variables.



Let's check Stationarity of the time series

Augmented Dickey-Fuller Test

Let's move to the next part of time series analysis. We will now check the stationarity of time series. When we say, a time series is stationary it means its statistical properties(mean, std dev e.t.c.) don't change over time. There are many standard tests for cheking stationarity of time series, we will be using ADF(Augmented dickey fuller Test).

ADF

The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary (has some time-dependent structure). The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary.

```
Null Hypothesis (H0): If failed to be rejected, it suggests the
time series has a unit root, meaning it is non-stationary. It has
some time dependent structure.
Alternate Hypothesis (H1): The null hypothesis is rejected; it
suggests the time series does not have a unit root, meaning it is
stationary. It does not have time-dependent structure.
```

We interpret this result using the p-value from the test. A p-value below a threshold (such as 5% or 1%) suggests we reject the null hypothesis (stationary), otherwise a p-value above the threshold suggests we fail to reject the null hypothesis (non-stationary).

```
p-value > 0.05: Fail to reject the null hypothesis (H0), the data
has a unit root and is non-stationary.
p-value <= 0.05: Reject the null hypothesis (H0), the data does
not have a unit root and is stationary.
```

We see that only Gross Adds is stationary while other have a unit root means non-stationarity.

	variable	ADF statistic	p-value
0	Falcon Average revenue per new customer - Falcon	-1.468396	0.549084
1	Falcon Average revenue per existing customer -...	-1.577903	0.494684
2	Falcon Gross Adds - Falcon(Norm)	-3.512264	0.007678
3	Falcon Net Migrations - Falcon(Norm)	-1.449689	0.558237

Auto-Correlation and Partial Auto-Correlation

ACF

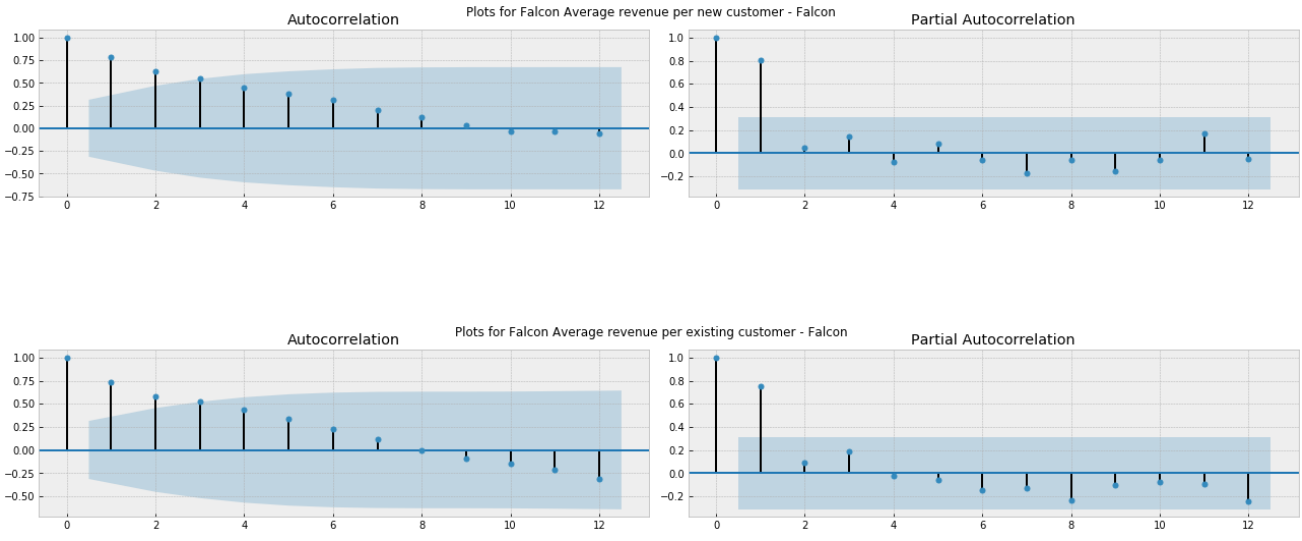
Autocoreelation gives us the amount of linear depenency between time series at index t and time series at indcies t+k or t-k. A zero auto correlation means that temporal dependencies are very small or negligible. A positive autocorrelation values simple means the present and future values move together in same direction while negative auto correlation implies the opposite.

