PROJECT REPORT

DAB 400- Supply Chain Analytics

Group 4

Group Members :

Amith John Varkey

Bhavya Vinod

Delta Joseph

Mohammad Hashim

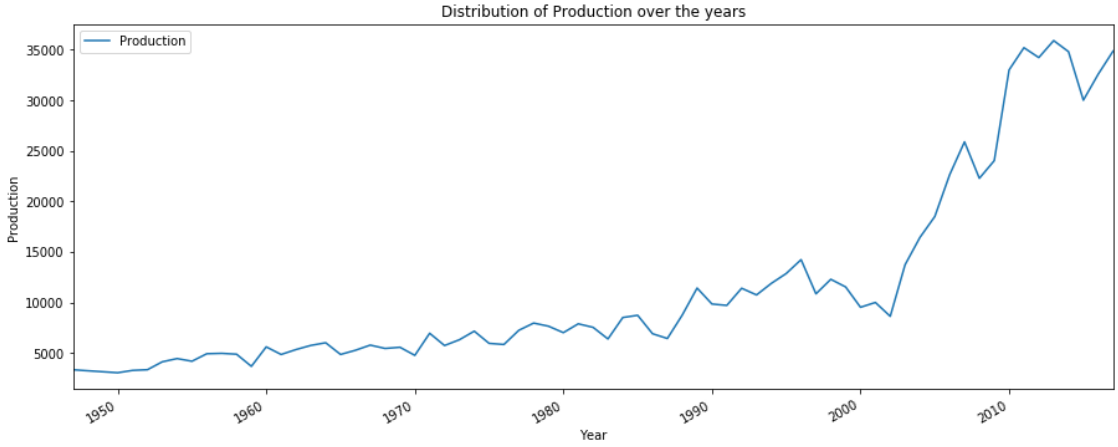
A Case Study on Forecasting Cotton Production in India

Introduction

This is a case study dealing with India's cotton Production, which is one among the largest producer in the world. Cotton Production is done in more than twelve states in India and state "Maharashtra" has the highest production among the twelve. The city" Mumbai" in Maharashtra is known as "**Cottonpolis of India**" .The crop is cultivated in more than 125 Hectares.

Here the data that we got is a time -series dataset. This is a yearly dataset from 1947 to 2017.The data consists of 69 rows and 5 features. The various features include Area, Production, Kg/Hectare and irrigation and also the Year column. Since this project is based on **Time Series Analysis** we are only going to deal with Year and Production column.

Time series deals with sequence of data points in order of time. It's often plotted in line graphs. Time series analysis involves analysing time series data to get meaningful insights from it. Moving average is the data point analysis, generating series of average of various subsets of the entire dataset. It is used in time series data to roll out short-term trends and high spot long term pattern.



So the above graph shows how India's cotton production was in the past 69 years. As you can see above the production was increasing over the years.

Methodology

First we found out whether the data is stationary or not by using the **Dicky Fuller test** and from the values we was able to understand that that data was not stationary as the test statistics value was greater than the 5% critical value.

In this project we have used normal prediction methods like moving average method and exponential smoothing techniques like simple exponential smoothing and double exponential smoothing technique. **Exponential smoothing** is the method of smoothing time series data. When there is no clear pattern, we can use exponential smoothing for forecast. In simple exponential data the previous observations have equal weightage, whereas it doesn't deal good with trends in data. Double exponential deals with trend component at each level.

We have also used different machine learning regression models like Linear regression, K-Neighbor regressor, Random forest regressor, Decision tree regressor, Gradient Boosting regression, Prophet model and also ARIMA model. We also used deep learning method like LSTM model in order to predict the values for cotton production. A short explanation for each models are as follows:

- **Linear regression** : It is a supervised machine learning algorithm which perform regression. The model finds the relation between independent variable and dependent variable and also used for forecasting.
- **Random forest Regressor**: It is a supervised machine learning model for classification and Regression. Problem. Using Random forest, we can build multiple decision trees and can compare their prediction all together a make an accurate prediction.
- **Decision Tree regressor**: It is used for building classification and regression. The model breaks data into subset and form tree structure. It is also a predictive modelling approach.
- **K- Nearest Neighbors** : It is an supervised machine leaning model used for classification and regression problem. KNN is used to estimate the data point seems in the group and also used estimate nearest point which is related.
- **Gradient Boosting Regression** : It is used for boosting week prediction model which makes use of decision trees.
- **Auto Regressive Integrated Moving Average (ARIMA)** : It describes a time series data upon the previous values and also to forecast the future data points. The application of the model is done when there is no stationarity in data points.
- **Prophet** : It is machine learning model used forecasting time series data. The model forecast performance and flags issue. It also compares simple and advanced forecasting methods and determine the performance.
- **LSTM**: Since problems of time series prediction is tough, recurrent neural networks are developed to tackle sequence dependence. LSTM (Long Short Term Memory) is used to handle long patterns or sequences in data, which can be easily trained using this method.

