|  |  |
| --- | --- |
| **Time Series Forecasting project report**  **By Shilpa Jha** |  |

## Executive Summary

The aim of the project is forecast temperature for each day in Delhi City ,India. The initial dataset has around 1576 periods starting from 1st January 2013 to 24th April 2017. The weather dataset usually follows annual seasonality; for example . Temperatures in May 2016 won’t be very much contrast from May 2017. Following this logic, the dataset is aggregated at monthly level and Mean temperature is calculate for each Month-Year combination. This also helps in reducing the noise and making the forecasting process smoother. The dataset shows strong annual seasonality along with moderate Trend and level component. The Maximum temperature is mainly a round May-July and temperature drops drastically during November -January months.

After aggregating the data ,the no. of periods shrinks to a count of only 52. Due to low count of records Splitting the dataset inro Training and Validation is not done and all models were developed on the entire dataset, and these are compared on the bases of accuracy metrics like MAPE and RMSE

Models like Regression-based models, advanced exponential smoothing models and, ARIMA seem to be good options to be tried for forecasting the temperature. The dataset also contains some other information like Humidity,Pressure,Wind Speed which may have effect on Temperatures. These relationships are considered while coming up with different types of models. Additionally, different parameters within all these models are tweaked and analyzed to get better forecasting accuracy . Model evaluation was based on the RMSE and MAPE accuracy metrics. ARIMA model outperforms all other models with MAPE of 1.5% .However other Models are also relatively performing well in forecasting .

Finally, the monthly forecasts are disaggregated to daily forecast by maintaining the daily profile for each date of a particular Month.The percent contribution of each day towards the mean monthly temperature is recorded for each Month . This weightage contribution is then used to forecast the daily temperature given the forecasted mean monthly temperature .

## Introduction

The dataset is from the Kaggle Website**.** [Daily Climate time series data | Kaggle](https://www.kaggle.com/sumanthvrao/daily-climate-time-series-data?select=DailyDelhiClimateTrain.csv). This dataset provides data from 1st January 2013 to 24th April 2017 in the city of Delhi, India. The 4 parameters here are meantemp, humidity, wind\_speed, meanpressure. The daily temperature is aggregated to monthly Mean temperatures ;So total no of periods is only 52 .All the eight steps of forecasting are implemented in this process and is shown below:

## Step 1: Define Goal

The primary goal of the project is to build a predictive model to forecast daily temperatures for the next 12 months. These monthly forecasts can then be further converted to daily temperatures as well. Along with this , the motive is to explore different components of time series and also study the relationship between Temperatures and other variables like Mean Pressure ,Humidity and Wind Speed.

## Step 2: Get data

The dataset is loaded into the software using read.csv() and stored in a variable weather.data

Text

Description automatically generated

## Step 3: Explore and Visualize Series

Foremost ,The relationship between Temperature and other variables need to be studied ,as this will give some clarity whether we can use any of those variables to forecast temperatures. Scatter Plots are created for each combination to explore the relationship.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

The above plots and correlation coefficients show that Mean Pressure and Wind Speed are not correlated to Mean temperature. However, We can see some moderate negative correlation between Temperature and Humidity and hence We will use this variable as a predictor to forecast temperature in Regression and Arima Models.

**Time Series Characteristics :**The STL components and ACF plots are shown below :

Chart, line chart, histogram

Description automatically generated

Chart

Description automatically generatedThe above plots show that the dataset has additive seasonality and an increasing trend over time . Overall Temperatures are increasing across the years. The significant values at Lag 1 ,2 and then dropping to 0 at 3 suggest that the dataset has a level component as well as an increasing trend component. Also, significant coefficients for lags 12 indicates the presence of strong seasonality

**Seasonal Variation and Monthly Variation in Temperatures**

Graphical user interface, chart

Description automatically generated

The above plot depicts so many minute details about the seasonality present .Some notable points are :

* April-June are the hottest months of most of the years
* Temperatures drop during October to February Months
* The highest temperature is recorded for year 2016 during the month of June

Chart

Description automatically generated

The above monthly plots are also exposing similar trends to that of Seasonal Plots.

**Stationarity test for the Dataset**

The ADF test(Augmented Dicky Fuller) test is utilized to find out if the time series is stationary or not. From the above exploration and plots it seems that the series is stationary .For this test ,The null Hypothesis states that Time series is nonstationary. Since the p value is less than the threshold ,Null Hypothesis can be rejected, and it can be concluded that the Time series is stationary.

The output of this test is useful as most of the models perform well on stationary time series and if the time series is nonstationary differencing can be done on original dataset .Additionally , it will be useful is selecting the d parameter used in Arima Model .Summary of the above test is shown below:

Text

Description automatically generated

## Step 4: Data Preprocessing

As discussed in the Introduction part , Aggregating the daily temperatures into average monthly temperatures is required in this dataset. The aggregation makes sense as weather forecasting usually follows annual seasonality and also smooth out the noise.

