Semester Project Report

Time series forecasting for climate data

Name: Barish Sarkhel
Reg no: 20201242
Semester : August 2023 (7th Semester)
Course code:DS4313
Supervisor: Dr Joy Merwin Monteiro

1 Objective:

In this project We learned different time series forecasting methods and implement them for forecasting of temperature data from across India. We tried to learn the strengths and weaknesses of each method. Our basic focus was on linear model and exponential smoothing model and we learned the skill of both categorical and quantitative forecast.

2 Introduction to the topics:

Here I will discuss the theory I learned and applied in this project. This includes about the informations of the time series prediction models and the error matrices.

2.1 Linear prediction model

Linear prediction time series models are one of the most simplest yet very useful type of model in time series analysis. In this case we take n number of previous observation for each date and do our liner prediction based on that data. This is computationally very easy. For Error analysis of the model MSE is used usually.

2.2 Exponential smoothing model

This an another type of time series prediction model which is designed to capture and emphasize recent trends while diminishing the impact of older observations. It gives different weight from large to small to the recent data to the old data and that is how it provide prediction with better accuracy then the linear prediction model usually.

2.3 Categorical Forecast

Categorical forecast is done for two or more classes. The threshold value is pre-determined in this forecast. Here I have done two class classification and taken the threshold value as the average of the temperature value of the whole dataset. The two classes were below average and above average temperature.

2.3.1 Contingency table

Contingency table is the table which is used to represent the output of the categorical variable. Table includes the probability of all four cases as shown in the Table 1.

Verifying analysis	Above Normal (Forecast)	Below Normal (Forecast)	
Above Normal	p_{aa}	p_{ba}	p_A^P
Below Normal	p_{ab}	p_{bb}	p_B^P
	p_A^F	p_B^F	1

Table 1: Contingency Table

2.3.2 Heidke skill score

A useful measure of the skill of a two-class categorical forecasting scheme is the Heidke skill score. The skill score is given by,

$$\frac{p_C - p_E}{1 - p_E}$$

Here $p_C = p_{aa} + p_{bb}$
and $p_E = p_a^P p_a^F + p_b^P p_b^F$
We can observe that the value of this skill score for a random prediction model is 0 and for perfect prediction model is 1 as the value of p_C is 1 in that case.

2.4 Quantitative forecast

In quantitative forecast we try to predict the actual value and the error matrices for quantitative forecast in dependent on the predicted forecast. Some of them are described here.

2.4.1 Forecast and Predictand as Bivariate Random Variable.

As we noted in the forecast/predictand pair (F; P) form a bivariate random variable with a joint density function f_{FP}. The conditional density functions $f_{F|P=p}$ and $f_{P|F=f}$ tell us something about the performance of the forecast. I computed this values for my models also by generating the histograms and calculating the the number of data points in that range for the particular data points in the observation and prediction dataset.

2.4.2 Correlation skill score

The correlation between the forecast F and the verifying observation P is called the correlation skill score and is given by

$$\frac{\mathrm{Cov}(F, P)}{\sqrt{\mathrm{Var}(F)\mathrm{Var}(P)}}$$

Clearly for a perfect forecast the value of this skill score will be 1.

2.4.3 Brier skill score

The Brier skill score is a measure of the skill of the forecast F relative to a reference forecast R of the same predictand P. The comparison is made on the basis of the mean square error of the individual forecasts. The Brier score is given by

$$\frac{S_{RP}^2 - S_{FP}^2}{S_{RP}^2}$$

By using this skill score we can compare between two models.

2.4.4 Mean squared error

The mean squared error is the expected (i.e.,long-term average) squared error which is defined by

$\sum_{i=1}^{n}(F-P)^2$

This is one of the most popular error metrics use to understand error in time series models.

3 Works done in the project:

3.1 Making models-

Time series modelling is a well know and well used topic to make model and prediction of the time series data. In this model I took the temperature data from Pune-Nagpur centre. The first target was to visualize data and then modelling them with the well known time series modelling processes. I mainly used linear modelling and exponential smoothing modelling and did the prediction.

3.2 Calculating error understanding the average model-

After that I focused on error part. I learned about categorical and quantitative errors associated with these models. There are different type of skill scores as discussed in the above section. I learned about them and applied them in the model and computed these values for the models I build.

After this I computed the average of the predictions of the model I had. In overall observation this average prediction perform better then the individual models. The final work was to find the reason why the average model predict better then the other models.

4 Methodology and results come in each part

4.1 Linear Prediction Model-

As I described I took the temperature data from the Nagpur-Pune station. Then I split the dataset in a way so that the linear prediction will be made by last data 3 years data of the same data. I split the dataset by taking last 1000 data points of test data. The prediction is shown in the image Figure 1.

4.2 Exponential Smoothing Prediction-

After building the Linear model I focused on exponential smoothing model. In this case also I put the training set and test set same as before. Tread and seasonality both I put as additive. The prediction I put in the Figure 2.

