**Daily Demand Forecasting**

1. **Main Objective**

Perform forecasting to urgent orders and non urgent orders using best model yielding lowest forecast error.

1. **Data Definition and Description**

This dataset provided by University of Calivornia Irvine machine learning repository that open public with educational purposes only. The data contained in the dataset was from real Brazilian company which has record of 60 days of daily orders. The data was inputted in 2017 late.

The dataset contain 60 training observations and 13 attributes (variables).

**Data description**

Week of the month (first week, second, third, fourth or fifth week : INT, value of encoded week order in one month.

Day of the week (Monday to Friday): INT,value of encoded workdays in the week

Non-urgent order: INT, amount of non urgent order

Urgent order: INT, amount of urgent order

Order type A: INT, no information

Order type B: INT, no information

Order type C: INT, no information

Fiscal sector orders : INT, no further information

Orders from the traffic controller sector : INT , no further information

Banking orders (1) : INT , no further information

Banking orders (2) : INT , no further information

Banking orders (3) : INT , no further information

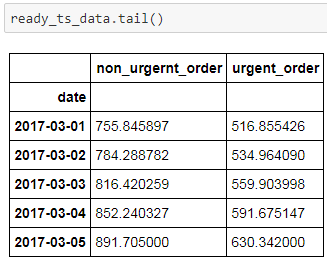
Target (Total orders) : INT , total orders accumulated

**Research Objective**

Built three model that can forecast the non urgent and urgent order with comparison and select the best one with elaboration of the choosen model.

1. **Pre-processing and Feature Engineering Plan**

* Importing required tools and library
* Inspect data and column type
* Inspect null values and fix with dropping them
* Renaming the column to suitable name
* Due to very limited datetime information, i’ll split the data set per month
* Group the data by week of month value
* Merge/concat the aggregated week value on rows. Hence yielding ten weeks
* Generate weekly datetime data with 10 weeks periods, add into the dataset as column
* Set the index to weekly datetime
* Resample the dataset from weekly to daily, yielding 64 obsevation
* Interpolate with spline method with degree 2

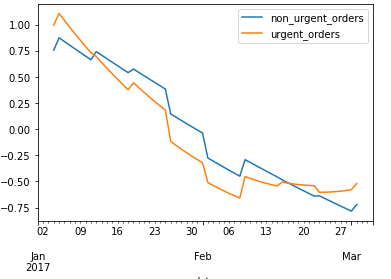
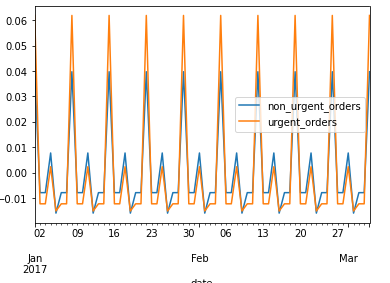
****

* Test the stationarity of the data with augmented dickey-fuller test

With H0 is the data is non stationary :

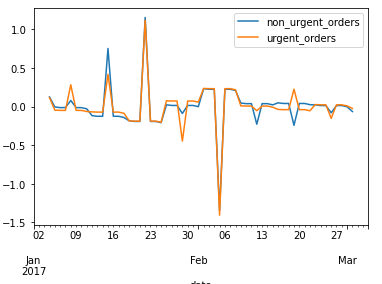
****

* Decompose the data, yielding trend, seasonality, and residual

****

**(Seasonality)**

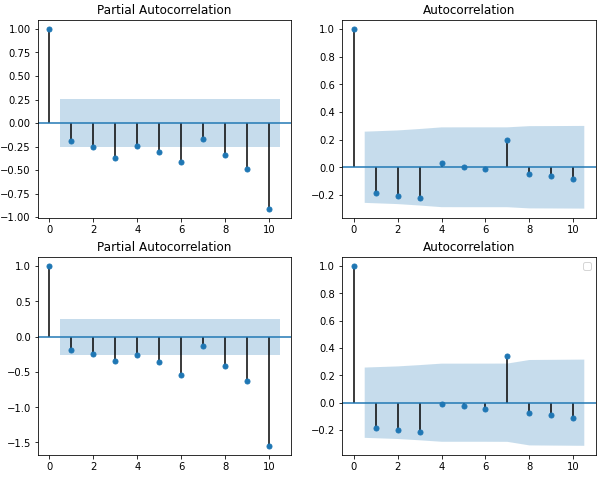
**(Trend)**

****

**(Residual)**

After decomposing

****

* since there is no difference between after and before log transformation, no need to perform log transformation.
* Prepare to forecast modeling, need to seek model component (AR(p), I(d),MA(q))
* Perform Auto Correlation Function and Partial Autocorrelation Function Plot to determine model component
* ARIMA modeling
* SARIMA modeling
* Comparison of ARIMA and SARIMA
* LSTM deep learning implementation preparation
* Create sequencing function
* Split data into train and validation set
* Build Neural network architecture

