Forecasting of Air Passengers using ARIMA Modeling

Sakshi Sharma, A.Jackulin Mahariba

Abstract*: Air passengers prediction is said to be the centre of gravity of the growth. With people on the move constantly, there is bound to be some dissatisfaction amongst the customers which could be due to various reason, varying from overbooking of flights to ground operations. This dissatisfaction can be controlled till a limit, in ballpark figuring. In the past, this has been done using various machine learning techniques. For this prediction, in this project, ARIMA Modeling is used which is a time series forecasting method, based on machine learning. To test the stationarity of the data, which is done using Dickey Fuller test. If the data is stationary, it is fit into the ARIMA Model. If the data isn't stationary, it is made stationary by differencing or by logarithmic transformation. The logarithmic method to make the data stationary. Once the data is stationary, using the Partial autocorrelation function and the autocorrelation function, values of p and q are found, which are required in the time series method. These values are then fit into the ARIMA Modeling and hence, the results are predicted. Upon the use and fitting of various models, the ARIMA(2,1,2) has been the best fit, having the least RMS and RMSE values.

I. INTRODUCTION

Prediction of air passengers is a crucial part of the execution of the operations on a ground level as well as the flying level. It helps in managing the workforce or resources required for the management, commerce, infrastructure required for accommodating the passengers comfortably, without causing any inconvenience or discomfort. The prediction will help in forming the base of what's required and what's not. Here, this is with respect to the the increase in infrastructure, work force, and commerce, to plan in such a way so that the developments can be used efficiently and aptly. For this prediction, in this paper, ARIMA Modeling is used which is a time series forecasting method, based on machine learning.

Machine learning is the science of programming in such a manner that the system doesn't have to individually programme, and based on the previous data, it can learn and perhaps improvise the suggestions and simultaneously increase the efficiency of the program. There are three different types of learning, namely, supervised, unsupervised and reinforcement learning.

In supervised learning, there is a an input and an output, where the output depends on the input. Seeing the trend through this, the machine learns and predicts. In unsupervised learning, it only has the input, and it has to predict the output depending on the parameters being used. In reinforcement learning, the training dataset on its own decided what to do to perform a given task.

In the paper, for ARIMA Modeling, which is a time series method, the stationarity of the data is tested, which is done using Dickey Fuller test. If the data is stationary, it is fit into the ARIMA Model.

Stationarity of the data is important for fitting it into an ARIMA Model. It is determined when both, mean and variance are constant in the data. The data should be stationary, so that it doesn't form a correlation even if does not exist. If the data isn't stationary, it is made stationary by differencing or by logarithmic transformation. Once the data is stationary, using the Partial autocorrelation function and the autocorrelation function, the values of p and q found, which are required in the time series method. These values are then fit into the ARIMA Modeling and hence, the results are predicted.

Liu Xia et al. use regression analysis and grey prediction method[1], which was proven efficient in the prediction.

YAN Kewu et al.compare the regression model of SVM with BPANN and linear regression, where the error of SVM was small and prediction was highly accurate.

The use of a hybrid research model [3] with two linear models and non-linear models, where they have used a combination of time series regression and ARIMAX, and the error has been analysed by NN and SVR. Hence, ARIMAX-TSR and TSR-NN's prediction is more accurate than TSR-SVR and ARIMAX-SVR's.

Juan Jose et al. demonstrate a model on the basis of Random Forest Algorithm for predicting delay in departure for a part of commercial airport for a period of time. The purpose of the study was to predict the delay at sir transport network level in a time horizon of 2 to 24 hours into the future. The results of the study had an error of 19%, when classifying for a delay of greater or lesser than 60n minutes, along with future horizon of 2 hours.

The uses Ration By Schedule method where it studies the tradeoffs between flights, when the delay is assigned pre-tactically [5]. The tradeoff analysis for this shows an "a posteriori" articulation of preferences.

HariBhaska Sankaranarayanan et al. have discussed the various factors on which the satisfaction depends, such as efficiency of ground operations, customer care and services. This paper uses LMT to estimate the passenger satisfaction, where the results are 80% accurate.

Bayesian estimation iteration model has been demonstrated [7],

Revised Manuscript Received on September 22, 2019.

