INDIAN INSTITUTE OF TECHNOLOGY PATNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

PROJECT REPORT

CS575 - APPLIED TIME SERIES ANALYSIS

BY - KOMAL PANDYA(2021CS17)

APRIL 2021

TABLE OF CONTENTS

DATASET USED FOR THE PROJECT -
https://www.statsmodels.org/devel/datasets/generated/sunspots.html

(Astronomical datasets consist of a wide range of exploration possibilities and multiple objectives can be fulfilled using a single dataset.)

OBJECTIVES -

1. **Stationarity Tests -**

   These are the tests used to check whether the time series is stationary or non-stationary using different measures.
   There are many methods (direct observations, residuals,etc.) to check whether a time series is stationary or non stationary. If the series is identified as non-stationary, then it has to be made stationary for modelling with different models like AutoRegressive and Moving Average. Also,many statistical models require the series to be stationary to make effective and precise predictions.
   The series can be made stationary using various methods like differencing the series, taking log of the values,etc.

2. **Modeling time series -**

   Time series are modeled to perform various tasks such as predicting future values, studying routine behaviour and identifying other important factors. There are various Models that are used for this, where each one of them exploits one property or others for making their decisions. THere are some models that require time series to be stationary, while for others, it's not mandatory to do so.

3. **Clustering -**

   It's a task of grouping similar data points together. This can be done by matching the properties of one data point with the other. IF they are found to be somewhat relative, they can be put into the same cluster. By clustering, their behavior under normal conditions can be deduced.

4. **Detecting Outliers-**

   Outliers are the data points that are unusual with respect to other data points of the series. Detection of those outliers are important as they introduce some irregularities in the dataset. Sometimes they affect the data analysis and statistical modeling. Therefore, detection of outliers help in finding any abnormal behaviour happening with the data series.

METHODOLOGY

1. STATIONARITY TESTS

- Data VIsualization -

  From the curve, it can be identified if data is stationary or non stationary(showing some trends, cyclicity, seasonality,etc.). This is a very simple method to check by observation, for stationarity, without imposing any statistical analysis on the data.

- Comparing Means -

  Another test is comparing statistical properties like mean,etc for different partitions of the same time series. If they are nearly constant, then the series is said to be stationary.

- ADF and KPSS tests-

  ADF test is used to determine the presence of unit root in the series, and hence helps in understanding if the series is stationary or not. The null and alternate hypothesis of this test are:

  Null Hypothesis: The series has a unit root.

  Alternate Hypothesis: The series has no unit root.

  If the null hypothesis is failed to be rejected, this test may provide evidence that the series is non-stationary.

  KPSS is another test for checking the stationarity of a time series. The null and alternate hypothesis for the KPSS test are opposite that of the ADF test.

  Null Hypothesis: The process is trend stationary.

  Alternate Hypothesis: The series has a unit root (series is not stationary).

  For both the tests, the statistics value and critical values are compared to make the decision.

2.**MODELING TIME SERIES-**

- ARIMA-

  After plotting the lag plot, the series was identified to be non-stationary. It was made stationary using differencing. From acf and pacf curves, the parameters(p,d,q) for AR, MA, ARIMA could be identified. For better results, parameters have been identified using GridSearchCV(with least MSE).

- Seasonal ARIMA-

  This ARIMA model was used to model the seasonal pattern occurring in the dataset.The criteria used here was 'least AIC'. For fitting the seasonal ARIMA model, the best parameters were chosen having least AIC value for the model, amongst others.

- Deep Learning Approaches-

  Statistical models need some constraints for modeling, like some require the data to be stationary. To capture the complete essence, deep learning techniques have been used here.

  1. *Deep Neural Network* - it's applied by considering time series as a linear model. {X(i) ...X(i+t)}~Y(i+t+1). In the format, it shows using t steps input time-series to predict the next step which is Y(i+t+1).

  2. *Recurrent Neural Network* - RNN takes time series values as sequences and every neuron is assigned to a fixed time stamp value. Hence at any given time value, only the neurons assigned to that time stamp can take part in their activation.

  3. *Long Short Term Memory* - LSTM is constructed using 4 main components - Input Gate(controlling which information should be added), Memory Cell(containing value that can be removed or refreshed), Output Gate(controlling selection of useful information) and Forget Gate(controlling the information not needed anymore). The dataset is first scaled and then LSTM was applied.

3. CLUSTERING-

- KMeans Clustering-

  It's the most widely known clustering algorithm. It involves assigning data points to clusters in an effort to minimize the variance within each cluster. The number of clusters have been decided using the 'Elbow Method'. It's then applied to the KMeans algorithm to visualize the clusters.

- Agglomerative Clustering -

  This is a bottom-up approach: each observation starts in its own **cluster**, and pairs of clusters are merged as one moves up the hierarchy.