Observations

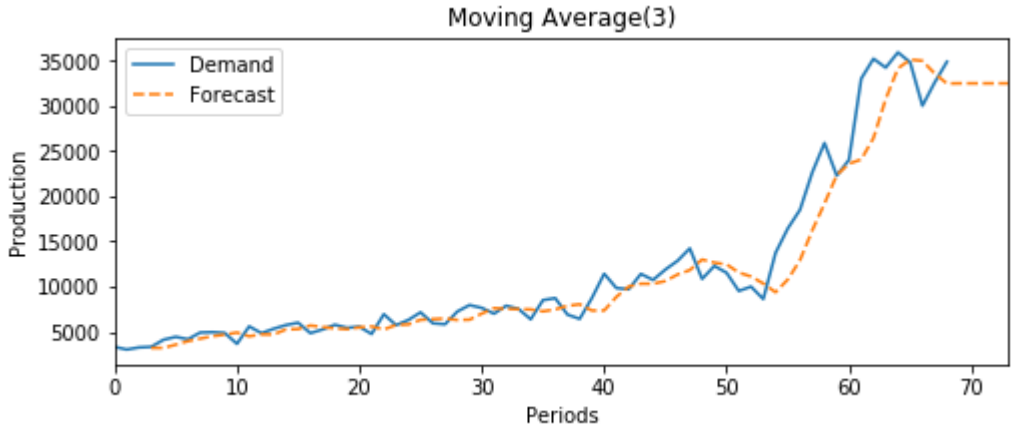
Here we have listed the graphs and also the values that we got for MAE and RMSE for different models.

Normal Prediction Methods

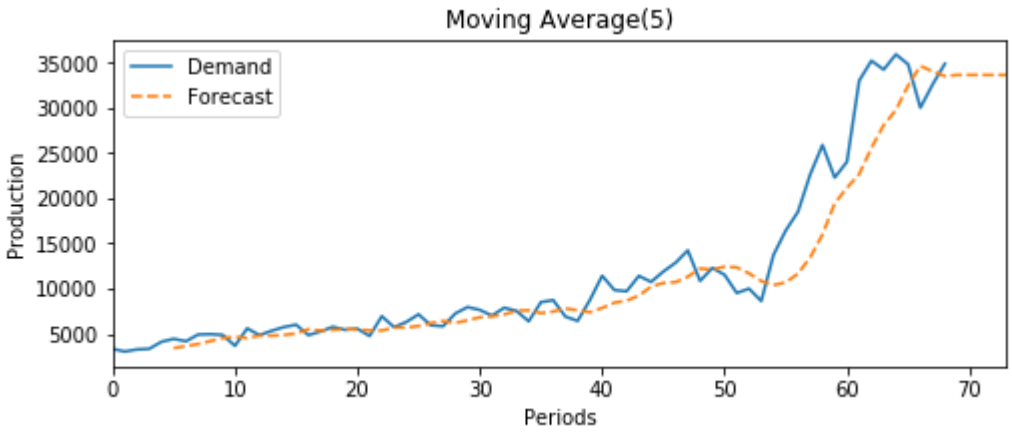
Moving Average method (MA) prediction Result

Here we have created a function which takes in the data frame, number of periods which needs to be predicted and also the value for n which is the number of years it takes in. So in this project we have done for 3 years Moving Average and also for 5 years Moving Average. So the graphs and result for these are shown below.