Text

Description automatically generated

After getting the mean values for each of the above columns, time series object is created for each of them .It will be used later while developing the models. The corresponding screenshot is shown below :

Text

Description automatically generated

## Step 5: Partition Series

As the dataset has very few time periods ,Partitioning is not done, and models are developed on the entire dataset.

## Step 6 & 7: Apply Forecasting & Comparing Performance

**PREDICTIBILITY TEST**

Before developing models, we should first test if the time series is predictable or not. To do so, there are two methods which can help in evaluating the same . These tests can save us a lot of efforts in case dataset is a random walk.

* **Method 1 :** . ARIMA(1, 0, 0) is an autoregressive (AR) model with order 1, no differencing, and no moving average model. Summary of the model is shown below:

Text

Description automatically generated

This model takes the immediately preceding values as predictors to forecast the future values. The typical equation of AR(1) model is y ***Yt* = *a* + *b*1*Yt*-1 + e*t***.Random Walk is a special case of AR(1) Model where b1 =1.It conveys that the difference between period t and t-1 is random, and We can say that the dataset is not predictable.

The equation of the above AR(1) model is y(t)= 24.62 +.8297 y(t-1) . Here Y(t-1) denotes the lag1 series.

As per the above screenshot We calculated the z value and corresponding p value for the same. The p value is .015(less than the threshold .05) which suggests that null hypothesis( H0 :b=1) can be rejected and hence the data is not random walk and the revenue can be predicted for future periods.

* **Method 2: Lag1** differenced series is calculated and the autocorrelation coefficients for different lags are examined .If most of the values are significant, we can say that it is worth to develop forecasting models for this series. ACF plot for the difference series is shown below :

Chart

Description automatically generated

As per the above plot ,All the correlation coefficients are above the threshold values which conveys that the dataset is highly correlated, and the past temperatures affect the current and future temperatures. Hence, we can say that the dataset is not a random walk and can be predicted

Now different Models can be explored, and the performance can be tested . Different flavors for each model type will be developed and tested on the dataset. Different Types of Models and their performances are discussed in the followings sections .

## Autoregressive Integrated Moving Average Models

The Autoregressive Integrated Moving Average (ARIMA) model is a well-known and one of topmost popular model that can be used for forecasting on any type of dataset irrespective of the presence of trend or seasonal components. This model can very well fit the dataset and can give accurate predictions. There are two ways in which model can be developed :

* Automated Parameter Selection by using auto.arima()
* Manual Parameter Selection using arima()

Arima model using auto.arima() is developed and the summary of the model is shown below

Text

Description automatically generated

This is a Seasonal Arima Model ; one of the best methods to forecast data with level, trend and seasonal Components. ARIMA(1,0,0)(1,1,0) is of the form ARIMA (p,d,q)(P,D,Q)[m] which has following definitions:

* p=1 : order 1 of Auto Regressive Model AR(1)
* d=0: order 0 no trend
* q=0 : no moving average error lags
* P=1 : order 1 of Auto Regressive Model AR(1) for seasonality component
* D=1 : order 1 differencing to remove linear trend
* Q=0: order 1 for moving average error lags
* m=12 : for annual seasonality

The residuals Plot is shown below. The expectation is to have normal distributed residuals, and these should not be autocorrelated.

The Ljung-Box test is implemented for all the models to explore the residuals. It is a statistical test to validate if residuals are autocorrelated. The null hypothesis states that “There is no autocorrleation in the residuals”. In the below case p value is above threshold ,so We can’t reject the null Hypothesis and hence Residuals from Auto Arima Model are just noise which signifies that this model is good enough to forecast future temperature values.

Graphical user interface, text

Description automatically generated

Chart

Description automatically generated

## Manual selection of the parameters

The ACF and PACF plots are shown below for this dataset :

Chart, histogram

Description automatically generated

The ACF and PACF plot along with the stationarity Test can help in selecting the values of p,d,q which is used in Arima Model .PACF is the indicator of Auto regression parameter ‘p’ while ACF is the indicator for Moving avg parameter ‘q’. Additionally, The Augmented Dicky-fuller test in the above section confirms that the dataset is stationary. Hence We can choose the value of d as 0.

In case of PACF First two lags are significant so ,We can start with the AR(2,0,0)(2,0,0) and then improve the model in future based on the residuals and other parameters. The seasonal component needs to be added in the model as the dataset shows annual seasonality.