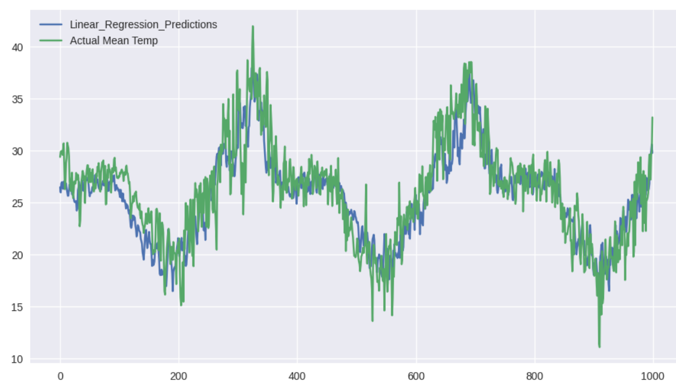

Figure 1: Linear Prediction

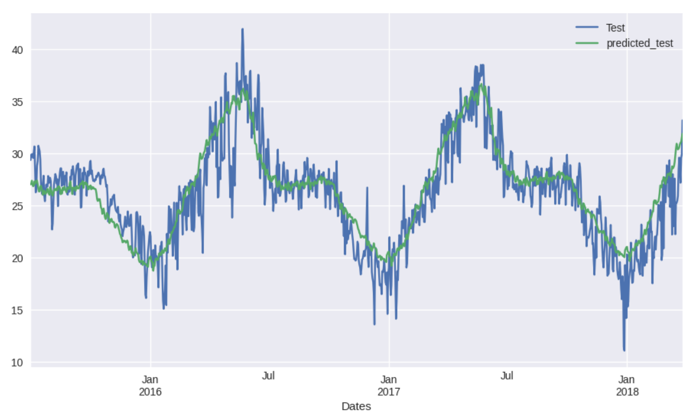

Figure 2: Exponential Smoothing Prediction

Forecast Name/Error Type	Lin_reg	Exp_smoothing	Average
MSE	6.2576	5.789	5.3177
RMSE	$\sqrt{6.2576}$	$\sqrt{5.789}$	$\sqrt{5.3177}$
Max_error	8.9752	10.9309	9.8142
Min_error	-8.9028	-8.0818	-8.4923
SD_error	2.4516	2.3691	2.3057
Error_median	-0.4802	0.2466	-0.158
Error_25 percentile	-2.004	-1.1351	-1.6134
Error_75 percentile	0.9744	1.8055	1.2354
MAE	1.9396	1.8233	1.7888
Var_obs	23.5081	23.5081	23.5081
Var_pred	16.9985	18.3039	17.1559
Var_error	6.0103	5.6128	5.3162
Co_variance	17.2654	18.1177	17.6916
P(f=20 to 25/ x=20 to 25)	0.6691	0.7063	0.7212
P(f=15 to 20/ x=20 to 25)	0.1822	0.0743	0.0892
P(f=25 to 30/ x=20 to 25)	0.145	0.2156	0.1858
Proportion of explained variable(Best=1)	0.7443	0.7612	0.7738
Co-relation skill score (best=1)	0.8637	0.8734	0.8809
Hedic skill score (best=1)	0.5841	0.6198	0.6182

Table 2: Error metrics values

4.3 Error metrics associated with these models-

After building the models I calculated different skill scores and compiled all of them in the Table 2. As I said earlier The average model perform quite better then these two models.

5 Relationship between the models through analysing the error metrics

The final goal of the project was to understand the relation between the two models by analysing the error metrics we got. Through an overview we can observe that the prediction we got from the average is performing well then both the models. Here we try to understand why such kind of phenomena happened.
Here I discussed what are the approaches I took to solve the problem and what are the outcomes I got from there.

5.1 First approach

Here I partition the test predictions in certain (in this case 10) and observe the error metrics are behave. To understand this I took the MSE as the error metrics. In the output there were some parts where linear model was predicting better and there were some part where exponential model was predicting better. So the average model was lying in the middle in most cases. Also it need to mentioned that there were some part where any of the either model was performing too bad but

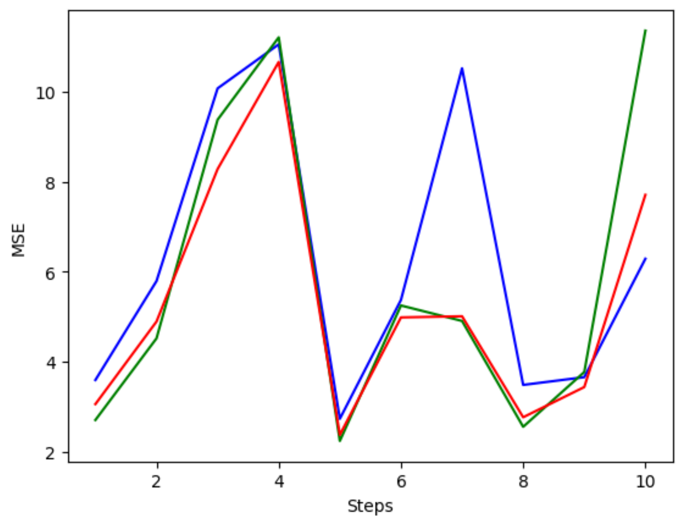

Figure 3: Step wise MSE

the average model reducing it in significant level. So by this simple observation we can understand that the average model basically nullifying the effect of the "bad" model in each part and as a result giving better result then any one of them. The plot is given in Figure 3

