WEBSITE TRAFFIC FORECASTING

ABOUT

Website Traffic Forecasting means forecasting traffic on a website during a particular period. It is one of the best use cases of Time Series Forecasting. If you want to learn how to forecast traffic on a website, I will take you through the task of Website Traffic Forecasting using Python. The dataset I am using for Website Traffic Forecasting is collected from the daily traffic data of thecleverprogrammer.com. It contains data about daily traffic data from June 2021 to June 2022.

ML MODEL ARIMA

ARIMA stands for autoregressive integrated moving average model and is specified by three order parameters: *(p, d, q).*

- **AR(*p*) Autoregression** – a regression model that utilizes the dependent relationship between a current observation and observations over a previous period. An auto regressive (*AR(p)*) component refers to the use of past values in the regression equation for the time series.
- **I(*d*) Integration** – uses differencing of observations (subtracting an observation from observation at the previous time step) in order to make the time series stationary. Differencing involves the subtraction of the current values of a series with its previous values d number of times.
- **MA(*q*) Moving Average** – a model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. A moving average component depicts the error of the model as a combination of previous error terms. The order *q* represents the number of terms to be included in the model.

Types of ARIMA Model

- **ARIMA:** Non-seasonal Autoregressive Integrated Moving Averages
- **SARIMA:** Seasonal ARIMA
- **SARIMAX:** Seasonal ARIMA with exogenous variables

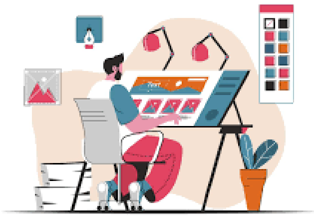
IMPLEMENTATION

Steps:

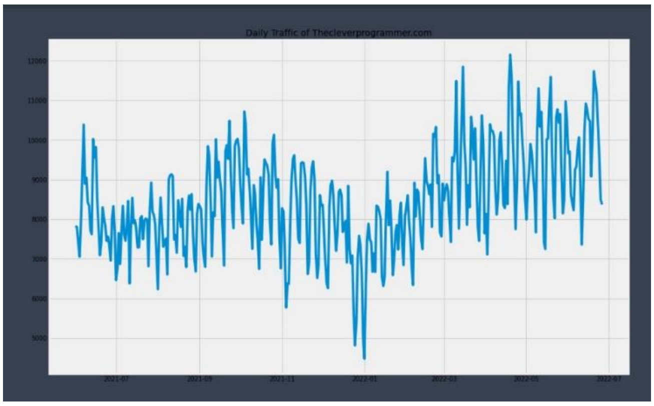
1. Data Pre-processing

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 391 entries, 0 to 390
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    391 non-null    datetime64[ns]
 1   Views   391 non-null    int64
dtypes: datetime64[ns](1), int64(1)
memory usage: 6.2 KB
None
```

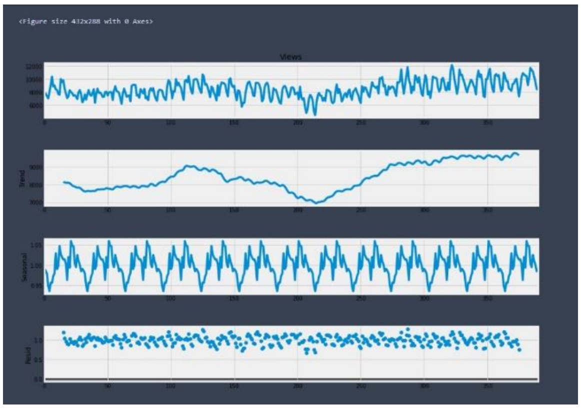
The Date time column was an object initially, so I converted it into a Datetime column. Now let's have a look at the daily traffic of the website:



2. Daily Traffic of the Website



3. Seasonal or Stationary? How to decide?



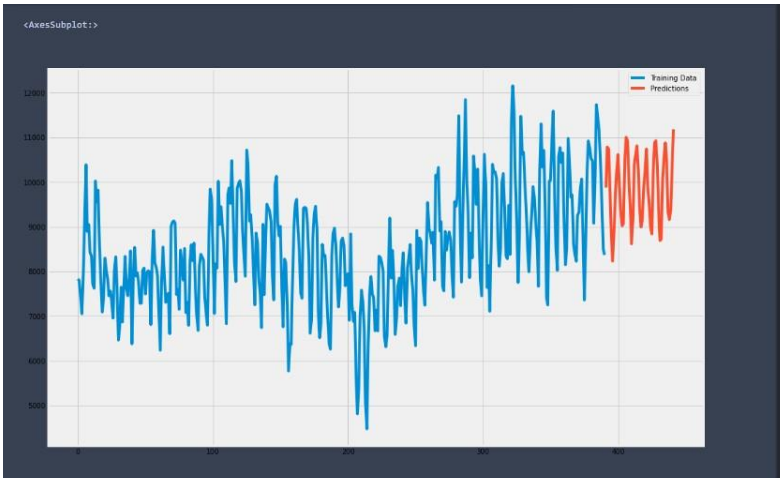
4. Fitting the Model