* **Model 1 : ARIMA(2,0,0)(2,0,0)-** Summary of the model is shown below :

Text

Description automatically generated

**Output of Ljung-Box test to validate if residuals are random is shown below :**

Graphical user interface, text

Description automatically generated

The Above Output and the plot shows that the Residuals are random which is a good sign and MAPE is around 4.7% less than that of Auto Arima Model.

We will try improving the model as MAPE is bit lower than that of Arima Model by tweaking certain parameters and then checking the MAPE and residuals distribution.

* **Model 2:Arima(2,0,0)(2,1,2)** – Summary and Residual plots of the model is shown below :

Graphical user interface, chart

Description automatically generated

A screen shot of a computer

Description automatically generated with low confidence

**Output of Ljung-Box test to validate if residuals are random is shown below :**

Graphical user interface, application

Description automatically generated

All the above plots and output suggest that Residuals are random, and We have achieved a lot of improvement in RMSE and MAPE values .Finally, we came up with the following 5 Arima Models:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **ARIMA(2,0,0)(2,0,0)** | **ARIMA(2,0,2),(2,1,2)** | **ARIMA(2,0,2),(2,1,2)**  **With external regressors** | **Auto Arima(1,0,0)(1,1,0)** | **Auto Arima(1,0,0)(1,1,0) with external regressor** |
| AIC | 218.44 | 147.77 | 129.76 | 140.48 | 126.44 |
| MAPE | 4.62 | 1.53 | 2.85 | 3.2 | 2.8 |

As per AIC values Arima Model with external regressors seems to outperform other models However It could be little tricky to forecast temperatures for future months as Humidity factor needs to be forecasted first and then used in this model. To prevent this complexity ,We won’t be considering these models but We do get a conclusion that including Humidity factor increases the forecast accuracy as compared to different order Models without external regressors

So considering this Auto Arima(2,0,2),(2,1,2) seems to be the most accurate model.

We could see the residual plots of that model . The residuals seem to be random which shows that these models have handled all the possible autocorrelation.

Chart

Description automatically generated

**Forecasting future 12 periods with this model :**

Chart, line chart

Description automatically generated

## Regression Based Models

Regression Models once again can be used for fitting all components of a time series. As per the nature of the dataset which has strong seasonality ,moderate trend ,level etc ,these models can be very much useful in forecasting the future temperatures. The impact of external regressors(humidity) is also explored in one of the models.

* **Model 1 :Regression Model with Linear Trend and Seasonlity** – Summary of the model is shown below :

Text

Description automatically generated

**Forecasting Accuracy of the above Model :**

Text

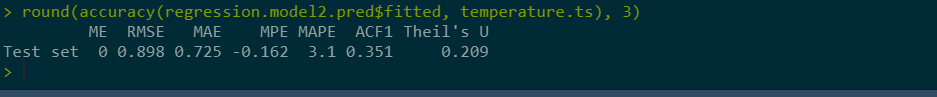
Description automatically generated with medium confidence

* **Model 2: Regression Model with Quadratic Trend and Seasonality-** Summary is shown below

Text

Description automatically generated

**Forecasting Accuracy of the above Model :**

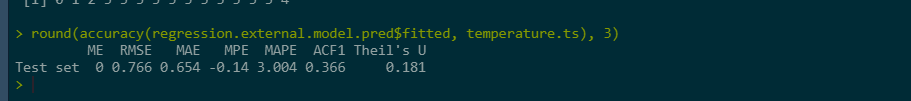


* **Model 3 : Regression Model with Quadritic Trend +Seasonality + External Regressors**

Text

Description automatically generated

**Forecasting Accuracy of the above Model :**



Three different flavors of Regression Model are shown below :

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Linear Trend+Season** | **Quad Trend +Season** | **Quad Trend+Season+External Vraibles** |
| MAPE | 3.3% | 3.1% | 3.04% |

As per the above metrics Regression Model with Quad Trend ,Seasonality and External variables seems to fit the dataset well.But the issue still persists with the external variables ;hence we will keep the Model with Quad Trend + Seasonality for consideration .

**Residual plot of the Model (Quad Trend +Season)**

Chart, line chart

Description automatically generated

The Regression Model doesn’t seem to handle the autocorrelation well ,as some of the coefficients are significant in the ACF plot. The solution to this issue to create a Two level Forecasting Model . We could use Auto Regressive models for handing the autocorrelation in residuals

**Model 4: Regression Model(Quad trend+Seasonality) +AR(3) Model for handling autocorrelation in residuals -**Summary is shown below

Text

Description automatically generated

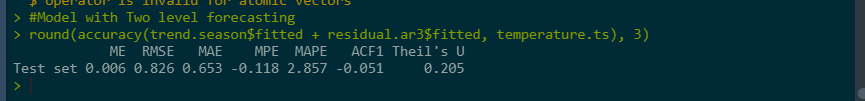
**Residual plot of of Above Model :**

Graphical user interface, application

Description automatically generated

Most of the autocorrelation coefficients are not significant and hence we can say that AR(3) model helps in handling the autocorrelation in the dataset.