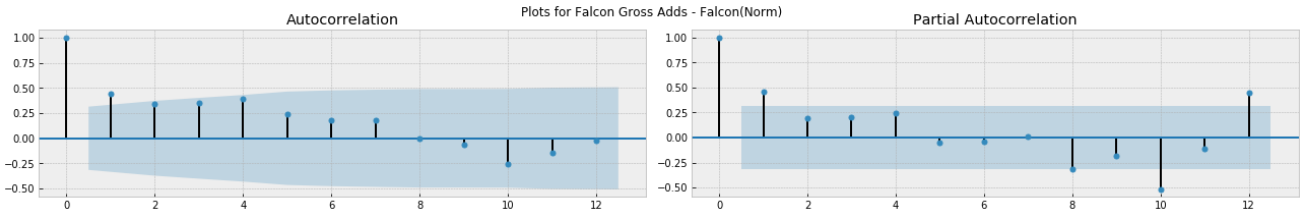
PACF

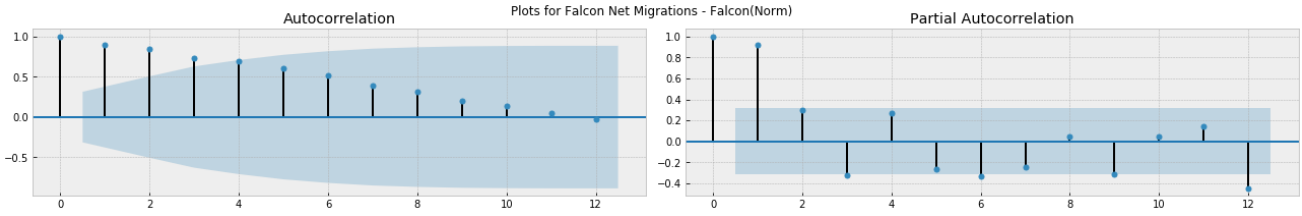
When we measure autocorrelation in time series between t and t+k, it gets influenced by time series at t+k-i. Therefore, autocorrelation is not the correct measure of the mutual correlation between x_t and x*(t+h) in the presence of the intermediate variables. Hence, it would erroneous to choose h in AR models based on autocorrelation. Partial autocorrelation solves this problem by measuring the correlation between x_t and x*(t+h) when the influence of the intermediate variables has been removed.
We plot the both auto correlation and partial auto correlation of the time series.

In the ACF plots, the lags which are outside of the shaded region are supposed to be significant and same is true of PCAF plots. It is evident from the plots that every series shows significant auto-correlation for lags=1, 2,3,.

Plots




Plots for Falcon Gross Adds - Falcon(Norm)
Autocorrelation
Partial Autocorrelation


Plots for Falcon Net Migrations - Falcon(Norm)
Autocorrelation
Partial Autocorrelation

SARIMAX

SARIMA: Seasonal Auto- Regressive Integrated Movig average model.

AR(auto-regressive) and **MA**(moving average) models are go-to tools for modeling time series especillay in business analytics. They are easy to intrepret which is important while making business decisions, since rationale behind those decision needs to be explained to management. They have another plus point that they work equally good with small amount of data like in our case.

Parameters in SARIMAX
The parameters of SARIMAX are explained below:

endog: array_like
The observed time-series process y

exog: array_like, optional
Array of exogenous regressors, shaped nobs x k.

order: iterable or iterable of iterables, optional
The (p,d,q) order of the model for the number of AR parameters, differences, and MA parameters. d must be an integer indicating the integration order of the process, while p and q may either be an integers indicating the AR and MA orders (so that all lags up to those orders are included) or else iterables giving specific AR and / or MA lags to include. Default is an AR(1) model: (1,0,0).

seasonal_order: iterable, optional
The (P,D,Q,s) order of the seasonal component of the model for the AR parameters, differences, MA parameters, and periodicity. D must be an integer indicating the integration order of the process, while P and Q may either be an integers indicating the AR and MA orders (so that all lags up to those orders are included) or else iterables giving specific AR and / or MA lags to include. s is an integer giving the periodicity (number of periods in season), often it is 4 for quarterly data or 12 for monthly data. Default is no seasonal effect.

trends: tr{'n','c','t','ct'} or iterable, optional
Parameter controlling the deterministic trend polynomial A(t). Can be specified as a string where 'c' indicates a constant (i.e. a degree zero component of the trend polynomial), 't' indicates a linear trend with time, and 'ct' is both. Can also be specified as an iterable defining the non-zero polynomial exponents to include, in increasing order. For example, [1,1,0,1] denotes a+bt+ct3

Parameter-Search: After analysizing the ACF and PACF plots, the maximum values of (p,d,q, P,D,Q) was fixed to 3 and minimum 1. For seasonality, we chose 3 values [3,6,12]. Then, we ran a grid search based parameters search to find the best parameters for each of the variable.

In the analysis above, we have seen that time series has some big anomalous points. We will use modified-Z score method(MAD) to remove these anomalous points and replace them with the next values.

MAD: Median Absolute Deviation also called Robust-Z score method
In this method, median of time series is calculated $\tilde{x}$. Then, median is subtracted from each value to get MAD. After that, modified-z score is calculated for the each point in the time series.

Then, a threshold is selected depending the amount of anomalous points. Any point greater than the threshold is removed and backfilled.
Using MAD, we got good results for all the variables.