1. **Model Selection and Training Results**

To achieve best forecast result that match the business objective, we will conduct three different model

**Modeling**

1. **ARIMA**

**Components :**

p = 1

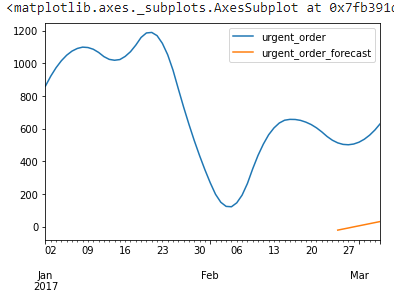
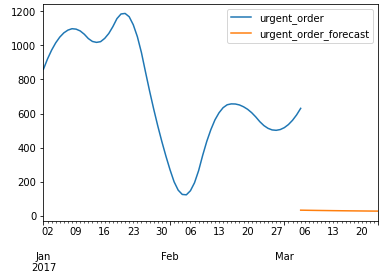
d = 1

q = 0

MSE = 337845.5

**On Validation**

**On Forecast**



1. **SARIMA**

**Components :**

p = 1

d = 1

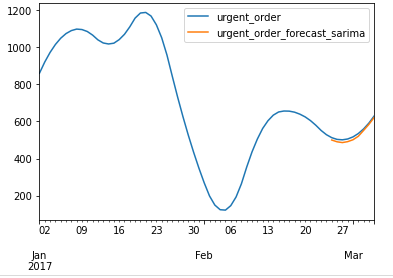
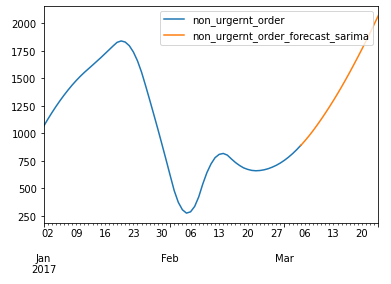
q = 0

seasonal = 7

MSE = 417

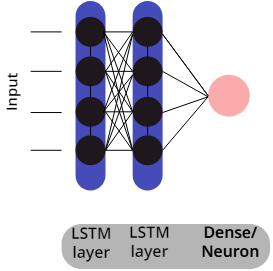
**On Validation**

**On Forecast**



1. **LSTM**

**Architecture :**



* + - Input layer : 3 sequence
    - LSTM layer 1 :
* Activation : ReLU
* Cell : 100
* Returned sequence
  + - LSTM layer 2 :
* Activation : ReLU
* Cell : 50
  + - Dense :
* Cell : 1
* Activation : none

**Neural network configuration**

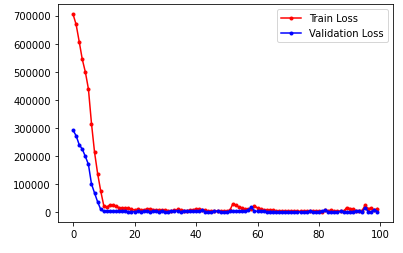
Optimizer : adam

Epoch : 100

Batch size = 10

Loss = MSE

Learning rate = 0.03



Score :

Train MSE : 12705

Validation MSE = 789

1. **Model Selection, Conclusion, and Summary**

Based on the model version and attempt, we would suggest to use SARIMA model 3 and choose it as the best model in this research with the result ;

* MSE = 417

From now on, this model will be mentioned as choosen model.

Due to best forecasting main objective, the model training will focus on how to forecast the timeseries without suffering overfit. From the chosen SARIMA model **we can conclude below insights that drive accuracy performace**:

1. The importance of correct p,d,q,m components will be main driver to the SARIMA
2. Trend, Seasonality, and Residual inspection with decomposing will help understand the observed timeseries
3. Autocorrelation plot take a huge part of deterimining model componets
4. Hopefully the **order will increase as the forecast shown for the next 19 days**

**Further Development**

1. Dataset flaw

* Observation is too small to perform forcasting
* Lack of time information
* Not suitable for neural network model that need a lot of observations

1. Futher possible research

With dataset open-sourcity and broader techniques, any researcher can expand this report or conduct different research purposes, exploratory data analysis,Timeseries analysis, and develop other suitable model. Highly encouraged to review author’s notebook and source code on this dataset.

* Need further experiment on deep learning model used like RNN, LSTM, or probably any model that suit this dataset
* Need further experiment on ‘best’ epoch number
* Need further experiment on another activation dunction
* Need more development on deep learning architecture
* Need to collect a lot more observation.
* Clarify datetime information

**References**

Author’s notebook and source code

<https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/bfb7f92c-ef74-4f5d-a01d-4175b8a6d862/view?access_token=8165ef321fd8186e11b0dd17bb4b95ff90f8226b53eee33b6a23b738aa3587b9>

datasets

<https://archive.ics.uci.edu/ml/datasets/Daily+Demand+Forecasting+Orders>