3 Year moving Average



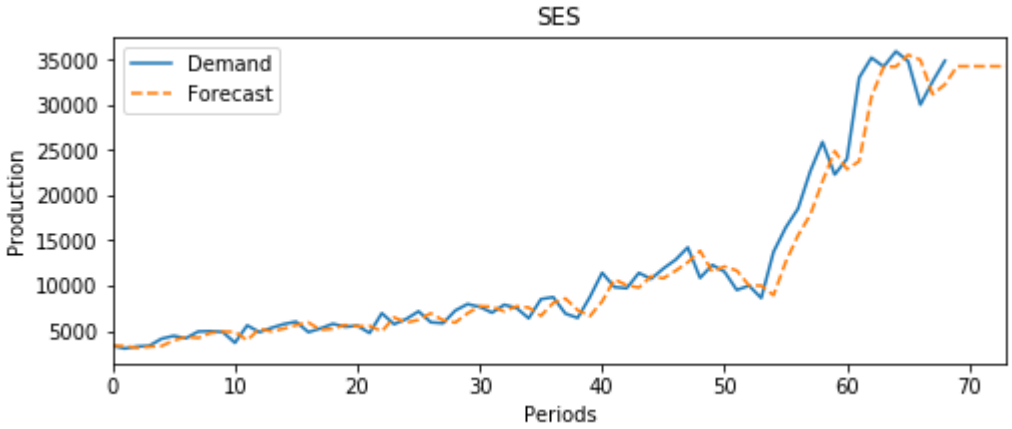
5 Year Moving Average



In the figure, we are showing MA of 3 years and 5 years of cotton production. The metrics which are used for evaluation are MAE and RMSE. MAE – 1624.08 and RMSE – 2583.21

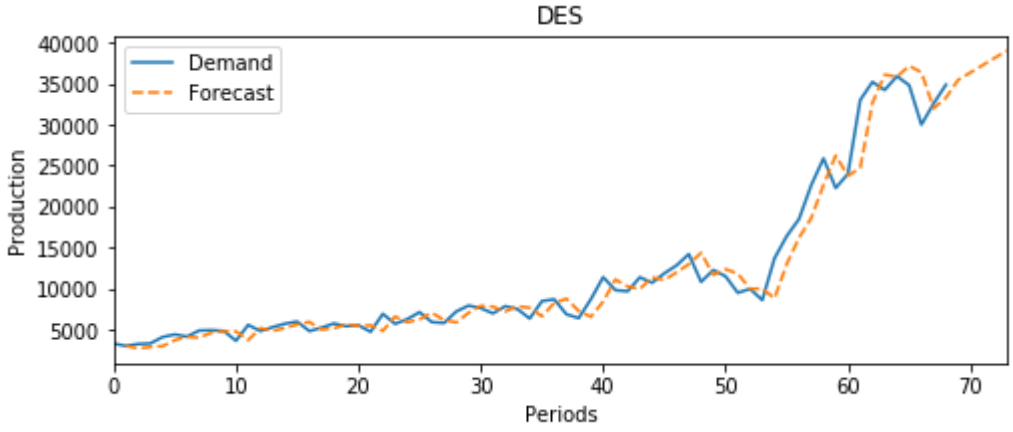
for 3 year Moving Average. MAE is 2073.3 and RMSE is 3249.59 for 5 year Moving Average.

Single Exponential Smoothing method prediction Result



Here we have created a function which takes in the value which needs to be predicted, number of periods which needs to be predicted and also the value of alpha .So from this figure, we get a clear picture of the plot with demand and forecasts, with $\alpha = 0.77$ the MAE value is 1385.42 and RMSE value is 2102.46. In the X- axis we have periods and Y- axis we have production. The demand and forecasting have a gradual increase until 50 and exponentially growing till 60 and starts to become stabilised after that.

Double Exponential Smoothing method

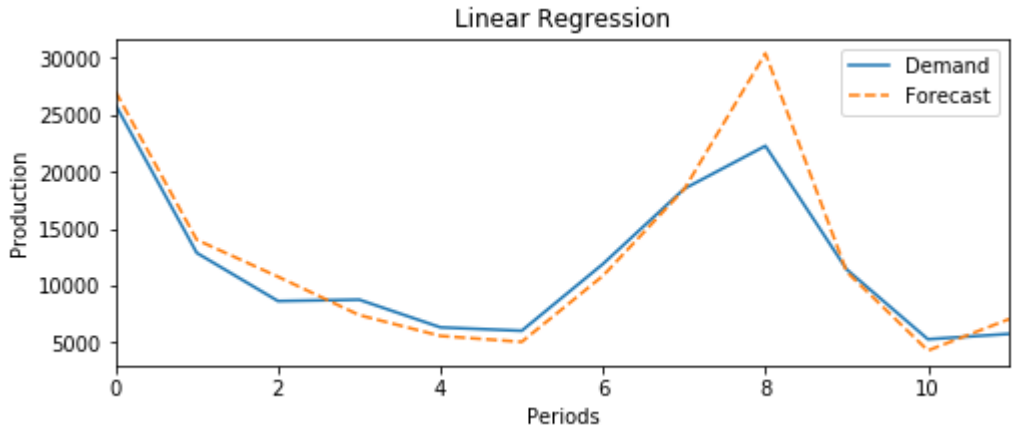


Here we have created a function which takes in the value which needs to be predicted, number of periods which needs to be predicted and also the value of alpha and also the value of beta. We have periods in X-axis and production in Y-axis, demand and forecasting is gradually increasing until 50 and shows an exponential growth after that. We used $\alpha = 0.81$ and $\beta = 0.07$, with that we got MAE value as 1388.81 and RMSE value as 2043.45.