Sakshi Sharma, SRM Institute of science and Technology
A.Jackulin Mahariba, SRM Institute of science and Technology

which is used for the prediction of the mode using decision tree.

The comparison of three learning methods: neural network, logistic regression and support vector machine has been performed, it shows that neural networks is the most optimised methods, as compared to logistic regression and support vector machine [8].

Mennatoallah Youssef et al. used electromagnetic propagation prediction tool for the prediction of the effects on the passengers and internal components. The results concluded that the internal components alter the propagation quite significantly.

The various models studied are ANN, MJLS, CART, LR, with the parameters being time, destinations and the times of the day [10]. MJLS gave least error with mean error of 4.7 for a time horizon of 2 hours. ANN performed well for the OD-pair delkay classification with mean error of 94% for 60 mins threshold and 2 hour time zone.

II. IMPLEMENTATION

Time series method is used for forecasting the results at constant intervals of time, where the data is a collection of numerical points in successive order. The time series can be decomposed, where there are different components based on various factors, and can be useful in the processing of data for more accurate results rather than just a trend. There are four types of decomposition as follows:

1. Trend: the increase or decrease in date over a long period of time.
2. Seasonal: seasonal factors which affect the time series, such as month.
3. Cyclic: the exhibition of rise and fall which isn't at a fixed frequency.
4. Noise: the observational variability which can't be explained.

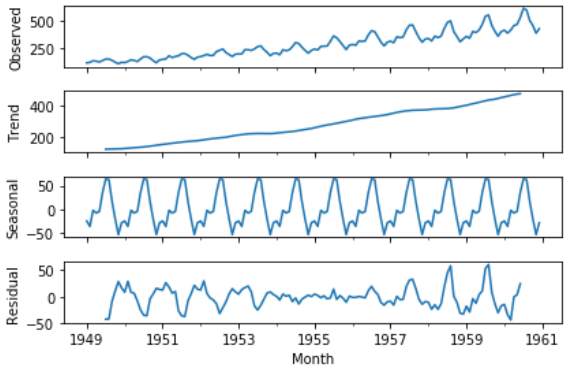

Fig 3. Additive models of non-stationary data

III. TESTING FOR STATIONARITY:

This data is analysed for using it in the prediction part, i.e. ARIMA Modeling. For a dataset to be fit into the ARIMA modeling, the data needs to be stationary, which can be done by checking for the following parameters:

1. The mean should be constant.
2. The variance should be constant.
3. Autocorrelation doesn't vary with time. For this reason, the testing of stationarity, the following methods exist:

1. Looking at the plot: if there is a lot of variation, such as in the trend and seasonality graphs, it is safe to assume that the data isn't stationary.

2. Augmented Dickey-Fuller Test: this test tests for the presence of the unit root, where null hypothesis is that the unit root is present. The lesser the value, stronger is the rejection of the null hypothesis.

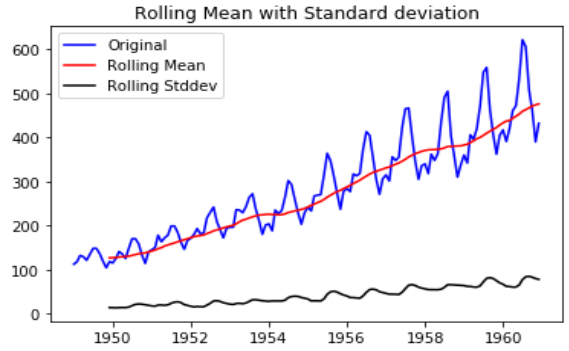

Fig 4. Rolling mean with Standard deviation with non-stationary data

In the graph above, the mean isn't constant, and hence it can concluded the data isn't stationary. If the test statistic value is lesser than the critical value, the data is stationary, and as seen above, the statistical value is more than the than critical value, and hence the null hypothesis is accepted, that the data is not stationary.

If suppose, when a time series doesn't follow any of the above mentioned conditions, it is said to be non-stationary, and then the data is made stationary. If the data isn't stationary, it is made stationary by using either of the methods:

1. Differencing: by taking the difference of two consecutive points, and hence the data becomes from (x1, x2, x3….xn) to (x1-2, x2-3, X3-4…..xn-n-1). This is repeated until the data is stationary.
2. Logarithms: Taking the logarithms of the data, and helps in the reduction of the divergence of the time series.