4. DETECTING OUTLIERS-

- Interquartile Range Method -

  This is a statistical based test to find the number of outliers in the series. Data is first divided into quartiles. A quartile divides the data into 3 points and 4 intervals. Interquartile range(IQR) is the difference between the third quartile and first quartile.(IQR = Q3-Q1). Outliers then can be defined as the observations that are below (Q1 − 1.5x IQR) or above (Q3 + 1.5x IQR).

- Standard Deviation Method-

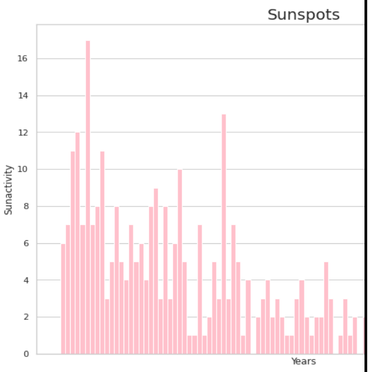
  Standard deviation refers to the spread of individual data points from the mean.If a data distribution is approximately normal then about 68% of the data values lie within one standard deviation of the mean and about 95% are within two standard deviations, and about 99.7% lie within three standard deviations. After measuring the upper bound and lower bound values of standard deviation, points not falling in range, can be considered as outliers.

- Z-Score Method -

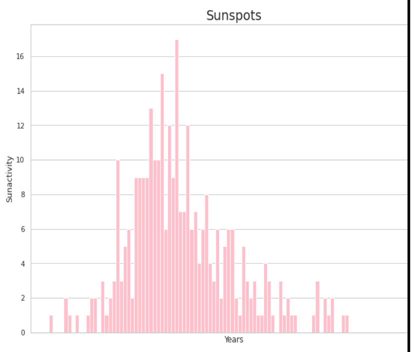
  The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. While calculating the Z-score, re-scale and center the data and look for data points which are too far from zero. These data points which are way too far from zero will be treated as the outliers. In most of the cases a threshold of 3 or -3 is used i.e if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

RESULTS AND ANALYSIS

1. **Stationarity Tests(Before)-**

Tests	Results	Analysis
1. VIsualization		The plot shows that the data is unevenly distributed and hence , it can be considered as a non-stationary series.
2. Mean	*Mean of partition 1* - 45.269903 *Mean of partition 2* - 62.150485	The significant difference in both means indicated non-stationarity of the dataset.
3. ADF and KPSS tests	*ADF* - ADF Statistics: -2.837780724938198 p-value: 0.05307642172812019 Critical Values: 1% -3.4523371197407404 5% -2.871222860740741 *KPSS* - KPSS Statistics: 0.6698662984667937 p-value: 0.01628488195756421 Critical Values: 10% 0.347 5% 0.463	For ADF, the value of ADF statistics is greater than 5% critical value, making the series non-stationary. FOr KPSS, the value of KPSS statistics is greater than 5% critical value, making the series non-stationary.

2. Stationarity Tests(After differencing the series)-

Tests	Results	Analysis
1. VIsualization	 	The plot shows that the data is now evenly distributed.This can be visualised from the curve which follows the gaussian distribution(bell-shaped curve) and hence , it can be considered as a stationary series now.
2. Mean	*Mean of partition 1 -* 0.392157 *Mean of partition 2 -* -0.262745	The means of both partitions are almost similar(centred around mean), having negligible difference, making the series to be stationary.
3. ADF and KPSS tests	*ADF -* ADF Statistics: -14.861663428129384 p-value: 1.715552423167133e-27 Critical Values: 1% -3.4523371197407404 5% -2.871222860740741 *KPSS -* KPSS Statistics: 0.047507174306859155 p-value: 0.1 Critical Values: 10% 0.347 5% 0.463 2.5% 0.574	For ADF, the value of ADF statistics is less than 5% critical value, making the series stationary. FOr KPSS, the value of KPSS statistics is less than 5% critical value, making the series stationary.

3. Modeling Time Series

- After finding the parameters(p,d,q) for AR(p), MA(q) and ARIMA(p,d,q) using GridSearch, the values obtained with least MSE are as follows -

  p - 9, d - 2, q - 3.

- For seasonal ARIMA, the parameters (P,D,Q,s) were found to be (2,2,2,2) respectively.

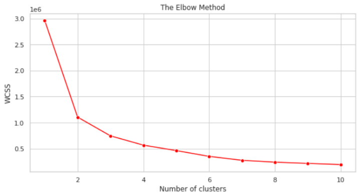
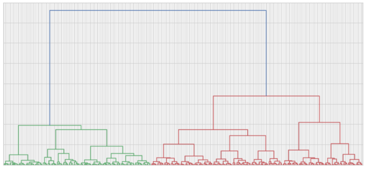
- The data was split into train and test sets with 2:1 ratio.