5.2 Second approach

Here I made a simple observation by the Figure 1 and Figure 2. We can see that the frequency of changing the temperature is quite high in the observation which is not coming in the models. Specially the point where there is certain unexpected high changes in temperature for some days. It is quite obvious by the way these models are prepared but this is a source of error. The average of the predicted models may getting the frequency little high due to the attachments of both the models and that may be reason it is giving little less error then them.

5.3 Third approach

In this approach I calculated the error value (i,e, (observed-predicted))for each of the point in both the models and plotted it here. We can made several observations from the plotting. The plot is given in Figure 4 (Blue is Linear prediction model and Green is exponential smoothing model). The error is high for both the model when the temperature of that point is either too high or too low. In between 22 or 23 and 30 degree temperature both the model performs really well. In the 15 to almost 20 degree part we can see linear model performs little bit better then the other. After 30 degree the scatters started increasing. In all these areas due to this ups and downs averaged out and the average model starts giving better prediction.

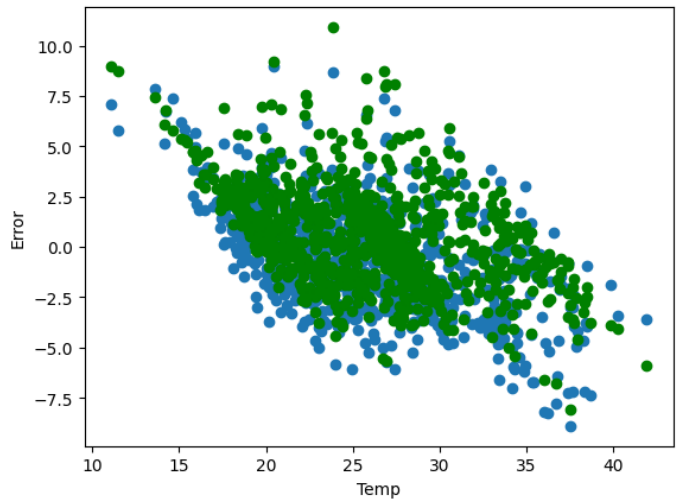

Figure 4: Error vs Temp

5.4 Fourth approach

Now I calculated the actual number of cases where linear prediction is giving more absolute error then exponential prediction and the vice versa case. I got out of my 1000 validation points 524 cases error in linear prediction model is high and in other 476 cases the other error is high. So this is quite clear that it is not the case that one model is performing extremely well then the others and in average model we are getting just the outcome of that phenomenon.

5.5 Fifth Approach

Now I extend my approach which I had intuitively in my third approach. I divided the dataset in 3 parts. There I made 4 lists for each part one for observation, one for linear prediction and one for exponential smoothing prediction and one for the average one. In the first part I took all the observation data which has value less then then 20 and its prediction values. Also did it for the case when the values are between 20 and 30 and those cases when the values are greater then 30. What I got I summarized in the Table 3

	Linear model	Exponential model	Average model
Less then 20	8.375	11.633	9.555
Between 20 and 30	4.363	4.669	3.994
More then 30	11.708	6.335	8.389

Table 3: MSE values

Now as we can see out basic understanding matches with the following results. In the both side cases of the middle the mse values are quite high then the middle case. Also the interesting point to note that in one case the exponential model error is quite high and in the other case the liner model error is quite high and both are averaged out in the in the average model. May be this is the reason why the average model is giving better results in the overall scenario.

As I discussed I tried to take different approaches to understand the reason behind the better performance of the average prediction model then the other two model. In each case I noted my observations and the reasons I found from the observations.

6 Conclusion

As I discussed in this semester project I tried to learn different models, their error and their relations along with the average prediction model. I took different approach to understand the error of the average prediction models. There are certain places where the work can be continued further such that getting many more models and to conclude a with some more strong mathematical argument along with the observational arguments. I will try to look into those points.

7 References

1. Storch, H. von, and Zwiers, F. W. (2003). Statistical Analysis in Climate Research. Cambridge University Press.
2.Hyndman, R.J., and Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3
3.Murphy Allen H, Brown Barbar G and Chen Yin-Sheng, 1989, Diagnostic Verification of Temperature Forecasts, American Meteorological Society