```
C:\python 3.1\lib\site-packages\statsmodels\tsa\statespace\sarimax.py:978: UserWarning: Non-invertible starting MA parameters found. Using
zeros as starting parameters.
  warn('Non-invertible starting MA parameters found.'
C:\python 3.1\lib\site-packages\statsmodels\base\model.py:604: ConvergenceWarning: Maximum Likelihood optimization failed to converge. Che
ck mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "

                                 SARIMAX Results
==========================================================================================
Dep. Variable:                              Views   No. Observations:                  391
Model:             SARIMAX(5, 1, 2)x(5, 1, 2, 12)   Log Likelihood               -3099.055
Date:                            Thu, 24 Nov 2022   AIC                           6228.110
Time:                                    17:26:13   BIC                           6287.134
Sample:                                         0   HQIC                          6251.536
                                            - 391
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7755      0.131      5.905      0.000       0.518       1.033
ar.L2         -0.7906      0.135     -5.837      0.000      -1.056      -0.525
ar.L3         -0.1421      0.171     -0.829      0.407      -0.478       0.194
ar.L4         -0.1938      0.152     -1.274      0.203      -0.492       0.104
ar.L5         -0.1462      0.138     -1.062      0.288      -0.416       0.124
ma.L1         -1.1783      0.089    -13.310      0.000      -1.352      -1.005
ma.L2          0.8952      0.075     11.984      0.000       0.749       1.042
ar.S.L12      -0.2486      4.303     -0.058      0.954      -8.681       8.184
ar.S.L24       0.0546      0.624      0.087      0.930      -1.169       1.278
ar.S.L36      -0.1816      0.272     -0.669      0.504      -0.714       0.351
ar.S.L48      -0.2075      0.876     -0.237      0.813      -1.924       1.509
ar.S.L60       0.0134      0.884      0.015      0.988      -1.718       1.745
ma.S.L12      -0.6957      4.304     -0.162      0.872      -9.131       7.739
ma.S.L24      -0.1124      3.478     -0.032      0.974      -6.930       6.705
sigma2      1.257e+06   1.65e-05    7.6e+10      0.000    1.26e+06    1.26e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 1.30
Prob(Q):                              0.98   Prob(JB):                         0.52
Heteroskedasticity (H):               1.05   Skew:                             0.14
Prob(H) (two-sided):                  0.80   Kurtosis:                         3.01
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.79e+27. Standard errors may be unstable.
```

5. Prediction of traffic for next 50 days

```
391     9876.249960
392    10784.402565
393    10741.832667
394     9865.688905
395     8781.939635
396     8230.529838
397     8931.974265
398     9696.989182
399    10283.824837
400    10614.788376
401     9881.974191
402     9352.863745
403     9019.906543
404     9071.109798
405    10512.177739
406    11003.050671
407    10920.348709
408    10093.767698
409     9432.806063
410     8618.436924
411     9178.692824
412    10364.140537
413    10621.282266
414    10812.063211
415    10269.353856
416     9424.469604
417     8992.843537
418     9155.052259
419     9902.925685
420    10246.994097
421    10739.876122
422     9908.288767
423     9519.429744
424     9014.893218
425     8839.941769
426    10165.034082
427    10876.260491
428    10925.804208
429    10397.342367
430     9430.235448
431     8688.542830
432     8725.235150
433    10082.548679
434    10546.733531
435    10879.087860
436    10466.283380
437     9322.827613
438     9160.891132
439     9361.106546
440    10313.171166
441    11180.724145
Name: predicted_mean, dtype: float64
```

6. Result



CONCLUSION

To summarize, a model has been created to predict web traffic for the next 50 days. The original data needed some clean up and some feature engineering as well. Deciding which method to use for solving this problem was a difficult and critical one because there are many techniques available which are popular for e.g. ARIMA, SARIMA, XGBoost, LightGBM, LSTM and libraries such as Facebook's prophet. In the end, it was a good decision to go with ARIMA as it provides a good groundwork for understanding other time series analysis techniques and is easier to implement than some other techniques.
The other challenge was to be able to visualize these results in concise manner to get a good overview how the models are performing. The plots for actual predictions and the errors painted a good picture of how the models perform.