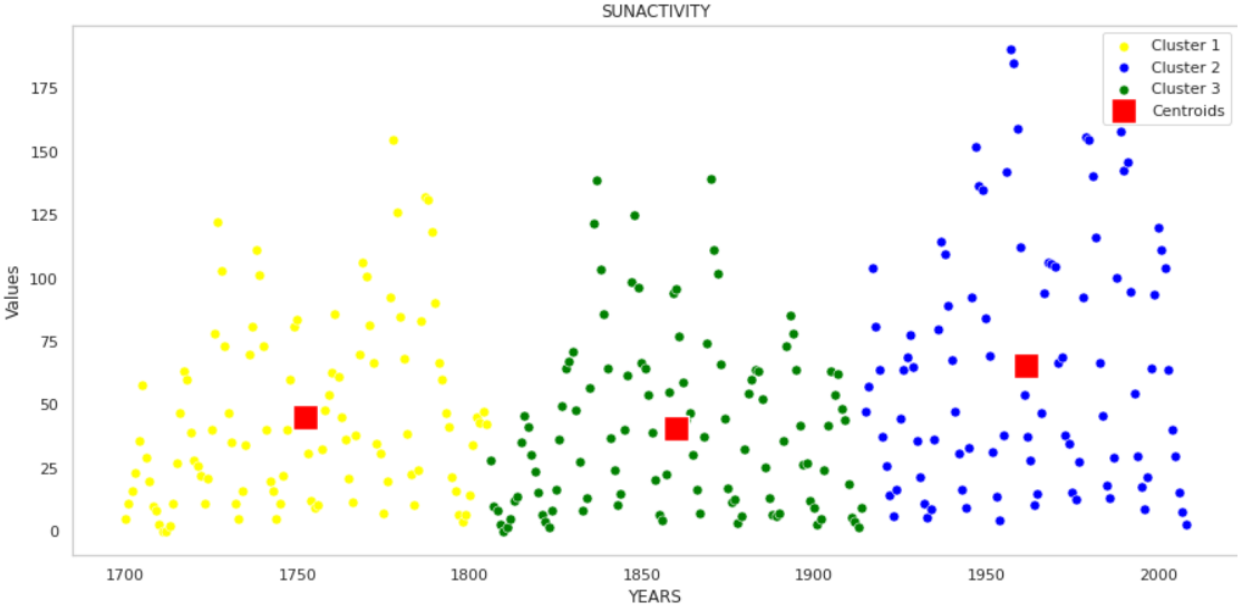
- The test MSE for each of them is shown here.

Model	MeanSquared Error
AR(9)	293.047
MA(3)	490.999
ARMA(9,3)	324.704
ARIMA(9,2,3)	300.941
SARIMA(2,2,2,2)	248.863
DNN	226.502
RNN	0.0144
LSTM	0.0121

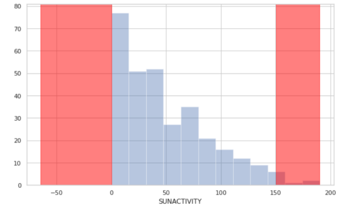
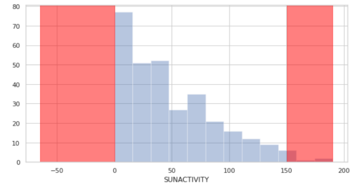
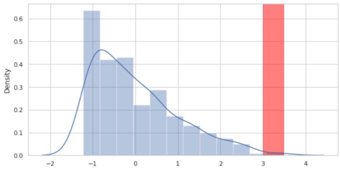
From the above table, certainly LSTM outperforms other models that were implemented.

4.CLustering -

ALgorithm	Basis	Analysis
KMeans		The number of clusters were identified to be 3 using the elbow method.
Agglomerative		The number of clusters were identified to be 3 using the dendrogram.



5. Outliers

Tests	Results	Analysis
IQR MEthod	**The IQR** is 53.8 **The lower bound value** is -64.69999999999999 **The upper bound value** is 150.5 	The number of outliers are 8, as calculated.
Std Deviation method	**The lower bound value** is -71.60568131066171 **The upper bound value** is 171.10988843040286 	The number of outliers are 2, as calculated.
Z-score Method (threshold of 3 used)		The number of outliers are 2, as calculated.

CONCLUSION

1. The SUNSPOTS dataset from statsmodels was initially found to be non stationary using various tests.

2. For models like AR and MA, the data needs to be stationary. Hence it was made stationary using differencing.

3. Various statistical and deep learning models were then applied to the data and mean squared error for each of them was reported.

4. The scatter plot of the original time series showed some possibility for clusters. Hence, clustering implementation was done using algorithms like KMEans and Agglomerative clustering.

5. As the data was having a high variance, there could be some outliers too. Approaches like IQR Method, Z-score method, etc were utilized to find and report the number of outliers.