**Accuracy Metrics of above Two -Level Forecasting Model**



Finally , All the regression basedmodels with their respective accuracy metrics are shown below :

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **MAPE** |
| Regression with Linear Trend and Seasonality | 0.929 | 3.308 |
| Regression with Quadratic Trend and Seasonality | 0.89 | 3.1 |
| Regression with Quadratic Trend + Seasonality and External Variables | 0.76 | 3.004 |
| Two Level Forecasting(Regression Model + AR(3) for residuals) | 0.82 | 2.85 |

In nutshell, We developed 5 different Regression Models and based on the accuracy metrics Two Level Forecasting Model comes out as the most accurate one with the MAPE of 2.85%.

**Forecasting for the future 12 periods using this model**

Chart, line chart

Description automatically generated

## Advanced Exponential Smoothing

* **Model 1 : Auto ets function**

Summary of the Holt Winter’s Model with automated selection of model options and parameter is shown below:

Text

Description automatically generated

**Accuracy metrics of the above Model :**

Text

Description automatically generated

This HW model has the (A, N, A) options, i.e., additive error, no trend, and additive seasonality. The optimal value for exponential smoothing constant (alpha) is 0.40, no smoothing constant for trend estimate (beta = 0), and smoothing constant for seasonality estimate (gamma) is 0.0001

**Residual Plot of the above model Residuals :**

Chart

Description automatically generated

Though Most of the correlation coefficients are not significant and Residuals seems to be random, We can still try improving the model by using Two Level Forecasting Model

* **Model 2 : Holt Winter’s Model(A,A,A) + AR(3) for residuals -**Summary of the above model is shown below

Text

Description automatically generated

**Summary of AR(3) model which handles the autocorrelation in Residuals**

Text

Description automatically generated

**Accuracy Metrics of the Combined Forecasting**

Text

Description automatically generated

Finally, We have two versions of Holt Winter’s Model .Out of these Two level Forecasting seems to give better results with MAP of 3.1%.

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **MAPE** |
| Holt winters Model (z,z,z) | 0.95 | 3.36 |
| Two Level Forecasting (Holt Winters for time series values and AR model for residuals)  Holt Winter’s(A,A,A)+AR(3) | 0.91 | 3.1 |

The Residual plot of both the Models is almost same ,but with Two level We achieved a slight improvement in MAPE values

**Forecasting future 12 periods with this model**

Chart, line chart

Description automatically generated

l

## Step 8: Implement Forecast

|  |  |  |
| --- | --- | --- |
| Best Model Per Methodology | RMSE | MAPE |
| ARIMA (2,0,2)(2,1,2) | 0.50 | 1.5% |
| Two Level Forecasting (Regression with Quad Trend +Seasonality ,AR(3) model for handling residual correlation | 0.82 | 2.85% |
| Two Level Forecasting (Holt-Winters(A,A,A) + AR(3) | 0.91 | 3.1% |

As seen from the above comparison table it is evident that the ARIMA(2,0,2)(2,1,2) is the most accurate model with MAPE of 1.5 % and RMSE of 0.5

**Forecasting for next 12 periods using ARIMA(2,0,2)(2,1,2)**

Text

Description automatically generated

Now these are mean temperature of each month; Our goal is to forecast for each day . These monthly forecasts need to be converted to daily forecast using following steps:

* We will use the most recent past year data to convert these aggregated forecasts to daily forecast values. For example, to calculate daily temperature values for May 2017 We will use the mean temperature of May 2016 and so on.
* We will maintain separate data Frames for each Month and will store the weightage of each day(daily Mean temperature/Monthly mean temp)
* To get Daily Forecast We will multiply this weightage value of each day with the forecasted mean Monthly Temperature

**Forecast daily temperature for May 2017**

A picture containing graphical user interface

Description automatically generated

**Forecast daily Temperature for Feb 2018**

A picture containing text, electronics, screenshot

Description automatically generated

## Conclusion

Out of several Models ARIMA model is giving the most accurate results . This makes lot of sense as this is one of the strongest models which can handle all the time series components along with autocorrelation in residuals.

## LIMITATIONS

* The external regressors variables are not included in the model to reduce the complexity of the forecasting. However, this could be included once we create a different model to predict the external regressors first and then use them as predictors to forecast temperature.
* The current dataset size is limited and hence the partitioning is not done . Partitioning is very much required as it prevent the overfitting of the dataset.
* Not all Models are explored and tested to forecast the values