Machine learning models

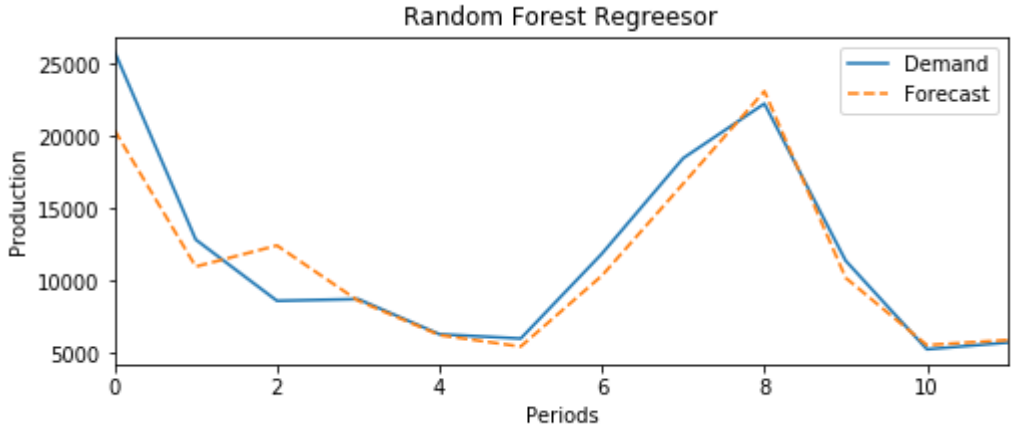
Here we have created 12 years of lag values in order to predict the current observation, and we have split the data into test and train with a test size of 0.2. Then we have fed these values into different machine learning regression models in order to predict the values. The result from these models are given below.

Linear regression



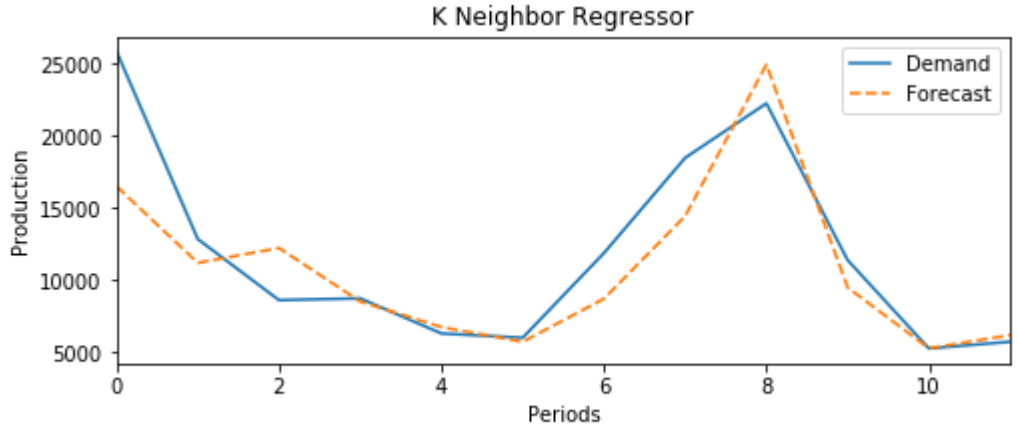
According to figure 5, we have periods in X-axis and production in Y-axis. Demand and forecasting is showing a gradual decrease until period 5 and it showed a steep increase till 8. Here, we have MAE value 1605.78 and RMSE value 2592.37.