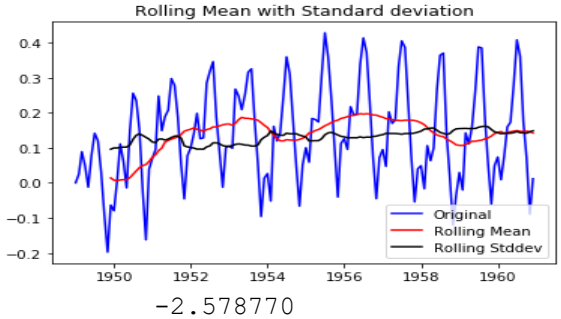

Fig. 5 Rolling mean with Standard deviation with stationary data

Logarithm is taken to stationarize the data, and hence there is a decrease in the values of y-axis. Taking the log and then plotting the graphs, the trend of the mean and variance is constant, which proves that the data is stationary. Null hypothesis is also checked, and since the test statistical value is lesser than critical value, the null hypothesis is rejected, and hence the data is stationary.

IV. PREDICTING VALUES

Fitting into ARIMA Model

ARIMA Model is a time series method of forecasting the future. The time series method, ideally uses a stationary dataset. If the data is not stationary, the data is stationarised, as seen above.

AutoRegression Integrated Moving Average is a time series model which is used for prediction of the trend for future. This is a combination of AR(AutoRegression) and MA(Moving Average) models. Each model has been represented separately, i.e. AutoRegression, Moving Average and the ARIMA model separately, where their combined effect are considered and individual effects on the forecast.

ARIMA Model can also be directly used to predict, but to leave no room for error, the AR and MA graphs are extrapolated as well, to check for any residual correlation in the series. There is no correlation, and hence ARIMA Modeling is done.

The values of p and q are for AR and MA respectively.

AutoRegression (AR)

AutoRegression is forecasting done based on past values only. Here, Yt is a function of different past values, with the order depending on the number of parameters. The number of parameters here are 2. As concluded earlier, the ACF and PACF graphs plotted are being used by estimation, so to reduce the error as much as possible, 1 and 2 are used as the values of AR.

The ARIMA forecasting equation (p,d,q) has been used. Since this is only AR, q is put as zero.

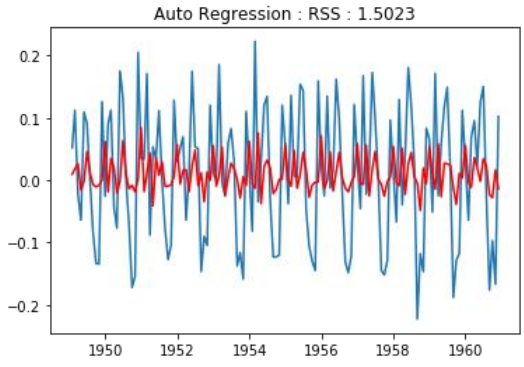

Fig 5.2.1. ARIMA(2,1,0) The above graph is for the ARIMA Model(2,1,2)

Values of p and q:

The values of p and q are important for the prediction for ARIMA Modeling. These values can be deduced using Autocorrelation and Partial Autocorrelation functions, tentatively. Autocorrelation function is the function of time series and the lags between itself, to check whetherthey are positively or negatively correlated. The value of q is determined by when the coefficient line touches the upper boundary of the confidence level in the graph.

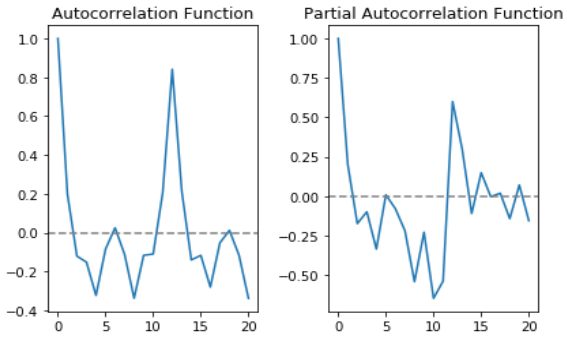

Fig 6. Autocorrelation and Partial Autocorrelations graphs

Partial Autocorrelation function is the correlation between the series and lag itself, where the correlation can't be explained by their mutual correlations with others. The value of p is determined by determined by when the coefficient line touches the upper boundary of the confidence level in the graph.

Here, there is a lot of variation in the graph, and hence values are estimated to be 1 and 2 for each, p and q.

Value of d:

The value of d is determined by the order of differencing. The order of differencing is the number of times the data has been differenced.

This is used for stationarity of the data. The entire data set is taken, and find the residue between two consecutive terms, and hence reducing the number of terms each time the differencing takes place. In this project, since there are 145 terms, hence after the differencing there will yet be 145 terms, since the first term is left as it is. Supposingly, the differencing is done twice, the number of terms would be 144.

Here, since differencing hasn't been done for stationarizing the data, and have used logarithmic method, the values have been assumed to be to be 0 and 1, with fitting the ARIMA model for each value of d.

Moving Average (MA)

Moving Average is the model which forecasts by taking the error terms only, which is gotten by regressing the term with its immediate past series. The data is divided into subsets and their average is taken, whose average again is taken. Plotting this helps in the removal of any residual autocorrelation.

We are using the same format as ARIMA, where substituting the value of p to be .

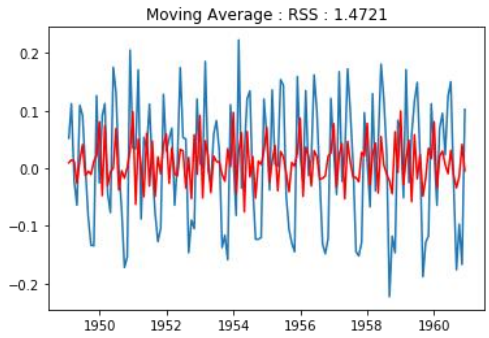

Fig 5.3.1. ARIMA(0,1,2) The above graph is for the ARIMA Model(2,1,2)

ARIMA Modeling

In the above section, the AR and MA graphs have been plotted with their respective values.

In this section, the graphs are plotted for the following models:

2,1,2

1,1,2

2,1,1

on top of the graphs, there are some values, called RSS.

RSS values are the residual sum of squares. It measures the variance of the data and the estimation model, which isn't explained by regression, i.e. the the error remaining between the dataset and the regression function. Smaller the RSS value, tighter fit is the model, i.e. more ideal is the model.

Along with the graphs, prediction values are also found. These prediction values are very small, since they are in log. Below are the graphs plotted of each of the models, where there is a very linear trend in the graph, which is due to the small differences in the terms, since they're in log.

Root Mean Square Error (RMSE) which is the root of the RSS value, which is more interpretable than the RSS values. It is a measure of the concentration of the data along the regressive line.

The values that have been predicted are in the log form and hence the log is taken. The values are in log since we take the logarithmic transformation to convert the data into stationary, which is later going to get converted to the error values.

```
Month
1949-02-01     0.009580
1949-03-01     0.017491
1949-04-01     0.027670
1949-05-01    -0.004521
1949-06-01    -0.023890
dtype: float64
```

Fig. 5.4.4 Differing log values

The above values are only for ARIMA(2,1,2).

A cumulative is taken of all the values, to reduce the total error, by extension reducing the error in the forecasting

Fig 5.4.5 cumulative log values

With these, by taking the exponent, there is a difference in the existing values and the predicted values, which have been plotted in the graphs below.

```
Month
1949-01-01    112.000000
1949-02-01    113.078122
1949-03-01    115.073417
1949-04-01    118.301998
1949-05-01    117.768375
dtype: float64
```

Fig. 5.4.6 New Values with differenced error

These values are the differing values for the existing dataset and the error between them and the predicted values.

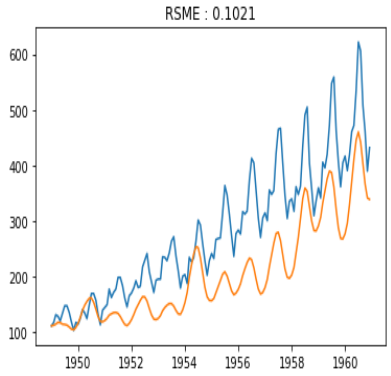

Fig 5.4.5. Error and original values graph with RSME in log form.

the actual data, where the yellow line represents the actual data and the blue line represents the error. The RSME value here is based on the log, where as in the figure below, it's based on the actual values, as whole numbers.