Random forest



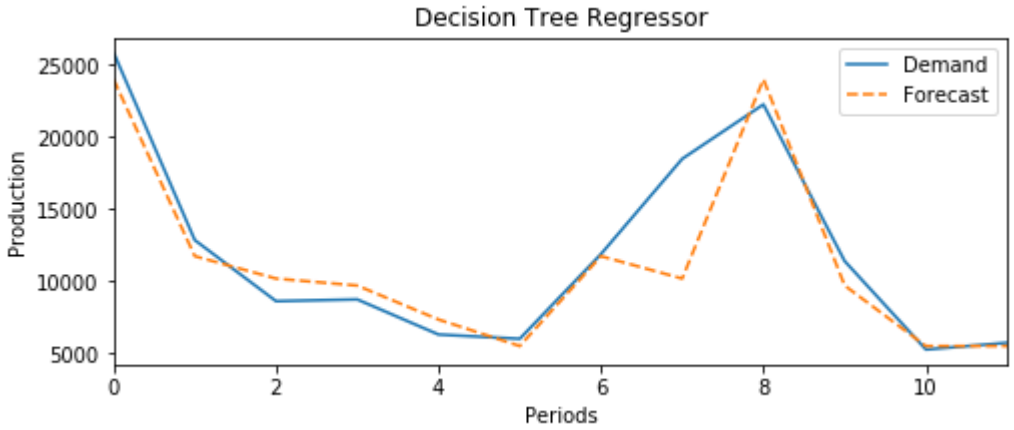
Here we have used grid search method in order to find the best hyper parameters to predict the model . We got best parameters as max_depth = 4 , n_estimators = 200 and max_feature= 10. We have periods in X-axis and production in Y-axis, here the fig is showing a steep decrease until 4 and then steep increase until 8. Here MAE value is 1472.78 and RMSE value is 2156.69.

K – neighbour regressor



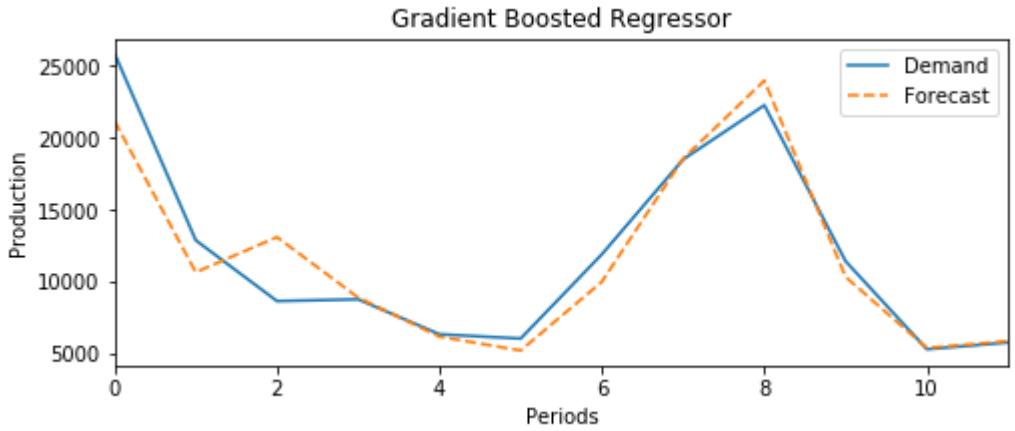
In the K Neighbour regressor, we got MAE value 2320.8 and RMSE value 3431.05

Decision tree regressor



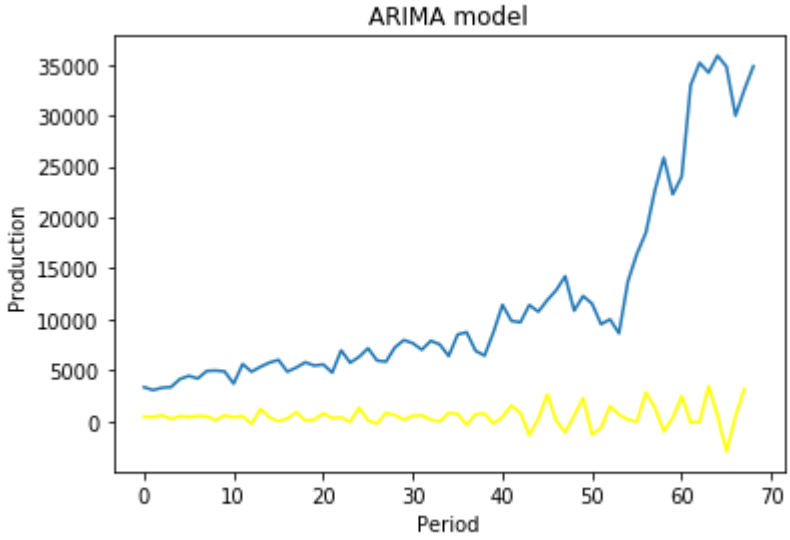
Here also we have used different values for max_depth parameter to see which value is giving the best model and the value of max_depth = 5 gave us best test accuracy. In the Decision regressor, we got MAE value 1617.46 and RMSE value 2653.87.