Prediction

The prediction is done on the basis of the past value, with a 95% confidence level.

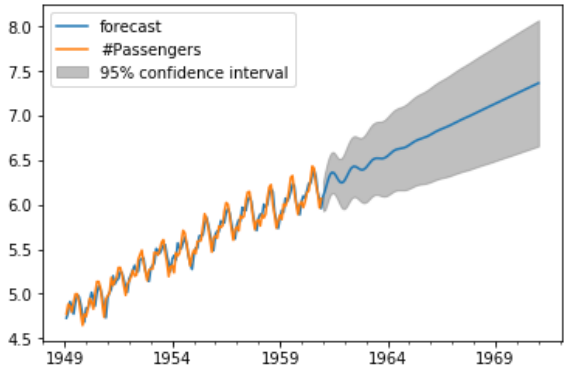

Fig 18. Prediction graph of (2,1,2)

The data from 1949-1960, and hence here it has been extrapolated for 1961-1969.

There is a linear trend, since the prediction is done only on the basis of the log values which were predcited.

The RSS and RMSE values have been shown in the table below, it can be concluded that the model with the least values of each is the most apt model.

Model	RSS	RMSE
2,1,2	1.0292	0.1021
2,1,1	1.1753	0.1021
1,1,2	1.2173	0.1037

Looking at the table above it can be concluded that the ARIMA Model (2,1,2) is the ideal model for the prediction.

V.DISCUSSIONS

As seen in literature survey, there are many similar proposed methods, except, there are drawbacks involved in it. A lot of these drawbacks are mostly due to high complexity and co-dependency of the variables involved.

In ARIMA Modeling, the is no variable codependency, and the complexity isn't that high.

VI. RESULT AND CONCLUSION

Upon the use and fitting of various models, the ARIMA(2,1,2) has been the best fit, having the least RMS and RMSE values. Looking at the table 6.1, it can be concluded that the ARIMA Model (2,1,2) is the ideal model for the forecasting. The existing studies have many proposed methods, with drawbacks involved in it, ue to high complexity and co-dependency of the variables. The above study was a remarkable example proving that ARIMA is one of the most independent and low costing model to use for the forecasting of the number of passengers travelling.

REFERENCES

1) Prediction for Air Route Passenger Flow Based on a Grey Prediction Model, Liu Xia,Yang Jie,Chen Lei, Cheng Ming-Rui
2) Prediction Study of Passenger flow of Airline Company Based on SVM Regression Method[J], Enterprise Economy,YAN Kewu, ZHU Jinfu
3) Hybrid Forecasting Model To Predict Air Passenger and Cargo in Indonesia, Ratna Sulistyowati, *Suhartono, Heri Kuswanto, Setiawan, Erni Tri Astuti
4) Characterization and prediction of air traffic delays. Transportation research Rebollo, Juan Jose; Balakrishnan, Hamsa; part C: Emerging technologies, v. 44, p. 231-241, 2014.
5) Arrival Trade-offs Considering Total Flight and Passenger Delays and Fairness, A. Montlaur, L. Delgado
6) An Exploratory Analysis for Predicting Passenger Satisfaction at Global Hub Airports using Logistic Model Trees, Hari Bhaskar Sankaranarayanan, Viral Rathod, Vishwanath BV
7) Analysis of Transfer Mode of Airport Passengers Based on Decision Tree, Huang Jia, Zhang Ning
8) Route Prediction in Air Travel Network Using Socio-Economic Factors and Learning Models, Sukrit Sriratanawilai, Supaporn Erjongmanee
9) Effects of Passengers and Internal Components on Electromagnetic Propagation Prediction inside Boeing Aircrafts, Mennatoallah Youssef, Linda Vahala
10) A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks, Karthik Gopalakrishnan, Hamsa Balakrishnan
11) https://www.kaggle.com/rakannimer/air-passengers#AirPassengers.csv