Gradient boosted



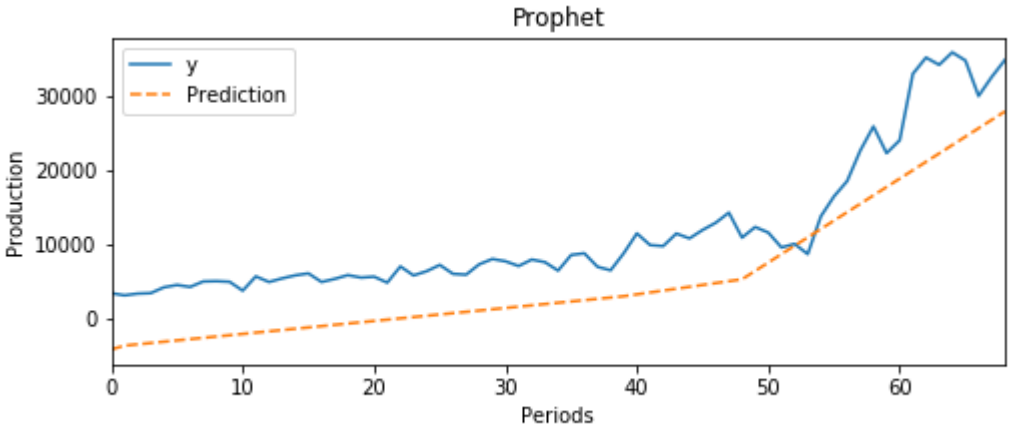
In Gradient boosted we used n_estimators as 500 and learning rate as 0.01 for building the model. Here in the graph we have periods in X-axis and production in Y-axis, its showing a gradual decrease until 5 and a steep increase until 8. Here, MAE is 1456.49 and RMSE is 2151.14. This model gave us the best results among the machine learning models.

ARIMA



In the ARIMA model, we got low aic values for p=3 and q =3. From this graph we can clearly see that the model was worse in predicting the values and that is the reason why we got worst values for MAE AND RMSE as MAE value 11149.0 and RMSE value 14625.0. ARIMA gave us the worst results among the machine learning models.

Prophet

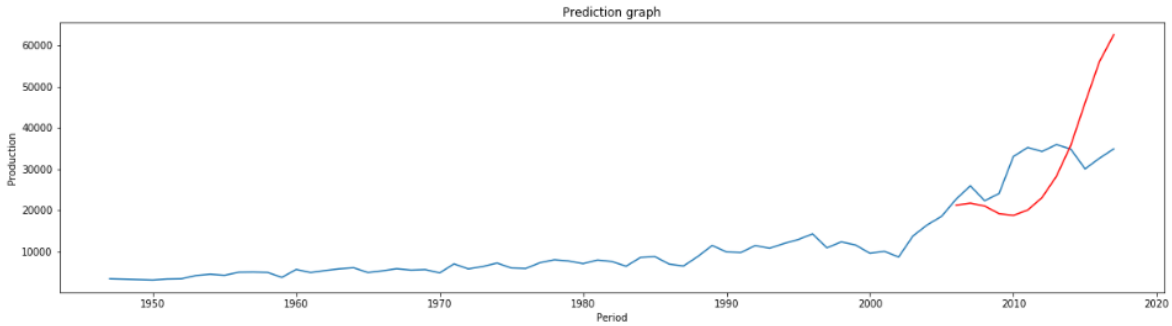


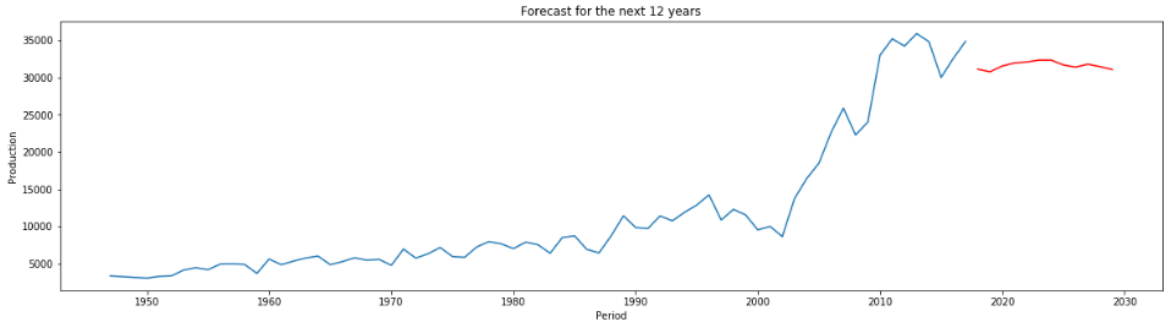
In the Prophet model, we got MAE value as 5180.32 and RMSE value as 5982.91. This model was almost working good in prediction but it was not having good values for MAE and RMSE as compared to other models.

LSTM

Here the LSTM model was built by splitting the data into test and train and these were scaled by using MinMax scaler function in order to make the values in the range of 0 to 1. These were then fed into the LSTM model for analysis. We performed LSTM for different values of epochs and also for different optimizers.

PART1- epochs =180, optimizer = adam

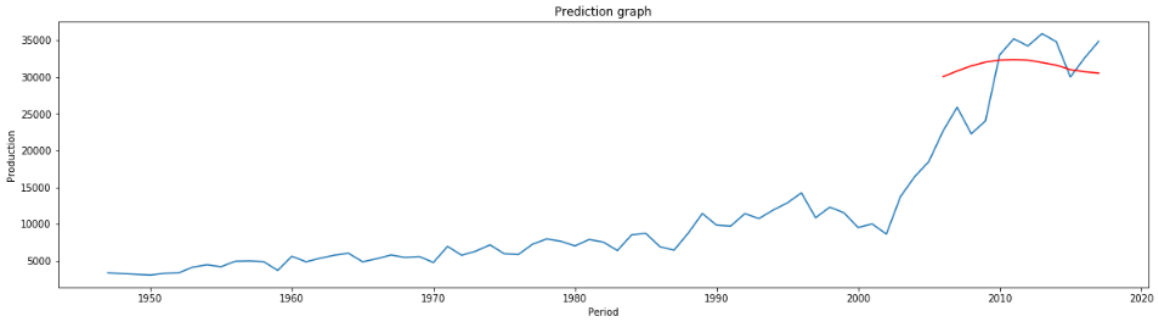


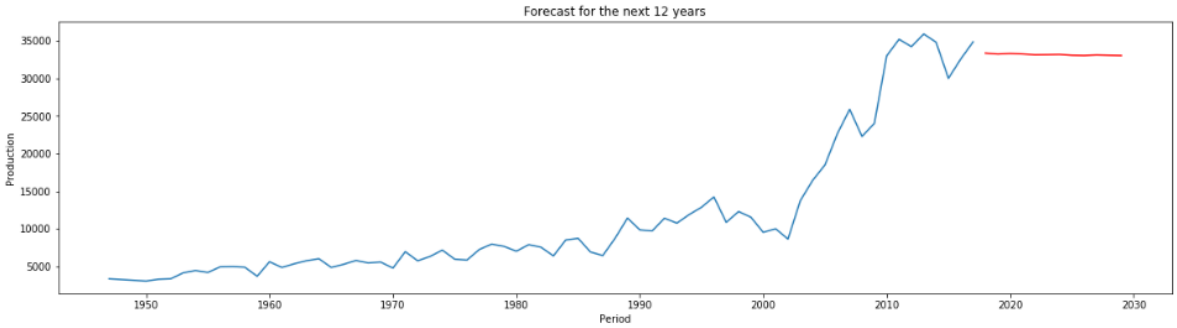


Here, the first graph shows the prediction of the LSTM model for the test data that we have created and how it is varied from the actual data whereas the second graph shows the forecast of LSTM model for the next 12 years.

For the LSTM model with epochs value as 180 and optimizer as adam , we got MAE value 10683.33 and RMSE value 13657.3.

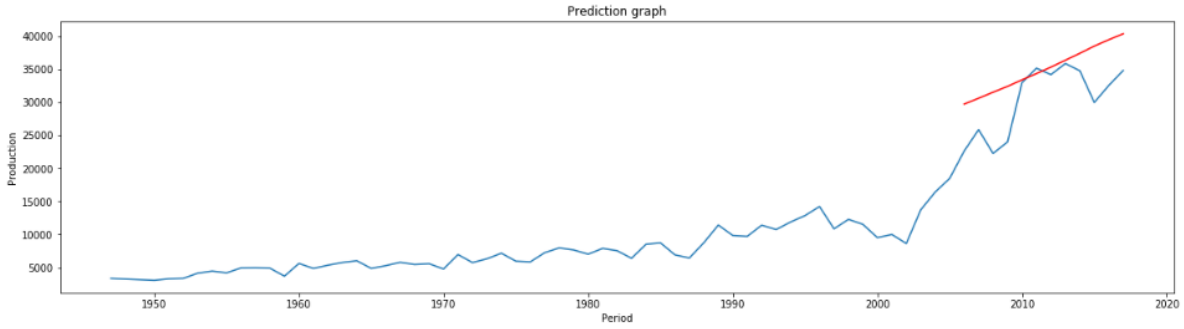
Part 2: epochs = 250 , optimizer = Adagrad

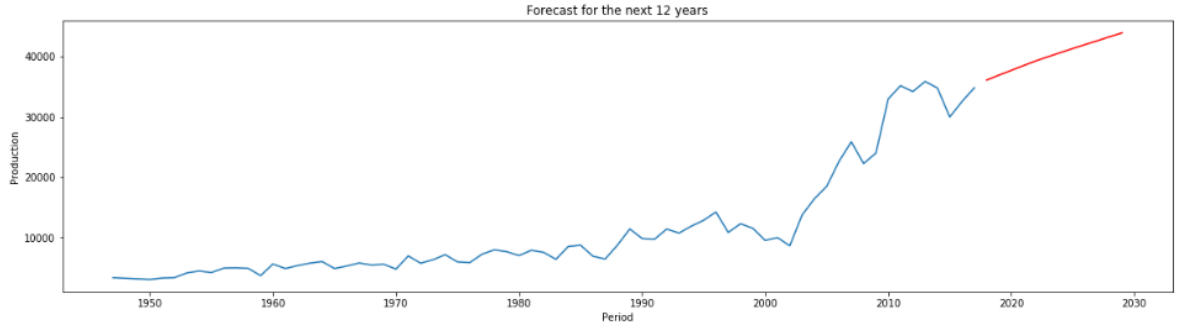




Here, the first graph shows the prediction of the LSTM model for the test data that we have created and how it is varied from the actual data whereas the second graph shows the forecast of LSTM model for the next 12 years. For the LSTM model with epochs -250, we got MAE value 4111.91 and RMSE value 4912.05.

Part 3 : epochs = 300 , optimizer = sgd





Here, the first graph shows the prediction of the LSTM model for the test data that we have created and how it is varied from the actual data whereas the second graph shows the forecast of LSTM model for the next 12 years. For the LSTM model with epochs -300, we got MAE value 4680.12 and RMSE value 5713.92.

Challenges

The main challenge that we faced was that, the data was a yearly data so we was not able to resample it to monthly data or weekly data and also we had data only from 1947 to 2017, so the data was not stationary and we were able to prove it from Augmented Dicky Fuller test, since the ADF test statistics value was greater than the 5% critical value.

Results

Models Implemented	MAE	RMSE
3 years Moving Average	1624.08	2583.21
5 years Moving Average	2073.3	3249.59
Simple Exponential Smoothing	1385.42	2102.46
Double Exponential Smoothing	1388.81	2043.45
--------Machine Learning Models----------		
Linear Regression	1605.78	2592.37
Random Forest Regressor	1472.78	2156.69
Decision Tree Regressor	1617.46	2653.87
K- Neighbor Regressor	2320.8	3431.05
Gradient Boosting Regressor	1456.49	2151.14
ARIMA	11149.0	14625.0
Prophet model	5180.32	5982.91
----------LSTM ----------		
Part 1 : epochs = 180, optimizer = Adam	10683.33	13657.3
Part 2 : epochs = 250, optimizer = Adagrad	4111.91	4912.05
Part 3: epochs = 300, optimizer = sgd	4680.12	5713.92

Table 1: Results from each models.

We used Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as the two metrics for evaluation of our different models that we have used for this Time Series Analysis. So, from the values in Table 1, we can conclude that Double Exponential method with alpha value as 0.81 and beta value as 0.07 was performing better than the normal prediction models whereas when we consider different machine learning models we can say that Gradient Boosted Regressor was performing better than any other regression models that we used to predict the model and also when we look at the LSTM model its much more clear that LSTM with epochs = 250 and optimizer as Adagrad was performing better than rest of the LSTM models since it was having least values for MAE and RMSE.

So when we consider all the models we can say that Double Exponential Smoothing method with alpha value as 0.81 and beta value as 0.07 was the best model among all the models that we have used , since it was having the least values for MAE and RMSE.

Reference

- Referred to all the projects and labs we have completed in class.
- https://stclairconnect-my.sharepoint.com/:b:/g/personal/w0735036_myscc_ca/EYcsU7qV9CBJg2R2h_4ncD4BAtiR96fJv274SqXw36_4CQ?e=HcfsUc
- Project Code: https://stclairconnect-my.sharepoint.com/:u:/g/personal/w0735036_myscc_ca/EeVqbkSwYoVGj6e2Ucbj-IwB8dm1bzhGYj1KgGUCUqpGKQ?e=6szSN1